



HAL
open science

Computational principles of adaptive coding in healthy and impaired reinforcement learning

Sophie Bavard

► **To cite this version:**

Sophie Bavard. Computational principles of adaptive coding in healthy and impaired reinforcement learning. Neuroscience. Université Paris sciences et lettres, 2021. English. NNT : 2021UPSLE009 . tel-04563060

HAL Id: tel-04563060

<https://theses.hal.science/tel-04563060>

Submitted on 29 Apr 2024

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

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

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École normale supérieure

**Computational principles of adaptive coding
in healthy and impaired reinforcement learning**

Soutenue par

Sophie BAVARD

Le 9 avril 2021

École doctorale n°158

**Cerveau, cognition,
comportement**

Spécialité

**Neurosciences
computationnelles**

Composition du jury :

Mathias PESSIGLIONE

Sorbonne University

Président

Claire GILLAN

University of Dublin

Rapporteur

Sebastian GLUTH

University of Hamburg

Rapporteur

Stefano PALMINTERI

ENS - PSL Research University

Directeur de thèse

Remerciements

En tout premier lieu, je souhaite remercier mon directeur, Stefano Palminteri, pour ta présence depuis mon premier jour de stage. Je ne peux malheureusement pas faire la liste exhaustive de tout ce que tu apportes en tant que directeur, mais si je devais ne retenir que quelques lignes, je te remercierais pour ta capacité d'écoute, ton soutien et tes encouragements, ta grande disponibilité, et toutes tes qualités scientifiques si inspirantes.

Je remercie ensuite les membres de mon jury, Claire Gillan et Sebastian Gluth, pour avoir accepté de lire et évaluer mon travail. Merci également à Mathias Pessiglione qui me fait l'honneur d'être membre de mon jury, et d'avoir suivi mon travail depuis plusieurs années.

Cette thèse n'aurait peut-être pas vu le jour sans le soutien financier de la MILDECA et l'EHESS, que je remercie vivement d'avoir cru en moi et en mon projet.

Pour toutes les discussions scientifiques que nous avons eu à divers moments au cours de cette thèse, je remercie chaleureusement Julie Grèzes, Maël Lebreton, Catherine Tallon-Baudry, et Valentin Wyart. Je remercie également Hernan Anllo, Peter Neri, Aldo Rustichini, et Jean-Christophe Vergnaud.

Je tiens ensuite à remercier toute l'équipe HRL, passée et présente : Anis, Germain (pour les gâteaux trop gras, trop sucrés, trop salés), Henri, Magda, Raggio (pour les délicieux gnocchis), et Sabine. Je remercie Basile, mon frère de thèse, pour les discussions nombreuses et nos débats scientifiques (et autres) ; enfin, je remercie Fabien, l'équipe ne serait pas la même sans toi (et de toute façon, je sais qu'on se suivra même après !).

Je tiens évidemment à remercier mes amis du LNC² : Vasilisa qui m'a beaucoup appris au labo comme en dehors, Charles pour les projets du futur (et les bières flottantes. . .), Héroïse pour ton aide et ton soutien (et les jeux de société), Felix pour toutes les coïncidences (sans tricher, bien sûr), Tahnee (met wie deel ik mijn passie voor planten en katten?), Morgan (pour les gâteaux, les chants, et surtout la bonne humeur). Merci à Adrian (pour l'abondance de café, indispensable à tout chercheur qui se respecte), Anne, Aurélien, Damiano, Julie, Jun, Rocco, Victor, pour tous les bons moments que nous avons partagés. Je remercie particulièrement Margaux et Clémence, c'était un plaisir de faire ma thèse à vos côtés (d'autant plus qu'on a maintenant une équipe

imbattable au Duel Quiz). Enfin, Marine, pour tout ce que tu fais pour nous tous au labo, mais surtout pour tout le reste (si tu crois que tu vas te débarrasser de nous comme ça, c'est pas si facile...)

J'ai également eu la chance de pouvoir compter sur mes amis : Elodie, Jonathan, Momo, Oli, Solène, Yoons. Je remercie Marc (mon fournisseur officiel de jeux vidéo), ainsi que mes compagnons d'infortune : Alan, Benjamin, Elliott et Elise. Enfin, merci à mes 4 compères : Loïs (el Steakos), Manu, le vieux Pew et Yasmine, merci de m'avoir accompagnée pendant ces 3 années.

Pour finir, mon plus grand merci va à ma famille, pour son soutien sans faille et pour tout le reste : merci à mes parents, à Mayaya, à Val, Cris et Vicky, à Ivan et Claire, à Raph et bien sûr à Juliette, *ma grande soeur préférée*.

Abstract

Reinforcement learning is a fundamental cognitive process operating pervasively, from our birth to our death. The core idea is that past experience gives us the ability of learning to improve our future choices in order to maximize the occurrence of pleasant events (rewards) and to minimize the occurrence of unpleasant events (punishments). Within the reinforcement learning framework, one of the most fundamental and timely questions is whether or not the values are learned and represented on an absolute or relative (i.e., context-dependent) scale. The answer to this question is not only central at the fundamental and theoretical levels, but also necessary to understand and predict why and how human decision-making often deviates from normative models, leading to sub-optimal behaviors as observed in several psychiatric diseases, such as addiction.

In an attempt to fill this gap, throughout the work carried out during this PhD, we developed existing models and paradigms to probe context-dependence in human reinforcement learning. Across two experiments, using probabilistic selection tasks, we showed that the choices of healthy volunteers displayed clear evidence for relative valuation, at the cost of making sub-optimal decisions when the options are extrapolated from their learning context, suggesting that economic values are rescaled as a function of the range of the available options. Moreover, results confirmed that this range-adaptation induces systematic extrapolation errors and is stronger when decreasing task difficulty. Behavioral analyses, model fitting and model simulations convergently led to the validation of a dynamically range-adapting model and showed that it is able to parsimoniously capture all the behavioral results. Our results clearly indicate that values are not encoded on an absolute scale in human reinforcement learning, and that this computational process has both positive and negative behavioral effects. In an attempt to explore the link to -an impairment of this process in reward-related psychiatric diseases, we performed a meta-analysis based on the valence bias observable in several pathologies. Preliminary results suggest that healthy volunteers learn similarly from rewards and punishments, whereas it is not the case for pathologies such as Parkinson's disease or substance-related disorders. In a large-scale experiment, coupled with a transnographic approach used in computational psychiatry, we found that the parameters of our model could not be directly linked with different dimensions of psychiatric symptoms, including obsessive compulsive disorders, social anxiety, and addiction. Further work will improve our modeling tools to better account for behavioral variance. In the long term, these analyses will potentially help to develop new tools to characterize phenotypes of several pathologies and behavioral disorders, as well as improve patients' treatment at the individual level.

General introduction to the manuscript

The notion of context-dependence in economic decision making emerged with experimental findings showing that our choices depend on the value of the alternative options. However, the investigation of context-dependence mostly focused on choices where options and prospects were fully described, and research in situations where the values have to be learned by trial-and-error has comparably neglected the notion of outcome context-dependence.

In the first chapter of this manuscript, I will review the state-of-the-art theoretical and experimental framework that motivated the research presented in this thesis. First, I will present the experimental background that led to investigating context-dependence in decision-making. Then, I will introduce the behavioral experiments which contributed to the definition of reinforcement learning. Subsequently, I will present the specific experimental modeling tools used in the reinforcement learning framework. Finally, I will address the specific aims of this thesis and provide a general outline of the different research questions.

In the second chapter, I will present two studies, in the form of scientific papers. Each paper will be briefly introduced with the specific aims and main findings of the study, and concluded with the limitations of our experiments and models.

In the third chapter, I will present ongoing work on the investigation of context-dependence in impaired reinforcement learning. Each section consists of introducing the current research questions, presenting preliminary results, and a brief discussion on the next steps that are to be followed.

In the fourth chapter, I will present some perspectives on the work that I conducted during this PhD. I will focus on future projects based on the findings of this thesis.

Finally, the appendices are mostly composed of supplemental studies in which I took part during my PhD. This includes four clinical studies, a perspective paper and a replication paper.

Table of Contents

- 1 Introduction** **1**
- 1.1 Context-dependence in decision-making 1
 - 1.1.1 Economic behaviors 2
 - 1.1.2 Utility theory 5
 - 1.1.3 Prospect theory 7
 - 1.1.4 Context-dependent neuronal activity 8
- 1.2 Behavioral reinforcement learning 13
 - 1.2.1 Animal conditioning in history 14
 - 1.2.2 Human reinforcement learning 18
 - 1.2.3 Neural reinforcement learning 19
- 1.3 Computational reinforcement learning 22
 - 1.3.1 Rescorla-Wagner model 25
 - 1.3.2 Temporal Difference learning 26
 - 1.3.3 Q-learning 28
 - 1.3.4 Action selection 29
 - 1.3.5 General method 31
- 1.4 Research questions 35
 - 1.4.1 Context-dependence in the general population 35
 - 1.4.2 Context-dependence in neuropsychiatric diseases 37
- 2 The paradoxical consequences of context-dependence in human reinforcement learning** **39**
- 2.1 Study 1: Bavard, Lebreton et al, 2018 39
 - 2.1.1 Introduction 39
 - 2.1.2 Article 39
 - 2.1.3 Conclusion 57

2.2	Study 2: Bavard et al, 2021	57
2.2.1	Introduction	57
2.2.2	Article	57
2.2.3	Conclusion	88
3	The multiple facets of reinforcement valence in neuropsychiatric diseases	89
3.1	A meta-analysis	89
3.1.1	Introduction	89
3.1.2	Methods	92
3.1.3	Preliminary results	93
3.1.4	Conclusion	95
3.2	A large-scale study	97
3.2.1	Introduction	97
3.2.2	Methods	98
3.2.3	Preliminary results	101
3.2.4	Conclusion	107
4	Discussion and perspectives	108
4.1	Contrasting adaptive coding and divisive normalization in human reinforcement learning	110
4.2	Assessing the role of working-memory in range-adapting learning	113
4.3	Cross-cultural study of the impact of contextual information in decision-making	115
4.4	"There are known knowns..."	118
	References	120
	Appendices	134
A	Estimation of the model evidence	135
B	Additional results	138
C	Assessing inter-individual differences with task-related functional neuroimaging	160
D	[Re] Adaptive properties of differential learning rates for positive and negative outcomes	170

Chapter 1

Introduction

1.1 Context-dependence in decision-making

In the famous Ebbinghaus illusion, two circles are placed next to each other. Larger circles surround the left one, while smaller circles surround the right one. If we look at Figure 1 and try to figure out which central circle is the biggest one, we might immediately see that the circle surrounded by smaller circles appears bigger than the one surrounded by bigger circles. Even if we objectively know that this is an illusion, somehow we cannot see the circles being of identical sizes. This simple optical illusion is an excellent indication that the subjective estimation of the size of an object might drastically be affected by its surroundings.

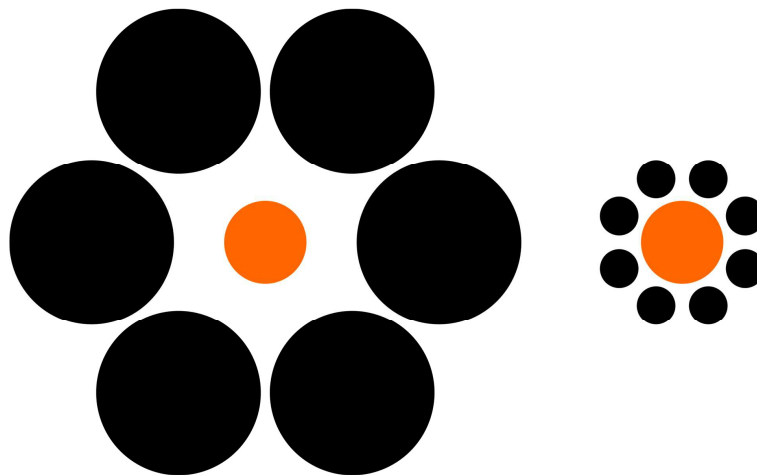


Figure 1. Ebbinghaus illusion, or *Titchener circles*. Which central circle is the biggest one? One might intuitively say that the one on the left is clearly smaller than the one on the right. In reality, as you might have guessed, both central circles are of identical objective size, indicating that our subjective size estimation is affected by the object's surroundings.

This is an illusion of relative size perception. The difference in size perception is due to the surrounding visual cues (larger or smaller surrounding circles), and the way the brain processes these visual cues. But beyond the brain’s visual system, can we find examples of biased perception in other domains of decision-making, such as economic decisions? Can we manipulate decisions by adding or removing surrounding components of the choice options? In this PhD, I focused on context-dependent decision making, and more specifically, the way context-dependence influences decisions toward deviations from optimality. By optimality, I mean minimizing the number of errors, and in this case, choosing the option with the highest objective value. For example, choosing the right circle as the smaller circle is objectively a wrong answer because the circles are of the same size, even if it seems like the right answer for our visual and decision-making systems. In an economic decision-making problem, the option with the highest objective value will be the option with the highest mathematical expectation. Therefore, choices deviating from optimality will be called *sub-optimal* or *irrational* choices, although I acknowledge that labeling them as *errors* is questionable. Indeed, one might argue that natural selection does not create organisms that follow economic theories, it creates decision makers that maximize some notion of fitness. Thus, my goal will be to focus on these choices and try to begin to understand some of the aspects of that fitness.

1.1.1 Economic behaviors

Most models of decision making assume that individuals have an ordered list describing their complete set of preferences. For example, a cherry might have a value of one, a banana a value of two, and an apple a value of four. Decisions are then made by comparing these options and choosing the one with the highest value (i.e., the apple). In this framework, some important principles must be followed:

- preferences should be *transitive*, with a consistent ranking of preference order. If an individual prefers apples to bananas and bananas to cherries, then the same individual should also prefer apples to cherries, that is:

$$\text{if } A > B \text{ and } B > C \text{ then } A > C \tag{1.1}$$

- decisions should be *independent of irrelevant alternatives* (IIA), which means that adding low-quality alternatives to a set of options should not influence the decisions (Luce 1959, Rieskamp et al. 2006). For example, when given a choice between apples, bananas, and cherries, the presence or absence of the cherry (the least preferred and therefore irrelevant option) in the choice set should not affect relative preferences between apples and bananas.

- preferences should be *invariant*, which means that the same options should produce the same decision, regardless of how the experimenter presents the options (Tversky and Kahneman 1986).

Taken together, these axioms predict consistent decision making: rational choice theory ignores how initial values are assigned to different options, but once they are assigned, decision makers should follow the principles. However, psychologists and behavioral economists have collected a wealth of evidence challenging these axioms: I will now present a few examples of violation of the principles of rational choice in human decision making.

Framing effect

Humans alter their choice depending on whether a purchase is framed as a loss or a gain. Among many examples of framing effect experiments, I chose to illustrate the framing effect with an experiment from Gächter and colleagues in 2009, where experimental economists registered for a conference in 2006. The price of the conference was of 145 dollars for early registration and 195 dollars for late registration. In a first version of the acceptance email, the price was framed as a discount of 50 dollars for early registration, whereas in a second version, it was framed as a penalty of 50 dollars for late registration. The price change of 50 dollars was the same for all participants but for half of them the price change was framed as a penalty. The way that the price change was presented to them affected the decision to register early or not. Indeed, among the participating PhD students, 67 % registered early in the first group when the change in registration price was a discount and 93 % registered early in the second group when the change in price was a penalty (interestingly, the effect did not occur for senior economists). This effect was originally published as the well-know "Asian disease problem" (Tversky and Kahneman 1981) where participants were asked to choose between 2 alternative programs to combat a disease, the first being framed a "people will live" and the second one framed as "people will die". These results fit a general pattern that losses or penalties affect our behavior more than gains or discounts, which is evidence for irrational behavior since the options have equal mathematical expectations but preferences are not invariant (Tversky and Kahneman 1981, Plous 1993, Druckman 2001, Gächter et al. 2009).

Decoy effect

Another instance of irrational decision making is the violation of the independence of irrelevant alternatives (IIA) mentioned above. When we choose between two options and then a third option is added, the third option might make us depreciate the original two options, but it

should not make us like one of the original options more. It also should not change how we compare the other two options to each other, as stated by the principle of IIA. To illustrate with the previous fruits example, if one prefers apples to bananas, the addition of cherries in the choice alternatives should not make one choose bananas over apples. However, the effect called decoy pricing takes advantage of the influence a third option can have on our perception of two other options. In an meta-analysis conducted by Heath and Chatterjee in 1995, participants had to choose whether they wanted to buy a smooth ride like a Rolls-Royce with worse gas mileage, or a rough ride like a Jeep with better gas mileage. When presented with both options, participants chose the Rolls-Royce 58% of the time and the Jeep 42% of the time. Then a third option, called a decoy, was presented: the decoy had as good of gas mileage as the Jeep but it was an even rougher ride. When this third option was presented, it actually made participants ignore the Rolls-Royce because they did not have a good comparison to make. Instead, participants compared the Jeep to the decoy as these were more similar and more easier to compare. This resulted in an increase in preference for the Jeep: participants chose the Jeep 70% of the time when they had a relatively worse but comparable decoy to compare it to (Heath and Chatterjee 1995). Another example is the attraction effect, in which adding a third alternative, which is clearly inferior to an option A but not to another option B, increases the probability of choosing A. In general, the addition of a third alternative can influence the choice between the two original options in many ways, depending on the third option's value, which suggests that alternatives are not independent (Huber et al. 1982, Heath and Chatterjee 1995).

Experience effect

Another way in which humans violate rational expectations is by changing behavior due to each individual's experience. Not only can the alternatives we are choosing from change our rating, but the options we have had in the past can also change our perception. In an experiment conducted by Simonson and Tversky in 1992, participants made decisions about buying car tires, based on past options. Participants were split into two groups: group A initially compared tire options that differed and how long they would last by 20,000 miles but differed in price by only 6 dollars, whereas group B initially compared tire options that differed and how long they would last by only 5,000 miles but differed in price by 24 dollars. This way, group A saw a big difference in quality for a small difference in price while group B saw a small difference in mileage for a big difference in price. Both of these groups were then given two tires to choose from, in a final choice set where the tires differed in quality by 10,000 miles and differed in price by 15 dollars. For group A, 10,000 miles was not worth the extra 15 dollars and they mostly chose the cheaper

option. For group B, 10,000 more miles for only 15 extra dollars seemed like a steal and they mostly chose the more expensive option. Even though both groups got to choose between the same tires, the options they previously saw influenced what they thought was a good deal versus what they thought was a ripoff (Simonson and Tversky 1992).

To account for these seemingly irrational choices, numerous amendments of the standard "rational choice" theory have been necessary to explain human behavioral biases, which I will briefly summarize in the next section.

1.1.2 Utility theory

Rational choice theory derives from the expected utility theory, first proposed in 1738 by the Swiss mathematician Daniel Bernoulli (1700-1782). Back then, it was known as moral expectation, as opposed to mathematical expectation, until the mid 20th century (Bernoulli 1738, 1954). When he was a resident of the eponymous Russian city, 38-year-old Bernoulli solved the well-known St. Petersburg paradox, introduced by his (mathematician) cousin, Nicolas Bernoulli (1687-1759), in 1713 (Montmort 1713). In this paradox, a casino offers a game to a single player, in which a fair coin is tossed at each stage. The initial stake begins at 2 dollars and is doubled every time heads appears. The first time tails appears, the game ends and the player wins whatever is in the pot. Thus, the player wins 2 dollars if the first toss is tails, 4 dollars if the tosses are head-tails, 8 dollars if the tosses are head-head-tails, and so on. Mathematically, the player wins 2^n dollars, where $n \in \mathbb{N}^*$ is the number of tosses. What would be a fair price to pay the casino for entering the game? To answer this, one needs to consider what would be the average payout: the player wins 2 dollars with probability $\frac{1}{2}$, 4 dollars with probability $\frac{1}{4}$, etc. Therefore, for an infinite number of stages, the expected value converges to infinity, because the sum grows without bounds. However, there is a discrepancy between what individuals seem willing to pay to enter the game and the infinite expected value. The classical resolution of the paradox involved the explicit introduction of a utility function, an expected utility hypothesis, and the presumption of diminishing marginal utility of money. For Daniel Bernoulli, what matters to the player is the utility, not the gain: utility is not only decreasing, but logarithmic, which means that doubling the gains actually means adding one unit of utility. In this case, utility takes only one finite, weak value, and paying a small bet to enter the game is actually a rational behavior. In Bernoulli's own words:

The determination of the value of an item must not be based on the price, but rather on the utility it yields... There is no doubt that a gain of one thousand ducats is

more significant to the pauper than to a rich man though both gain the same amount.

(Bernoulli 1738)

For each possible event, the change in utility will be weighted by the probability of that event occurring, describing *risk-averse* behaviors (Figure 2).

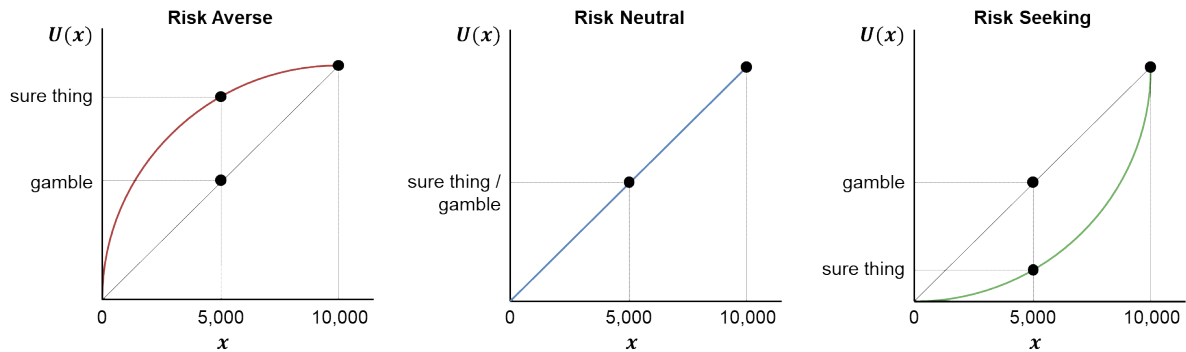


Figure 2. Graphical representation risk preferences in expected utility theory. The values on the x -axis are arbitrary and represent, for example, monetary outcomes. A risk-averse individual has a concave utility function and prefers the sure thing over the gamble. A risk-neutral individual does not care about risk. The utility derived from the gamble and the sure thing are the same and the utility function is a straight line. A risk-seeking individual has a convex utility function and prefers the gamble over the sure thing.

Expected utility theory was further developed by John von Neumann and Oskar Morgenstern in 1947. Their work describes utility as an index of "usefulness" and assumes that decision makers attempt to maximize their expected utility. Therefore, individuals should prefer options that offer the highest utility, weighted by the probability of acquiring the outcome. From the axioms described in the previous section, expected utility theory predicts how *rational actors* should behave (Von Neumann and Morgenstern 1947). Therefore, expected utility theory represents a normative theory of choice, because it describes what a rational actor should do to achieve a norm of behavior, namely maximize utility. However, as we just saw in the previous examples, expected utility often does a poor job at predicting how humans actually behave (Thaler 1992, Camerer et al. 2004). The **context**, including the decision maker's previous experiences, the set of available options when they make their decisions, and the framing of these options, has a pervasive influence on human decision making (Tversky and Kahneman 1981, Simonson and Tversky 1992, Kahneman and Tversky 2000). Of note, these context-dependent types of behavior have also been observed in non-human species, such as monkeys (Chen et al. 2006) and starlings (Marsh and Kacelnik 2002) for the framing effect, and hummingbirds (Bateson et al. 2002), honeybees and gray jays (Shafir et al. 2002) for violations of IIA.

1.1.3 Prospect theory

Prospect theory was proposed by Daniel Kahneman and Amos Tversky in 1979 and developed until 1992, for which Kahneman won the Nobel Memorial Prize in Economic Sciences in 2002. Prospect theory examines the same core concepts as utility theory, however it includes the individuals' *reference-point* in regards to decision-making: it is about the individuals' gains and losses rather than utility or usefulness of their wealth. In simple terms, we dislike losing more than we like winning. Prospect theory goes on to explain why individuals might not always be risk-averse when faced with bad outcomes: individuals become risk seeking in hopes of receiving the better outcome (Kahneman and Tversky 1979).

In prospect theory, Kahneman and Tversky address two key additions to utility theory. First, utility theory does not take into account where the individuals started from and how it will feel to shift from that point of view. For example, two individuals A and B own one million dollars and three million dollars respectively; the next day, they both end up with two million dollars. According to utility theory, on day two they should be equally happy. However if we look at the numbers, individual A is probably much happier than individual B. Prospect theory can predict this difference, as it takes into account that individual A won one million dollars whereas individual B lost one million dollars. Second, prospect theory takes into account that people are not entirely rational: individuals do not make decisions based solely upon which choice has more utility, but also upon which choice is less aversive or causes them less loss. Prospect theory starts with the concept of *loss aversion*, an asymmetric form of risk aversion, from the observation that people react differently between potential losses and potential gains. Thus, people make decisions based on the potential gains or losses relative to their specific situation (the reference-point), rather than in absolute terms; this is referred to as reference-point dependence.

- Faced with a risky choice leading to gains, individuals are risk-averse, preferring solutions that lead to a lower expected utility but with a higher certainty (concave value function, Figure 3).
- Faced with a risky choice leading to losses, individuals are risk-seeking, preferring solutions that lead to a lower expected utility as long as it has the potential to avoid losses (convex value function, Figure 3).

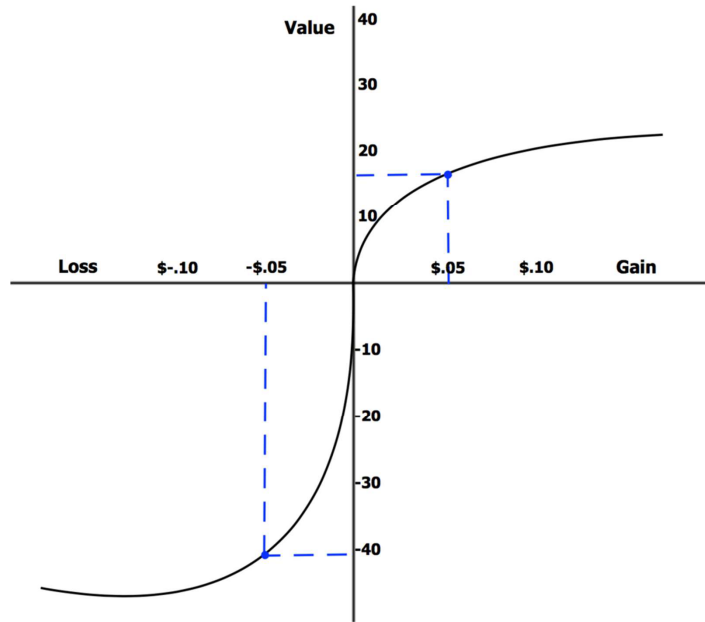


Figure 3. Graphical representation of prospect theory. The curve in the upper right represents gains while the curve in the lower left represents losses. The decline on the left is steeper than the incline on the right, indicating that losses are more salient than gains.

1.1.4 Context-dependent neuronal activity

The nature of valuation and decision processes is thus of fundamental interest to researchers at the intersection of psychology, neuroscience, and economics. Until now, we have discussed some theories of representation of value in human behavior, but as mentioned in section 1.1.2, context-dependent decision making has been observed in non-human species, allowing us to study the representation of value in the brain itself and investigate how value is instantiated, in the activity of neurons and neural circuits. Motivated by economic models of choice, a growing number of neuroscientific studies have demonstrated that it is in fact the subjective rather than objective value of rewards that best correlates with reward-related activity in the brain (Kable and Glimcher 2007, Rangel and Hare 2010). Although the concept of utility in economic models of choice is not attached to any particular unit of measure, the neural representation of value is instantiated via actual spiking rates. As a result, many different possible neural representations of value will be consistent with a given set of choice data; for example, two systems whose value representations are $V_1 = 10, V_2 = 20$ spikes/sec and $V_1 = 100, V_2 = 200$ spikes/sec, would produce identical behavioral choice preferences. Thus, behaviorally generated models of value only provide limited constraints on how neural systems represent values (Louie and Glimcher 2012). Does the brain represent action values in absolute terms, independent of the other available options, or in relative terms?

Divisive normalization

In monkey lateral intraparietal cortex (LIP), a parietal region responsive to both visual stimuli and saccadic eye movements, neuronal activity is strongly modulated by the value associated with a saccade. To investigate the different forms of value representation observed in the brain, with a focus on primate electrophysiology, Louie and colleagues quantified LIP responses in a two-target task, in which the response field (RF) target value was held constant and the extra-RF target value was explicitly varied (Figure 4, top panel). The RF target value is labeled V_{in} and the value of the alternatives (extra-RF targets) is labeled V_{out} . The neuronal activity showed three main interesting results. First, when a RF target is presented, the activity elicited by RF target onset is modulated by the value of the alternatives, with larger V_{out} magnitudes leading to greater suppression. Second, when no RF target is presented, the activity is also suppressed with context-dependence, with larger V_{out} magnitudes driving activity further below baseline activity levels (Figure 4). Finally, these results are consistent with a model of divisive normalization:

$$R_i \propto \frac{V_i + \beta}{\sigma^2 + \sum_j V_j} \quad (1.2)$$

where the activity of a neuron R_i is dependent on both the value of the RF target V_i and the sum of the alternative targets V_j , the empirical parameter β models the suppression below the baseline rate and σ^2 is an empirical semi-saturation constant (Heeger 1992, Louie et al. 2011).

By investigating the different forms of value representation observed in the brain with a focus on primate electrophysiology, Louie and colleagues showed that context-dependent behaviors exhibited by monkeys, such as decoy effect and violation of IIA, have similar patterns in the neuronal activity of the visual system. More recently, Webb and colleagues formalized a divisive normalization model which shapes the substitution patterns that violate IIA (Webb et al. 2020b). Let $\mathbf{v} = [v_1, \dots, v_N] \in \mathbb{R}_+^N$ be an input vector of N alternatives. The transformation of the valuation of each alternative i in a choice set at the time of decision is:

$$z_i(\mathbf{v}) = \frac{v_i}{\|\mathbf{v}\|_\beta} \quad (1.3)$$

where $\|\mathbf{v}\|_\beta$ is the β -norm of vector \mathbf{v} :

$$\|\mathbf{v}\|_\beta = \left(\sum_{n=1}^N v_n^\beta \right)^{\frac{1}{\beta}} \quad (1.4)$$

If $\beta = 1$, the 1-norm is equal to the sum of the elements and its graphical representation in \mathbb{R}^2 is a line; if $\beta = 2$, the 2-norm is the square root of the sum of the squared elements

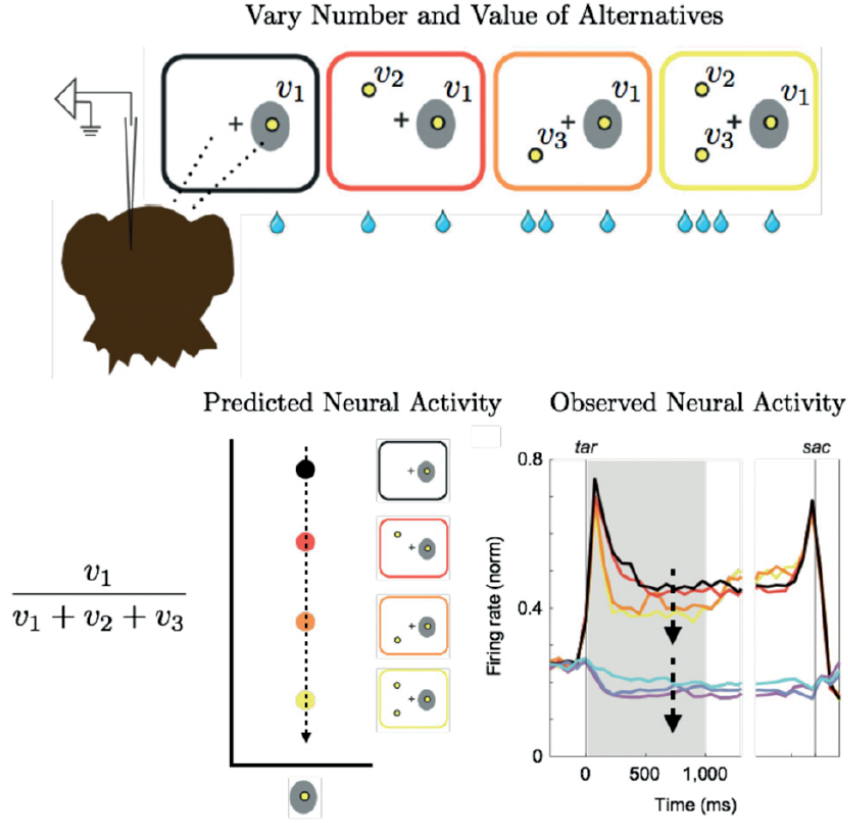


Figure 4. Spatial context dependence in LIP value coding. Top: Different value conditions in an oculomotor saccade task. Monkeys were presented with a target array of one, two, or three peripheral targets associated with different reward magnitudes. The value of the response field (RF) target was constant, whereas the value context varied with the number and reward magnitude of extra-RF targets. Bottom: Population average activity. Both target-driven (black-yellow) and baseline (cyan-purple) activity exhibit suppression by the presence of extra-RF targets. (Figure adapted from Louie et al. 2011 in Webb et al. 2020b)

and its representation in \mathbb{R}^2 is a quarter circle, etc. When $\beta = \infty$, the uniform norm is the maximum of the elements and its representation in \mathbb{R}^2 is a quarter square. For example, consider binary choices in vectors \mathbf{v} and $\mathbf{v}' \in \mathbb{R}_+^2$. The proportionate scaling implemented by divisive normalization depends drastically on the value of β (Figure 5). In their model, Webb and colleagues generalize the simple divisive normalization function by adding the saturation parameter σ and a weight ω , which determines the contribution of other alternatives to the normalization:

$$z_i(\mathbf{v}) = \frac{v_i}{\sigma + \omega \left(\sum_n v_n^\beta \right)^{\frac{1}{\beta}}} \quad (1.5)$$

If $\omega = 0$, there is no normalization. This yields a bounded valuation $z_i(\mathbf{v})$ with a relative

relationship between alternatives. The choice is then performed by comparing options, with a probability that depends on the distribution of an error term η . Webb and colleagues have implemented the model in a previous data set from Louie and colleagues (Louie et al. 2013), and show that the divisive normalization model captures an important component of the variance and captures the sample choice probabilities for all set sizes (Webb et al. 2020b).

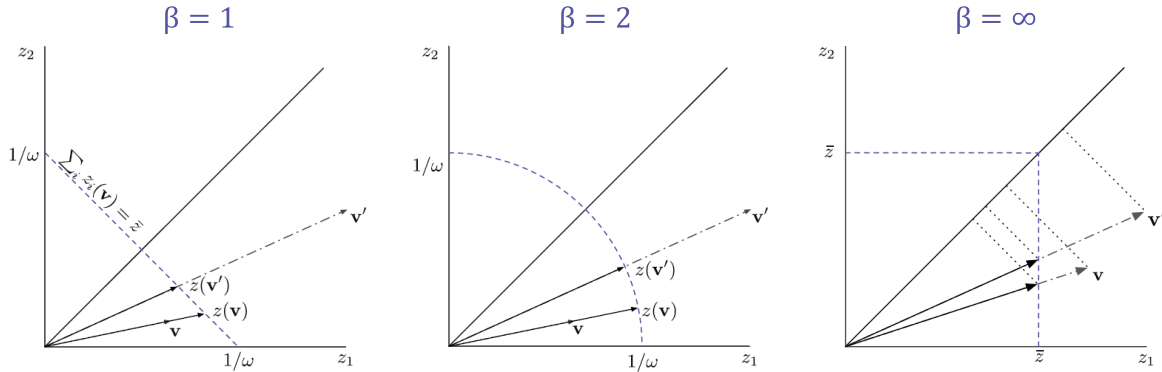


Figure 5. Proportionate scaling implemented by normalization. Representation of the β -norm and divisive normalization from equation 1.5, for $\beta = 1$ (left), $\beta = 2$ (middle), $\beta = \infty$ (right). For two vectors $\mathbf{v}, \mathbf{v}' \in \mathbb{R}_+^2$, the relative sizes of the normalized vectors depends on β . Figure adapted from Webb et al. 2020b.

Range adaptation

The results described in the previous section suggest that at least some parietal circuits involved in decision making reflect a normalization process across the available choice options, but one important issue is how contextual value coding varies in different brain areas performing different value-related processing. In a study published in *Nature* in 2006, Padoa-Schioppa and Assad showed that orbitofrontal (OFC) neurons encode a goods-based representation of value. In a series of following studies (including but not limited to Padoa-Schioppa and Assad 2008, Padoa-Schioppa 2009, 2013, Rustichini et al. 2017), Padoa-Schioppa and colleagues use the same task to investigate whether those value representations are dependent on the other available rewards in a choice situation. As in the original demonstration of value coding by OFC neurons, monkeys chose between pairs of varying amounts of juices A, B and C, that could be ranked by relative preference order (when offered in equal amounts, $A > B > C$). In this task, monkeys displayed transitivity, as in equation 1.1, indicating that the different rewards could be compared on a common value scale, enabling the examination and comparison of the different neural value representations. As in the original publication, the authors found three general types of response, which they labeled *offer value* (the presented value of a specific reward type), *chosen value* (the value of the selected option in a given trial, regardless of type), and *taste* (received reward type).

The distribution of possible reward sizes for a given juice type were fixed for each neuron, but varied across neurons. For example, one neuron may have been recorded with B rewards ranging from 0 to 2 (in equivalent units of juice A, determined by behavior), whereas a separate neuron was recorded with B rewards ranging from 0 to 10. To examine value-based adaptation, the authors examined whether, across the population of OFC neurons, firing rates depended on the range of the offer values. They proposed a model, where the firing rate ϕ of a neuron encoding the offer value or the chosen value and the encoded value V , is formulated as follows:

$$\phi = \phi_0 + \Delta\phi \cdot \frac{V - V_0}{\Delta V} \quad (1.6)$$

where $\phi_0 = c_0 + c_1 V_0$ is the baseline activity with parameters c_0 and c_1 representing, respectively, the intercept and the slope of the encoding, $\Delta\phi = c_1 \cdot \Delta V$ is the activity range, and V_0 is the minimum value available. Under this model, the slope of the relationship between firing rate and value would decrease as the possible value range increases, and the maximum of the value range should be represented by the same firing rate in different value-range conditions (Figure 6, top panels). When the mean population firing rates were split by value range, OFC activity showed a clear adaptation to the locally experienced range of values, for both offer value and chosen value responses (Figure 6, bottom panels).

The results of the previous examples suggest that contextual modulation plays an important role in determining the neural coding of value in multiple brain circuits. As such, one might hypothesize that relative value coding provides a possible link between decision-making circuits and context-dependent valuation. Indeed, contextual modulation, such as divisive normalization or range adaptation, can alter the relative distance between the mean firing rates that represent different actions. If we go back to the decoy effect, consider choosing between three options – two high-value items and one low-value distractor item. Under a relative value coding system, such as described previously in the parietal cortex, the mean firing rates representing the values of each option will be divisively scaled by the total value of all alternatives. Similarly to neurons in the visual system that adapt to components of the stimulus feature distribution, OFC responses show modulation by values encountered over a longer timescale. This adaptation is sensitive to multiple distribution components, including the mean, range, and variance of the recent value signals. Thus, adaptive processes in value storage areas may produce a more efficient neural representation of value for use in downstream decision processes (Kobayashi et al. 2010, Rangel and Clithero 2012, Soltani et al. 2017, Zimmermann et al. 2018, Conen and Padoa-Schioppa 2019).

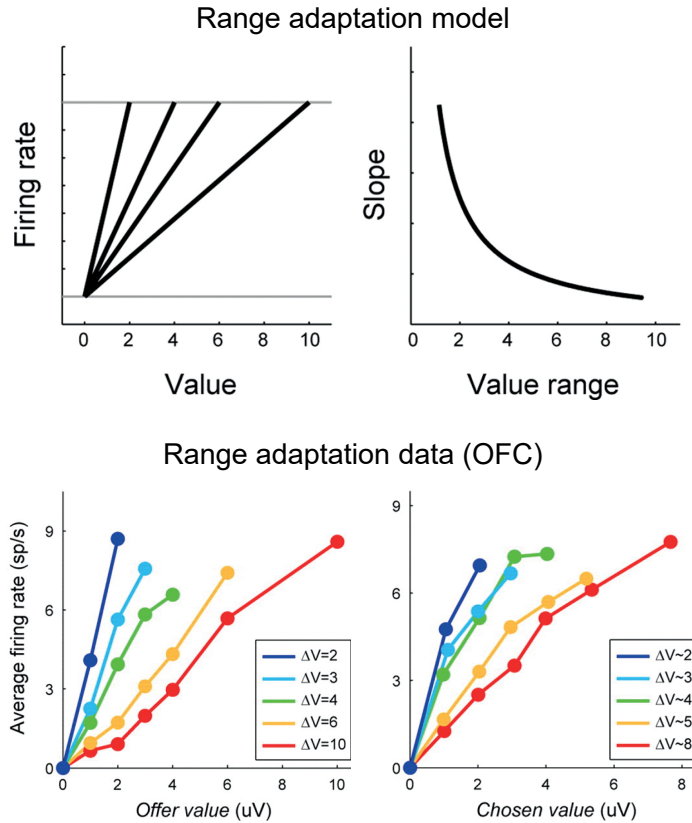


Figure 6. Context-dependence in orbitofrontal (OFC) value coding. Top: predictions of the range adaptation model in value coding neurons. The key assumption is that the range of neural activity remains constant across different behavioral value conditions. Bottom: Range adaptation in OFC neuronal activity. The two panels show average OFC activity in two different types of value-coding neurons, color-coded by the range of experienced values (plotted as normalized unit value). OFC population activity adapts to the range of possible values, indicating that such activity is sensitive to the context. Figure adapted from [Padoa-Schioppa 2009](#).

1.2 Behavioral reinforcement learning

When we think about learning, we often picture students in a classroom or a lecture hall, books open on their desk, listening intently to a teacher or professor in the front of the room. But in psychology, learning is defined as a long-lasting change in behavior as a result of experience. How is a new skill learned? This question has been fascinating scientists, from the first behavioral psychologists to ourselves today. One of the first to have published influential results on the subject was E. Thorndike ([Thorndike 1898](#)), whose major work shed new light on the associations made by the individual, and therefore is called *connectionism theory*. In the beginning of the 20th century, I. Pavlov ([Pavlov 1927](#)) developed one the 2 main types of conditioning, *classical conditioning*, by showing evidence for automatic responses in the learning process. A few years

later, B.F. Skinner (Skinner 1938) made some great progress in the 2nd main type, *operant conditioning*, by investigating learned behavior. The research on learning and conditioning kept going since then, and everything started with this discovery:

When behaviors change, learning happens.

1.2.1 Animal conditioning in history

Classical conditioning

Russia, 1927. More than 20 years after winning the Nobel Prize in recognition of his work on the physiology of digestion, the 78-year-old Ivan Pavlov (1849-1936) was pursuing his research on the gastric system of dogs by establishing connections in the ducts of the salivary glands, in order to carry out experiments on the nature of these glands. Over the years, Pavlov paid special attention to the phenomenon of what he called *psychic secretion*, which is caused by food stimuli at a distance from the animal. A series of experiments caused Pavlov to reject the subjective interpretation of *psychic* salivary secretion, but also to conclude that a reflex, though not a permanent but a temporary or conditioned one, was involved. In the experiment that led to one of the most fundamental discoveries in behavioral psychology, Pavlov delivered food (unconditioned stimulus, US) to hungry dogs right after a tone presentation (conditioned stimulus, CS). At the beginning of the experiment, the dogs produced saliva only at the delivery of the food (unconditioned response, UR). After repeating the tone-food (CS-US) pairings a sufficient number of times, he observed that the dogs began to salivate before the food was delivered, at the exact time of the tone presentation (conditioned response, CR)(Figure 7). Pavlov published his work on classical conditioning in 1927 (Pavlov 1927). The same year, the already famous writer H.G. Wells wrote an essay about Pavlov for *The New York Times Magazine*. After reading Well's article, a 23-year-old B.F. Skinner discovered Pavlov and became his biggest fan. According to Skinner's autobiography, he used to carry an autographed picture of Pavlov around, and later grew to be one of History's most influential behavioral psychologist himself.

Operant conditioning

United States of America, 1898. One of the first to publish influential results on operant conditioning was the 34-year-old Edward Thorndike (1874-1949). During his PhD, he build an ingenious puzzle box from which a cat could only escape by operating latches. Even though he believed that cats cannot stand being confined and would try to escape the box for the very need

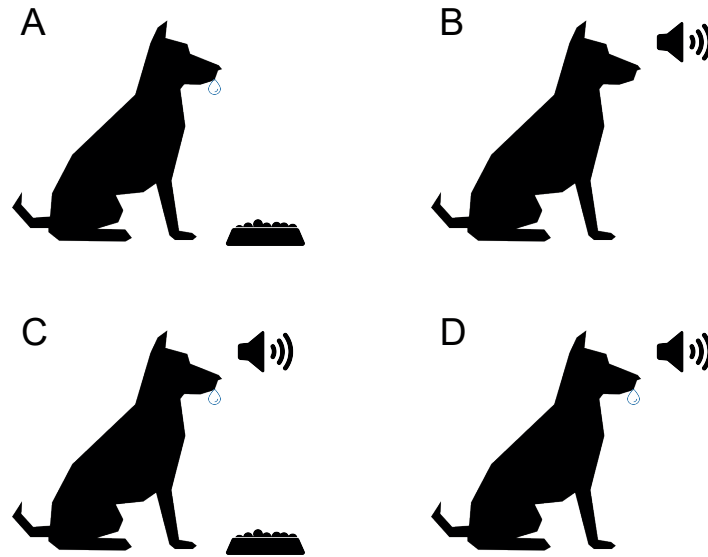


Figure 7. Classical conditioning: Pavlov's dog experiment. (A) Before learning, the dog displays an unconditioned response and salivates when the food is presented. (B) Before learning, the dog displays no conditioned response and does not salivate at the tone presentation. (C) During learning, both the food and the tone are presented concomitantly and the dog salivates. (D) After learning, the dog displays a conditioned response and salivates at the tone presentation, even with no food presented.

of being free, Thorndike placed a food platter outside the box to increase the cat's motivation, so that the cat could only get the food if it escaped from the box. This is the fundamental difference between Pavlov's and Thorndike's experimental setups: in classical conditioning, a reward is delivered regardless of the animal's behavior, whereas in operant conditioning, the reward's delivery depends on a behavioral action. When the cat was trapped in the puzzled box for the first time, there was no evidence for insight or cleverness, and the successful actions appeared to occur by chance. But after several times of being trapped in the box, the cat could resolve the puzzle faster and faster (Figure 8). This decrease of latency to escaping the puzzle box and getting the food occurred by trial-and-error: if an action brings a reward, Thorndike believed that this action becomes stamped into the mind. In other words, behavior changes because of its consequences. Thorndike called this the *Law of Effect*: of several responses made to the same situation, those which bring satisfaction will be more firmly connected with the situation; those which bring discomfort will have their connections with the situation weakened. Thorndike later noted that the greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond (Thorndike 1911).

Decades later, in the same country, Burrhus F. Skinner (1904-1990) discovers Pavlov's work on classical conditioning and becomes a pioneer of modern behaviorism. Building on Thorndike's

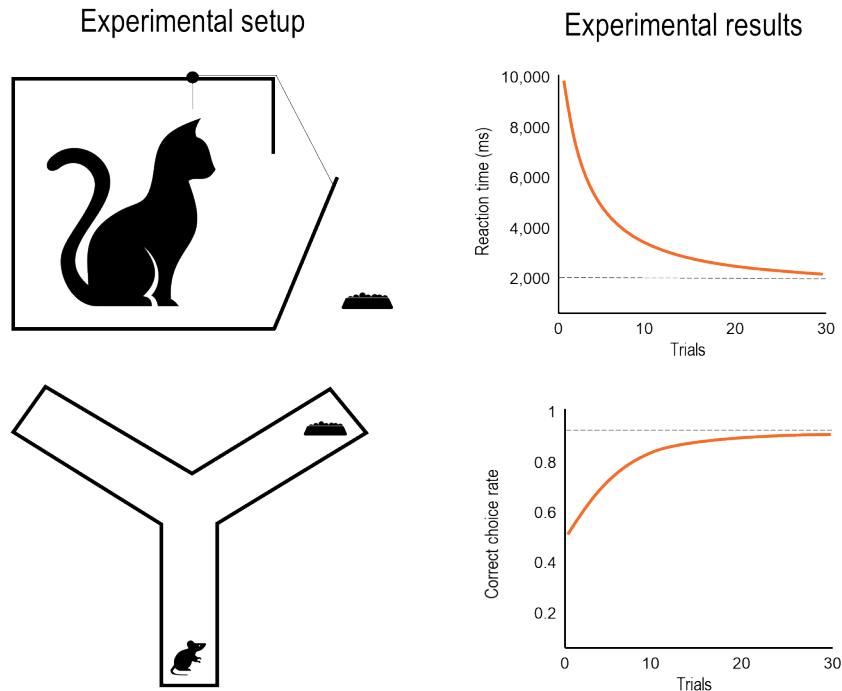


Figure 8. Operant conditioning. Schematic representations of typical learning curves of a cat escaping a puzzle box (top) and a mouse in a Y-maze (bottom) obtained from classical instrumental conditioning experimental setups. By trial-and-error, the cat learns to escape the box faster and faster; the mouse learns to choose the most rewarding arm more often.

work, his major contribution to operant conditioning was the invention of an operant conditioning chamber, aka the Skinner box. The box was composed of an electrified grid, a food dispenser, a speaker and a cue light; there were 2 levers inside the box. Using this setup, the experimenter can investigate classical (speaker, lights) as well as operant (levers) conditioning with different species, usually rodents. The structure of the Skinner box allows to study different types of learning (Figure 9):

- positive reinforcement. The rodent is in the box, presses the lever, receives food = increase of specific behavior by adding reward
- negative reinforcement. The rodent is in the box and receives electric shocks, presses the lever, the shocks disappear = increase of specific behavior by deleting punishment
- positive punishment. The rodent is in box, presses the lever, receives an electric shock = decrease of specific behavior by adding punishment
- negative punishment. The rodent is in the box and receives food, presses the lever, the food disappears = decrease of specific behavior by deleting reward

Using this operant conditioning chamber to strengthen behavior, he considered the probability of response to be the most effective measure of response strength (Skinner 1938). This led to the dominance of response rate as the dependent variable of operant learning.

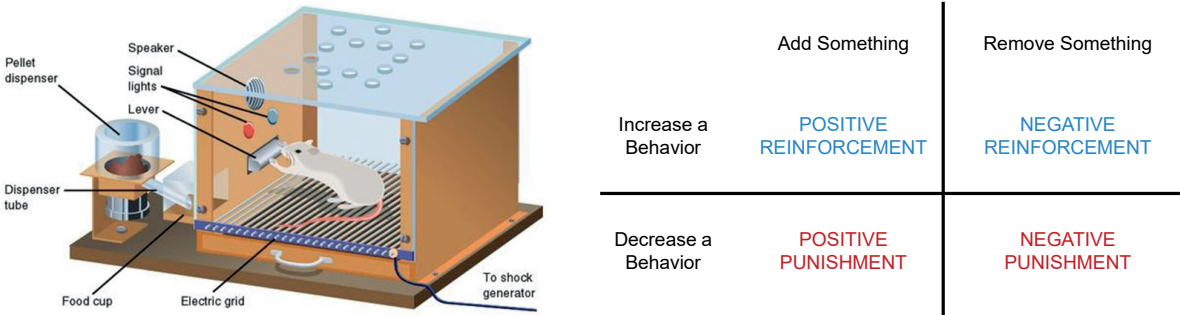


Figure 9. Operant conditioning: Skinner’s box. Operant chambers have at least one operandum (or "manipulandum"), often two or more, that can automatically detect the occurrence of a behavioral response or action. Typical operanda for primates and rats are response levers; if the animal presses the lever, the opposite end moves and closes a switch that is monitored by a computer or other programmed device. Typical operanda for pigeons and other birds are response keys with a switch that closes if the bird pecks at the key with sufficient force. The other minimal requirement of a conditioning chamber is that it has a means of delivering a primary reinforcer (a reward, such as food, etc) or unconditioned stimulus like food (usually pellets) or water. It can also register the delivery of a conditioned reinforcer, such as an LED signal as a "token" (Jackson and Hackenberg 1996).

The fundamental work of Thorndike and Skinner tells us that different rates of reinforcement imply different rates of responses. This notion is called the *matching law*: it has been observed in behavioral learning that animals tend to match their response rate to the earned reinforcement rates. For example, if two response alternatives A and B are offered to an animal, the ratio of response rates to A and B equals the ratio of reinforcements yielded by each response:

$$\frac{Resp_A}{Resp_A + Resp_B} = \frac{Reinf_A}{Reinf_A + Reinf_B} \tag{1.7}$$

The matching law was first formulated by 31-year-old Richard J. Herrnstein (1930-1994) following an experiment with pigeons on concurrent variable interval schedules. Pigeons were presented with two buttons in a Skinner box, each of which led to varying rates of food reward. The pigeons tended to peck the button that yielded the greater food reward more often than the other button, and the ratio of their rates to the two buttons matched the ratio of their rates of reward on the two buttons (Herrnstein 1961). The experiment was performed on a small group of 3 pigeons, but allowed to expose relative response rates, and with them, to detect hints of

relative learning in animals. Starting from this influential result, some economic models now assume that the primary determinant of choice behavior is the relative value of rewards, such as normalization models as mentioned in section 1.1.1 and equation 1.2.

1.2.2 Human reinforcement learning

In the same reasoning as Thorndike aiming at increasing the cat's motivation by adding food to the equation, one might ask if different types of rewards can lead to different behaviors. Investigating reinforcement learning in humans might require other rewards than food pellets, because human motivations might differ from animal motivations.

In general, motivation is defined as the process that initiates, guides, and maintains goal-oriented behaviors. In the framework of a short-term human reinforcement learning study, rewards usually come as primary rewards (e.g., food or erotic outcomes) or secondary rewards (e.g., monetary outcomes). In a meta-analysis published in 2013, Sescousse and colleagues showed that those three rewards robustly engaged a common brain network, although with some variations in the intensity and location of peak activity. The observation of money-specific responses in different areas supported the idea that abstract secondary rewards are represented in evolutionary more recent brain regions. Their results indicate that the computation of experienced reward value does not only recruit a core "reward system" but also reward type-dependent brain structures (Sescousse et al. 2013). Therefore, a reward is actually a composite or complex process containing several psychological components that correspond to distinguishable neurobiological mechanisms. The major components of reward and their subdivisions include:

- *liking*: the actual pleasure component or hedonic impact of a reward
- *wanting*: motivation for reward, which makes the animal approach reward and avoid punishment
- *learning*: associations, representations, and predictions about future rewards based on past experiences, as described above

These different aspects are mediated by partly dissociable brain substrates. Within each reward component, there are further subdivisions and levels, including both conscious and non-conscious processing (Berridge and Kringelbach 2008). The challenge in the liking aspect is that it is very difficult to access such subjective "pleasure" states in experimental work, particularly in animals. In humans, one can simply ask participants to verbally report or rate their subjective pleasure (O'Doherty 2014). However, results about brain reward systems derived from animal studies

versus human studies typically produce conclusions that are similar and complementary, at least for mechanisms of core pleasure reactions ([Berridge and Kringelbach 2008](#)).

1.2.3 Neural reinforcement learning

During the last two decades, neuroscientific research has provided robust findings about the way reinforcement learning processes are implemented in the human brain. I will now briefly describe some neural pathways involved in the decision-making process. This section will be kept short, because the different projects of my PhD are mainly behavioral and computational. Among the four different types of neuromodulators involved in the process of decision-making in animals, namely acetylcholine, norepinephrine, serotonin and dopamine, the latter is believed to modulate reinforcement learning processes. Dopamine is a monoamine neurotransmitter, a term that refers to its chemical structure and the fact that it is derived from an amino acid. To synthesize dopamine, the amino acid tyrosine is converted to L-DOPA, then L-DOPA is decarboxylated to form dopamine. There are several areas of the brain where dopamine neurons are concentrated. The largest are the substantia nigra and ventral tegmental area in the midbrain. Other areas include the hypothalamus, olfactory bulb, and retina. There are several major dopamine pathways that carry dopamine from these areas of concentration to other parts of the brain ([Kandel et al. 2000](#)). Some of the largest are:

- the mesostriatal or nigrostriatal pathway, which stretches from the substantia nigra to the striatum
- the mesolimbic pathway, which stretches from the ventral tegmental area to the nucleus accumbens and other limbic structures
- the mesocortical pathway, which stretches from the ventral tegmental area throughout the cerebral cortex.

The function of dopamine will vary depending on the neural pathway. In the nigrostriatal pathway, more dopamine leads to more movement and less dopamine leads to less movement. Thus, more movement can be observed in chorea such as in Huntington's disease, tics such as in Tourette syndrome or OCD, or athetosis that can be seen in cerebral palsy. Less movement can be observed in Parkinson's disease, or side effects of antipsychotics. Another dopaminergic function in the mesolimbic and mesocortical pathways is to modulate the mood or the reward. An increase of dopamine can correlate with euphoria, psychosis, hallucinations, schizophrenia. This pathway is involved in both classical and operant conditioning described in

previous sections. The intake of drugs such as cocaine or methamphetamine lead to an increase of dopamine, hence heightened mood, and this is why the behavior is reinforced. On the other hand, dopamine shortage correlate with anhedonia, lack of pleasure, and therapeutics effects of antipsychotics (Nestler et al. 2009, Ikemoto 2010).

The fundamental role of dopamine in reinforcement learning was identified by Schultz, Dayan, and Montague, in 1997, in a key paper published in *Science*. Using electrophysiological recordings in primates during a classical conditioning task, they showed that midbrain dopaminergic neurons encoded the difference between the reward that is obtained and the reward that is expected. In the task, Schultz and colleagues delivered some juice (reward, R) to a monkey after the presentation of a tone (conditioned stimulus, CS). At first, the activity of dopaminergic neurons increased after the delivery of the juice. After learning however, not only did the phasic dopaminergic activity occur after the tone presentation instead of after the reward delivery, but they also observed a decrease of dopaminergic activity when the reward was omitted (Figure 10). Therefore, for the first time, there was neural evidence for a signed prediction error: an unpredicted reward generates phasic dopaminergic activity (positive prediction error), a fully predicted reward generates no phasic response (no prediction error), and the omission of a predicted reward generates a dip in the tonic dopaminergic activity (negative prediction error)(Figure 10, Schultz et al. 1997). This fundamental discovery was, back then, even more striking and impactfull since it perfectly fitted a mathematical formulation developed many years earlier by Rescorla and Wagner (Rescorla and Wagner 1972), which is the starting point of the next section on computational reinforcement learning.

Evidence of the prediction error representation from non-human primate electrophysiology was strengthened by functional magnetic resonance imaging (fMRI) data in humans. For example, O’Doherty and colleagues scanned human participants while performing classical and instrumental tasks and found neural correlates of the reward prediction error in the ventral striatum, a subcortical region that receives a lot of projections of dopaminergic neurons (O’Doherty et al. 2004). In humans, assessing directly midbrain areas such as the VTA is a challenge because it is a small and deep region. However, using high-resolution fMRI, D’Ardenne and colleagues were able in 2008 to retrieve blood-oxygen-level-dependent (BOLD) signal and showed that the VTA reflects the positive reward prediction error (D’Ardenne et al. 2008).

Once the involved brain regions have been identified, another approach to investigate the link between dopaminergic neurons and reward prediction error is to directly modify the firing of these neurons and look at the behavior. If the activity of dopaminergic neurons is altered,

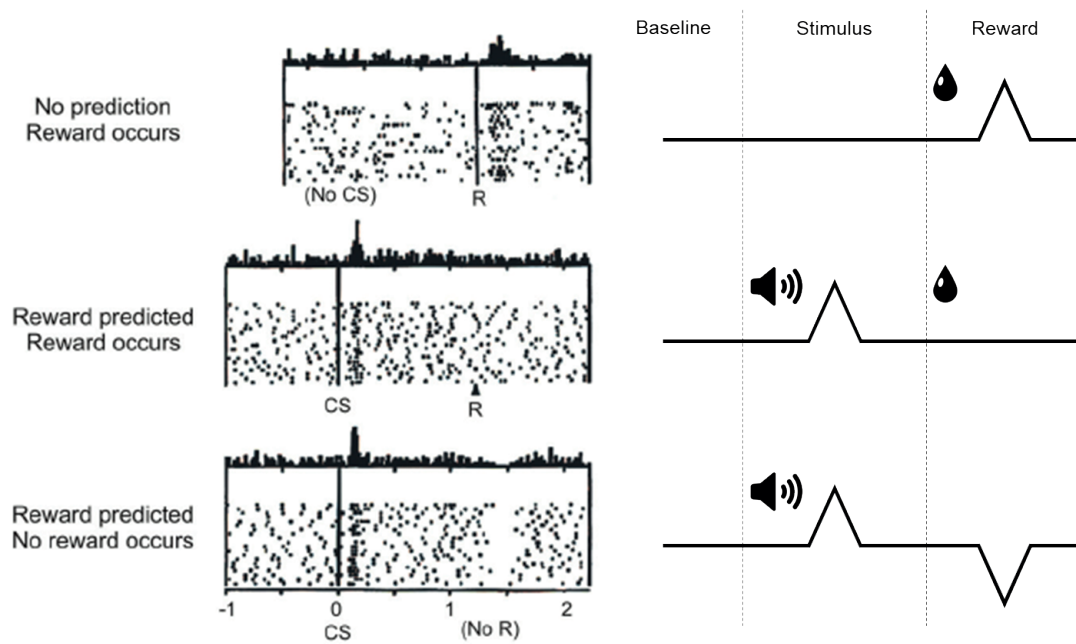


Figure 10. Dopamine as a reward prediction error signal. Temporal raster plots representing individual responses of dopaminergic neurons in different phases of a classical conditioning procedure and cumulative activity. Dopaminergic neurons deviated from their phasic activity. **Top:** before learning, dopaminergic neurons augmented their activity when the (unpredicted) reward occurred. **Middle:** after learning, dopaminergic neurons augmented their activity when the tone occurred, and not when the (predicted) reward occurred. **Bottom:** after learning, dopaminergic neurons reduced their activity when the predicted reward was omitted. Left panels reproduced from [Schultz et al. 1997](#). Each dot represents one neuron firing. CS: conditioned stimulus, R: reward

what are the consequences at the behavioral level? The best way of having access to altered dopaminergic neurons in humans is to study reinforcement learning tasks in patients who have specific lesions in the involved brain areas ([Vaidya et al. 2019](#)). The lesions can be mechanistic, such as brain injuries, or the consequence of a pathology, such as Parkinson's disease. In a neuropsychological study published in 2004, Frank and colleagues showed evidence for causal implications of dopamine modulation in human reinforcement learning. They administrated an instrumental learning task to a cohort of Parkinson's disease patients medicated ("ON") or unmedicated with levodopa ("OFF"), a precursor of dopamine, used as treatment in Parkinson's disease. The results showed that the patients OFF medication were impaired in learning from positive outcomes, whereas patients ON medication were impaired in learning from negative outcomes ([Frank et al. 2004](#)). This fundamental result is consistent with the idea that conditioning is driven by dopaminergic prediction errors. To conclude, the study of patients with neuropsychological pathologies or brain lesions allows us to draw conclusions on brain mechanisms via

dysfunctional behavior.

Finally, another approach to investigate the role of dopamine in human reinforcement learning is to study the behavioral effects of dopamine-modulating drugs. To this aim, in a pharmacological study published in *Nature* in 2006, Pessiglione and colleagues administered different types of treatment to different groups of healthy volunteers. The treatment was either a dopamine enhancer (levodopa), a dopamine blocker (haloperidol), or a placebo. The results showed that the reward prediction error is correlated with ventral striatum activity, and that the dopamine treatments modified the amplitude of these signals: levodopa amplified prediction errors correlates and haloperidol blunted them. Moreover, these medications affected learning performances according to their neural effects, suggesting a causal role of dopamine modulation in human reinforcement learning.

1.3 Computational reinforcement learning

The idea that we learn by interacting with our environment is probably the first to occur to us when we think about the nature of learning. Whether a toddler is finding out how to walk, a teenager is going to school, or an adult is choosing the best way to drive to work, all of these example situations require an agent (or *learner*) interacting with its environment by taking actions and receiving feedback. In our daily life, we are all acutely aware of how our environment responds to what we do, and we seek to influence what happens through our behavior. Learning from interaction is a fundamental idea underlying nearly all theories of learning and intelligence. As such, reinforcement learning is an area which takes its origin from machine learning and is concerned with how these agents take actions in an environment in order to maximize the notion of cumulative reward (Sutton and Barto 1998). The reinforcement signal that the agent receives is a numerical reward, which encodes the success of an action's outcome, and the agent seeks to learn to select actions that maximize the accumulated reward over time. Several academic disciplines have contributed to reinforcement learning models and most notably optimal control (Bellman 1958) and experimental psychology of conditioning (Rescorla 1988).

Reinforcement learning is one of three basic machine learning paradigms, alongside supervised learning and unsupervised learning, each paradigm being differentiated on the basis of how the learner is supposed to interact with the environment (Alpaydm 2004, Dayan and Abbott 2005).

Supervised learning is the machine learning paradigm in which a supervisor provides the learner with examples of correct behavior. The learner infers a function from *labeled* training data

consisting of a set of training examples (Russell et al. 2010, Mohri et al. 2018). Each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations.

Unsupervised learning refers to the machine learning task of finding hidden structures and patterns in *unlabeled* data. Since the data has not been labeled, classified or categorized, instead of responding to feedback, the agent identifies commonalities in the data and reacts based on the presence or absence of such commonalities in each new piece of data. This approach helps detect anomalous data points that do not fit into either group. This distinguishes unsupervised learning from supervised learning and reinforcement learning.

Reinforcement learning differs from supervised learning in not needing labeled input/output pairs to be presented, and in not needing sub-optimal actions to be explicitly corrected. Instead the focus is on finding a balance between exploration and exploitation (Kaelbling et al. 1996). To obtain a lot of reward, a reinforcement learning agent must prefer actions that it has tried in the past and found to be effective in producing reward. But to discover such actions, it has to try actions that it has not selected before. The agent has to exploit what it already knows in order to obtain reward, but it also has to explore in order to make better action selections in the future. The dilemma is that neither exploitation nor exploration can be pursued exclusively without failing at the task. The agent must try a variety of actions and progressively favor those that appear to be best. On a stochastic task, each action must be tried many times to reliably estimate its expected reward. The exploration/exploitation dilemma has been intensively studied by mathematicians for many decades (Ghemawat and Costa 1993, Benner and Tushman 2003, Cohen et al. 2007, Wilson et al. 2014), but won't be our main focus in the next sections.

In the reinforcement learning framework, the environment is usually outlined as a finite Markov decision process (Howard 1960, Wiering and Otterlo 2012). The agent and environment interact at each discrete time steps of a time sequence (Werbos 1992, Bertsekas and Tsitsiklis 1996). At each time step t , the agent has some representation of the environment's state s and takes an action a available in the current state s . At the next time step, as a consequence of its action a , the agent receives a numerical reward r and moves into a new state s' . Both the reward r and the probability of moving to the new state s' depend on the action a , taken in state s (Figure 11, Sutton and Barto 1998).

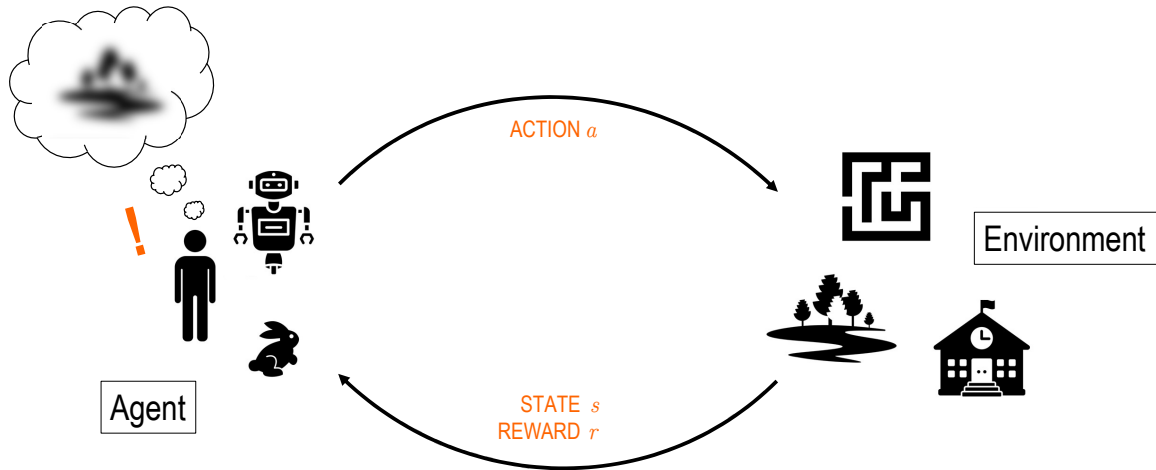


Figure 11. The reinforcement learning framework. In the standard loop architecture, at each discrete time step, the agent perceives the environment’s state s and performs an action a . The environment evolves to a new state and the agent receives a reward r . The exclamation point represents the update performed by the agent when receiving information from the environment. Figure adapted from [Sutton and Barto 1998](#)

Therefore, the basic reinforcement learning framework is defined by:

- an action space A : all the possible moves that the agent can make
- a state space S : the current situation returned by the environment
- a transition probability P : for all action $a \in A$ and state $s \in S$, $P(s, a, s')$ gives the probability to evolve into state s' given the agent performed action a in state s
- a reward function R : for all action $a \in A$ and state $s \in S$, $R(s, a, s')$ gives the immediate return after transition from s to s' with action a .

The agent’s goal is to build an optimal policy, given the probability to perform action a in state s , that maximizes cumulative reward. As I briefly mentioned in the previous section, responses of dopaminergic neurons as well as their target projections (e.g., in the ventral striatum and medial prefrontal cortex) align with the prediction error signal derived from several already existing models ([Schultz 1998](#), [Hollerman and Schultz 1998](#), [O’Doherty et al. 2003](#), [Eshel et al. 2015](#)). I will now describe some of these models, while keeping in mind that this is not an exhaustive enumeration of reinforcement learning algorithms, but mostly models that have most frequently been adapted to experimental neuroscience research and particularly in our experimental studies ([Daw and Doya 2006](#), [McClure and D’Ardenne 2009](#)).

1.3.1 Rescorla-Wagner model

Now that we have seen classical and operant conditioning as well as some reinforcement learning descriptions, we can return to Pavlov’s original experiments and we might have some questions. We might think, what if the dog was looking at Pavlov before the food was presented? Why would the dog not salivate to Pavlov’s presence? Or based on what we know about operant conditioning, one might wonder, what if the dog wagged his tail just before the food was presented? Wouldn’t this serve as reinforcement for this behavior of wagging the tail and so the dog would now wag his tail all the time? The way that the contingency model of classical conditioning tries to answer this is by focusing on the food as a reliable predictor. So when asked the question of why is the bell causing the salivation and why did the dog not salivate to other stimuli that were also in the room, one idea is that the food being presented was reliably predicted by the bell. In other words, during the conditioning process, the bell was always followed by the food, it became the *reliable predictor*, whereas other stimuli that were also in the room were not as reliable. The second reason why the bell is going to be more likely to be associated with the food is that it is *salient*, meaning that it captures the attention. Building on this idea that the association is learned with the surprise of the reward, Rescorla and Wagner developed the well know Rescorla-Wagner model of classical conditioning (Rescorla and Wagner 1972). The model aims at measuring the changed conditioned properties of stimuli from one trial to the next. On a learning trial in which two stimuli A and B are followed by an US, according to the Rescorla-Wagner model, the rules for change in associative strength of A and B are:

$$\Delta V_A = \alpha_A \cdot \beta(\lambda - \bar{V}) \quad (1.8)$$

$$\Delta V_B = \alpha_B \cdot \beta(\lambda - \bar{V})$$

$$\bar{V} = V_A + V_B \quad (1.9)$$

where V_A is the strength of the gradient due to prior learning, ΔV_A represents the changed conditioned properties of stimulus A, \bar{V} is the sum of the gradient strengths of all stimuli present (it is assumed that learning to a given stimulus is influenced by the associative strength of all stimuli present), $\alpha_A \in [0, 1]$ is the salience (or associative value) of stimulus A, $\beta \in [0, 1]$ is the intensity (or significance) of the US (learning rate parameter determined by the vigor of the goal response) and $\lambda \in [0, 1]$ is the magnitude of the goal event. Therefore, the difference $(\lambda - \bar{V})$ represents the upper limit of the associability of the US. In other words, on any given trial the current global associative strength \bar{V} is compared with λ and the difference is treated like an error to be corrected. This happens by producing a change in associative strength ΔV accordingly: this is an error-correction model.

The Rescorla-Wagner model is a very influential model to explain behavior in humans and other animals in conditioning tasks (Miller et al. 1995, Siegel and Allan 1996, Bouton 2007). The widespread influence of this model stems from its capacity to explain behavioral features in a simple manner. Among the successfully explained behavioral features, we can mention for instance the blocking effect, where an association between two stimuli is impaired if, during the conditioning process, the conditioned stimulus is presented together with a second conditioned stimulus that has already been associated with the unconditioned stimulus (Kamin 1967). However, reinforcement learning as formulated above consists of a trial-by-trial update, not sensitive to temporal blocks within learning. It is thus agnostic to possible higher-order structures of the environment in which learning occurs, which can be a limitation of this model. Another limitation lies in the inability of the model to handle within-trial temporal effects such as Inter Stimulus Interval effects where the temporal delay between stimuli affects the associative strength (Davis 1970, Buonomano et al. 2009), or primacy effects where the first items of a sequence are better remembered (Healy et al. 2000). More specifically, although it explains a large collection of behavioral data, the Rescorla-Wagner model does not take into account second-order conditioning (i.e., associating the first conditioned stimulus with a second stimulus) and assumes that a conditioning trial is a discrete temporal object (Niv and Schoenbaum 2008). The Rescorla-Wagner model is a special case of a larger class of the reinforcement learning models, where the rewards are delayed in time so the agent has to predict the total cumulative but discounted reward. A variant of this model uses an algorithm known as temporal difference (TD) learning (Schultz et al. 1997, Sutton and Barto 1990).

1.3.2 Temporal Difference learning

The term Temporal Difference (TD) learning algorithm was first used by Richard S. Sutton back in 1988 and has been extensively developed by Sutton and Barto (Sutton 1988, Sutton and Barto 1990, 1998) as an extension of the Rescorla-Wagner model in the sense that the core learning rule is an error-correction rule. If we picture an agent driving a car, in the Rescorla-Wagner model, the agent would have to wait for the car to crash multiple times to learn not to crash the car, which can be a long and painful process. In TD-learning, concurrently to the agent needing information at each turn about how not to crash the car, the model will make updates at every step and will be able to use it to solve both continuous and episodic tasks. Studying the results of Inter Stimulus Interval in eye-blink conditioning in rabbits, Sutton and Barto formulated TD-learning as follows. Let us consider an agent traveling to a sequence of states and actions

during T time steps. Let R_t be the discounted sum of all the rewards in the current state:

$$R_t = r_{t+1} + \gamma r_{t+2} + \dots + \gamma^{T-1} r_T \quad (1.10)$$

where r_{t+1} is the immediate reward, and $\gamma \in [0, 1]$ is the discount factor, powered to allow more significance to more recent rewards and discount more heavily in the future. If the value of the current state $V(s_t)$ depends on the complete return, i.e., the cumulative future reward R_t expected from this state s_t , then we can estimate the value of a state with an error-correction term:

$$V(s_t) \leftarrow V(s_t) + \alpha \cdot (R_t - V(s_t)) \quad (1.11)$$

where $\alpha \in [0, 1]$ is a learning rate parameter to adjust how much of that error will be updated. If $\alpha = 0$, the agent does not learn anything at all. If $\alpha = 1$, the agent drastically only considers the most recent information. Of note, a higher learning rate does not necessarily mean better learning or higher performance (Buduma and Locascio 2017). Similarly to the Rescorla-Wagner model, the term $R_t - V(s_t)$ is a reward prediction error, i.e., the difference between the complete return and the predicted one. Moreover, if $V(s_t)$ correctly predicts the complete return R_t , the reward prediction error (and therefore, the update) will be zero, meaning that the algorithm has found the final value for V . In the TD(0) algorithm described by Sutton and Barto, instead of using the accumulated sum of discounted rewards R_t , we only look at the immediate reward r_{t+1} , plus the discount of the estimated value of only one time step ahead $V(s_{t+1})$:

$$V(s_t) \leftarrow V(s_t) + \alpha \cdot (r_{t+1} + \gamma V(s_{t+1}) - V(s_t)) \quad (1.12)$$

The TD(0) has a higher bias than the previous equation because it's making estimates from estimates instead of estimates from seeing an entire sequence. Yet, this tends to have lower variance. This is especially useful for very long sequences, or for continuous tasks, since the algorithm does not need to wait for the entire sequence to be over before calculating the returns and update the value. We can finally define the TD-error δ as:

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \quad (1.13)$$

which is again a reward prediction error. The error signal can be used to reinforce actions leading to better states of the environment (in terms of future predicted rewards) and punish those leading to worse states (Niv and Schoenbaum 2008). Without considering any other previously visited states, we assign a new state value to one state by performing:

$$V(s_t) \leftarrow V(s_t) + \alpha \cdot \delta_t \quad (1.14)$$

This procedure has been shown to converge to an accurate value function, providing a solution to the prediction problem. The core idea of TD-learning is that we adjust predictions to match other, more accurate, predictions about the future. The TD-learning formalism was then extended to incorporate action learning and can be used almost unaltered to address the estimation of state-action-value instead of state-value, i.e., the expected return when choosing a given action in a given context. In comparison to these formulations, such as SARSA (State Action Reward State Action) algorithm or Q-learning algorithm (Watkins and Dayan 1992), the TD-learning algorithm directly updates the reward value of states, rather than state-action pairs, based on discrete periods of time between the CS and US. Therefore, it is more commonly used in Pavlovian conditioning experiments (Sutton and Barto 1998, O’Doherty et al. 2003). Moreover, the TD-learning prediction error can be modified to cope with instrumental-conditioning scenarios such as actor–critic and advantage learning models (O’Doherty et al. 2004).

1.3.3 Q-learning

Q-learning is a reinforcement learning algorithm that seeks to find the best action to take given the current state. It was first introduced by Chris Watkins in 1989 (Watkins 1989), during his PhD that he labeled *Learning from delayed rewards*. The "Q" in Q-learning stands for *quality*, which in this case represents how useful a given action is in gaining some future reward. Q-learning was then further developed until convergence proof was presented by Watkins and Dayan in 1992 (Watkins and Dayan 1992). The goal of Q-learning is to find the optimal policy by learning the optimal Q values for each state-action pair, stored in a Q-matrix $Q(s, a)$. Before learning begins, the Q-matrix Q is initialized to a fixed value (chosen by the experimenter). Then, at each time step t , the agent chooses an action a_t , receives a reward r_t , enters a new state s_{t+1} . The difference with TD-learning is that the transition from old state s_t to new state s_{t+1} now depends on both the previous state s_t and the selected action a_t , whereas in TD-learning the action was left unspecified. Although, the same rule applies to approximate Q , using the weighted average of the previous value and the new information:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \cdot \left(r_t + \gamma \cdot \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right) \quad (1.15)$$

where r_t is the reward received when moving from state s_t to state s_{t+1} , α is the learning rate, γ is the discount factor, and $\max_a Q(s_{t+1}, a)$ is the maximum reward that can be obtained from state s_{t+1} . If the algorithm is able to sample all the available actions in all states a sufficient number of times, it will find the optimal value function. Considering the example of Thorndike’s cat, the animal’s choices were reinforced towards operating latches, because the value of this

action became higher compared to door scratching or meowing.

1.3.4 Action selection

Q-learning is a reinforcement learning algorithm that provides a way to obtain accurate estimates of action values, and therefore the policy can only be based on action-values estimates. However, accurate estimation of action-values depends on the possibility of sampling sufficiently all the available actions. Therefore, the decision (or action selection) rule should include the possibility to explore all the options, all the while trying to maximize the cumulative reward by choosing the actions with the highest estimated value, an issue that I already mentioned in the introduction of this section as the exploration/exploitation trade-off. In the reinforcement learning framework, how does the agent select the action?

An agent interacts with the environment in several ways. One of them is to use the Q-matrix as a reference and view all possible actions for a given state. The agent then selects the action based on the maximum value of those actions. This is known as *exploiting* since we use the information we have available to us to make a decision. Another way to take action is to act randomly. This is called *exploring* (randomly). Instead of selecting actions based on the maximum future reward, the agent selects an action at random. Acting randomly is important because it allows the agent to explore and discover new states that otherwise may not be selected during the exploitation process. Among the different action selection rules known in reinforcement learning models, the most widely used in Q-learning are:

- *greedy* policy: the agent always chooses the action associated with the highest expected reward, i.e., $\max_a Q(s, a)$. While this policy might sometimes be referred to as "optimal policy" because it exploits the current knowledge to maximize immediate rewards, it can be disadvantageous in dynamically changing, probabilistic environments, because it gives no space for exploration choices, and therefore can lead to inaccurate action value estimations.
- ϵ -*greedy* policy: the agent balances exploitation and exploration by adding a random component in the action selection rule. For some $0 \leq \epsilon \leq 1$, the agent chooses the action associated with the highest expected reward with probability $(1 - \epsilon)$, and otherwise, the agent chooses an action at random with probability ϵ . Note that for $\epsilon = 0$, the agent only exploits (greedy policy) and for $\epsilon = 1$, the agent only explores (full random). The value of ϵ can even be reduced over time, thus shifting the emphasis from exploration to exploitation. However, one limitation of the ϵ -greedy policy is that when it explores the non-optimal actions, it chooses equally between all alternatives, meaning that it is equally

likely to choose the worst possible action than the second-best action, which might be unsatisfactory.

- *softmax* rule: the agent chooses an action with some probability based on the actions' relative expected reward. The exponential function is applied to each action value, and the values are then normalized by dividing by the sum of all the exponential, ensuring that the sum of the components is 1. When actions do not differ in their value estimates, choices are equiprobable. Thus, the probability of choosing option a in state s is given by:

$$P(s, a) = \frac{e^{\beta Q(s, a)}}{\sum_{a'} e^{\beta Q(s, a')}} \quad (1.16)$$

where $\beta > 0$, is called the *inverse temperature* (because it comes from statistical thermodynamics) and determines the steepness of the softmax S-shaped curve (Figure 12). Low inverse temperatures ($\beta \rightarrow 0$) cause the actions to be all (nearly) equiprobable. High inverse temperatures ($\beta \rightarrow +\infty$) cause a greater difference in selection probability for actions that differ in their value estimates, i.e., the probability to choose the action associated with the highest expected reward tends to 1, and the rule tends to a greedy policy. The term "softmax" comes from Bridle in 1990, but the formula appears to have first been proposed by Luce in 1959, as the action selection rule respects the IIA axiom (see Section 1.1.1, Luce 1959, Bridle 1990).

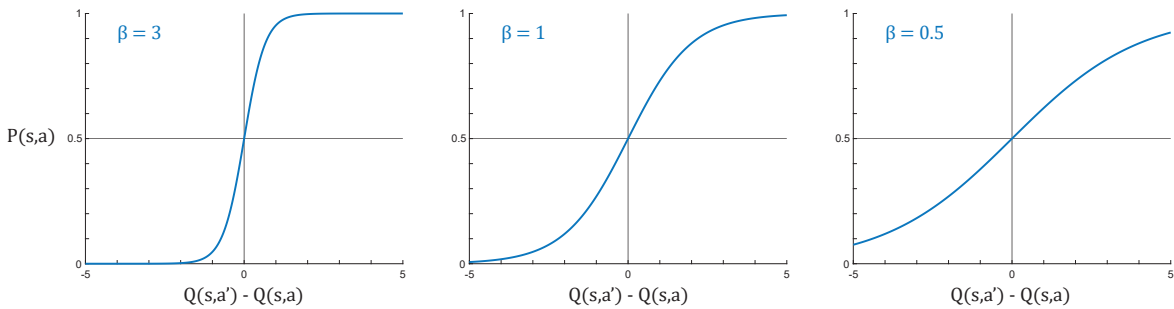


Figure 12. Effects of different inverse temperatures on choice probability in the softmax function. The example is shown for a binary choice: in this case, the softmax depends on the difference between the values of the two actions a and a' . The probability of selecting action a , $P(s, a)$, is 0.5 when the actions have equal values; it increases as their difference gets bigger and decreases as it gets smaller. High inverse temperatures result in an abrupt sigmoid function, while low inverse temperatures result in a softer S-shaped curve. For $\beta = 0$, $P(s, a) = 0.5$, independently of the action values.

Whether softmax action selection or ε -greedy action selection is better is unclear and may depend on the task and on human factors, since both methods have only one parameter that must be

set. Yet, in all models developed during my PhD, I only used classical or modified versions of the softmax selection rule.

1.3.5 General method

The goal of computational modeling in behavioral science is to use precise mathematical models to make better sense of behavioral data. In the case of my PhD, the behavioral data comes in the form of choices, but can also be reaction times, eye movements, or other easily observable behaviors, and even neural data. As described in this section, models come in the form of mathematical equations that link the experimentally observable variables (e.g., stimuli, outcomes, past experiences) to behavior in the immediate future. In this sense, computational models instantiate different algorithmic hypotheses about how behavior is generated. Keeping in mind the famous aphorism "*All models are wrong, but some are useful*" (Box 1976), a sufficient amount of data can often prove that a model is not "true". By the same reasoning, if a model is made considerably complex to fit a specific data set, it won't be applicable to any other data set (this is referred to as *overfitting*); if the model is too general, it won't be able to explain the whole variability of any data set. Therefore, cognitive modeling must rely not only on a comparison between various models, but also on absolute falsification criteria (Palminteri et al. 2017b). In practice, the general method used in this PhD includes (but is not limited to) *parameter estimation*, *model comparison*, and *model falsification*. Other approaches, such as Bayesian hierarchical modeling (Allenby et al. 2005) or Bayesian inference (Bishop 2006), are based on the same hypotheses but might use different methods, and won't be extensively discussed here.

Parameter estimation

Model parameters can characterize a variety of scientifically interesting quantities, from how quickly participants can learn (Behrens et al. 2007) to how sensitive they are to different rewards and punishments (Tom et al. 2007). Each model M has a set of free parameters θ , which can be of various sizes, and which will be optimized for each participant separately. For example, the Q-learning model described in section 1.3.3 has a set of 2 free parameters $\theta = (\alpha, \beta)$, namely the learning rate and the inverse temperature. For simplicity, the next equation should be put in a context of a binary choice between left L and right R options. Considering that each option is chosen via a probabilistic softmax selection rule as in equation 1.16, the probability of a whole data set D of T time steps (i.e., a whole sequence of choices $a = a_1, \dots, a_T$ given the rewards $r = r_1, \dots, r_T$) is the product of the choices' probabilities:

$$\prod_t P(a_t = L \mid Q_t(L), Q_t(R)) \quad (1.17)$$

Note that the terms Q_t in the softmax are determined by the actions a_1, \dots, a_{t-1} and rewards r_1, \dots, r_{t-1} on the $t - 1$ trials prior to t . This product constitutes the *likelihood* function $P(D \mid \theta^*, M)$, which represents the probability of obtaining the data set D with model M and set of parameters θ^* . We can estimate the optimal free parameters $\theta^* = (\alpha, \beta)$ by maximum likelihood (Figure 13). In practice, to avoid computation of very high numbers, the logarithm is applied to the product to transform it into a sum, easier to estimate. In this case, we refer to it as the *log-likelihood*:

$$\log P(D \mid \theta^*, M) = \sum_t \log P(a_t = L \mid Q_t(L), Q_t(R)) \quad (1.18)$$

Thus, from the experimental data observed for each participant, we obtain a value for the log-likelihood, associated with a set of parameters, which are the best-fitting parameters. The maximum log-likelihood was obtained with this set of parameters, meaning that the highest probability of observing this data set was given by this set of parameters.

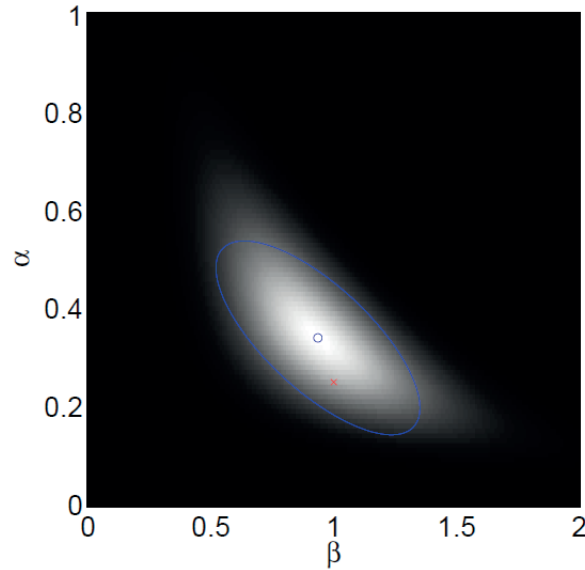


Figure 13. Maximum likelihood estimation. Likelihood surface for simulated reinforcement learning data, as a function of two free parameters. Lighter colors denote higher data likelihood. The maximum likelihood estimate is shown as the blue "o" surrounded by an ellipse of one standard error (a region of about 90% confidence); the true parameters from which the data were generated are denoted by the red "x". Figure from [Daw 2011](#)

Model comparison

To compare the "goodness" of two different models, one might simply compare their respective log-likelihood, since it is a measure of how well the model fits the data. However, this comparison would be biased towards overfitting, mentioned earlier: a model more complex, i.e., with more parameters, would win over a simpler model, even if it not generalizable to other data sets. One approach to model selection is to pick the candidate model with the highest probability given the data, regardless of the set of parameters. To determine how well a model fits the data, one might consider its *posterior probability*, which is the conditional probability of the model M being "true" after observing some data set D , $P(M | D)$. According to Bayes' theorem, or Bayes' rule (Bayes and Price 1763):

$$P(M | D) = P(D | M) \cdot \frac{P(M)}{P(D)} \quad (1.19)$$

where:

- $P(M | D)$ is the posterior, the degree of belief in M after having accounted for data D
- $P(M)$ is the prior, the initial degree of belief in M
- $\frac{P(D | M)}{P(D)}$ is the support that the data D provides for model M

We know that D is fixed and we wish to consider the impact of D having been observed on our belief in M . Therefore, $P(D)$ is also fixed and we can write:

$$P(M | D) \propto P(D | M) \cdot P(M) \quad (1.20)$$

The key quantity, $P(D | M)$ is called the *model evidence* and represents the probability of model M generating data D . Importantly, this expression does not make reference to any particular parameter settings, since in asking how well a model predicts data, we are not given any particular parameters. This is why the likelihood examined above, $P(D | M, \theta)$, is inflated by the number of free parameters: in asking how well a model predicts a data set, it is a fallacy, having seen the data, to retrospectively choose the parameters that would have best fit it. This overstates the ability of the model to predict the data set. Comparing models according to $P(D | M)$, instead, avoids overfitting. In the literature, model evidence is also referred to as evidence, marginal likelihood, or integrated likelihood. This is because, to evaluate the model evidence, one might integrate it over all the possible sets of parameters of model M :

$$P(D | M) = \int_{\theta} P(D, \theta | M) d\theta \quad (1.21)$$

Then, the quantity is evaluated using Laplace approximation ([Laplace 1820](#)), which combines Taylor expansion and the Gaussian integral. The logarithm of the conditional probability $\log P(D, \theta | M)$ is approximated via a multivariable case of Taylor expansion to the second order around the maximum, i.e., the optimal set of parameters, θ^* , of size d . Then the formulation of the Gaussian integral is used to approximate the exponential. To see the demonstration in details, see [Appendix A](#).

$$\log P(D | M) \approx \log P(D | \theta^*, M) - \frac{d}{2} \log n \quad (1.22)$$

This final formulation, an approximation of the log model evidence, is known as the Bayesian Information Criterion (BIC, [Schwarz 1978](#)) and is widely used in model comparison and model selection. Although it is based on a lot of assumptions, and therefore is to be used with caution, we note that the BIC depends on $\log P(D | \theta^*, M)$, which is exactly the log likelihood obtained when optimizing the free parameters. The BIC also has the advantage of avoiding overfitting because it is penalized for the number of parameters d ([Bishop 2006](#), [Claeskens and Hjort 2007](#), [Bhat and Kumar 2010](#), [Daw 2011](#)).

Model simulation

Although models can be compared using a quantitative measure as explained above, it is important to evaluate the qualitative performance of the models as well. Not only can the models' performance be qualitatively compared, it gives an objective measure of the "goodness" and the generalizability of the models. [Palminteri, Wyart, and Koechlin](#) have argued in 2017 that the simulation of candidate models is necessary to falsification and therefore to supporting the specific claims about cognitive function made by the vast majority of cognitive modeling studies. Indeed, the ability of a candidate model to generate a behavioral effect of interest is rarely assessed, although it is an absolute falsification criterion ([Palminteri et al. 2017b](#)).

In practice, using the set of parameters obtained for each participant after the parameter estimation phase, we can create new, simulated data, as if the algorithm was performing the task for each participant. Comparing the simulated variables (in my case, the choice patterns) to the collected behavioral data, allows us to confirm or infirm the tested hypotheses about participants' strategies. Model simulations can be "one-step ahead" predictions, i.e., the probability of choice given the actual participant's history of past choices and outcomes, or "generative"

simulations, i.e., playing the task ex-novo. If the task design allows it, one can also optimize the parameters on half of the trials, and perform model simulations on the other half, with the parameters estimated on the first half. This approach, known as *cross-validation*, has the advantage of testing the model’s generalizability, because its predictions are validated out-of-sample. For examples of model falsification, see Figure 14.

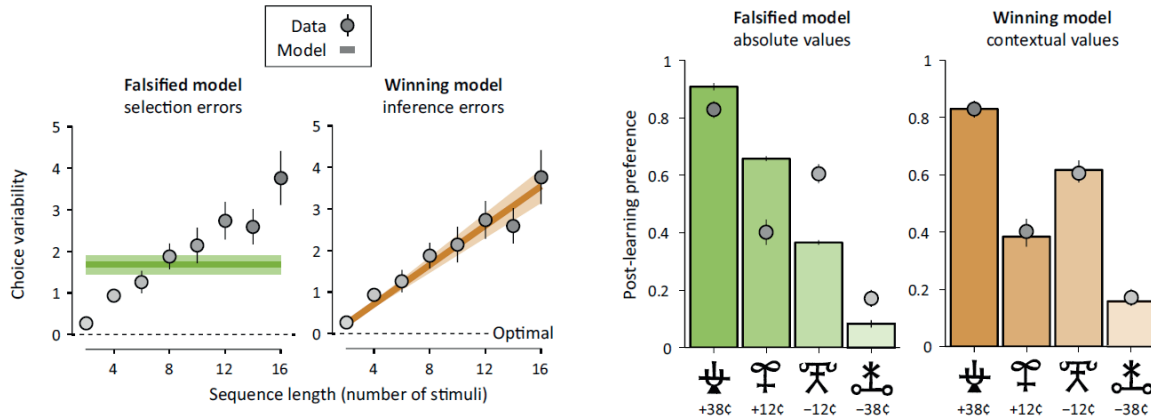


Figure 14. Concrete examples of model falsification. Left panels: observed (grey dots) and model simulated (colored lines) choice variability in a probabilistic inference task as a function of the sequence length. Right panels: observed (grey dots) and model simulated (colored bars) post-learning preference as a function of the stimulus value. Figure adapted from Palminteri et al. 2017b.

1.4 Research questions

In the previous sections, I gave some examples of context-dependence in economic behaviors and decision making, I described reinforcement learning behaviors in animals and humans, and I outlined some reinforcement learning models widely used in value-based decision making. Throughout the work carried out during this PhD, I developed existing models further to study context-dependence in human reinforcement learning. In the first part of this manuscript, I will present my work on modeling context-dependence in healthy volunteers; this includes two first-author papers published in 2018 and in 2021. In the second part, I will present ongoing work on context-dependence in impaired individuals; this includes a meta-analysis focusing on clinical papers and a large-scale experiment using a transnosographic approach. Additional results, including four experiments performed by patients with Huntington’s Disease, Parkinson’s Disease, Major Depressive Disorder, and brain lesions, are presented in Appendix B.

1.4.1 Context-dependence in the general population

Reference dependence can be defined as the evaluation of outcomes as gains or losses relative to a temporal or spatial reference point, such as the context, and is one of the fundamental

principles of prospect theory and behavioral economics (Kahneman and Tversky 1979, Köszegi and Rabin 2006). Yet, only recently have theoretical and experimental studies in animal and human investigated this reference dependence in reinforcement learning (Palminteri et al. 2015, Klein et al. 2017, Rigoli et al. 2018). These studies have notably revealed that reference dependence can significantly improve learning performances in contexts of negative valence (i.e., loss avoidance), but at the cost of generating post-learning inconsistent preferences. In a paper published in 2015, Palminteri and colleagues show that, based on the principle behind two-factor theory, successful avoidance is reframed as a positive outcome, because it is computed relative to the value of its choice context (Kim et al. 2006, Palminteri et al. 2015).

In addition to this valence reference dependence, evidence suggests that our sensitivity to sensory stimuli or monetary amounts is not the same across different ranges of magnitude (Bernoulli 1738, Fechner 1860), which is in line with the description of neuronal range adaptation described in section 1.1.4 (Carandini and Heeger 2011). In the reinforcement learning framework, the notion of context is embodied in the notion of state. Therefore, behavioral and neural manifestations of context-dependence could be achieved by (or reframed as) state-dependent processes.

In the first publication, we hypothesized that in human reinforcement learning, the trial-by-trial learning of option and action values is concurrently affected by reference-point centering and range adaptation. To test this hypothesis and investigate the computational basis of such state-dependent learning, we adapted a validated reinforcement learning paradigm (Palminteri et al. 2015, 2016) to include orthogonal manipulations of outcome valence and outcome magnitude.

If range adaptation is an automatic consequence of how the brain adapts its response to the distributions of the available outcomes, factors that facilitate the identification of these distributions should make it more pronounced. This would translate into a bigger difference between the objective option values and their corresponding subjective values, which is a counter-intuitive prediction in the context of reinforcement learning. Indeed, this is in striking contrast with the intuition embedded in virtually all learning algorithms, that making a learning problem easier (by facilitating the identification of the outcome distributions) should lead to more accurate and objective internal representations.

In the second study, we aim at testing this hypothesis, while concomitantly gaining a better understanding of range adaptation at the computational level. Using an online-based experiment with a similar task on a large sample of healthy participants, we varied this paradigm in eight different versions where we manipulated the task difficulty in complementary ways.

1.4.2 Context-dependence in neuropsychiatric diseases

Computational psychiatry aims at describing the relationship between the brain's neurobiology, its environment, and mental symptoms, in computational terms. Through this approach, it may improve psychiatric classification and the diagnosis and treatment of mental illness, as well as unite many levels of description in a mechanistic and rigorous fashion, while avoiding biological reductionism and artificial categorisation (Montague et al. 2012, Adams et al. 2015).

One current issue of the Diagnostic and Statistical manual of Mental disorders (DSM-5) classification is the purely categorical diagnoses, as it seems that the current categories are not valid at the clinical (Van Os et al. 1999) or genetic (Cross-Disorder Group of the Psychiatric Genomics Consortium 2013) levels. A more dimensional system would not classify a person with psychosis as just one of "schizophrenic" or "bipolar" or "schizoaffective", but might instead score them on scales of manic and depressive mood symptoms, positive and negative psychotic symptoms, and cognitive impairment (Adams et al. 2015). Likewise, the category "schizophrenic" includes individuals with very heterogeneous profiles, which hardens the task of coming with a reliable treatment for a patient. Moreover, different categories are very correlated: individuals with depressive disorders tend to be also anxious, and vice versa (American Psychiatric Association 2013).

Computational psychiatry can accommodate and inform both categorical and dimensional approaches, each driven by data. For example, one might find that depressed participants and healthy controls differ dimensionally on a certain parameter derived from a certain computational model (Kumar et al. 2008). Alternatively, one might find evidence that different models are used by distinct groups (i.e., different possible categories) to perform the same task. For example, patients with schizophrenia with high or low negative symptoms (Gold et al. 2012), or those with remitted psychosis and controls (Moutoussis et al. 2011). More generally, computational psychiatry allows to assess the evidence for competing theories formally, for instance using Bayesian model comparison as described in section 1.3.5. Identifying computational categories and dimensions in this way ought to improve both psychiatric nosology (Brodersen et al. 2011) and the targeting and monitoring of treatments.

In the first project, we performed a meta-analysis on (categorical) clinical studies investigating the difference between the performance when seeking rewards and avoiding punishments. We refer to this effect as the *valence bias*. Based on two fundamental publications (Frank et al. 2004, Pessiglione et al. 2006), we screened around 2500 papers and found around 120 publications that

take account of the effect, to explore the valence bias among the different categories.

In the second project, we used a more dimensional approach to assess the valence bias in regards to inter-individual differences. Using a task design combining the key features of the tasks used in [Frank et al. 2004](#) and [Pessiglione et al. 2006](#), we applied a factor analysis to a large dataset from an online-based experiment ([Gillan and Daw 2016](#)), such as the analyses performed in a paper by Gillan and colleagues, published in *Elife* in 2016. Similarly to [Gillan et al. 2016](#), we were able to bring out dimensional factors, but our data did not allow us to correlate dimensions with the parameters of a reinforcement model developed in the team ([Maia and Frank 2011](#), [Palminteri et al. 2015, 2017a](#)).

Chapter 2

The paradoxical consequences of context-dependence in human reinforcement learning

2.1 Study 1: Bavard, Lebreton et al, 2018

2.1.1 Introduction

The aim of this study was to implement a model encoding two crucial features of context-dependent valuation, reference point dependence and range adaptation, in a reinforcement learning task manipulating outcome valence and outcome magnitude. Over two experiments, results show that context-dependent valuation emerges progressively over the task time. Our data show that, while being locally adaptive (for instance in negative valence and small magnitude contexts), context-dependent valuation comes at the cost of seemingly irrational choices, when options are extrapolated out from their original contexts.



2.1.2 Article

ARTICLE

DOI: 10.1038/s41467-018-06781-2

OPEN

Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences

Sophie Bavard ^{1,2,3}, Maël Lebreton^{4,5,6}, Mehdi Khamassi^{7,8}, Giorgio Coricelli^{9,10} & Stefano Palminteri ^{1,2,3}

In economics and perceptual decision-making contextual effects are well documented, where decision weights are adjusted as a function of the distribution of stimuli. Yet, in reinforcement learning literature whether and how contextual information pertaining to decision states is integrated in learning algorithms has received comparably little attention. Here, we investigate reinforcement learning behavior and its computational substrates in a task where we orthogonally manipulate outcome valence and magnitude, resulting in systematic variations in state-values. Model comparison indicates that subjects' behavior is best accounted for by an algorithm which includes both reference point-dependence and range-adaptation—two crucial features of state-dependent valuation. In addition, we find that state-dependent outcome valuation progressively emerges, is favored by increasing outcome information and correlated with explicit understanding of the task structure. Finally, our data clearly show that, while being locally adaptive (for instance in negative valence and small magnitude contexts), state-dependent valuation comes at the cost of seemingly irrational choices, when options are extrapolated out from their original contexts.

¹Laboratoire de Neurosciences Cognitives Computationnelles, Institut National de la Santé et Recherche Médicale, 29 rue d'Ulm, 75005 Paris, France.

²Département d'Etudes Cognitives, Ecole Normale Supérieure, Paris 75005, France. ³Institut d'Etudes de la Cognition, Université de Paris Sciences et Lettres, Paris 75005, France. ⁴CREED lab, Amsterdam School of Economics, Faculty of Business and Economics, University of Amsterdam, Roetersstraat 11, Amsterdam 1018 WB, The Netherlands. ⁵Amsterdam Brain and Cognition, University of Amsterdam, Amsterdam 1018 WB, The Netherlands. ⁶Swiss Centre for Affective Sciences, University of Geneva, 24 rue du Général-Dufour, Geneva 1205, Switzerland. ⁷Institut des Systèmes Intelligents et Robotiques, Centre National de la Recherche Scientifique, 4 place Jussieu, 75005 Paris, France. ⁸Institut des Sciences de l'Information et de leurs Interactions, Sorbonne Universités, 3 rue Michel-Ange, Paris 75794, France. ⁹Department of Economics, University of Southern California, Los Angeles, CA 90007, USA. ¹⁰Centro Mente e Cervello, Università di Trento, corso Bettini 21, Rovereto 38068, Italy. These authors contributed equally: Sophie Bavard, Maël Lebreton. Correspondence and requests for materials should be addressed to S.P. (email: stefano.palminteri@ens.fr)

In everyday life, our decision-making abilities are solicited in situations that range from the most mundane (choosing how to dress, what to eat, or which road to take to avoid traffic jams) to the most consequential (deciding to get engaged, or to give up on a long-lasting costly project). In other words, our actions and decisions result in outcomes, which can dramatically differ in terms of affective valence (positive vs. negative) and intensity (small vs. big magnitude). These two features of the outcome value are captured by different psychological concepts— affect vs. salience—and by different behavioral and physiological manifestations (approach/avoidance vs. arousal/energization levels)^{1–3}.

In ecological environments, where new options and actions are episodically made available to a decision-maker, both the valence and magnitude associated with the newly available option and action outcomes have to be learnt from experience. The reinforcement-learning (RL) theory offers simple computational solutions, where the expected value (product of valence and magnitude) is learnt by trial-and-error, thanks to an updating mechanism based on prediction error correction^{4,5}. RL algorithms have been extensively used during the past couple of decades in the field of cognitive neuroscience, because they parsimoniously account for behavioral results, neuronal activities in both human and non-human primates, and psychiatric symptoms induced by neuromodulatory dysfunction^{6–10}.

However, this simple RL model is unsuited to be used as is in ecological contexts^{11,12}. Rather, similarly to the perceptual and economic decision-making domains, growing evidence suggests that reinforcement learning behavior is sensitive to contextual effects^{13–16}. This is particularly striking in loss-avoidance contexts, where an avoided-loss (objectively an affectively neural event) can become a relative reward if the decision-maker has frequently experienced losses in the considered environment. In that case, the decision-maker's knowledge about the reward distribution in the recent history or at a specific location, affects her perception of the valence of outcomes. Reference-dependence, i.e., the evaluation of outcomes as gains or losses relative to a temporal or spatial reference point (context), is one of the fundamental principles of prospect theory and behavioral economics¹⁷. Yet, only recently have theoretical and experimental studies in animal and human investigated this reference-dependence in RL^{18–20}. These studies have notably revealed that reference-dependence can significantly improve learning performances in contexts of negative valence (loss-avoidance), but at the cost of generating post-learning inconsistent preferences^{18,19}.

In addition to this valence reference-dependence, another important contextual effect that may be incorporated in ecological RL algorithms is range adaptation. At the behavioral level, it has long been known that our sensitivity to sensory stimuli or monetary amounts is not the same across different ranges of intensity/magnitude^{21,22}. These findings have recently paralleled with the description of neuronal range adaptation: in short, the need to provide efficient coding of information in various ranges of situations entails that the firing rate of neuron adapts to the distributional properties of the variable being encoded²³. Converging pieces of evidence have recently confirmed neuronal range-adaptation in economic and perceptual decision-making, although its exact implementation remains debated^{24–27}.

Comparatively, the existence of behavioral and neural features of range-adaptation has been less explored in RL, where it could critically affect the coding of outcome magnitude. In the RL framework the notion of context, which is more prevalent in the economic or perception literatures, is embodied in the notion of state. In the RL framework the environment is defined as a collection of discrete states, where stimuli are encountered, decisions are made and outcomes are collected. Behavioral and neural

manifestations of context-dependence could therefore be achieved by (or reframed as) state-dependent processes.

Here, we hypothesized that in human RL, the trial-by-trial learning of option and action values is concurrently affected by reference-point centering and range adaptation. To test this hypothesis and investigate the computational basis of such state-dependent learning, we adapted a well-validated RL paradigm^{19,28}, to include orthogonal manipulations of outcome valence and outcome magnitude.

Over two experiments we found that human RL behavior is consistent with value-normalization, both in terms of state-based reference-dependence and range-adaptation. To better characterize this normalization process at the algorithmic level, we compared several RL algorithms, which differed in the extent and in the way they implement state-dependent valuation (reference-dependence and range adaptation). In particular, we contrasted models implementing full, partial or no value normalization²⁹. We also evaluated models implementing state-dependent valuation at the decision stage (as opposed to the outcome evaluation stage) and implementing marginally decreasing utility (as proposed by Bernoulli²²). Overall, the normalization process was found to be partial, to occur at the valuation level, to progressively arise during learning and to be correlated with explicit understanding of the task structure (environmental). Finally, while being optimal in an efficient coding perspective, this normalization leads to irrational preference when options are extrapolated out from their original learning context.

Results

Behavioral paradigm to challenge context-dependence. Healthy subjects performed two variants of a probabilistic instrumental learning task with monetary rewards and losses. In those two variants, participants saw at each trial a couple of abstract stimuli (options), which were probabilistically paired with good or bad outcomes, and had to select the one they believed would be most beneficial for their payoff. The options were always presented in fixed pairs, which defined stable choice contexts. These contexts were systematically manipulated, so as to implement a 2×2 factorial design across two qualities of the option outcomes: outcome valence (reward or loss) and outcome magnitude (big: 1€; or small: 10c). In all contexts, the two options were associated with different, stationary, outcome probabilities (75% or 25%). The 'favorable' and 'unfavorable' options differ in their net expected value. The favorable option in the reward and big magnitude context is paired with a reward of 1€ with probability 75%, while the unfavorable option only 25% of the time. Likewise, the favorable option in the loss and small magnitude context is paired with a loss of 10 cents with probability 25%, while the unfavorable option 75% of the time (Fig. 1). Subjects therefore had to learn to choose the options associated either with highest reward probability or those associated with lowest loss probability. After the last learning session, subjects performed a transfer test in which they were asked to indicate the option with the highest value, in choices involving all possible binary combinations—that is, including pairs of options that had never been associated during the task. Transfer test choices were not followed by feedback, to not interfere with subjects' final estimates of option values. In the second variant of the experiment, an additional factor was added to the design: the feedback information about the outcomes (partial or complete) was manipulated to make this variant a $2 \times 2 \times 2$ factorial design. In the partial context, participants were only provided with feedback about the option they chose, while in the complete context, feedback about the outcome of the non-chosen option was also provided.

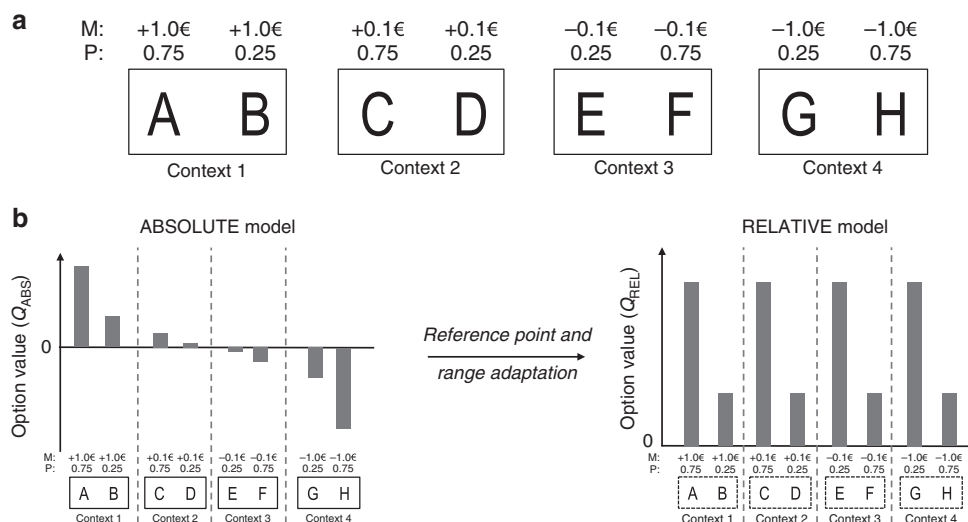


Fig. 1 Experimental design and normalization process. **a** Learning task with four different contexts: reward/big, reward/small, loss/small, and loss/big. Each symbol is associated with a probability (P) of gaining or losing an amount of money or magnitude (M). M varies as a function of the choice contexts (reward seeking: +1.0€ or +0.1€; loss avoidance: -1.0€ or -0.1€; small magnitude: +0.1€ or -0.1€; big magnitude: +1.0€ or -1.0€). **b** The graph schematizes the transition from absolute value encoding (where values are negative in the loss avoidance contexts and smaller in the small magnitude contexts) to relative value encoding (complete adaptation as in the RELATIVE model), where favorable and unfavorable options have similar values in all contexts, thanks to both reference-point and range adaptation

Table 1 Correct choice rate of the learning sessions as a function of task factors in Experiments 1, 2 and both experiments

	Experiment 1 ($N = 20$)		Experiment 2 ($N = 40$)		Both experiments ($N = 60$)	
	F-val	P-val	F-val	P-val	F-val	P-val
Val	0.002	0.969	0.285	0.597	0.167	0.684
Inf	-	-	7.443	0.0095**	-	-
Mag	4.872	0.0398*	4.267	0.0456*	9.091	0.00378**
Val × Inf	-	-	1.037	0.315	-	-
Val × Mag	4.011	0.0597	0.08	0.779	1.755	0.19
Inf × Mag	-	-	0.006	0.939	-	-
Val × Inf × Mag	-	-	0.347	0.559	-	-

** $P < 0.01$; * $P < 0.05$, t-test

Outcome magnitude moderately affects learning performance.

In order to characterize the learning behavior of participants in our tasks, we first simply analyzed the correct response rate in the learning sessions, i.e., choices directed toward the most favorable stimulus (i.e., associated with the highest expected reward or the lowest expected loss). In all contexts, this average correct response rate was higher than chance level 0.5, signaling significant instrumental learning effects ($T(59) = 16.6, P < 0.001$). We also investigated the effects of our main experimental manipulations (outcome valence (reward/loss), outcome magnitude (big/small), and feedback information (partial/complete, Experiment 2 only)) (Table 1). Because there was no significant effect of the experiment (i.e., when explicitly entered as factor ‘Experiment’: $F(59) = 0.96, P > 0.3$), we pooled the two experiments to assess the effects of common factors (outcome valence and magnitude). Replicating previous findings¹⁹, we found that the outcome valence did not affect learning performance ($F(59) = 0.167, P > 0.6$), and that feedback information significantly modulated learning in Experiment 2 ($F(39) = 7.4, P < 0.01$). Finally, we found that the outcome magnitude manipulation,

which is a novelty of the present experiments, had a significant effect on learning performance ($F(59) = 9.09, P < 0.004$); Post-hoc test confirmed that across both experiments subjects showed significantly higher correct choice rate in the big-magnitude compared with the small-magnitude contexts ($T(59) > 3.0, P < 0.004$), and similar correct choice rate in the reward compared to the losses contexts ($T(59) = 0.41, P > 0.13$).

Transfer test choices do not follow expected values.

Following the analytical strategy used in previous studies^{18,19}, we next turned to the results from the transfer test, and analyzed the pattern of correct choice rates, i.e., the proportion of choices directed toward the most favorable stimulus (i.e., associated with the highest expected reward or the lowest expected loss). Overall, the correct choice rate in the transfer was significantly higher than chance, thus providing evidence of significant value transfer and retrieval ($T(59) > 3.0, P < 0.004$). We also analyzed how our experimental factors (outcome valence (reward/loss), outcome magnitude (big/small) and option favorableness (i.e., being the symbol the most favorable of its pair during the learning sessions) influenced the choice rate per symbol. The choice rate per symbol is the average frequency with which a given symbol is chosen in the transfer test, and can therefore be taken as a measure of the subjective preference for a given option. Consistent with significant value transfer and retrieval, the ANOVA revealed significant effects of outcome valence ($F(59) = 76, P < 0.001$) and option correctness ($F(59) = 203.5, P < 0.001$) indicating that—in average—symbols associated with favorable outcomes were preferred compared to symbols associated with less favorable ones (Table 2). However, and in line with what we found in simpler contexts^{19,28}, the analysis of the transfer test revealed that option preference did not linearly follow the objective ranking based on their absolute expected value (probability(outcome) × magnitude (outcome)). For example, the favorable option of the reward/small context was chosen more often than the less favorable option of the reward/big context (0.71 ± 0.03 vs. 0.41 ± 0.04 ; $T(59) = 6.43, P < 0.0001$). Similarly, the favorable option of the loss/small magnitude context was chosen more often than the less favorable option of the reward/small context (0.42 ± 0.03 vs. 0.56

Table 2 Symbol choice rate of the transfer test as a function of task factors and option correctness in Experiments 1, 2 and both experiments

	Experiment 1 (N = 20)		Experiment 2 (N = 40)		Both experiments (N = 60)	
	F-val	P-val	F-val	P-val	F-val	P-val
Valence	33.42	1.43e-05***	43.78	7.23e-08***	76	3.38e-12***
Favorableness	57.66	3.6e-07***	149.5	6.46e-15***	203.5	<2e-16***
Magnitude	2.929	0.103	4.225	0.0466*	0.525	0.472
Val × Fav	4.039	0.0589	6.584	0.0142*	10.8	0.00171**
Val × Mag	11.68	0.00289**	3.565	0.0665	11.55	0.00122**
Fav × Mag	10.8	0.00388**	0.441	0.51	4.131	0.0466*
Val × Fav × Mag	8.241	0.00979**	1.529	0.224	7.159	0.00964**

***P < 0.001; *P < 0.05; **P < 0.01; t-test

± 0.03; $T(59) = 2.88$, $P < 0.006$). Crucially, while the latter value inversion reflects reference-point dependence, as shown in previous studies^{19,28}, the former effect is new and could be a signature of a more global range-adaptation process. To verify that these value inversions were not only observed at the aggregate level (i.e., were not an averaging artifact), we analyzed the transfer test choice rate for each possible comparison. Crucially, analysis of the pairwise choices confirm value inversion also for direct comparisons.

Delineating the computational hypothesis. Although these overall choice patterns appear puzzling at first sight—since they would be classified as “irrational” from the point of view of the classical economic theory based on absolute values³⁰—we previously reported that similar seemingly irrational behavior and inconsistent results could be coherently generated and explained by state-dependent RL models. To hypothesize this reasoning, we next turned to computational modeling to provide a parsimonious explanation of the present results.

To do so, we fitted the behavioral data with several variations of standard RL models (see Methods). The first model is a standard Q-learning algorithm, referred to as ABSOLUTE. The second model is a modified version of the Q-learning model that encodes outcomes in a state-dependent manner:

$$R_{REL,t} = \frac{R_{ABS,t}}{|V_t(s)|} + \max\left\{0, \frac{-V_t(s)}{|V_t(s)|}\right\} \quad (1)$$

where the state value $V(s)$ is initialized to 0, takes the value of the first non-zero (chosen or unchosen) outcome in each context s , and then remains stable over subsequent trials. The first term of the question implements range adaptation (divisive normalization) and the second term reference point-dependence (subtractive normalization). As a result, favorable/unfavorable outcomes are encoded in a binary scale, despite their absolute scale. We refer to this model as RELATIVE, while highlighting here that this model extends and generalizes the so-called “RELATIVE model” employed in a previous study, since the latter only incorporated a reference-point-dependence subtractive normalization term, and not a range adaptation divisive normalization term¹⁹.

The third model, referred to as HYBRID, encodes the reward as a weighted sum of an ABSOLUTE and a RELATIVE reward:

$$R_{HYB,t} = \omega * R_{REL,t} + (1 - \omega) * R_{ABS,t} \quad (2)$$

The weight parameter (ω) of the HYBRID model quantifies at the individual level the balance between absolute ($\omega = 0.0$) and relative value encoding ($\omega = 1.0$).

The fourth model, referred to as the UTILITY model, implements the economic notion of marginally decreasing subjective utility^{17,22}. Since our task included only two non-zero outcomes, we implemented the UTILITY model by scaling the big magnitude outcomes ($|1\epsilon|$) with a multiplicative factor ($0.1 < v < 1.0$).

Finally, the fifth model, referred to as the POLICY model, normalizes (range adaptation and reference point correction) values at the decision step (i.e., in the softmax), where the probability of choosing ‘a’ over ‘b’ is defined by

$$P_t(s, a) = \frac{1}{1 + e^{\left(\frac{Q_t(s,b) - Q_t(s,a)}{Q_t(s,b) + Q_t(s,a)}\right) \frac{1}{\beta}}} \quad (3)$$

Model comparison favors the HYBRID model. For each model, we estimated the optimal free parameters by likelihood maximization. The Bayesian Information Criterion (BIC) was then used to compare the goodness-of-fit and parsimony of the different models. We ran three different optimization and comparison procedures, for the different phases of the experiments: learning sessions only, transfer test only, and both tests. Thus we obtained a specific fit for each parameter and each model in the learning sessions, transfer test, and both.

Overall (i.e., across both experiments and experimental phases), we found that the HYBRID model significantly better accounted for the data compared to the RELATIVE, the ABSOLUTE, the POLICY, and the UTILITY models (HYB vs. ABS $T(59) = 6.35$, $P < 0.0001$; HYB vs. REL $T(59) = 6.07$, $P < 0.0001$; HYB vs. POL $T(59) = 6.79$, $P < 0.0001$; HYB vs. UTY $T(59) = 2.72$, $P < 0.01$). This result was robust across experiments and across experimental sessions (learning sessions vs. transfer test) (Table 3). In the main text we focus on discussing the ABSOLUTE and the RELATIVE models, which are nested within the HYBRID and therefore represent extreme cases (absent or complete) of value normalization. We refer to the Supplementary Methods for a detailed analysis of the properties of the POLICY and the UTILITY models (Supplementary Figure 1), and additional model comparison (Supplementary Table 1).

Model simulations falsify the ABSOLUTE and RELATIVE models. Although model comparison unambiguously favored the HYBRID model, we next aimed to falsify the alternative models, using simulations³¹. To do so, we compared the correct choice rate in the learning sessions to the model predictions of the three main models (ABSOLUTE, RELATIVE, and HYBRID). We generated for each model and for each trial t the probability of choosing the most favorable option, given the subjects’ history of choices and outcomes, using the individual best-fitting sets of

Table 3 BICs as a function of the dataset used for parameter optimization (Learning sessions, Transfer test or Both) and the computational model

	Experiment 1 (N = 20)			Experiment 2 (N = 40)			Both experiments (N = 60)		
	Learning sessions (nt = 160)	Transfer test (nt = 112)	Both (nt = 272)	Learning sessions (nt = 160)	Transfer test (nt = 112)	Both (nt = 272)	Learning sessions (nt = 160)	Transfer test (nt = 112)	Both (nt = 272)
ABSOLUTE (df = 2/3)	179.8 ± 5.9	113.6 ± 5.7	295.1 ± 9.9	190.9 ± 5.9	126.9 ± 4.1	325.4 ± 6.5	187.2 ± 3.8	122.4 ± 3.4	315.3 ± 5.6
RELATIVE (df = 2/3)	193.3 ± 4.5	135.8 ± 5.1	329.6 ± 8.0	185.1 ± 5.6	121.1 ± 4.0	306.0 ± 7.3	187.9 ± 4.0	126.0 ± 3.3	313.9 ± 5.7
HYBRID (df = 3/4)	178.3 ± 6.0	109.3 ± 5.0	284.6 ± 9.1	181.5 ± 5.8	105.8 ± 4.1	290.5 ± 8.0	180.5 ± 4.3	106.9 ± 3.2	288.5 ± 6.1
POLICY (df = 2/3)	185.4 ± 6.9	123.7 ± 6.3	311.0 ± 12.2	190.1 ± 4.9	139.4 ± 3.9	334.6 ± 6.5	188.5 ± 3.9	134.2 ± 3.4	326.7 ± 6.0
UTILITY (df = 3/4)	173.9 ± 6.5	107.5 ± 6.3	282.2 ± 10.8	183.4 ± 5.6	123.1 ± 4.5	310.1 ± 7.1	180.2 ± 4.3	117.9 ± 3.8	300.8 ± 6.2

Nt, number of trials; df, degree of freedom

parameters. Concerning the learning sessions, we particularly focused on the magnitude effect (i.e., the difference in performance between big and small magnitude contexts). As expected, the ABSOLUTE model exacerbates the observed magnitude effect (simulations vs. data, $T(59) = 5.8$, $P < 0.001$). On the other side, the RELATIVE model underestimates the actual effect (simulations vs. data, $T(59) = 3.0$, $P < 0.004$). Finally (and unsurprisingly), the HYBRID model manages to accurately account for the observed magnitude effect ($T(59) = 0.93$, $P > 0.35$) (Fig. 2a, b). We subsequently compared the choice rate in the transfer test to the three models' predictions. Both the ABSOLUTE and the RELATIVE models failed to correctly predict choice preference in the transfer test (Fig. 2c). Crucially, both models failed to predict the choice rate of intermediate value options. The ABSOLUTE model predicted a quite linear option preference, predicting that the transfer test choice rate should be highly determined by the expected utility of the options. On the other side, the RELATIVE model's predictions of the transfer test option preferences were uniquely driven by the option context-dependent favorableness. Finally, choices predicted by the HYBRID model accurately captured the observed option preferences by predicting both an overall correlation between preferences and expected utility and the violation of the monotony of this relation concerning intermediate value options (Figs. 2d, 3). To summarize, and similarly to what was observed in previous studies^{18,19,29}, choices in both the learning and transfer test could not be explained by assuming that option values are encoded in an absolute manner, nor by assuming that they are encoded in a fully context-dependent manner, but are consistent with a partial context dependence. In the subsequent sections we analyze the factors that affect value contextualization both within and between subjects.

Relative value encoding emerges during learning. Overall we found that a weighted mixture of absolute and relative value encoding (the HYBRID model) better explained the data compared to the “extreme” ABSOLUTE or RELATIVE models. However, this model comparison integrates over all the trials, leaving open the possibility that, while on average subjects displayed no neat preference for either of the two extreme models, this result may arise from averaging over different phases in which one of the models could still be preferred. To test this hypothesis, we analyzed the trial-by-trial likelihood difference between the RELATIVE and the ABSOLUTE model. This quantity basically measures which model better predicts the data in a given trial: if positive, the RELATIVE model better explains the data, if negative, the ABSOLUTE model does. We submitted the trial-by-trial likelihood difference during a learning session to

a repeated measure ANOVA with ‘trial’ (1:80) as within-subject factor. This analysis showed a significant effect of trial indicating that the evidence for the RELATIVE and the ABSOLUTE model evolves over time ($F(79) = 6.2$, $P < 2e-16$). Post-hoc tests revealed two big clusters of trials with non-zero likelihood difference: a very early cluster (10 trials from the 4th to the 14th) and a very late one (17 trials from the 62nd to the 78th). To confirm this results, we averaged across likelihood difference in the first half (1:40 trials) and in the second half (41:80 trials). In the first half we found this differential to be significantly negative, indicating that the ABSOLUTE model better predicted subjects' behavior ($T(59) = 2.1$, $P = 0.036$). In contrast, in the second half we found this differential to be significantly positive, indicating that the RELATIVE model better predicted subjects' behavior ($T(59) = 2.1$, $P = 0.039$). Furthermore, a direct comparison between the two phases also revealed a significant difference ($T(59) = 3.9$, $P = 0.00005$) (Fig. 4a, b). Finally, consistent with a progressively increasing likelihood of the RELATIVE compared the ABSOLUTE model during the learning sessions, we found that the weight parameter (ω) of the HYBRID model obtained from the transfer test (0.50 ± 0.05) was numerically higher compared to that of the learning sessions (0.44 ± 0.05) (Table 4).

Counterfactual information favors relative value learning. The two experiments differed in that in the second one (Experiment 2) half of the trials were complete feedback trials. In complete feedback trials, subjects were presented with the outcomes of both the chosen and the forgone options. In line with the observation that information concerning the forgone outcome promotes state-dependent valuation both at the behavioral and neural levels^{18,32}, we tested whether or not the presence of such “counterfactual” feedbacks affects the balance between absolute and relative value learning. To do so, we compared the negative log-likelihood difference between the RELATIVE and the ABSOLUTE model separately for the two experiments. Note that since the two models have the same number of free parameters, they can be directly compared using the log-likelihood. In Experiment 2 (where 50% of the trials were “complete feedback” trials) we found this differential to be significantly positive, indicating that the RELATIVE model better fits the data ($T(39) = 2.5$, $P = 0.015$). In contrast, in Experiment 1 (where 0% of the trials were “complete feedback” trials), we found this differential to be significantly negative, indicating that the ABSOLUTE model better fits the data ($T(19) = 2.9$, $P = 0.001$). Furthermore, a direct comparison between the two experiments also revealed a significant difference ($T(58) = 3.9$, $P = 0.0002$) (Fig. 4c). Accordingly, we also found the weight parameter (ω) of

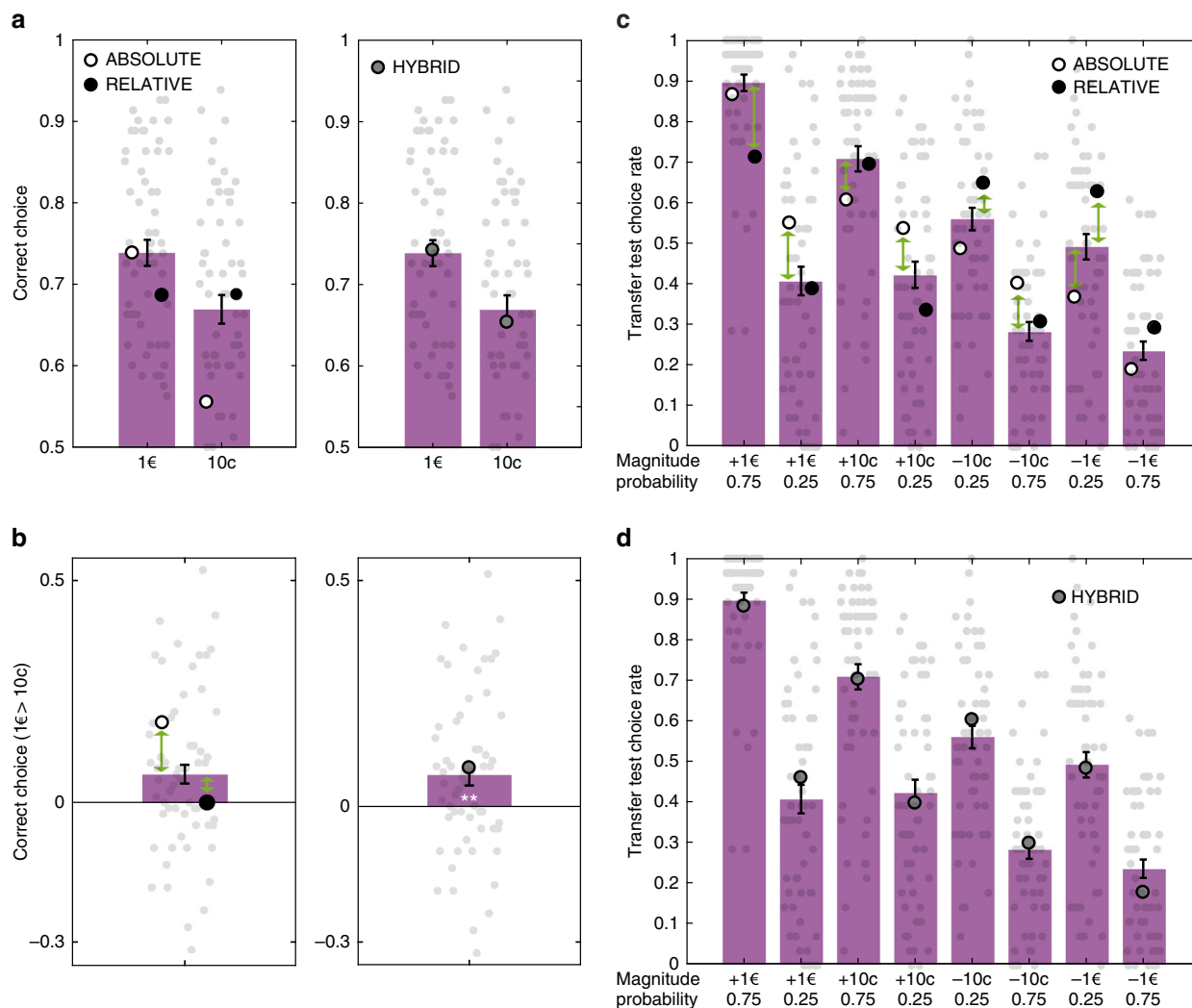


Fig. 2 Behavioral results and model simulations. **a** Correct choice rate during the learning sessions. **b** Big magnitude contexts' minus small magnitude contexts' correct choice rate during the learning sessions. **c** and **d** Choice rate in the transfer test. Colored bars represent the actual data. Big black (RELATIVE), white (ABSOLUTE), and gray (HYBRID) dots represent the model-predicted choice rate. Small light gray dots above and below the bars represent individual subjects ($N = 60$). White stars indicate significant difference compared to zero. Error bars represent s.e.m. $**P < 0.01$, t -test. Green arrows indicate significant differences between actual and predicted choices at $P < 0.001$, t -test

the HYBRID model to be significantly higher in Experiment 2 compared to Experiment 1 ($T(58) = 2.8$, $P = 0.007$) (Fig. 4d). Finally, consistently with reduced relative value learning, we found that the correct choice difference between the 1€ and the 0.1€ contexts in Experiment 1 (mean: $+0.10$; range: $-0.24/+0.51$) was 189.5% of that observed in Experiment 2 (mean: $+0.05$; range: $-0.32/+0.40$).

Explicit grasp of task structure links to relative valuation. In our learning protocol the fact that options were presented in fixed pairs (i.e., contexts) has to be discovered by subjects, because the information was not explicitly given in the instructions and the contexts were not visually cued. In between the learning and the transfer phases subjects were asked whether or not they believed that options were presented in fixed pairs and how many pairs there were (in the second session). Concerning the first question (“fixed pairs”), 71.7% of subjects responded correctly. Concerning the second question (“pairs number”), 50.0% of subjects responded correctly and the average number of pairs was 3.60 ± 0.13 , which significantly underestimated the true value (four: $T(59) = 3.0$, $P = 0.0035$). To test whether or not the explicit

knowledge of the subdivision of the learning task in discrete choice contexts was correlated with the propensity to learn relative values, we calculated the correlation between the number of correct responses in the debriefing (0, 1, or 2) and the weight parameter (ω) of the HYBRID model. We found a positive and significant correlation ($R^2 = 0.11$, $P = 0.009$) (direct comparison of the weight parameter (ω) between subjects with 0 vs. 2 correct responses in the debriefing: $T(37) = 2.8$, $P = 0.0087$) (Fig. 4e). To confirm this result, we ran the reciprocal analysis, by splitting subjects into two groups according to their weight parameter and we found that subjects with $\omega > 0.5$ had a significantly higher number of correct responses in the debriefing compared to subjects with $\omega < 0.5$ ($T(58) = 3.0$, $P = 0.0035$) (Fig. 4f).

Rational and irrational consequences of relative valuation. Previous behavioral analyses, as well as model comparison results, showed that a mixture of relative and absolute value learning (the HYBRID model) explained subjects' behavior. In particular, during the learning sessions, subjects displayed a correct choice difference between the 1€ and the 0.1€ contexts smaller than that predicted by the ABSOLUTE model. During the transfer test, the

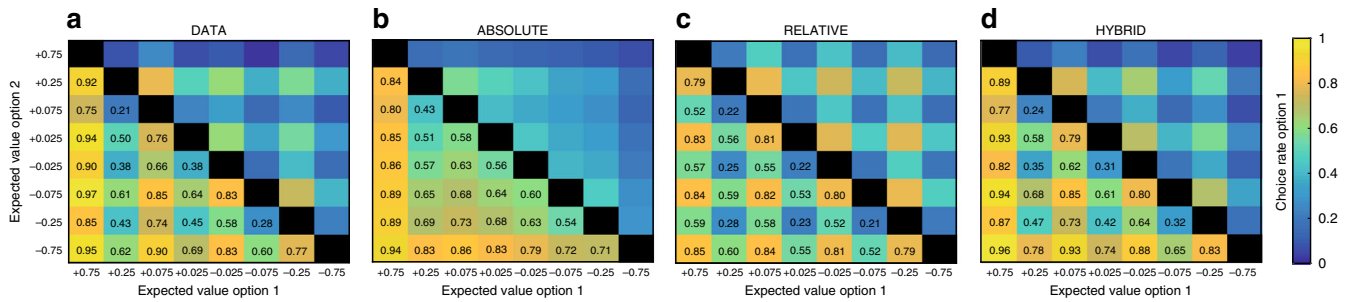


Fig. 3 Transfer test behavioral results and model simulations. Colored map of pairwise choice rates during the transfer test for each symbol when compared to each of the seven other symbols, noted here generically as ‘option 1’ and ‘option 2’. Comparisons between the same symbols are undefined (black squares). **a** Experimental data, **b** ABSOLUTE model, **c** RELATIVE model, and **d** HYBRID model

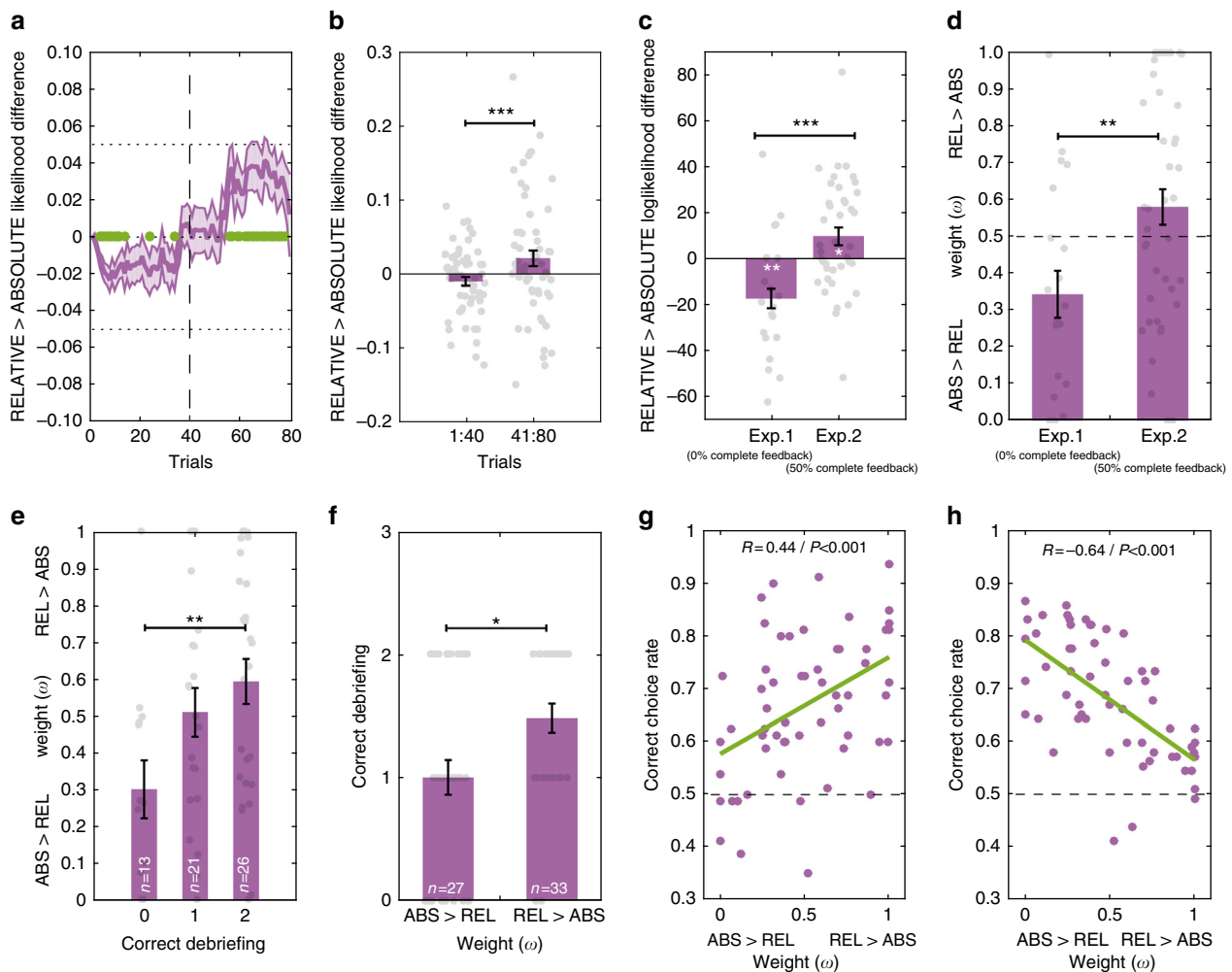


Fig. 4 Computational properties and behavioral correlates of value normalization. **a** Likelihood difference (from model fitting) between the RELATIVE and the ABSOLUTE models over the 80 trials of the task sessions for both experiments ($N = 60$). A negative likelihood difference means that the ABSOLUTE model is the best-fitting model for the trial and a positive likelihood difference means that the RELATIVE model is the best-fitting model for the trial. Green dots: likelihood difference significantly different from 0 ($P < 0.05$, t -test). **b** Likelihood difference between the RELATIVE and the ABSOLUTE models over the first part of the task (40 first trials) and the last part (40 last trials) for both experiments. **c** Likelihood difference between the RELATIVE and the ABSOLUTE models for the two experiments. A negative likelihood difference means that the ABSOLUTE model is the best-fitting model for the experiment and a positive likelihood difference means that the RELATIVE model is the best-fitting model for the experiment. **d** Subject-specific free parameter weight (ω) comparison for the two experiments. **e** Subject-specific free parameter weight (ω) as a function of correct debriefing for the two questions (“fixed pairs” and “number of pairs”). **f** Debriefing as a function of the weight parameter. Small light gray dots above and below the bars in **a–f** represent individual subjects ($N = 60$). **g** and **h** Correct choice rate as a function of subjects’ weight parameter in the learning sessions and the transfer test for both Experiments 1 and 2. One dot corresponds to one participant ($N = 60$); green lines represent the linear regression calculations. Error bars represent s.e.m. $***P < 0.001$, $**P < 0.01$, $*P < 0.05$, t -test

Table 4 Model parameters of the HYBRID model as a function of the dataset used for parameter optimization (learning sessions, transfer test or Both) and the computational model

	Experiment 1 (N = 20)			Experiment 2 (N = 40)			Both experiments (N = 60)		
	Learning sessions	Transfer test	Both	Learning sessions	Transfer test	Both	Learning sessions	Transfer test	Both
β	0.15 ± 0.04	0.12 ± 0.03	0.09 ± 0.02	0.30 ± 0.11	0.13 ± 0.04	0.17 ± 0.04	0.25 ± 0.08	0.13 ± 0.03	0.15 ± 0.03
α_F	0.25 ± 0.06	0.30 ± 0.08	0.14 ± 0.04	0.23 ± 0.04	0.34 ± 0.07	0.20 ± 0.04	0.24 ± 0.04	0.33 ± 0.05	0.18 ± 0.03
α_C	—	—	—	0.16 ± 0.04	0.25 ± 0.05	0.16 ± 0.03	—	—	—
ω	0.29 ± 0.07	0.34 ± 0.06	0.34 ± 0.06	0.52 ± 0.06	0.58 ± 0.06	0.58 ± 0.05	0.44 ± 0.05	0.50 ± 0.05	0.50 ± 0.04

response pattern indicated, consistent with the RELATIVE model, “correct” options with lower expected utility were often preferred to “incorrect” options with higher expected utility. To formally test the hypothesis that relative value learning is positively associated with correct choice in the learning phase (i.e., rational) and negatively associated with correct choice (i.e., choice of the option with the highest absolute value) in the transfer phase (i.e., irrational), we tested the correlation between correct choice rates in these two phases and the weight parameter (ω), which quantifies the balance between the ABSOLUTE ($\omega = 0.0$) and RELATIVE models ($\omega = 1.0$). Consistent with this idea we found a positive and significant correlation between the weight parameter and the correct choice rate in the 0.1€ contexts ($R^2 = 0.19$, $P = 0.0005$) and a negative and significant correlation between the same parameter and the correct choice rate in the transfer test ($R^2 = 0.42$, $P = 0.00000003$) (Fig. 4g, h). This means that, the better a subject was at picking the correct option during the learning phase (rational behavior), the least often she would pick the option with the highest absolute value during the test phase (irrational behavior).

Discussion

In the present paper, we investigated state-dependent valuation in human reinforcement learning. In particular, we adapted a task designed to address the reference-dependence¹⁹ to include an additional manipulation of the magnitude of outcomes, in order to investigate range-adaptation²⁶. In the learning sessions, analyses of behavioral data showed that the manipulation of outcome magnitude had a significant effect on learning performance, with high-magnitude outcomes inducing better learning compared to low-magnitude outcomes. On the contrary, and in line with what we reported previously¹⁹, the manipulation of outcome valence had no such effect. In the transfer test, participants exhibited seemingly irrational preferences, sometimes preferring options that had objectively lower expected values than other options. Crucially, these irrational preferences are compatible with state-dependent valuation.

State-dependent (or context-dependent) valuation has been ascribed to a large number of different behavioral, neural and computational manifestations¹⁶. Under this rather general umbrella, reference-dependence and range-adaptation constitute two specific, and in principle dissociable, mechanisms: on the one hand, reference-dependence is the mechanism through which, in a context where monetary losses are frequent, loss avoidance (an affective neural event) is experienced as a positive outcome. On the other hand, range-adaptation is the mechanism through which, in contexts with different outcome magnitudes (i.e.,

different affective saliency), high-magnitude and low-magnitude outcomes are experienced similarly.

In order to formally and quantitatively test for the presence of these two components of state-dependent valuation in our experimental data, we used computational modeling. Our model space included two ‘extreme’ models: the ABSOLUTE and the RELATIVE models. The ABSOLUTE model learns the context-independent—absolute—value of available options. In contrast, the RELATIVE model implements both reference-dependence and range-adaptation (‘full’ adaptation²⁹). These two ‘extreme’ models predict radically different choice patterns in both the learning sessions and the transfer test. While the ABSOLUTE model predicts a big effect of outcome magnitude in the learning sessions and rational preferences in the transfer test, the RELATIVE model predicts no magnitude effect and highly irrational preferences in the transfer test. Specifically, according to the RELATIVE model, the choices in the transfer test are not affected by the outcome valence or by the outcome magnitude, but dominated by options’ context-dependent favorableness factor. Comparison between model simulations and experimental data falsified both models³¹, since in both the learning sessions and in the transfer test, subjects performance lied in between the predictions of the ABSOLUTE and RELATIVE models. To account for this pattern we designed a HYBRID model. The HYBRID model implements a trade-off between the absolute and relative learning modules, which is governed by an additional free parameter (‘partial adaptation’²⁹). Owing to this partial adaptation, the HYBRID model accurately accounts for the performance in the learning sessions and for the preferences expressed in the transfer test, including the preference inversion patterns.

Using model comparison, we attempted to provide a specific description of the process at stake in our task, and ruled out alternative accounts of normalization. Crucially, normalization can be implemented as an adaptation over time of the valuation mechanism to account for the distribution of option values encountered in successive choices, or as a time-independent decision mechanism limited to the values of options considered in one choice event^{24,33}. In the present case, model comparison favored the HYBRID model, which implements a time-adapting value normalization against the POLICY model, which implements a time-independent decision normalization. This result derives from the fact that during the learning sessions, the POLICY model uses a divisive normalization at the moment of choice to level the learning performance in different contexts (e.g. big and small magnitudes), while still relying on learning absolute values²⁵. Therefore, these absolute values cannot produce the seemingly irrational preferences observed in the transfer test.

The idea that the magnitude of available outcomes is somewhat rescaled by decision-makers is the cornerstone of the concept of utility²². In economics, this magnitude normalization is considered a stable property of individuals, and typically modeled with a marginally decreasing utility function whose parameters reflect individual core preferences^{34,35}. This approach was implemented in the UTILITY model, present in our model space. However, this model did not provide a satisfactory account of the behavioral data, and hence was not favored by the model-comparison approach. Similarly to the case of the POLICY model, this result derives from the fact that the UTILITY model cannot account for the emergence of reference-dependence, which is necessary to produce preference reversals between the symbols of opposite valence in the transfer test. Crucially, correct choice rate during the learning sessions were equally well predicted by the UTILITY and the HYBRID models, thus highlighting the importance of using a transfer test, where options are extrapolated from original contexts, to challenge computational models of value learning and encoding^{19,36,37}.

Overall, our model comparison (based on both goodness-of-fit criteria and simulation-based falsification) favored the HYBRID model, which indicates that the pattern of choices exhibited by our subjects in the learning sessions and in the transfer test is most probably the result of a trade-off between absolute and relative values. In the HYBRID model, this trade-off was implemented by a subject-specific weight parameter (ω), which quantified the relative influence of the normalized vs. absolute value-learning modules. A series of subsequent analyses revealed that several relevant factors affect this trade-off. First, we showed using an original trial-by-trial model comparison that the trade-off between absolute value-learning and normalized value learning implemented by the HYBRID model is progressive and gradual. This is an important novelty compared to previous work which only suggested such progressivity by showing that value rescaling was dependent of progressively acquired feedback information¹⁹. Note that learning normalized value ultimately converges to learning which option of a context is best, regardless of its valence or relative value compared to the alternative option. Second, and in line with the idea that information concerning the forgone outcome promotes state-dependent valuation^{18,32}, we also found that the relative weight of the normalized-value learning module (ω) increased when more information was available (counterfactual feedback). Finally, individuals whose pattern of choices was indicative of a strong influence of the normalized value learning module (i.e., with higher ω) appeared to have a better understanding of the task, assessed in the debriefing. Future research, using larger sample sizes and more diversified cohorts, will indicate whether or not the weight parameter (and therefore the value contextualization process) is useful to predict real life outcomes in terms of socio-economics achievements and psychiatric illness.

Overall, these findings suggest that value normalization is the results of a 'high-level'—or 'model-based'—process through which outcome information is not only used to update action values, but also to build an explicit representation of the embedding context where outcomes are experienced. Consistent with this interpretation, value normalization has recently been shown to be degraded by manipulations imposing a penalty for high-level costly cognitive functions, such as high memory load conditions in economic decision-making tasks³⁸. One can also speculate that value contextualization should be impaired under high cognitive load³⁹ and when outcome information is made unconscious⁴⁰. Future research using multi-tasking and visual masking could address these hypotheses⁴¹. An additional feature of the design suggests that this value normalization is an active process. In our paradigm the different choice contexts were

presented in an interleaved manner, meaning that a subject could not be presented with the same context more than a few times in a row. Therefore, contextual effects could not be ascribed to slow and passive habituation (or sensitization) processes.

Although the present results, together with converging evidence in economics and psychology, concordantly point that state-dependent valuation is needed to provide a satisfactory account of human behavior, there is still an open debate concerning the exact implementation of such contextual influences. In paradigms where subjects are systematically presented with full feedback information, it would seem that subjects simply encode the difference between obtained and forgone outcome, thus parsimoniously achieving full context-dependence without explicitly representing and encoding state value^{18,32}. However, such models cannot be easily and effectively adapted to tasks where only partial feedback information is available. In these tasks, context-dependence has been more efficiently implemented by assuming separate representational structures for action and state values which are then used to center action-specific prediction errors^{19,20}. In the present paper, we implemented this computational architecture in the HYBRID model, which builds on a partial adaptation scheme between an ABSOLUTE and a RELATIVE model. Although descriptive by nature, such hybrid models are commonly used in multi-step decision-making paradigms, e.g., to implement trade-offs between model-based and model free learning^{42–44}, because they allow to readily quantify the contributions of different learning strategies, and to straightforwardly map to popular dual-process accounts of decision-making^{45,46}. In this respect, future studies adapting the present paradigm for functional imaging will be crucial to assess whether absolute and relative (i.e., reference-point centered and range adapted) outcome values are encoded in different regions (dual valuation), or whether contextual information is readily integrated with outcome values in a single brain region (partial adaptation). However, it should be noted that previous studies using similar paradigms, consistently provided support for the second hypothesis, by showing that contextual information is integrated in a brain valuation system encompassing both the ventral striatum and the ventral prefrontal cortex, which therefore represent 'partially adapted' values^{19,20,29}. This is corroborated by similar observations from electrophysiological recordings of single neurons in monkeys^{26,27,47,48}.

As in our previous study^{19,28}, we also manipulated outcome valence in order to create 'gain' and 'loss' decision frames. While focusing on the results related to the manipulation of outcome magnitude, which represented the novelty of the present design, we nonetheless replicated previous findings indicating that subjects perform equally well in both decision frames and that this effect is parsimoniously explained assuming relative value encoding. This robust result contradicts both standard reinforcement principles and behavioral economic results. In the context of animal learning literature, while Thorndike's famous law of effect parsimoniously predicts reward maximization in a 'gain' decision frame, it fails to explain punishment minimization in the 'loss' frame. Mower elegantly formalized this issue⁴⁹ ('how can a shock that is not experienced, i.e., which is avoided, be said to provide [...] a source of [...] satisfaction?') and proposed the two-factor theory that can be seen as an antecedent of our relative value-learning model. In addition, the gain/loss behavioral symmetry is surprising with respects to behavioral economic theory because it contradicts the loss aversion principle¹⁷. In fact, if 'losses loom larger than gains', one would predict a higher correct response rate in the 'loss' compared to the 'gain' domain in our task. Yet, such deviations to standard behavioral economic theory are not infrequent when decisions are based on experience rather than description⁵⁰, an observation referred to as the "experience/

description gap”^{51,52}. While studies of the “experience/description gap” typically focus on deviations regarding attitude risky and rare outcomes, our and other groups’ results indicate that a less documented but nonetheless—robust instance of the experience/description gap is precisely the absence of loss aversion^{3,53}.

To conclude, state-dependent valuation, defined as the combination of reference-point dependence and range-adaptation, is a double-edged sword of value-based learning and decision-making. Reference-point dependence provides obvious beneficial behavioral consequences in punishment avoidance contexts and range-adaptation allows to perform optimally when decreasing outcome magnitudes. The combination of these two mechanisms (implemented in the HYBRID model) is therefore accompanied with satisfactory learning performance in all proposed contexts. However, these beneficial effects on learning performance are traded-off against possible suboptimal preferences and decisions, when options are extrapolated from their original context. Crucially, our results show that state-dependent valuation remains only partial. As a consequence, subjects under-performed in the learning sessions relative to full context-dependent strategies (RELATIVE model), as well as in the transfer test relative to absolute value strategies (ABSOLUTE model). These findings support the idea that bounded rationality may not only arise from intrinsic limitations of the brain computing capacity, but also from the fact that different situations require different valuation strategies to achieve optimal performance. Given the fact that humans and animals often interact with changing and probabilistic environments, apparent bounded rationality may simply be the result of the effort for being able to achieve a good level of performance in a variety of different contexts. These results shed new light on the computational constraints shaping everyday reinforcement learning abilities in humans, most-likely set by evolutionary forces to optimally behave in ecological settings featuring both changes and regularities³⁶.

Methods

Experimental subjects. We tested 60 subjects (39 females; aged 22.3 ± 3.3 years). Subjects were recruited via Internet advertising in a local mailing-list dedicated to cognitive science-related activities. We experienced no technical problems, so we were able to include all 60 subjects. Experiment 1 included 20 subjects. The sample size was chosen based on previous studies. Experiment 2 included 40 subjects: we doubled the sample size because Experiment 2 involved a more complex design with an additional factor (see below). The research was carried out following the principles and guidelines for experiments including human participants provided in the declaration of Helsinki (1964, revised in 2013). The local Ethical Committee approved the study and subjects provided written informed consent prior to their inclusion. To sustain motivation throughout the experiment, subjects were given a bonus dependent on the actual money won in the experiment (average money won: 3.73 ± 0.27 , against chance $T(59) = 13.9$, $P < 0.0001$).

Behavioral protocol. Subjects performed a probabilistic instrumental learning task adapted from previous imaging and patient studies¹⁹. Subjects were first provided with written instructions, which were reformulated orally if necessary. They were explained that the aim of the task was to maximize their payoff and that seeking monetary rewards and avoiding monetary losses were equally important. For each experiment, subjects performed two learning sessions. Cues were abstract stimuli taken from the Agathodaimon alphabet. Each session contained four novel pairs of cues. The pairs of cues were fixed, so that a given cue was always presented with the same other cue. Thus, within sessions, pairs of cues represented stable choice contexts. Within sessions, each pair of cues was presented 20 times for a total of 80 trials. The four cue pairs corresponded to the four contexts (reward/big magnitude, reward/small magnitude, loss/big magnitude, and loss/small magnitude). Within each pair, the two cues were associated to a zero and a non-zero outcome with reciprocal probabilities (0.75/0.25 and 0.25/0.75). On each trial, one pair was randomly presented on the left and the right side of a central fixation cross. Pairs of cues were presented in a pseudo-randomized and unpredictable manner to the subject (intermixed design). The side in which a given cue was presented was also pseudo-randomized, such that a given cue was presented an equal number of times in the left and the right of the central cue. Subjects were required to select between the two cues by pressing one of the corresponding two buttons, with their left or right thumb, to select the leftmost or the rightmost cue, respectively, within a

3000 ms time window. After the choice window, a red pointer appeared below the selected cue for 500 ms. At the end of the trial, the cues disappeared and the selected one was replaced by the outcome (“+1.0€”, “+0.1€”, “0.0€”, “−0.1€” or “−1.0€”) for 3000 ms. In Experiment 2, in the complete information contexts (50% of the trials), the outcome corresponding to the unchosen option (counterfactual) was displayed. A novel trial started after a fixation screen (1000 ms, jittered between 500 and 1500 ms). After the two learning sessions, subjects performed a transfer test. This transfer test involved only the eight cues (2^4 pairs) of the last session, which were presented in all possible binary combinations (28, not including pairs formed by the same cue) (see also ref. ¹⁸). Each pair of cues was presented four times, leading to a total of 112 trials. Instructions for the transfer test were provided orally after the end of the last learning session. Subjects were explained that they would be presented with pairs of cues taken from the last session, and that all pairs would not have been necessarily displayed together before. On each trial, they had to indicate which of the cues was the one with the highest value by pressing on the buttons as in the learning task. Subjects were also explained that there was no money at stake, but encouraged to respond as they would have if it were the case. In order to prevent explicit memorizing strategies, subjects were not informed that they would have to perform a transfer test until the end of the second (last) learning sessions. Timing of the transfer test differed from that of the learning sessions in that the choice was self-paced and in the absence of outcome phase. During the transfer test, the outcome was not provided in order not to modify the option values learned during the learning sessions. Between the learning sessions and the transfer test subjects were interviewed in order to probe the extent of their explicit knowledge of the task’s structure. More precisely the structured interview assessed: (1) whether or not the subjects were aware about the cues being presented in fixed pairs (choice contexts); (2) how many choice contexts they believed were simultaneously present in a learning session. The experimenter recorded the responses, but provided no feedback about their correctness in order to not affect subjects’ performance in the transfer test.

Model-free analyses. For the two experiments, we were interested in three different variables reflecting subjects’ learning: (1) correct choice rate (i.e., choices directed toward highest expected reward or the lowest expected loss) during the learning task of the experiment. Statistical effects were assessed using multiple-way repeated measures ANOVAs with feedback valence, feedback magnitude, and feedback information (in Experiment 2 only) as within-subject factors; (2) correct choice rate during the transfer test, i.e., choosing the option with the highest absolute expected value (each symbol has a positive or negative absolute expected value, calculated as Probability(outcome) \times Magnitude(outcome)); and (3) choice rate of the transfer test (i.e., the number of times an option is chosen, divided by the number of times the option is presented). The variable represents the value attributed to one option, i.e., the preference of the subjects for each of the symbols. Transfer test choice rates were submitted to multiple-way repeated measures ANOVAs, to assess the effects of option favorableness (being the most advantageous option of the pair), feedback valence and feedback magnitude as within-subject factors. In principle, probabilistic designs like ours the theoretical values (i.e., imposed by design) of the contexts and options may not correspond to the outcomes experienced by subjects. To verify that our design-based categories used in the ANOVAs analyses were legitimated, we checked the correlation between the theoretical and the empirical values of the outcomes. The results indicate that there was no systematic bias ($R > 0.99$; and $0.9 < \text{slope} < 1.2$). Post-hoc tests were performed using one-sample t -tests. To assess overall performance, additional one-sample t -tests were performed against chance level (0.5). Correct choice rates from the learning test meet a normal distribution assumption (Kolmogorov–Smirnov test: $K(60) = 0.087$, $P > 0.72$; Lilliefors test: $K(60) = 0.087$, $P > 0.30$), as well as correct choice rates from the transfer test (Kolmogorov–Smirnov test: $K(60) = 0.092$, $P > 0.65$; Lilliefors test: $K(60) = 0.092$, $P > 0.22$). All statistical analyses were performed using Matlab (www.mathworks.com) and R (www.r-project.org).

Model space. We analyzed our data with extensions of the Q-learning algorithm^{4,54}. The goal of all models was to find in each choice context (or state) the option that maximizes the expected reward R .

At trial t , option values of the current context s are updated with the Rescorla–Wagner rule⁵:

$$\begin{aligned} Q_{t+1}(s, c) &= Q_t(s, c) + \alpha_c \delta_{c,t} \\ Q_{t+1}(s, u) &= Q_t(s, u) + \alpha_u \delta_{u,t} \end{aligned} \quad (4)$$

where α_c is the learning rate for the chosen (c) option and α_u the learning rate for the unchosen (u) option, i.e., the counterfactual learning rate. δ_c and δ_u are prediction error terms calculated as follows:

$$\begin{aligned} \delta_{c,t} &= R_{c,t} - Q_t(s, c) \\ \delta_{u,t} &= R_{u,t} - Q_t(s, u) \end{aligned} \quad (5)$$

δ_c is updated in both partial and complete feedback contexts and δ_u is updated in the complete feedback context only (Experiment 2, only).

We modeled subjects' choice behavior using a softmax decision rule representing the probability for a subject to choose one option a over the other option b :

$$P_t(s, a) = \frac{1}{1 + e^{\frac{Q_t(s,b) - Q_t(s,a)}{\beta}}} \quad (6)$$

where β is the temperature parameter. High temperatures cause the action to be all (nearly) equi-probable. Low temperatures cause a greater difference in selection probability for actions that differ in their value estimates⁴.

We compared four alternative computational models: the ABSOLUTE model, which encodes outcomes in an absolute scale independently of the choice context in which they are presented; the RELATIVE model which encodes outcomes on a binary (correct/incorrect) scale, relative to the choice context in which they are presented⁵⁵; the HYBRID model, which encodes outcomes as a weighted sum of the absolute and relative value; the POLICY model, which encodes outcome in an absolute scale, but implements divisive normalization in the policy.

ABSOLUTE model. The outcomes are encoded as the subjects see them as feedback. A positive outcome is encoded as its "real" positive value (in euros) and a negative outcome is encoded as its "real" negative value (in euros): $R_{ABS,t} \in \{-1.0\text{€}, -0.1\text{€}, 0.0\text{€}, 0.1\text{€}, 1.0\text{€}\}$.

RELATIVE model. The outcomes (both chosen and unchosen) are encoded on a context-dependent correct/incorrect relative scale. The model assumes the effective outcome value to be adapted to the range of the outcomes present in a given context. The option values are no longer calculated in an absolute scale, but relatively to their choice context value: in the delta-rule, the correct option is updated with a reward of 1 and the incorrect option is updated with a reward of 0. To determine the context of choice, the model uses a state value $V(s)$ stable over trials, initialized to 0, which takes the value of the first non-zero (chosen or unchosen) outcome in each context s .

$$R_{REL,t} = \frac{R_{ABS,t}}{|V_t(s)|} + \max\left\{0, \frac{-V_t(s)}{|V_t(s)|}\right\} \quad (7)$$

Thus, the outcomes (chosen and unchosen) are now normalized to a context-dependent correct/incorrect encoding: $R_{REL,t} \in \{0, 1\}$. The chosen and unchosen option values and prediction errors are updated with the same rules as in the ABSOLUTE model.

HYBRID model. At trial t the prediction errors of the chosen and unchosen options are updated as a weighted sum of the absolute and relative outcomes:

$$R_{HYB,t} = \omega * R_{REL,t} + (1 - \omega) * R_{ABS,t} \quad (8)$$

where ω is the individual weight. At each trial t , the model independently encodes both outcomes as previously described and updates the final HYBRID outcome:

$$R_{HYB,t} = \begin{cases} R_{ABS,t} & \text{if } \omega = 0 \\ R_{REL,t} & \text{if } \omega = 1 \end{cases}$$

The chosen and unchosen option values and prediction errors are updated with the same rules as in the ABSOLUTE model. If the RELATIVE model is conceptually similar to a policy-gradient algorithm, because it does not encode cardinal option values but only context-dependent ordinal preferences, the HYBRID model is reminiscent of a recently proposed model that features an interaction between a Q-learning and an actor-critic^{56,57}.

UTILITY model. We also considered a fourth UTILITY model, which implements the economic notion of marginally decreasing subjective utility at the outcome encoding step^{17,22}. The big magnitude outcomes ($|R| = 1$) are re-scaled with a multiplicative factor $0.1 < v < 1.0$:

$$R_{UTY,t} = v * R_{ABS,t} \text{ if } |R| = 1 \quad (9)$$

POLICY model. Finally, we considered a fifth POLICY model that encodes option values as the ABSOLUTE model and normalizes them in the softmax rule, i.e., at the decision step only^{25,26,47}:

$$P_t(s, a) = \frac{1}{1 + e^{\frac{Q_t(s,b) - Q_t(s,a)}{Q_t(s,b) + Q_t(s,a)} \frac{1}{\beta}}} \quad (10)$$

Additional computational hypotheses are addressed (and rejected) in the Supplementary Methods.

Model fitting, comparison, and simulation. Specifically for the learning sessions, transfer test, and both, we optimized model parameters, the temperature β , the factual learning rate α_f , the counterfactual learning rate α_c (in Experiment 2 only) and the weight ω (in the HYBRID model only), by minimizing the negative log likelihood LL_{max} using Matlab's *fmincon* function, initialized at starting points of 1 for the temperature and 0.5 for the learning rates and the weight. As a quality check we replicated this analysis using multiple starting points and this did not change the results (Supplementary Table 2). We computed at the individual level the BIC using, for each model, its number of free parameters d_f (note that the Experiment 2 has an additional parameter α_c) and the number of trials n_{trials} (note that this number of trials varies with the optimization procedure: learning sessions only, 160, transfer test only, 112, or both, 272):

$$BIC = 2 * LL_{max} + \log(n_{trials}) * d_f \quad (11)$$

Model estimates of choice probability were generated trial-by-trial using the optimal individual parameters. We made comparisons between predicted and actual choices with a one-sample t -test and tested models' performances out of the sample by assessing their ability to account for the transfer test choices. On the basis of model-estimate choice probability, we calculated the log-likelihood of learning sessions and transfer test choices that we compared between computational models. Finally, we submitted the model-estimate transfer-test choice probability to the same statistical analyses as the actual choices (ANOVA and post-hoc t -test; within-simulated data comparison) and we compared modeled choices to the actual data. In particular, we analyzed actual and simulated correct choice rates (i.e., the proportions of choices directed toward the most advantageous stimulus) and compared transfer-test choices for each symbol with a sampled t -test between the behavioral choices and the simulated choices.

Code availability. All custom scripts have been made available from Github repository <https://github.com/sophiebavard/Magnitude>. Additional modified scripts can be accessed upon request.

Data availability

Data that support the findings of this study are available from Github repository <https://github.com/sophiebavard/Magnitude>.

Received: 5 April 2018 Accepted: 26 September 2018

Published online: 29 October 2018

References

- Guitart-Masip, M., Duzel, E., Dolan, R. & Dayan, P. Action versus valence in decision making. *Trends Cogn. Sci.* **18**, 194–202 (2014).
- Knutson, B., Katovich, K. & Suri, G. Inferring affect from fMRI data. *Trends Cogn. Sci.* **18**, 422–428 (2014).
- Yechiam, E. & Hochman, G. Losses as modulators of attention: review and analysis of the unique effects of losses over gains. *Psychol. Bull.* **139**, 497–518 (2013).
- Sutton, R. S. & Barto, A. G. Reinforcement learning: an introduction. *IEEE Trans. Neural Netw.* **9**, 1054–1054 (1998).
- Rescorla, R. A. & Wagner, A. R. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. *Class. Cond. II Curr. Res. Theory* **2**, 64–99 (1972).
- Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
- O'Doherty, J. et al. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452–454 (2004).
- Frank, M. J., Seeberger, L. C. & O'Reilly, R. C. By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science* **306**, 1940–1943 (2004).
- Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J. & Frith, C. D. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* **442**, 1042–1045 (2006).
- Palminteri, S. et al. Pharmacological modulation of subliminal learning in Parkinson's and Tourette's syndromes. *Proc. Natl Acad. Sci. USA* **106**, 19179–19184 (2009).
- McNamara, J. M., Trimmer, P. C. & Houston, A. I. The ecological rationality of state-dependent valuation. *Psychol. Rev.* **119**, 114–119 (2012).
- Pompilio, L. & Kacelnik, A. Context-dependent utility overrides absolute memory as a determinant of choice. *Proc. Natl Acad. Sci. USA* **107**, 508–512 (2010).
- Bar, M. Visual objects in context. *Nat. Rev. Neurosci.* **5**, 617–629 (2004).
- Schwartz, O., Hsu, A. & Dayan, P. Space and time in visual context. *Nat. Rev. Neurosci.* **8**, 522–535 (2007).
- Kahneman, D. & Tversky, A. Choices, values, and frames. *Am. Psychol.* **39**, 341–350 (1984).

16. Louie, K. & De Martino, B. Chapter 24—The neurobiology of context-dependent valuation and choice. in *Neuroeconomics*, 2nd edn (eds. Glimcher, P. W. & Fehr, E.) 455–476 (Academic Press, San Diego, CA, 2014).
17. Kahneman, D. & Tversky, A. Prospect theory: an analysis of decision under risk. *Econ. J. Econ. Soc.* **47**, 263–291 (1979).
18. Klein, T. A., Ullsperger, M. & Jocham, G. Learning relative values in the striatum induces violations of normative decision making. *Nat. Commun.* **8**, 16033 (2017).
19. Palminteri, S., Khamassi, M., Joffily, M. & Coricelli, G. Contextual modulation of value signals in reward and punishment learning. *Nat. Commun.* **6**, 8096 (2015).
20. Rigoli, F., Friston, K. J. & Dolan, R. J. Neural processes mediating contextual influences on human choice behaviour. *Nat. Commun.* **7**, 12416 (2016).
21. Fechner, G. T. *Elemente der psychophysik*. (Leipzig, Breitkopf und Härtel, 1860).
22. Bernoulli, D. *Specimen Theoriae Novae de Mensura Sortis* (1738).
23. Carandini, M. & Heeger, D. J. Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* **13**, 51–62 (2011).
24. Webb, R., W. Glimcher, P. & Louie, K. Rationalizing context-dependent preferences: divisive normalization and neurobiological constraints on choice. *SSRN Electron. J.* <https://doi.org/10.2139/ssrn.2462895> (2014).
25. Yamada, H., Louie, K., Tymula, A. & Glimcher, P. W. Free choice shapes normalized value signals in medial orbitofrontal cortex. *Nat. Commun.* **9**, 162 (2018).
26. Padoa-Schioppa, C. Range-adapting representation of economic value in the orbitofrontal cortex. *J. Neurosci.* **29**, 14004–14014 (2009).
27. Rustichini, A., Conen, K. E., Cai, X. & Padoa-Schioppa, C. Optimal coding and neuronal adaptation in economic decisions. *Nat. Commun.* **8**, 1208 (2017).
28. Palminteri, S., Kilford, E. J., Coricelli, G. & Blakemore, S.-J. The computational development of reinforcement learning during adolescence. *PLoS Comput. Biol.* **12**, e1004953 (2016).
29. Burke, C. J., Baddeley, M., Tobler, P. N. & Schultz, W. Partial adaptation of obtained and observed value signals preserves information about gains and losses. *J. Neurosci.* **36**, 10016–10025 (2016).
30. Neumann, J. von & Morgenstern, O. *Theory of Games and Economic Behavior*. (Princeton University Press, Princeton, NJ, 1953).
31. Palminteri, S., Wyart, V. & Koehlin, E. The importance of falsification in computational cognitive modeling. *Trends Cogn. Sci.* **21**, 425–433 (2017).
32. Li, J. & Daw, N. D. Signals in human striatum are appropriate for policy update rather than value prediction. *J. Neurosci.* **31**, 5504–5511 (2011).
33. Rangel, A. & Clithero, J. A. Value normalization in decision making: theory and evidence. *Curr. Opin. Neurobiol.* **22**, 970–981 (2012).
34. Fox, C. R. & Poldrack, R. A. Appendix—prospect theory and the brain. in *Neuroeconomics*, 2nd edn (eds. Glimcher, P. W. & Fehr, E.) 533–567 (Academic Press, San Diego, CA, 2014).
35. Pedroni, A. et al. The risk elicitation puzzle. *Nat. Hum. Behav.* **1**, 803–809 (2017).
36. Kacelnik, A. Tools for thought or thoughts for tools? *Proc. Natl Acad. Sci. USA* **106**, 10071–10072 (2009).
37. Wimmer, G. E. & Shohamy, D. Preference by association: how memory mechanisms in the hippocampus bias decisions. *Science* **338**, 270–273 (2012).
38. Holper, L. et al. Adaptive value normalization in the prefrontal cortex is reduced by memory load. *eNeuro* ENEURO.0365-17.2017, <https://doi.org/10.1523/ENEURO.0365-17.2017> (2017).
39. Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A. & Daw, N. D. Working-memory capacity protects model-based learning from stress. *Proc. Natl Acad. Sci. USA* **110**, 20941–20946 (2013).
40. Ogmen, H., Breitmeyer, B. G. & Melvin, R. The what and where in visual masking. *Vision Res.* **43**, 1337–1350 (2003).
41. Pessiglione, M. et al. How the brain translates money into force: a neuroimaging study of subliminal motivation. *Science* **316**, 904–906 (2007).
42. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
43. Gläscher, J., Daw, N., Dayan, P. & O’Doherty, J. P. States versus Rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010).
44. Lessaint, F., Sigaud, O., Flagel, S. B., Robinson, T. E. & Khamassi, M. Modelling individual differences in the form of pavlovian conditioned approach responses: a dual learning systems approach with factored representations. *PLoS Comput. Biol.* **10**, e1003466 (2014).
45. Evans, J. S. B. T. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.* **59**, 255–278 (2008).
46. Kahneman, D. A perspective on judgment and choice: mapping bounded rationality. *Am. Psychol.* **58**, 697–720 (2003).
47. Louie, K., LoFaro, T., Webb, R. & Glimcher, P. W. Dynamic divisive normalization predicts time-varying value coding in decision-related circuits. *J. Neurosci.* **34**, 16046–16057 (2014).
48. Louie, K., Khaw, M. W. & Glimcher, P. W. Normalization is a general neural mechanism for context-dependent decision making. *Proc. Natl Acad. Sci. USA* **110**, 6139–6144 (2013).
49. Herzberg, F. *The Motivation to Work* (Wiley, New York, NY, 1959).
50. Ariely, D., Huber, J. & Wertenbroch, K. When do losses loom larger than gains? *J. Mark. Res.* **42**, 134–138 (2005).
51. Camilleri, A. & Newell, B. *Within-subject Preference Reversals in Description- and Experience-based Choice*. 449–454 (Cognitive Science Society, Austin, TX, 2009).
52. Hertwig, R. & Erev, I. The description-experience gap in risky choice. *Trends Cogn. Sci.* **13**, 517–523 (2009).
53. Ludvig, E. A. & Spetch, M. L. Of black swans and tossed coins: is the description-experience gap in risky choice limited to rare events? *PLOS ONE* **6**, e20262 (2011).
54. Watkins, C. J. C. H. & Dayan, P. Q-learning. *Mach. Learn.* **8**, 279–292 (1992).
55. Vlaev, I., Chater, N., Stewart, N. & Brown, G. D. A. Does the brain calculate value? *Trends Cogn. Sci.* **15**, 546–554 (2011).
56. Gold, J. M. et al. Negative symptoms and the failure to represent the expected reward value of actions: behavioral and computational modeling evidence. *Arch. Gen. Psychiatry* **69**, 129–138 (2012).
57. Dayan, P. & Abbott, L. F. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (Massachusetts Institute of Technology Press, Cambridge, MA, 2001).

Acknowledgements

Emmanuel Noblins and Alexander Salvador provided help for data collection. S.P. is supported by an ATIP-Avenir grant (R16069JS) Collaborative Research in Computational Neuroscience ANR-NSF grant (ANR-16-NEUC-0004), the Programme Emergence (s) de la Ville de Paris, and the Fondation Fyssen. S.B. is supported by MILDECA (Mission Interministérielle de Lutte contre les Drogues et les Conduites Addictives) and the EHESS (Ecole des Hautes Etudes en Sciences Sociales). M.L. is supported by an NWO Veni Fellowship (Grant 451-15-015) and a Swiss National Fund Ambizione grant (PZ00P3_174127). The Institut d’Etudes de la Cognition is supported financially by the LabEx IEC (ANR-10-LABX-0087 IEC) and the IDEX PSL* (ANR-10-IDEX-0001-02 PSL*). The funding agencies did not influence the content of the manuscript.

Author contributions

S.P. and G.C. designed the task. S.P. performed the experiments. S.B., M.L., and S.P. analyzed the data. S.B., M.L., S.P., and M.K. wrote the manuscript. All authors interpreted the results, commented, and approved the final version of the manuscript.

Additional information

Supplementary Information accompanies this paper at <https://doi.org/10.1038/s41467-018-06781-2>.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at <http://npj.nature.com/reprintsandpermissions/>

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



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

© The Author(s) 2018

Supplementary Information

Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences

Bavard et al.

Supplementary Methods

Model simulations of the POLICY and the UTILITY models

We analyzed the generative performances of the POLICY model: similarly to the RELATIVE model, the POLICY model underestimates the difference between the big and the small magnitude contexts (simulations vs. data, $T(59)=2.9$, $P<0.006$). When considering the transfer test, the POLICY model predicts a linear pattern, because, despite the normalization process within the softmax function, option values remain encoded in an absolute scale. Paradoxically, whereas in the learning sessions the POLICY model predicts a behavior compatible with the RELATIVE model (i.e., no magnitude effect), in the transfer test it predicts a behavior consistent with the ABSOLUTE model (i.e., no value inversion) (**Supplementary Fig. 1 a-c**).

We also analyzed the generative performances of the UTILITY model: similarly to the HYBRID model, the UTILITY model is able to perfectly capture the size of the magnitude effect in the learning sessions (simulation vs. data, $T(59)=0.2$, $P>0.80$). Accordingly, the quality of fit (BIC) difference between these two models was not different when considering the learning sessions alone (HYB vs. UTY, $T(59)=0.2$, $P>0.84$, **Table 3**). However, when considering the transfer test, the UTILITY model unsurprisingly also predicted linear patterns (similar to the ABSOLUTE model), and failed to predict the value inversion between the intermediate options (**Supplementary Fig. 1 d-f**). Accordingly, the quality of fit (BIC) difference between the HYBRID and the UTILITY models was significantly different when considering the transfer sessions alone (HYB vs. UTY, $T(59)=3.3$, $P<0.002$, **Table 3**) (**Supplementary Fig. 1 d-f**).

Additional model comparison: the SEPARATE and the ABS-AC models

The fourth model, referred to as the SEPARATE model, encodes range adaptation and reference-point dependence separately with 2 respective additional free parameters ρ and π . The model describes an absolute value encoding when both parameters are set to 0 and a relative value encoding when both parameters are set to 1 :

$$R_{\text{SEP},t} = (1 - \rho) * R_{\text{ABS},t} + \rho * \frac{R_{\text{ABS},t}}{|V_t(s)|} + \pi * \max \left\{ 0, \frac{-V_t(s)}{|V_t(s)|} \right\}$$

We analyzed the generative performances of the SEPARATE model, which encodes range adaptation and reference-point dependence separately. Coherently, the model behaves similarly to the HYBRID model and captures both the magnitude effect in the learning sessions (simulation vs. data, $T(59)=1.3$, $P>0.18$) and the behavioral patterns when considering the transfer test (**Supplementary Fig. 1 g-i**). However, by increasing its complexity with two additional free parameters, the quality of fit (BIC) difference between the HYBRID and the SEPARATE model was significantly different (HYB vs. SEP $T(59)=5.42$, $P<0.0001$, **Supplementary Table 1**) in favor of the HYBRID model. In addition, we retrieved a significant correlation between the ρ and the π parameter ($R=0.31$, $P<0.02$), partially explaining the fact that a model with the two processes governed by only one parameter is more parsimonious.

We considered a fifth model, referred to as the ABS-AC model, is a mixture between a standard Q-learning algorithm (similar to the ABSOLUTE model) and an actor-critic algorithm¹. State values, changing over trials, are updated as a function of prediction errors using the delta-rule, such as Q-values in the ABSOLUTE model. Prediction errors in the critic are also used to adjust weights in the actor. Then "hybrid" Q-values are computed and an additional weighting free parameters makes the balance between the two mechanisms :

$$Q_{\text{A-AC},t}(s, a) = w_{\text{A-AC}} * Q_{\text{ABS},t}(s, a) + (1 - w_{\text{A-AC}}) * Q_{\text{AC},t}(s, a)$$

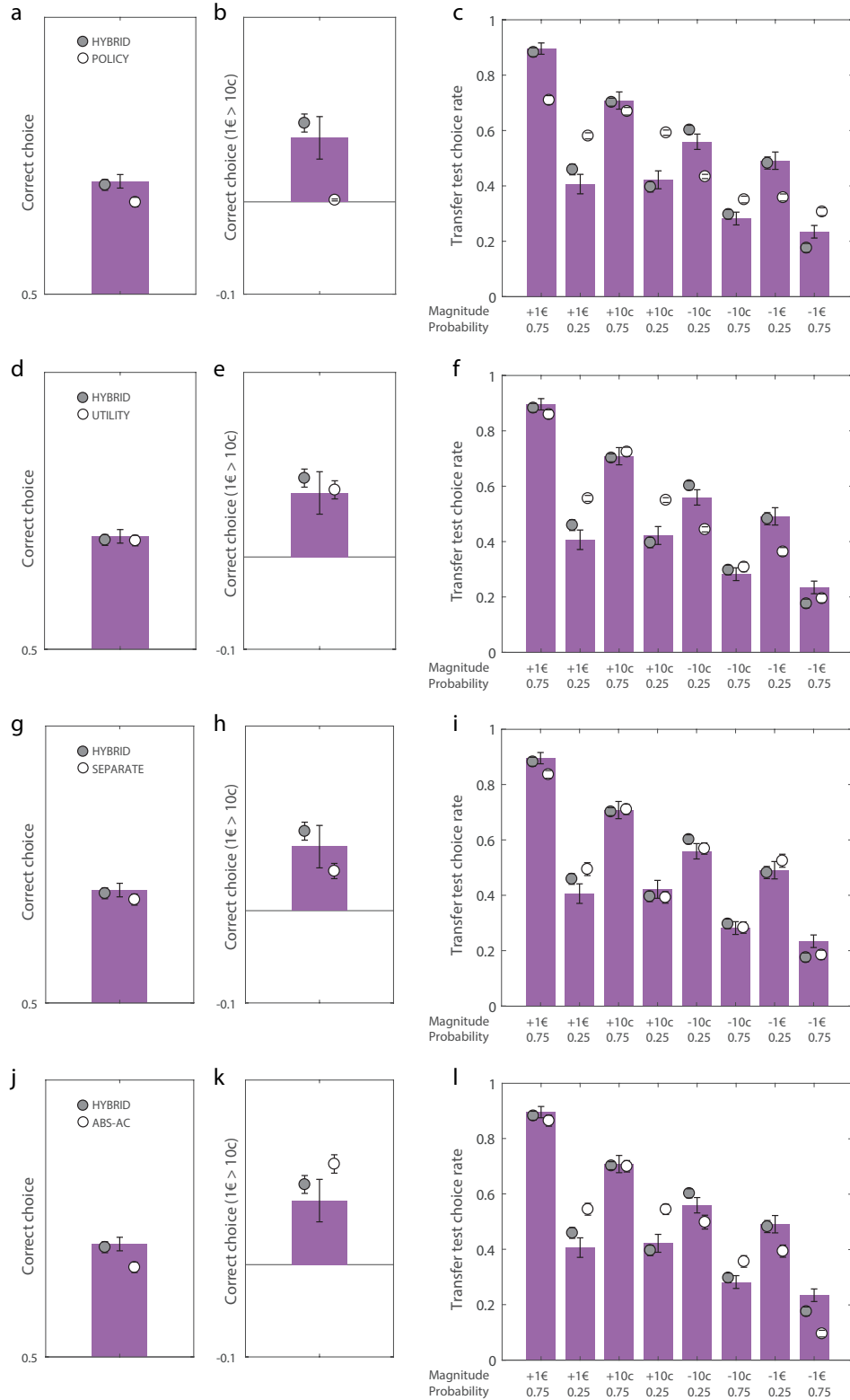
with Q_{ABS} the option value updated with the ABSOLUTE (Q-learning) value encoding and Q_{AC} the actor-critic option value updated as follows : $Q_{\text{AC}}(s, a) \leftarrow Q_{\text{AC}}(s, a) + \alpha_{\text{AC}} * (R_{\text{ABS}} - V(s))$, with $V(s)$ the state value at each trial. Action choices are computed using a softmax decision rule, by replacing individual contributions of each model by the mixture value.

To understand why relative model comparison favours the HYBRID model, we analyzed the generative performances of the ABS-AC model: the model doesn't perform as well as participants in the big magnitude context. As a result, it overestimates the difference of performance between magnitude contexts

in the learning phase and fails to match the global performance level. When extrapolating options the transfer test, the model doesn't successfully capture the value inversion and predicts a behavior consistent with absolute value encoding (**Supplementary Fig. 1 j-l**). Accordingly, the quality of fit (BIC) difference between the HYBRID and the ABS-AC models was significantly different (HYB vs. A-AC $T(59)=4.80$, $P<0.0001$, **Supplementary Table 1**).

Supplementary References

¹ Gold, J. M. et al. Negative symptoms and the failure to represent the expected reward value of actions: behavioral and computational modeling evidence. *Arch. Gen. Psychiatry* 69, 129–138 (2012).



Supplementary Figure 1: Behavioral results and model simulations of Experiment 1 and Experiment 2 pooled together. **a, d, g, j** Correct choice rate during the learning sessions. **b, e, h, k** Big magnitude context's minus small magnitude context's correct choice rate during the learning sessions. **c, f, i, l** Choice rate in the transfer test. Colored bars represent the actual data; grey dots (HYBRID) and white dots represent the model-simulated data; error bars represent s.e.m.

	Experiment 1 (N=20)			Experiment 2 (N=40)			Both experiments (N=60)		
	Learning sessions (nt=160)	Transfer test (nt=112)	Both (nt=272)	Learning sessions (nt=160)	Transfer test (nt=112)	Both (nt=272)	Learning sessions (nt=160)	Transfer test (nt=112)	Both (nt=272)
HYBRID (df=3/4)	178.3±6.0	109.3±5.0	284.6±9.1	181.5±5.8	105.8±4.1	290.5±8.0	180.5±4.3	106.9±3.2	288.5±6.1
SEPARATE (df=4/5)	197.9±4.4	115.9±5.1	314.5±7.4	190.7±5.6	109.6±4.4	300.6±7.6	192.8±4.0	111.7±3.4	305.2±5.7
ABS-AC (df=5/5)	189.1±7.0	127.8±5.7	308.2±9.8	195.3±5.4	124.8±4.5	314.8±7.4	193.2±4.3	125.8±3.5	312.6±5.9

Supplementary Table 1: BICs as a function of the dataset used for parameter optimization (Learning sessions, Transfer test or Both) and the computational model. nt: number of trials; df: degree of freedom.

	Experiment 1 (N=20)			Experiment 2 (N=40)			Both experiments (N=60)		
	Learning sessions (nt=160)	Transfer test (nt=112)	Both (nt=272)	Learning sessions (nt=160)	Transfer test (nt=112)	Both (nt=272)	Learning sessions (nt=160)	Transfer test (nt=112)	Both (nt=272)
ABSOLUTE (df=2/3)	179.8±5.9	113.6±5.7	295.1±9.4	190.6±4.7	125.2±4.2	324.2±6.4	187.0±3.7	121.3±3.4	315.5±5.5
RELATIVE (df=2/3)	193.6±4.6	136.5±5.1	329.3±8.4	184.7±5.6	119.0±4.1	303.6±7.6	187.7±4.0	124.8±3.4	312.2±6.0
HYBRID (df=3/4)	178.3±6.0	107.5±5.1	284.6±9.1	181.0±5.7	103.2±4.0	288.2±8.0	180.1±4.3	104.6±3.2	287.0±6.1
POLICY (df=2/3)	185.4±6.9	121.3±5.8	308.0±11.8	189.5±4.8	135.5±3.7	333.0±6.4	188.1±3.9	130.7±3.3	323.3±5.9
UTILITY (df=3/4)	173.9±6.5	107.4±6.3	282.2±10.8	182.8±5.5	122.2±4.4	308.4±7.1	179.9±4.3	117.3±3.7	299.6±6.1
SEPARATE (df=4/5)	196.7±4.4	115.0±5.3	312.5±7.7	189.2±5.4	107.7±4.3	299.4±7.4	191.7±3.9	110.4±3.3	303.7±5.6
ABS-AC (df=5/5)	183.3±7.3	127.7±5.7	300.7±10.2	193.0±5.3	120.1±4.5	312.5±7.2	190.3±4.2	122.8±3.6	309.1±5.9

Supplementary Table 2: BICs as a function of the dataset used for parameter optimization (Learning sessions, Transfer test or Both) and the computational model using multiple starting points (5 different random initializations per parameter, model and subject). nt: number of trials; df: degree of freedom.

2.1.3 Conclusion

In this paper, we have presented a satisfactory model which implements the trade-off between context-independent and context-dependent valuation, with a participant-specific weight parameter which quantifies the relative influence of the normalized vs. absolute value learning modules. Using an original trial-by-trial model comparison, we showed that the trade-off between absolute and normalized value learning, implemented by the model, is progressive and gradual over the task. However, although the model allows to readily quantify the contributions of different learning strategies, it remains descriptive by nature. Therefore, further work is needed to implement a model which would dynamically implement context-dependent valuation over the task, which is one of the aims of Study 2.

2.2 Study 2: Bavard et al, 2021

2.2.1 Introduction

In this study, we aimed at testing the paradoxical relation between range adaptation and performance in a large sample of participants performing variants of a reinforcement learning task, where we manipulated outcome magnitude and task difficulty. Our results replicated previous findings and confirmed that range adaptation induces extrapolation errors and is stronger when decreasing task difficulty. We proposed a dynamic version of the previous model, a range-adapting model, and show that it is able to parsimoniously capture all the behavioral results, including re-analyses on the previous dataset.

2.2.2 Article

PSYCHOLOGICAL SCIENCE

Two sides of the same coin: Beneficial and detrimental consequences of range adaptation in human reinforcement learning

Sophie Bavard^{1,2,3}, Aldo Rustichini⁴, Stefano Palminteri^{1,2,3*}

Evidence suggests that economic values are rescaled as a function of the range of the available options. Although locally adaptive, range adaptation has been shown to lead to suboptimal choices, particularly notable in reinforcement learning (RL) situations when options are extrapolated from their original context to a new one. Range adaptation can be seen as the result of an adaptive coding process aiming at increasing the signal-to-noise ratio. However, this hypothesis leads to a counterintuitive prediction: Decreasing task difficulty should increase range adaptation and, consequently, extrapolation errors. Here, we tested the paradoxical relation between range adaptation and performance in a large sample of participants performing variants of an RL task, where we manipulated task difficulty. Results confirmed that range adaptation induces systematic extrapolation errors and is stronger when decreasing task difficulty. Last, we propose a range-adapting model and show that it is able to parsimoniously capture all the behavioral results.

INTRODUCTION

In the famous Ebbinghaus illusion, two circles of identical size are placed near to each other. Larger circles surround one, while smaller circles surround the other. As a result, the central circle surrounded by larger circles appears smaller than the central circle surrounded by smaller circles, indicating that the subjective perception of size of an object is affected by its surroundings.

Beyond perceptual decision-making, a wealth of evidence in neuroscience and in economics suggests that the subjective economic value of an option is not estimated in isolation but is highly dependent on the context in which the options are presented (1, 2). The vast majority of neuroeconomic studies of context-dependent valuation in human participants considered situations where subjective values are triggered by explicit cues, that is, stimuli whose value can be directly inferred, such as lotteries or snacks (3–5). However, in a series of recent papers, we and other groups demonstrated that contextual adjustments also permeate reinforcement learning situations in which option values have to be inferred from the history of past outcomes (6–8). We showed that an option, whose small objective value [for example 7.5 euro cents (c)] is learned in a context of smaller outcomes, is preferred to an option whose objective value (25c) is learned in a context of bigger outcomes, thus providing an equivalent of the Ebbinghaus illusion in economic choices. Similar observations in birds suggest that this is a feature of decision-making broadly shared across vertebrates (9, 10).

Although (as illustrated by the Ebbinghaus example) value context dependence may lead to erroneous or suboptimal decisions, it could be normatively understood as an adaptive process aimed at rescaling the neural response according to the range of the available options. Specifically, it could be seen as the result of an adaptive coding process aiming at increasing the signal-to-noise ratio by a

system (the brain) constrained by the fact that behavioral variables have to be encoded by finite firing rates. In other terms, such range adaptation would be a consequence of how the system adjusts and optimizes the function associating the firing rate to the objective values, to set the slope of the response to the optimal value for each context (11, 12).

If range adaptation is a consequence of how the brain automatically adapts its response to the distributions of the available outcomes, then factors that facilitate the identification of these distributions might make more pronounced its behavioral consequences, because of the larger difference between the objective option values (context-independent or absolute) and their corresponding subjective values (context-dependent or relative). This leads to a counterintuitive prediction in the context of reinforcement learning. This prediction is in notable contrast with the intuition embedded in virtually all learning algorithms that making a learning problem easier (in our case, by facilitating the identification of the outcome distributions) should, if anything, lead to more accurate and objective internal representations. In the present study, we aim at testing this hypothesis while, at the same time, gaining a better understanding of range adaptation at the algorithmic level.

To empirically test this hypothesis, we build on previous research and used a task featuring a learning phase and a transfer phase (6). In the learning phase, participants had to determine, by trial and error, the optimal option in four fixed pairs of options (contexts), with different outcome ranges. In the transfer phase, the original options were rearranged in different pairs, thus creating new contexts. This setup allowed us to quantify learning (or acquisition) errors during the first phase and transfer (or extrapolation) errors during the second phase. Crucially, the task contexts were designed such that the correct responses (that is, choice of options giving a higher expected value) in the transfer phase were not necessarily correct responses during the learning phase. We varied this paradigm in eight different versions where we manipulated the task difficulty in complementary ways. First, some of the experiments (E3, E4, E7, and E8) featured complete feedback information, meaning that participants were informed about the outcome of the

Copyright © 2021
The Authors, some
rights reserved;
exclusive licensee
American Association
for the Advancement
of Science. No claim to
original U.S. Government
Works. Distributed
under a Creative
Commons Attribution
NonCommercial
License 4.0 (CC BY-NC).

¹Laboratoire de Neurosciences Cognitives et Computationnelles, Institut National de la Santé et Recherche Médicale, 29 rue d'Ulm, 75005 Paris, France. ²Ecole normale supérieure, 29 rue d'Ulm, 75005 Paris, France. ³Université de Recherche Paris Sciences et Lettres, 60 rue Mazarine 75006 Paris, France. ⁴University of Minnesota, 1925 4th Street South 4-101, Hanson Hall, Minneapolis, MN, USA.

*Corresponding author. Email: stefano.palminteri@ens.fr

forgone option. This manipulation reduces the difficulty of the task by resolving the uncertainty concerning the counterfactual outcome (that is the outcome of the unchosen option). Complete feedback information has been repeatedly shown to improve learning performance (8, 13). Second, some of the experiments (E5, E6, E7, and E8) featured a block (instead of interleaved) design, meaning that all the trials featuring one context were presented in a row. This manipulation reduces task difficulty by reducing working memory demand and has also been shown to improve learning performance (14). Last, in some of the experiments (E2, E4, E6, and E8), feedback was also provided in the transfer phase, thus allowing us to assess whether and how the values learned during the learning phase can be revised.

Analysis of choice behavior provided support for the counterintuitive prediction and indicated that acquisition error rate in the learning phase is largely dissociable from extrapolation error rate in the transfer phase. Critically (and paradoxically), error rate in the transfer phase was higher when the learning phase was easier. Accordingly, the estimated deviation between the objective values and the subjective values increased in the complete feedback and block design tasks. The deviation was corrected only in the experiments with complete feedback in the transfer phase.

To complement choice rate analysis, we developed a computational model that implements range adaption as a range normalization process, by tracking the maximum and the minimum possible reward in each learning context. Model simulations parsimoniously captured performance in the learning and the transfer phase, including the suboptimal choices induced by range adaptation. Model simulations also allowed us to rule out alternative interpretations of our results offered by two prominent theories in psychology and economics: habit formation and risk aversion (15, 16). Model comparison results were confirmed by checking out-of-sample likelihood as a quantitative measure of goodness of fit.

RESULTS

Experimental protocol

We designed a series of learning and decision-making experiments involving variants of a main task. The main task was composed of two phases: the learning and the transfer phase. During the learning phase, participants were presented with eight abstract pictures, organized in four stable choice contexts. In the learning phase, each choice context featured only two possible outcomes: either 10/0 points or 1/0 point. The outcomes were probabilistic (75 or 25% probability of the nonzero outcomes), and we labeled the choices contexts as a function of the difference in expected value between the most and the least rewarding option: $\Delta EV = 5$ and $\Delta EV = 0.5$ (Fig. 1A). In the subsequent transfer phase, the eight options were rearranged into new choice contexts, where options associated with 10 points were compared to options associated with 1 point [see (7, 10) for similar designs in humans and starlings]. The resulting new four contexts were also labeled as a function of the difference in expected value between the most and the least rewarding option: $\Delta EV = 7.25$, $\Delta EV = 6.75$, $\Delta EV = 2.25$, and $\Delta EV = 1.75$ (Fig. 1B). In our between-participants study, we developed eight different variants of the main paradigm where we manipulated whether we provided trial-by-trial feedback in the transfer phase (with/without), the quantity of information provided at feedback (partial: only the outcome of the chosen option is shown/complete: both outcomes

are shown), and the temporal structure of choice contexts presentation (interleaved: choice contexts appear in a randomized order/block: all trials belonging to the same choice contexts are presented in a row) (Fig. 1C). All the experiments implementing the above-described experimental protocol and reported in the Results section were conducted online ($n = 100$ participants in each experiment); we report in the Supplementary Materials the results concerning a similar experiment realized in the laboratory.

Overall correct response rate

The main dependent variable in our study was the correct response rate, i.e., the proportion of expected value-maximizing choices in the learning and the transfer phase (crucially our task design allowed to identify an expected value-maximizing choice in all choice contexts). In the learning phase, the average correct response rate was significantly higher than chance level 0.5 [0.69 ± 0.16 , $t(799) = 32.49$, $P < 0.0001$, and $d = 1.15$; Fig. 2, A and B]. Replicating previous findings, in the learning phase, we also observed a moderate but significant effect of the choice contexts, where the correct choice rate was higher in the $\Delta EV = 5.0$ compared to the $\Delta EV = 0.5$ contexts (0.71 ± 0.18 versus 0.67 ± 0.18 ; $t(799) = 6.81$, $P < 0.0001$, and $d = 0.24$; Fig. 2C) (6).

Correct response rate was also higher than chance in the transfer phase (0.62 ± 0.17 , $t(799) = 20.29$, $P < 0.0001$, and $d = 0.72$; Fig. 2, D and E), but it was also strongly modulated by the choice context ($F_{2,84,2250.66} = 271.68$, $P < 0.0001$, and $\eta^2 = 0.20$, Huynh-Feldt corrected). In the transfer phase, the $\Delta EV = 1.75$ choice context is of particular interest, because the expected value maximizing option was the least favorable option of a $\Delta EV = 5.0$ context in the learning phase, and conversely, the expected value minimizing option was the most favorable option of a $\Delta EV = 0.5$ context of the learning phase. In other words, a participant relying on expected values calculated on a context-independent scale will prefer the option with $EV = 2.5$ ($EV_{2.5}$ option) compared to with $EV = 0.75$ ($EV_{0.75}$ option). On the other side, a participant encoding the option values on a fully context-dependent manner (which is equivalent to encode the rank between two options in a given context) will perceive the $EV_{2.5}$ option as less favorable compared to the $EV_{0.75}$ option. Therefore, preferences in the $\Delta EV = 1.75$ context are diagnostic of whether values are learned and encoded on an absolute or relative scale. Crucially, in the $\Delta EV = 1.75$ context, we found that participants' average correct choice rate was significantly below chance level (0.42 ± 0.30 , $t(799) = -7.25$, $P < 0.0001$, and $d = -0.26$; Fig. 2F), thus demonstrating that participants express suboptimal preferences in this context, i.e., they do not choose the option with the highest objective expected value.

Between-experiments comparisons: Learning phase

In this section, we analyze the correct response rate as a function of the experimental factors manipulated across the eight experiments (the quantity of feedback information, which could be either partial or complete; the temporal structure of choice context presentation, which could be block or interleaved; and whether feedback was provided in the transfer phase). In the Results section, we report the significant results, but please see Tables 1 and 2 for all results and effect sizes.

First, we analyzed the correct choice rate in the learning phase (Fig. 2B). As expected, increasing feedback information had a significant effect on correct choice rate in the learning phase ($F_{1,792} =$

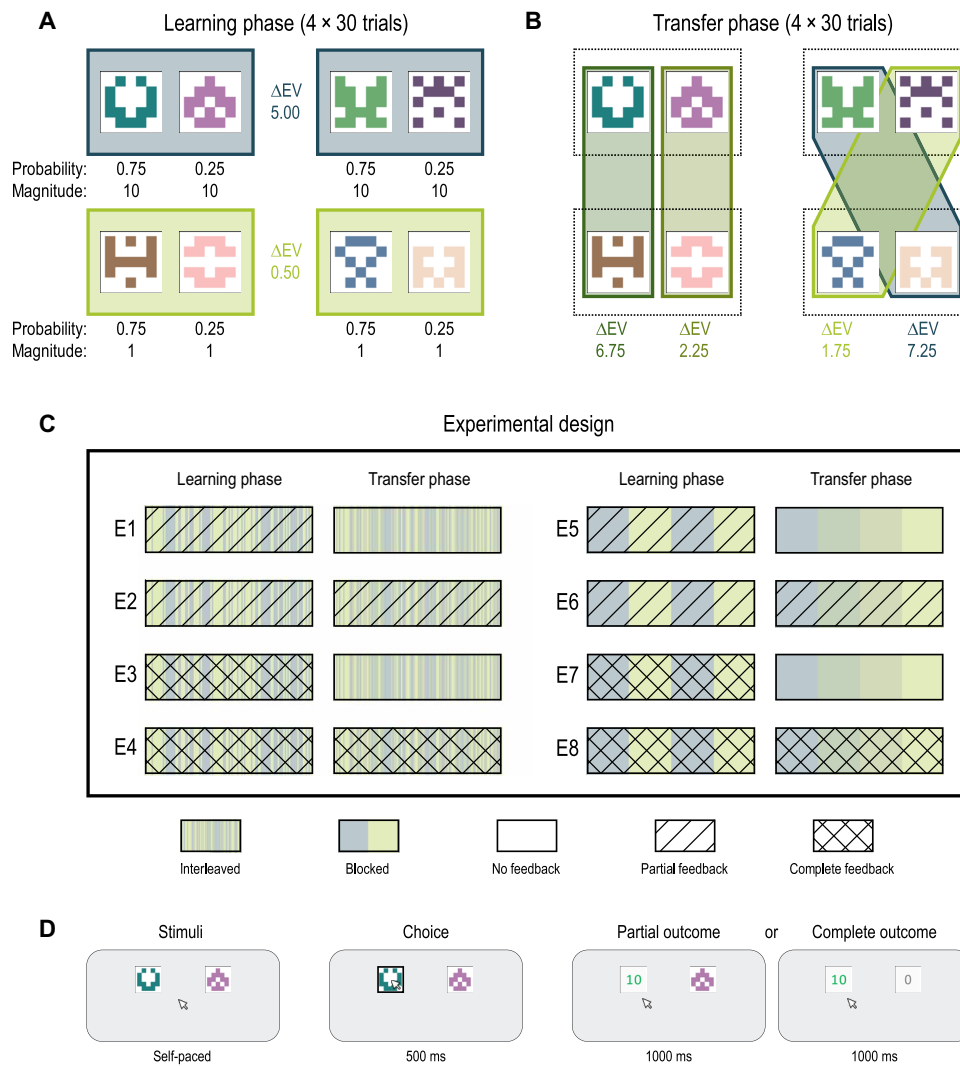


Fig. 1. Experimental design. (A) Choice contexts in the learning phase. During the learning phase, participants were presented with four choice contexts, including high magnitude ($\Delta EV = 5.0$ contexts) and low magnitude ($\Delta EV = 0.5$ contexts). (B) Choice contexts in the transfer phase. The four options were rearranged into four new choice contexts, each involving both the 1- and the 10-point outcome. (C) Experimental design. The eight experiments varied in the temporal arrangement of choice contexts (interleaved or block) and the quantity of feedback in the learning phase (partial or complete) and the transfer phase (without or with feedback). (D) Successive screens of a typical trial (durations are given in milliseconds).

55.57, $P < 0.0001$, and $\eta_p^2 = 0.18$); similarly, performance in the block design experiments was significantly higher ($F_{1,792} = 87.22$, $P < 0.0001$, and $\eta_p^2 = 0.25$). We found a significant interaction between feedback information and task structure, reflecting that the difference of performance between partial and complete feedback was higher in block design ($F_{1,792} = 5.05$, $P = 0.02$, and $\eta_p^2 = 0.02$). We found no other significant main effect nor double or triple interaction (Table 1).

We also analyzed the difference in performance between the $\Delta EV = 5.0$ and $\Delta EV = 0.5$ choice contexts across experiments (Fig. 2C). We found a small but significant effect of temporal structure, the differential being smaller in the block compared to interleaved experiments ($F_{1,792} = 7.71$, $P = 0.006$, and $\eta_p^2 = 0.01$), and found no other significant main effect nor interaction.

To sum up, as expected (8, 13, 14), increasing feedback information and clustering the choice contexts had a beneficial effect on correct response rate in the learning phase. Designing the choice

contexts in blocks also blunted the difference in performance between the small ($\Delta EV = 0.5$) and big ($\Delta EV = 5.0$) magnitude contexts.

Between-experiments comparisons: Transfer phase

We then analyzed the correct choice rate in the transfer phase (Fig. 2E). Expectedly, showing trial-by-trial feedback in the transfer phase led to significantly higher performance ($F_{1,792} = 137.18$, $P < 0.0001$, and $\eta_p^2 = 0.07$). Increasing feedback information from partial to complete also had a significant effect on transfer phase correct choice rate ($F_{1,792} = 22.36$, $P < 0.0001$, and $\eta_p^2 = 0.01$). We found no significant main effect of task structure in the transfer phase (see Table 1).

We found a significant interaction between feedback information and the presence of feedback in the transfer phase, showing that the increase in performance due to the addition of feedback information is higher when both outcomes were displayed during the learning phase ($F_{1,792} = 20.18$, $P < 0.0001$, and $\eta_p^2 = 0.01$). We

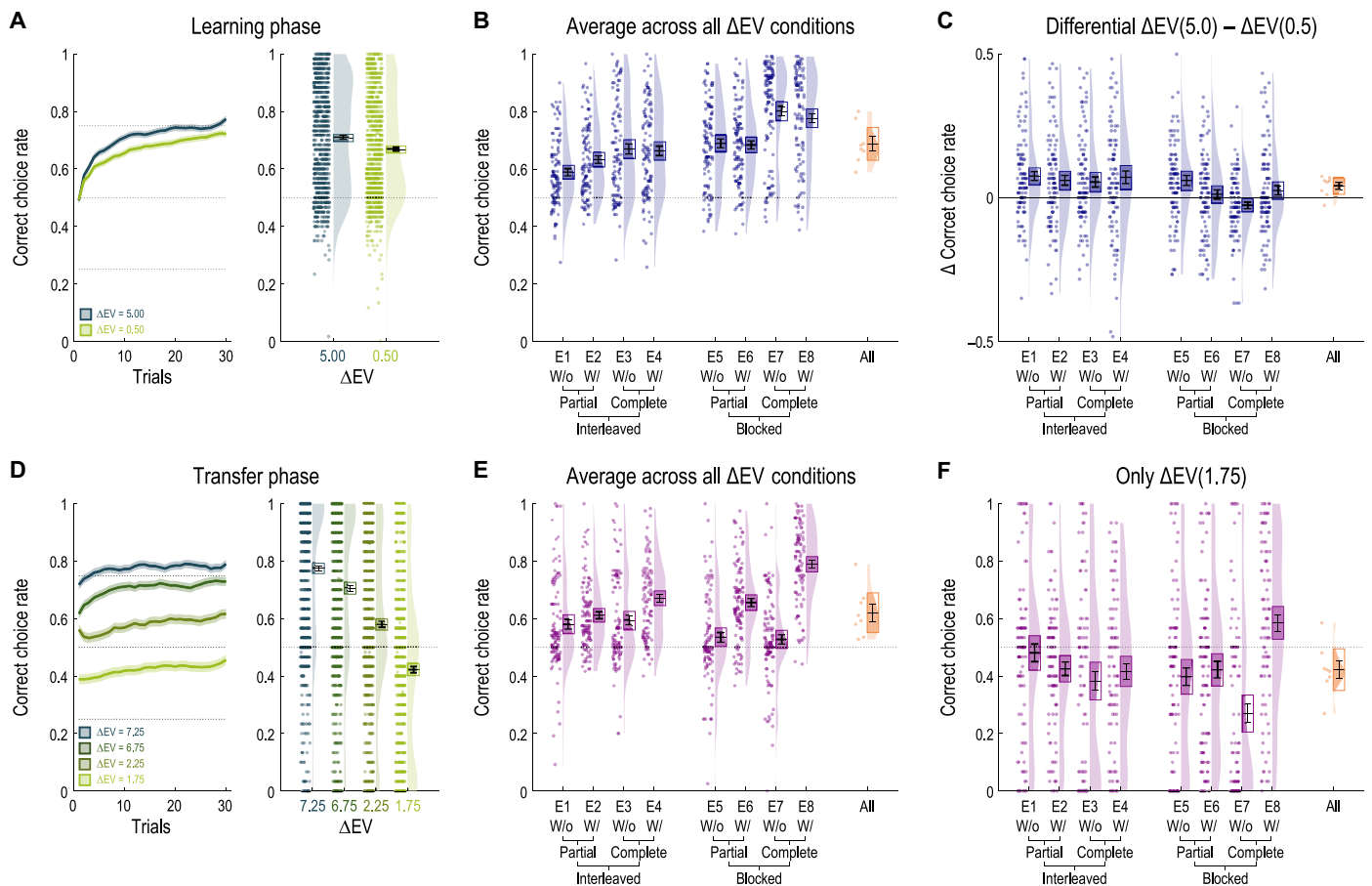


Fig. 2. Behavioral results. (A) Correct choice rate in the learning phase as a function of the choice context ($\Delta EV = 5.0$ or $\Delta EV = 0.5$). Left: Learning curves. Right: Average across all trials ($n = 800$ participants). (B) Average correct response rate in the learning phase per experiment (in blue: one point per participant) and meta-analytical (in orange: one point per experiment). (C) Difference in correct choice rate between the $\Delta EV = 5.0$ and the $\Delta EV = 0.5$ contexts per experiment (in blue: one point per participant) and meta-analytical (in orange: one point per experiment). (D) Correct choice rate in the transfer phase as a function of the choice context ($\Delta EV = 7.25$, $\Delta EV = 6.75$, $\Delta EV = 2.25$, or $\Delta EV = 1.75$). Left: Learning curves. Right: Average across all trials ($n = 800$ participants). (E) Average correct response rate in the transfer phase per experiment (in pink: one point per participant) and meta-analytical (in orange: one point per experiment). (F) Correct choice rate for the $\Delta EV = 1.75$ context only (in pink: one point per participant) and meta-analytical (in orange: one point per experiment). In all panels, points indicate individual average, areas indicate probability density function, boxes indicate 95% confidence interval, and error bars indicate SEM.

also found a significant interaction between transfer feedback and task structure, reflecting that the increase in performance due to the addition of feedback information was even higher in block design ($F_{1,792} = 42.22$, $P < 0.0001$, and $\eta_p^2 = 0.02$). Last, we found a significant triple interaction between feedback information, the presence of feedback in the transfer phase, and task structure ($F_{1,792} = 5.02$, $P = 0.03$, and $\eta_p^2 = 0.003$). We found no other significant double interaction. We also separately analyzed the correct choice rate in the $\Delta EV = 1.75$ context (Fig. 2F). Overall, the statistical effects presented a similar pattern as the correct choice rate across all conditions (see Table 2), indicating that overall correct choice rate and the correct choice rate in the key comparison $\Delta EV = 1.75$ provided a coherent picture. Furthermore, comparing the $\Delta EV = 1.75$ to chance level (0.5) revealed that participants, overall, significantly expressed expected value minimizing preferences in this choice context. Crucially, the lowest correct choice rate was observed in the experiment featuring complete feedback, clustered choice contexts (i.e., block design), and no feedback in the transfer phase [E7; 0.27 ± 0.32 , $t(99) = -7.11$, $P < 0.0001$, and $d = -0.71$]; the addition of feedback in the transfer

phase reversed the situation, because the only experiment where participants expressed expected value maximizing preference was E8 [0.59 ± 0.29 , $t(99) = 2.96$, $P = 0.0038$, and $d = 0.30$].

Between-phase comparison

We found a significant interaction between the phase (learning or transfer) and transfer feedback (without/with) on correct choice rate ($F_{1,792} = 82.30$, $P < 0.0001$, and $\eta_p^2 = 0.09$). This interaction is shown in Fig. 3 and reflects the fact that while adding transfer feedback information had a significant effect on transfer performance ($F_{1,792} = 137.18$, $P < 0.0001$, and $\eta_p^2 = 0.05$; Fig. 3, A and B), it was not sufficient to outperform learning performance (with transfer feedback: learning performance 0.69 ± 0.16 versus transfer performance 0.68 ± 0.15 , $t(399) = 0.89$, $P = 0.38$, and $d = 0.04$; Fig. 3B).

Last, close inspection of the learning curves revealed that in experiments where feedback was not provided in the transfer phase (E1, E3, E5, and E7), correct choice rates (and therefore option preferences) were stationary (Fig. 3, A and B). This observation rules out the

Table 1. Statistical effects of the ANOVA on the choice rate as a function of task factors. LF, learning feedback (complete/partial); TF, transfer feedback (with/without); BE, block effect (block/interleaved); PE, phase effect (learning/transfer), DFn, degrees of freedom numerator, DFd, degrees of freedom denominator, F-val, Fisher value; Diff, value of the difference between the two conditions (main effects only); η_p^2 , portion of variance explained. * $P < 0.05$, ** $P < 0.01$, and *** $P < 0.001$.

	DFn	DFd	Learning performance			Transfer performance			Overall performance					
			F-val	Diff	η_p^2	F-val	Diff	η_p^2	F-val	Diff	η_p^2			
LF - learning feedback; complete > partial	1	792	55.57	***	0.079	0.18	22.36	***	0.050	0.01	61.68	***	0.064	0.11
TF - transfer feedback; with > without	1	792	0.04		0.0021	0.00	137.18	***	0.12	0.07	58.11	***	0.063	0.10
BE - block effect; block > interleaved	1	792	87.22	***	0.099	0.25	1.53		0.013	0.00	46.82	***	0.056	0.08
PE - phase effect; learning > transfer	1	792	-		-	-	-		-	-	103.07	***	0.067	0.12
LF x TF	1	792	2.61			0.01	20.18	***		0.01	3.33			0.01
LF x BE	1	792	5.05	*		0.02	1.66			0.00	5.20	*		0.01
TF x BE	1	792	2.43			0.01	42.22	***		0.02	9.89	**		0.02
LF x PE	1	792	-		-	-	-		-	-	4.97	*		0.01
TF x PE	1	792	-		-	-	-		-	-	82.30	***		0.09
BE x PE	1	792	-		-	-	-		-	-	42.09	***		0.05
LF x TF x BE	1	792	0.55			0.00	5.02	*		0.00	3.65			0.01
LF x TF x PE	1	792	-		-	-	-		-	-	23.37	***		0.03
LF x BE x PE	1	792	-		-	-	-		-	-	0.61			0.00
TF x BE x PE	1	792	-		-	-	-		-	-	40.58	***		0.05
LF x TF x BE x PE	1	792	-		-	-	-		-	-	1.39			0.00

Table 2. Participants' age and correct choice rate as a function of experiments and task factors.

	Experiment 1 (n = 100)		Experiment 2 (n = 100)		Experiment 3 (n = 100)		Experiment 4 (n = 100)		Experiment 5 (n = 100)		Experiment 6 (n = 100)		Experiment 7 (n = 100)		Experiment 8 (n = 100)		Total (n = 800)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Age	30.48	10.70	27.23	8.30	32.01	10.51	31.57	9.80	33.04	10.48	28.46	10.20	28.73	9.89	28.84	9.60	30.06	10.10
% Correct																		
Learning phase	0.59	0.12	0.63	0.13	0.67	0.17	0.66	0.16	0.69	0.15	0.68	0.14	0.80	0.17	0.78	0.16	0.69	0.16
ΔEV = 5.0	0.63	0.16	0.66	0.17	0.70	0.19	0.70	0.20	0.72	0.17	0.69	0.17	0.79	0.19	0.79	0.18	0.71	0.18
ΔEV = 0.5	0.55	0.13	0.60	0.14	0.64	0.19	0.63	0.19	0.66	0.17	0.68	0.14	0.81	0.17	0.76	0.18	0.67	0.18
Transfer phase	0.58	0.17	0.61	0.12	0.59	0.16	0.67	0.13	0.54	0.16	0.66	0.14	0.53	0.16	0.79	0.14	0.62	0.17
ΔEV = 7.25	0.67	0.28	0.76	0.22	0.75	0.29	0.85	0.19	0.66	0.30	0.84	0.18	0.76	0.31	0.93	0.14	0.77	0.26
ΔEV = 6.75	0.64	0.29	0.68	0.26	0.70	0.31	0.81	0.19	0.62	0.32	0.76	0.27	0.55	0.37	0.89	0.16	0.71	0.30
ΔEV = 2.25	0.54	0.27	0.58	0.19	0.54	0.34	0.61	0.28	0.47	0.32	0.60	0.18	0.54	0.36	0.76	0.22	0.58	0.29
ΔEV = 1.75	0.48	0.30	0.43	0.23	0.38	0.33	0.42	0.27	0.40	0.31	0.42	0.28	0.27	0.32	0.59	0.29	0.42	0.30

possibility that reduced performance in the transfer phase was induced by progressively forgetting the values of the options (in which case we should have observed a nonstationary and decreasing correct response rate).

In conclusion, comparison between the learning and the transfer phase reveals two interrelated and intriguing facts: (i) Despite the fact that the transfer phase happens immediately after an extensive learning phase, performance is, if anything, lower compared to the

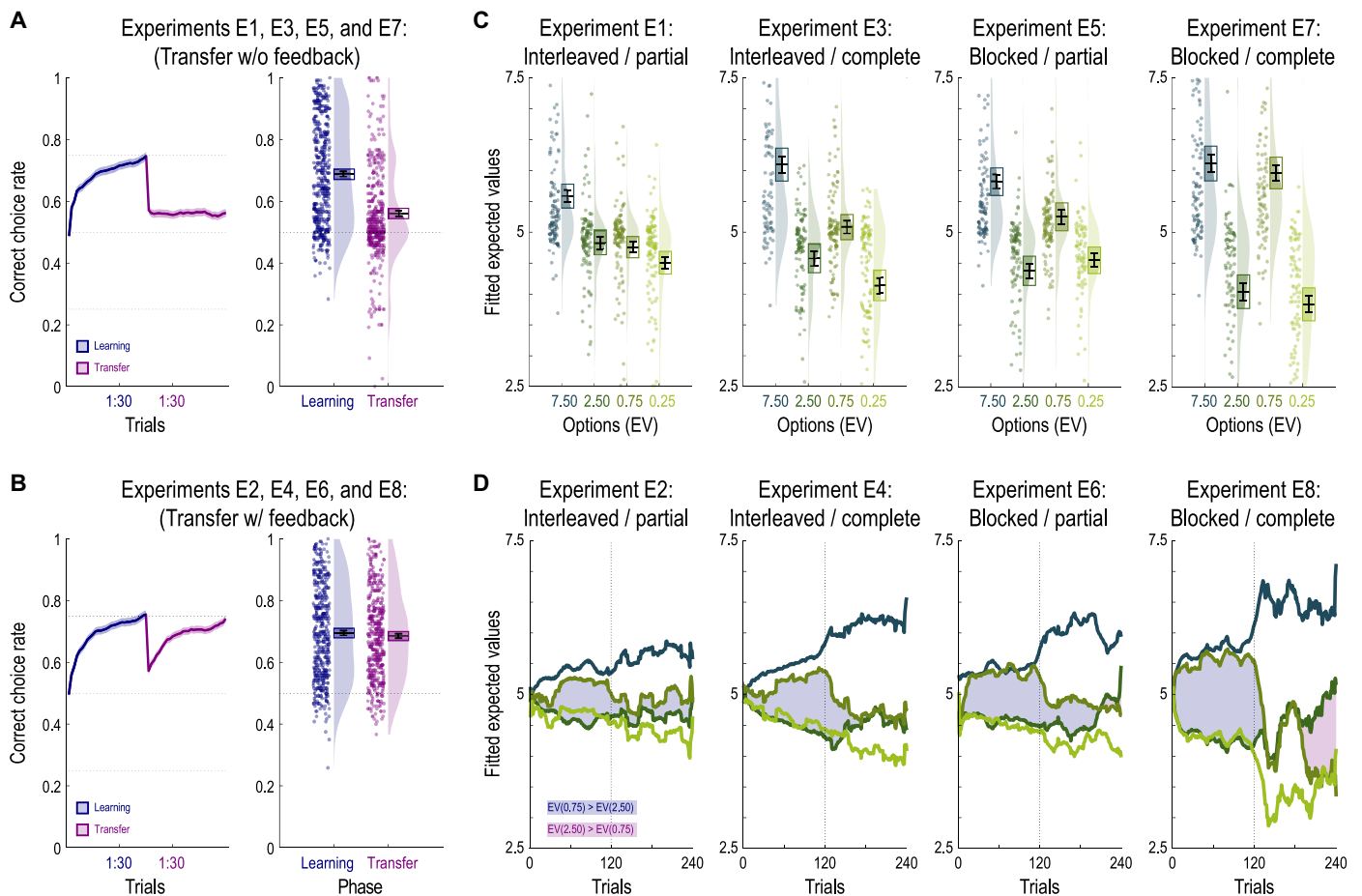


Fig. 3. Learning versus transfer phase comparison and inferred option values. (A and B) Average response rate in the learning (blue) and transfer (pink) phase for experiments without (A) and with (B) trial-by-trial transfer feedback. Left: Learning curves. Right: average across all trials. (C) Average inferred option values for the experiments without trial-by-trial transfer feedback (E1, E3, E5, and E7). (D) Trial-by-trial inferred option values for the experiments with trial-by-trial transfer feedback (E2, E4, E6, and E8). In all panels, points indicate individual average, areas indicate probability density function, boxes indicate 95% confidence interval, and error bars indicate SEM.

learning phase; (ii) factors that improve performance (by intrinsically or extrinsically reducing task difficulty) in the learning phase have either no (feedback information) or a negative (task structure) impact on the transfer phase performance.

Inferred option values

To visualize and quantify how much observed choices deviate from the experimentally determined true option values, we treated the four possible subjective option values as free parameters. More precisely, we initialized all subjective option values to their true values (accordingly, we labeled the four possible options as follows: $EV_{7.5}$, $EV_{2.5}$, $EV_{0.75}$, and $EV_{0.25}$), and fitted their values, as if they were free parameters, by maximizing the likelihood of the observed choices. We modeled choices using the logistic function (for example, options $EV_{2.5}$ and $EV_{0.75}$)

$$P(EV_{2.5}) = \frac{1}{1 + e^{(V(EV_{0.75}) - V(EV_{2.5}))}} \quad (1)$$

So that if a participant chose indifferently between the $EV_{2.5}$ and the $EV_{0.75}$ option, their fitted values would be very similar: $V(EV_{2.5}) \approx V(EV_{0.75})$. Conversely, a participant with a sharp (optimal) preference

for $EV_{2.5}$ over $EV_{0.75}$ would have different fitted values: $V(EV_{2.5}) > V(EV_{0.75})$. In a first step, in the experiments where feedback was not provided in the transfer phase (E1, E3, E5, and E7), we optimized a set of subjective values per participant.

Consistent with the correct choice rate results described above, we found a value inversion of the two intermediary options ($EV_{2.5} = 4.46 \pm 1.2$, $EV_{0.75} = 5.26 \pm 1.2$, $t(399) = -7.82$, $P < 0.0001$, and $d = -0.67$), which were paired in the $\Delta EV = 1.75$ context (Fig. 3C). The differential was also strongly modulated across experiments ($F_{3,396} = 18.9$, $P < 0.0001$, and $\eta_p^2 = 0.13$; Fig. 3C) and reached its highest value in E7 (complete feedback and block design).

As a second step, in the experiments where feedback was provided in the transfer phase (E2, E4, E6, and E8), we optimized a set of subjective values per trial. This fit allows us to estimate the trial-by-trial evolution of the subjective values over task time. The results of this analysis clearly show that suboptimal preferences progressively arise during the learning phase and disappear during the transfer phase (Fig. 3D). However, the suboptimal preference was completely corrected only in E8 (complete feedback and block design) by the end of the transfer phase.

The analysis of inferred option values clearly confirms that participants' choices do not follow the true underlying monotonic

ordering of the objective option values. Furthermore, it also clearly illustrates that in choice contexts that are supposed to facilitate the learning of the option values (complete feedback and block design), the deviation from monotonic ordering, at least at the beginning of transfer phase, is paradoxically greater. Monotonicity was fully restored only in E8, where complete feedback was provided in the transfer phase.

Computational formalization of the behavioral results

To formalize context-dependent reinforcement learning and account for the behavioral results, we designed a modified version of a standard model, where option-dependent Q values are learnt from a range-adapted reward term. In the present study, we implemented range adaptation as a range normalization process, which is one among other possible implementations (17). At each trial *t*, the relative reward, $R_{RAN,t}$ is calculated as follows

$$R_{RAN,t} = \frac{R_{OBJ,t} - R_{MIN,t}(s)}{R_{MAX,t}(s) - R_{MIN,t}(s) + 1} \tag{2}$$

where *s* is the decision context (i.e., a combination of options) and R_{MAX} and R_{MIN} are state-level variables, initialized to 0 and updated at each trial *t* if the outcome is greater (R_{MAX}) or smaller (R_{MIN}) than its current value. In the denominator “+1” is added, in part, to prevent division by zero (even if this could also easily be avoided by adding a simple conditional rule) and, mainly, to make the model nest a simple Q-learning model. $R_{OBJ,t}$ was the objective obtained reward, which in our main experiments could take the following values: 0, +1, and +10 points. Thus, because in our task, the minimum possible outcome is always zero, $R_{MIN,t}$ update was omitted while fitting the first eight experiments (but included in a ninth dataset analyzed below). On the other side, R_{MAX} will converge to the maximum outcome value in each decision context, which in our task is either 1 or 10 points. In the first trial, $R_{RAN} = R_{OBJ}$ [because $R_{MAX,0}(s) = 0$], and in later trials, it is progressively normalized between 0 and 1 as the range value $R_{MAX}(s)$ converges to its true value. We refer to this model as the RANGE model, and we compared it to a benchmark model (ABSOLUTE) that updates option values based the objective reward values (note that the ABSOLUTE is nested within the RANGE model).

For each model, we estimated the optimal free parameters by likelihood maximization. We used the out-of-sample likelihood to compare goodness of fit and parsimony of the different models (Table 3). To calculate the out-of-sample likelihood in the learning phase, we performed the optimization on half of the trials (one $\Delta EV = 5.0$ and one $\Delta EV = 0.5$ context) in the learning phase, and the best-fitting parameters in this first set were used to predict choices in the remaining half of trials. In the learning phase, we found that the RANGE model significantly outperformed the ABSOLUTE model [out-of-sample log-likelihood LL_{RAN} versus LL_{ABS} , $t(799) = 6.89$, $P < 0.0001$, and $d = 0.24$; Table 3]. To calculate the out-of-sample likelihood in the transfer phase, we fitted the parameters on all trials of the learning phase, and the best-fitting parameters were used to predict choices in the transfer phase. Thus, the resulting likelihood is not only out-of-sample but also cross-learning phase. This analysis revealed that the RANGE model outperformed the ABSOLUTE model [out-of-sample log-likelihood LL_{RAN} versus LL_{ABS} , $t(799) = 8.56$, $P < 0.0001$, and $d = 0.30$].

To study the behaviors of our computational model and assess the behavioral reasons underlying the out-of-sample likelihood

Table 3. Quantitative model comparing. Values reported here represent out-of-sample log-likelihood after twofold cross-validation. Comparison to the RANGE model: *** $P < 0.001$; ** $P < 0.01$; ⁵ $P < 0.08$.

Model	Learning phase	Transfer phase
ABSOLUTE	$-42.74 \pm 1.27^{***}$	$-161.19 \pm 11.41^{***}$
RANGE	-37.72 ± 0.96	-96.79 ± 4.79
HABIT	-36.68 ± 0.91	-104.62 ± 6.01^5
UTILITY	$-36.31 \pm 0.53^{***}$	$-104.94 \pm 5.24^{**}$

results, we simulated the two models (using the individual best-fitting parameters) (18). In the learning phase, only the RANGE model managed to reproduce the observed correct choice rate. Specifically, the ABSOLUTE model predicts very poor performance in the $\Delta EV = 0.5$ context [ABSOLUTE versus data, $t(799) = -16.90$, $P < 0.0001$, and $d = 0.60$; RANGE versus data, $t(799) = -1.79$, $P = 0.07$, and $d = -0.06$; Fig. 4A].

In the transfer phase, and particularly in the $\Delta EV = 1.75$ context, only the RANGE model manages to account for the observed correct choice rate, while the ABSOLUTE model fails (ABSOLUTE versus data, $t(799) = 13.20$, $P < 0.0001$, and $d = 0.47$; RANGE versus data, $t(799) = 0.36$, $P = 0.72$, and $d = 0.01$; Fig. 4, C and D). In general, the ABSOLUTE model tends to overestimate the correct choice rate in the transfer phase.

In addition to looking at the qualitative choice patterns, we also inferred the subjective option values from the RANGE model simulations. The RANGE model was able to perfectly reproduce the subjective option value pattern that we observed in the data, specifically the violation of monotonic ranking (Fig. 4E) and their temporal dynamics (Fig. 4F).

Ruling out habit formation

One of the distinguishing behavioral signatures of the RANGE model compared the ABSOLUTE one is the preference for the sub-optimal option in the $\Delta EV = 1.75$ context. Because the optimal option in the $\Delta EV = 1.75$ context is not often chosen during the learning phase (where it is locally suboptimal), it could be argued that this result arises from taking decisions based on a weighted average between their absolute values and past choice propensity (a sort of habituation or choice trace). To rule out this interpretation, we fitted and simulated a version of a HABIT model, which takes decisions based on a weighted sum of the absolute Q values and a habitual choice trace (16, 19). The habitual choice trace component is updated with an additional learning rate parameter that gives a bonus to the selected action. Decisions are taken comparing option-specific decision weights D_t

$$D_t(s, c) = (1 - \omega) * Q_t(s, c) + \omega * H_t(s, c) \tag{3}$$

where at each trial *t*, state *s*, and chosen option *c*, ω is the arbiter, *Q* is the absolute Q value, and *H* is the habitual choice trace component. The weight ω is fitted as an additional parameter (for $\omega = 0$, the model reduces to the ABSOLUTE model) and governs the relative influence of each controller.

We found that the HABIT model, similarly to the ABSOLUTE model, fails to perfectly match the participants’ behavior, especially in the $\Delta EV = 0.5$ and $\Delta EV = 1.75$ contexts (Fig. 5A). In the learning

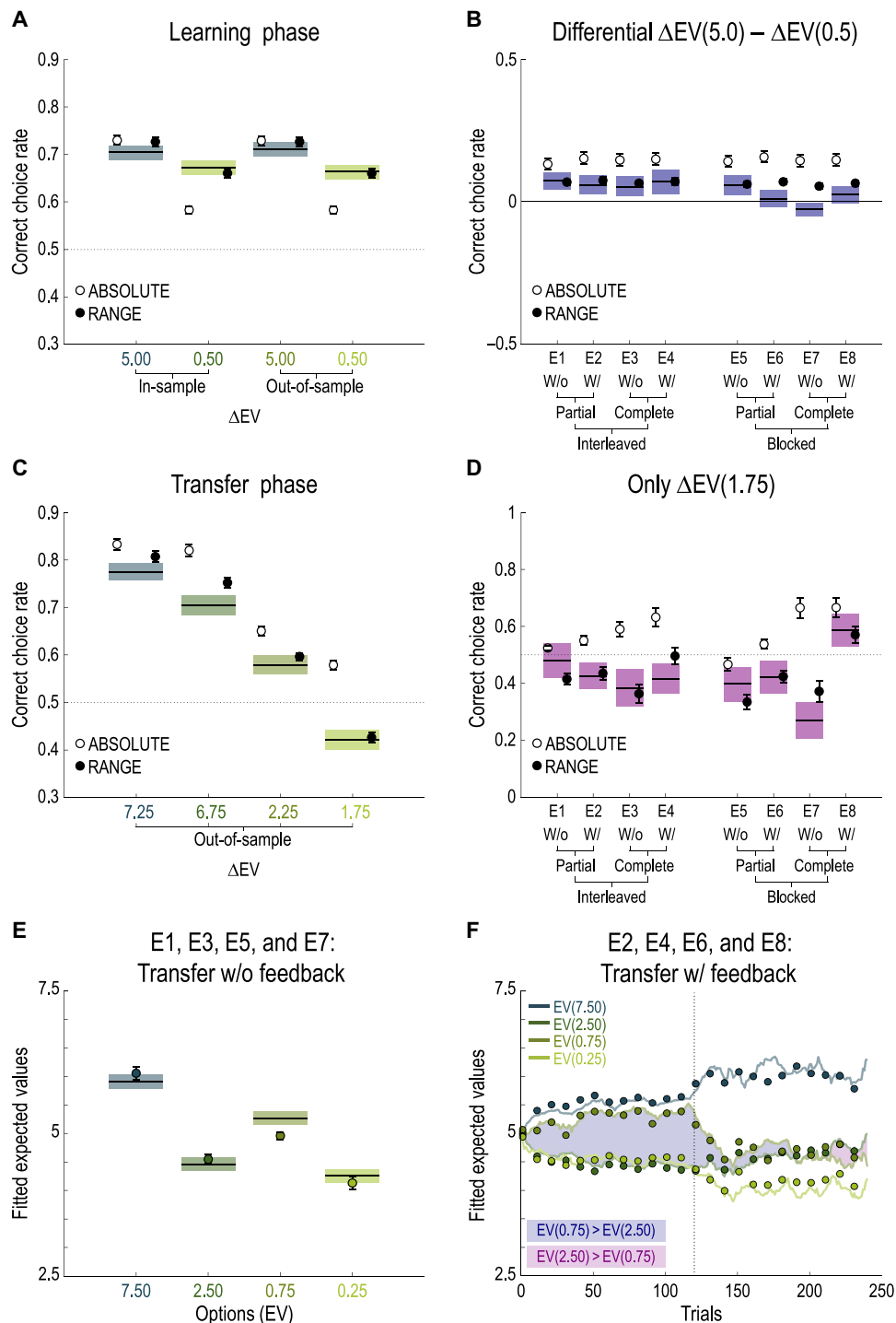


Fig. 4. Model comparison. Model simulations of the ABSOLUTE and the RANGE models (dots) superimposed on the behavioral data (boxes indicated the mean and 95% confidence interval) in each context. **(A)** Simulated data in the learning phase were obtained with the parameters fitted in half the data (the $\Delta EV = 5.0$ and the $\Delta EV = 0.5$ contexts on the leftmost part of the panel) of the learning phase. **(B)** Data and simulations of the correct choice rate differential between high-magnitude ($\Delta EV = 5.0$) and low-magnitude ($\Delta EV = 0.5$) contexts. **(C)** Simulated data in the transfer phase were obtained with the parameters fitted in all the contexts of the learning phase. **(D)** Data and simulations in the context $\Delta EV = 1.75$ only. **(E)** Average inferred option values for the behavioral data and simulated data (colored dots: RANGE model) for the experiments without trial-by-trial feedback in the transfer phase. **(F)** Trial-by-trial inferred option values for the behavioral data and simulated data (colored dots: RANGE model) for the experiments with trial-by-trial feedback in the transfer phase. As in Fig. 3D, here, the curves indicate trial-by-trial fit of each inferred option value.

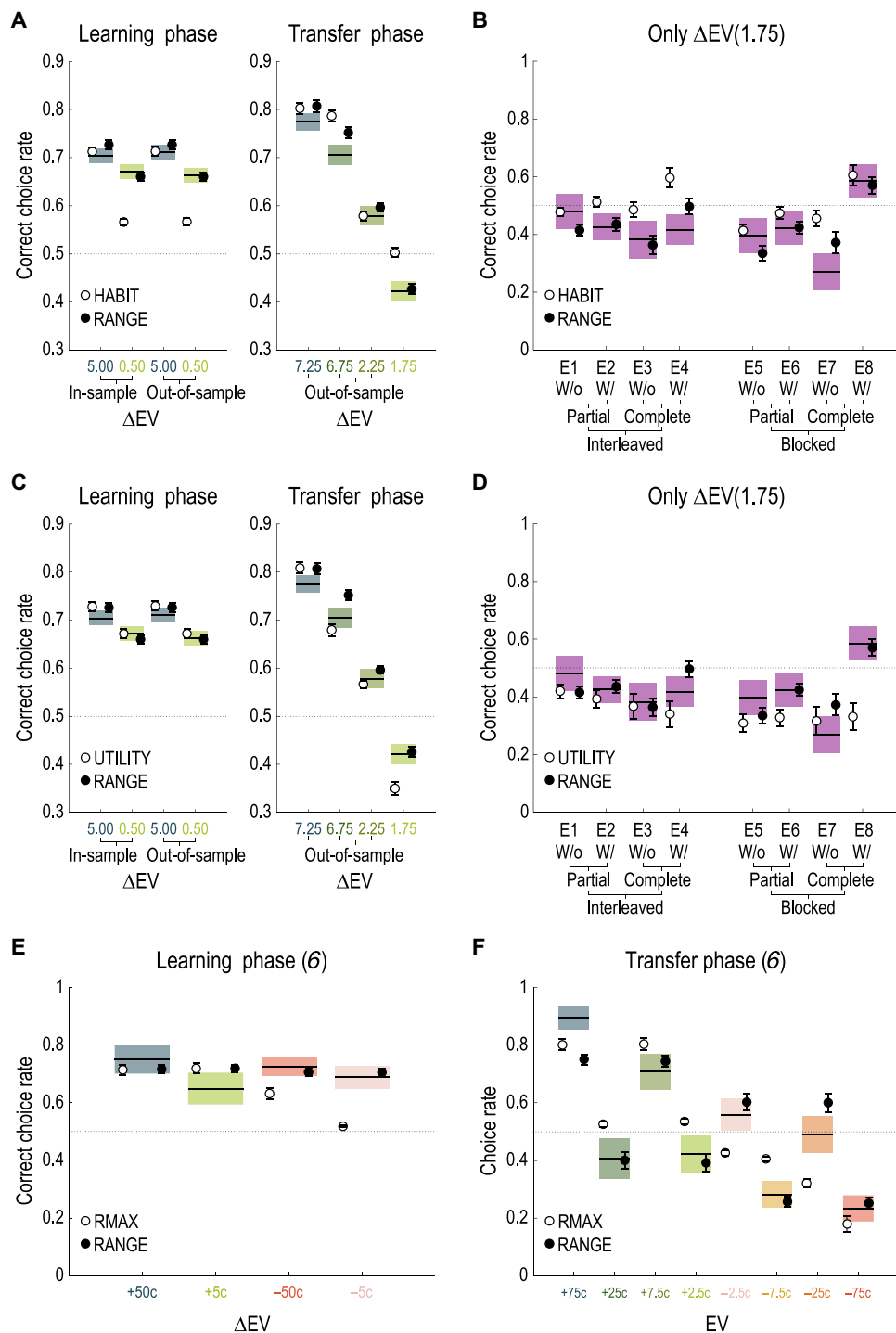


Fig. 5. Ruling out alternative models and validation in an additional experiment. Model simulations of the HABIT, the UTILITY, and the RANGE models (dots) over the behavioral data (mean and 95% confidence interval) in each context. **(A and C)** Simulated data in the learning phase were obtained with the parameters fitted in half the data (the $\Delta EV = 5.0$ and the $\Delta EV = 0.5$ contexts on the leftmost part of the panel) of the learning phase. Simulated data in the transfer phase were obtained with the parameters fitted in all the contexts of the learning phase. **(B and D)** Data and simulations in the context $\Delta EV = 1.75$ only. **(E and F)** Behavioral data from Bavard *et al.* (6). Comparing the full RANGE model to its simplified version RMAX in the learning phase (correct choice rate per choice context) and in the transfer test (choice rate per option). This study included both gain-related contexts (with +1€, +0.1€, and 0.0€ as possible outcomes) and loss-related contexts (with -1€, -0.1€, and 0.0€ as possible outcomes) in the learning phase. Choice rates in the transfer phase are ordered as a function of decreasing expected value as in (6).

phase, the addition of a habitual component is not enough to cope for the difference in option values, and therefore, the model simulations in the transfer phase fail to match the observed choice pattern (Fig. 5B). This is because the HABIT model encodes values on an absolute scale and does not manage to develop a strong preference for the correct response in the $\Delta EV = 0.5$ context, in the first place (Fig. 5A). Thus, it does not carry a choice trace strong enough to overcome the absolute value of the correct response in the $\Delta EV = 1.75$ context (Fig. 5B; fig. S2, A and B; and Table 3). Quantitative model comparison between the RANGE and the HABIT model capacity to predict the transfer phase choices, numerically favored the RANGE model reaching marginal statistical significance [out-of-sample log-likelihood LL_{RAN} versus LL_{HAB} , $t(799) = 1.77$, $P = 0.07$, and $d = 0.05$; Table 3]. To summarize, a model assuming absolute value encoding coupled with a habitual component could not fully explain observed choices in both the learning and transfer phase.

Ruling out diminishing marginal utility

One of the distinguishing behavioral signatures of the RANGE model is that it predicts very similar correct choice rates in the $\Delta EV = 5.00$ and the $\Delta EV = 0.50$ contexts compared to the behavioral data, while both the ABSOLUTE and the HABIT predict a huge drop in performance in the $\Delta EV = 0.50$ that directly stems from its small difference in expected value. It could be argued that this result arises from the fact that expected utilities (and not expected values) are learned in our task. Specifically, a diminishing marginal utility parameter would blunt differences in outcome magnitudes and would suppose that choices are made by comparing outcome probabilities. The process could also explain the preference for the suboptimal option in the $\Delta EV = 1.75$ context, because the optimal option in the $\Delta EV = 1.75$ context is rewarded (10 points) only the 25% of the time, while the suboptimal option is rewarded (1 point) 75% of the time. To rule out this interpretation, we fitted and simulated a UTILITY model, which updates Q value-based reward utilities calculated from absolute reward as follows

$$R_{UTI,t} = (R_{OBJ,t})^{\nu} \quad (4)$$

where the exponent ν is the utility parameter ($0 < \nu < 1$, for $\nu = 1$ the model reduces to the ABSOLUTE model). We found an empirical average value of $\nu = 0.32$ (± 0.01 SEM).

We found that the UTILITY model, similarly to the RANGE model, captures quite well the participants' behavior in the learning phase (Fig. 5C). However, concerning the transfer phase (especially the $\Delta EV = 1.75$ context), it fails to capture the observed pattern (Fig. 5, C and D). Additional analyses suggest that this is specifically driven by the experiments where the feedback was provided during the transfer phase (Fig. 5D). The static nature of the UTILITY fails to match the fact that the preferences in the $\Delta EV = 1.75$ context can be reversed by providing complete feedback (fig. S2, C and D). Quantitative model comparison showed that the RANGE model also outperformed the UTILITY model in predicting the transfer phase choices [out-of-sample log-likelihood LL_{RAN} versus LL_{UTI} , $t(799) = 3.21$, $P = 0.001$, and $d = 0.06$; Table 3]. To summarize, a model assuming diminishing marginal utilities could not fully explain observed choices in the transfer phase.

Suboptimality of range adaptation in our task

The RANGE model is computationally more complex compared to the ABSOLUTE model, as it presents additional internal variables

(R_{MAX} and R_{MIN}), which are learnt with a dedicated parameter. Here, we wanted to assess whether this additional computational complexity really paid off in our task.

We split the participants according to the sign of out-of-sample likelihood difference between the RANGE and the ABSOLUTE model: If positive, then the RANGE model better explains the participant's data ($RAN > ABS$), if negative, the ABSOLUTE model does ($ABS > RAN$). Reflecting our overall model comparison result, we found more participants in the $RAN > ABS$, compared to the $ABS > RAN$ category ($n = 545$ versus $n = 255$).

We found no main effect of winning model on overall (both phases) performance [$F_{1,798} = 0.03$, $P = 0.87$, and $\eta_p^2 = 0$]. We found that while RANGE encoding is beneficial and allows for better performances in the learning phase, it leads to the worst performance in the transfer phase [$F_{1,798} = 187.3$, $P < 0.0001$, and $\eta_p^2 = 0.19$; Fig. 6A]. In other terms, in our task, it seems that the learning phase and the transfer phase are playing the game tug of war: When performance is pulled in favor of the learning phase, this will be at the cost of the transfer phase (and vice versa).

A second question is whether overall in our study, behaving as a RANGE model turns out to be economically advantageous. To answer this question, we compared the final monetary payoff in the real data, following the simulations using the participant-level best-fitting parameters. Consistently with the task design, we found that the monetary outcome was higher in the transfer phase than in the learning phase [transfer gains $M = 2.16 \pm 0.54$, learning gains $M = 1.99 \pm 0.35$, $t(799) = 8.71$, $P < 0.0001$, and $d = 0.31$]. Crucially, we found that the simulation of the RANGE model induces significantly lower monetary earnings (ABSOLUTE versus RANGE, $t(799) = 19.39$, $P < 0.0001$, and $d = 0.69$; Fig. 6B). This result indicates that despite being locally adaptive (in the learning phase), in our task, range adaptation is economically disadvantageous, thus supporting the idea that it is the consequence of an automatic, uncontrolled process.

Validation of range adaptation in previous dataset

The first eight experiments only featured positive outcomes (in addition to 0). Because, in our model, the state-level variables (R_{MAX} and R_{MIN}) are initialized to 0, R_{MAX} converges to the maximum outcome value in each choice context, while R_{MIN} remains 0 in every trial and choice context. This setup is therefore not ideal to test the full normalization rule that we are proposing here. To obviate this limitation, we reanalyzed a ninth dataset ($n = 60$) from a previously published study on a related topic (6). Crucially, in addition to manipulating outcome magnitude ("10c" versus "1€", similar to our learning phase), this study also manipulated the valence of the outcomes (gain versus loss). This latter manipulation allows to assess situations where the value of R_{MIN} should change and converge to negative values, thus allowing us to compare the full range normalization rule to its simplified version

$$\frac{R_{OBJ} - R_{MIN}}{R_{MAX} - R_{MIN}} \text{ versus } \frac{R_{OBJ}}{R_{MAX}}$$

We note that in this ninth dataset outcomes can take both negative and positive values: -1€ , -0.1€ , 0.0€ , $+0.1\text{€}$, and $+1.0\text{€}$. We later refer to the simplified version of the model as the RMAX model. Model simulations show that while the RMAX model can capture the learning and transfer phase patterns for the gain-related options, it fails to do so for the loss-related options (Fig. 5, E and F). In the

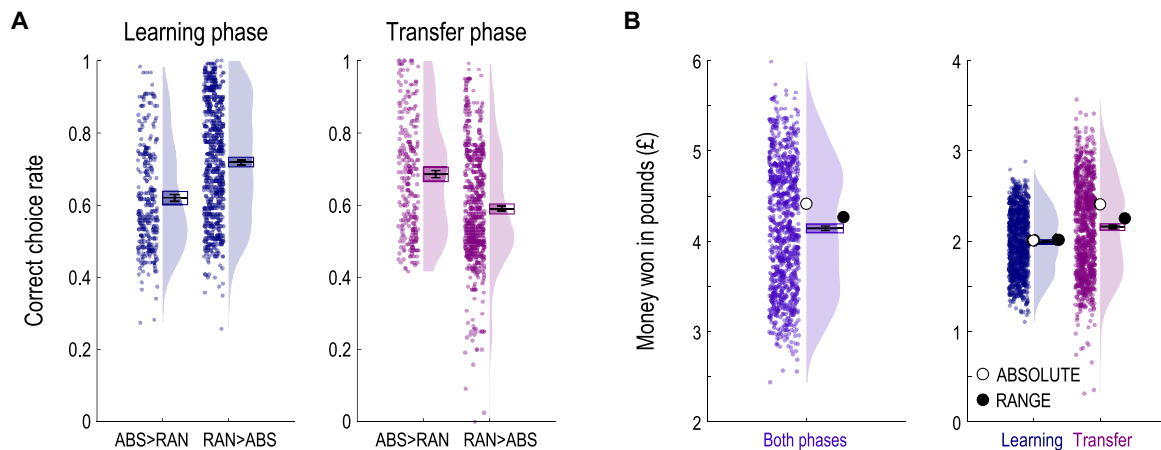


Fig. 6. The financial cost of relative value learning. (A) Correct choice rate in the learning phase (blue) and the transfer phase (pink). Participants are split as a function of the individual difference in out-of-sample log-likelihood between the ABSOLUTE and the RANGE models. ABS > RAN participants are better explained by the ABS model (positive difference, $n = 255$). RAN > ABS participants are better explained by the RAN model (negative difference, $n = 545$). (B) Actual and simulated money won in pounds over the whole task (purple), the learning phase only (blue), and the transfer phase only (pink). Points indicate individual participants, areas indicate probability density function, boxes indicate confidence interval, and error bars indicate SEM. Dots indicate model simulations of ABSOLUTE (white) and RANGE (black) models.

loss-related contexts (where the maximum possible outcome is 0) outcome value normalization can only rely on R_{MIN} . Because the RMAX model does not take into account R_{MIN} , it is doomed to encode loss-related outcomes on an objective scale.

On the other hand, by updating both R_{MAX} in the gain contexts and R_{MIN} in the loss contexts, the RANGE model can normalize outcomes in all contexts and is able to match participants' choice preferences concerning both loss-related and gain-related options in the learning and the transfer phases (Fig. 5, E and F). To conclude, this final analysis is consistent with the idea that range adaptation takes the form of a range normalization rule, which takes into account both the maximum and the minimum possible outcomes.

DISCUSSION

In the present paper, we investigated context-dependent reinforcement learning, more specifically range adaptation, in a large cohort of human participants tested online over eight different variants of a behavioral task. Building on previous studies of context-dependent learning, the core idea of the task is to juxtapose an initial learning phase with fixed pairs of options (featuring either small or large outcomes) to a subsequent transfer phase where options are rearranged in new pairs (mixing up small and large outcomes) (6, 7, 10). In some experiments, we directly reduced task difficulty by reducing outcome uncertainty by providing complete feedback. In some experiments, we indirectly modulated task difficulty by clustering in time the trials of a given contexts, therefore reducing working memory demand. Last, in some experiments, feedback was also provided in the transfer phase.

Behavioral findings

As expected, correct choice rate in the learning phase was higher when the feedback was complete, which indicates that participants integrated the outcome of the forgone option when it is presented (8, 14). Also expectedly, in the learning phase, participants displayed a higher correct choice rate when the trials of a given context were blocked together, indicating that reducing working memory demands

facilitate learning (15). Replicating previous findings, we also found that, overall, correct response rate was slightly but significantly higher in the big magnitude contexts ($\Delta EV = 5.0$), but the difference was much smaller compared to what one would expect assuming unbiased value learning and representation [as showed by the ABSOLUTE model simulations (6)]: a pattern consistent with a partial range adaptation. The outcome magnitude-induced difference in correct choice rate was significantly smaller and not different from zero in block experiments (full adaptation), thus providing a first suggestion that reducing task difficulty increases range adaptation. Despite learning phase performance being fully consistent with our hypothesis, the crucial evidence comes from the results of the transfer phase. Overall correct response rate pattern in the transfer phase did not follow that of the learning phase. Complete feedback and block design factors have no direct beneficial effect on transfer phase performance. In fact, the worst possible transfer phase performance was obtained in a complete feedback and block experiment. This was particularly notable in the $\Delta EV = 1.75$ condition, where participants significantly preferred the suboptimal option and, again, the worst score was obtained in a complete feedback and block design experiment. Crucially, we ruled out that the comparably low performance in the transfer phase was due to having forgotten the value of the options. Because the transfer phase is, by definition, after the learning phase, although very unlikely (the two phases were only a few seconds apart), it is conceivable that a drop in performance is due to the progressive forgetting of the option values. Two features of the correct choice rate curves allowed to reject this interpretation: (i) Correct choice rate abruptly decreases just after the learning phase; (ii) when feedback is not provided, the choice rate remains perfectly stable with no sign of regression to chance level. On the other side, i.e., when feedback was provided in the transfer phase, the correct choice rate increased to reach (on average) the level reached at the end of the learning phase. The results are therefore consistent with the idea that in the transfer phase, participants express context-dependent option values acquired during the learning phase, which entails a first counterintuitive phenomenon: Even if the transfer phase is performed immediately after the learning

phase, the correct choice rate drops. This is due to the rearrangement of the options in new choice contexts, where options that were previously optimal choices (in the small magnitude contexts) become suboptimal choices. We also observed a second counterintuitive phenomenon: Factors that increase performance during the learning phase (i.e., increasing feedback information and reducing working memory load) paradoxically further undermined transfer phase correct choice rate. The conclusions based on these behavioral observations were confirmed by inferring the most plausible option values based on the observed choices, where we could compare the objective ranking of the options to their subjective estimation. The only experiment where we observed an almost monotonic ranking was the partial feedback/interleaved experiment, even if we observed no significant difference between the $EV = 2.5$ and the $EV = 0.75$ options. In all the other experiments, the $EV = 0.75$ option was valued more compared to the $EV = 2.5$ option, with the highest difference observed in the complete feedback/block design. Thus, in notable opposition with the almost universally shared intuition that reducing task difficulty should lead to more accurate subjective estimates; here, we present a clear instance where the opposite is true.

Computational mechanisms

The observed behavioral results were satisfactorily captured by a parsimonious model (the RANGE model) that instantiated a dynamic range normalization process. Specifically, the RANGE model learns in parallel context-dependent variables (R_{MAX} and R_{MIN}) that are used to normalize the outcomes. The variables R_{MAX} and R_{MIN} are learnt incrementally, and the speed determines the extent of the normalization, leading to partial or full range adaptation as a function of a dedicated free parameter: the contextual learning rate. Developing a new model was necessary, as previous models of context-dependent reinforcement learning did not include range adaptation and focused on different dimensions of context dependence (reference point centering and outcome comparison) (7, 8). The model also represents an improvement over a previous study where we instantiated partial range adaptation assuming a perfect and innate knowledge about the outcome ranges and a static hybridization between relative and absolute outcome values (6).

One limitation is that in the present formulation R_{MAX} and R_{MIN} can only grow and decrease, respectively. This is a feature that is well suited for our task, which features static contingencies, but may not correspond to many other laboratory-based and real-life situations, where the outcome range can drift over time. This limitation could be overcome by assuming, for example, that R_{MAX} is also updated at a smaller rate when the observed outcome is smaller than the current R_{MAX} (the opposite could be true for R_{MIN}). Last, we note that our model applied to the main eight experiments (where R_{MIN} was irrelevant) can also be seen as a special case of a divisive normalization process [temporal normalization (20)]. To verify the relevance of the full range normalization rule, we re-analyzed a previous dataset involving negative outcomes, where we were able to show that both the R_{MAX} and R_{MIN} were important to explain the full spectrum of the behavioral results. However, we acknowledge that additional functional forms of normalization could and should be considered in future studies to settle the issue of the exact algorithmic implementation of outcome normalization. Last, it is worth noting that range normalization has been shown to perform poorly in explaining context-dependent decision-making in other (i.e., not reinforcement learning) paradigms (17, 21, 22),

opening to the possibility that the normalization algorithm is different in experience-based and description-based choices. Future research contrasting different outcome ranges and multiple-option tasks are required to firmly determine which functional forms of normalization are better suited for both experience-based and description-based choices (23).

We compared and ruled out another plausible computational interpretation derived from learning theory (24, 25). Specifically, we considered a habit formation model (16). We reasoned that our transfer phase results (and particularly the value inversion in the $\Delta EV = 1.75$ context) could derive from the participants choosing on the basis of a weighted average between objective values and past choice propensities. In the learning phase, the suboptimal option in the $\Delta EV = 1.75$ context ($EV = 0.75$) was chosen more frequently than the optimal option ($EV = 2.5$). However, model simulations showed that the HABIT model was not capable to explain the observed pattern. In the learning phase, the HABIT model, just like the ABSOLUTE model, did not develop a preference for the $EV = 0.75$ option strong enough to generate a habitual trace sufficient to explain the transfer phase pattern. Beyond model simulation comparisons, we believe that this interpretation could have been rejected on the basis of a priori arguments. The HABIT model can be conceived as a way to model habitual behavior, i.e., responses automatically triggered by stimulus-action associations. However, both in real life and laboratory experiments, habits have been shown to be acquired over time scales (days, months, and years) order of magnitudes bigger compared to the time frame of our experiments (26, 27). It is even debatable whether in our task participants developed even a sense of familiarity toward the (never seen before) abstract cues that we used as stimuli. The HABIT model can also be conceived as a way to model choice hysteresis, sometimes referred to as choice repetition or perseveration bias, that could arise from a form of sensory-motor facilitation, where recently performed actions become facilitated (19, 28). However, in our case the screen position of the stimuli was randomized in a trial-by-trial basis and most of the experiments involved interleaved design, thus precluding any strong role for sensory-motor facilitation-induced choice inertia.

We also compared and ruled out a plausible computational interpretation derived from economic theory (29). Since the pioneering work of Daniel Bernoulli [1700 to 1782 (30)], risk aversion is explained by assuming diminishing marginal utility of objective outcomes. At the limit, if diminishing marginal utility was applied in our case, then the utility of 10 points could be perceived as the utility of 1 point. In this extreme scenario, choices would be only based on the comparison between the outcome probabilities. This could explain most aspects of the choice pattern. The UTILITY model did a much better job compared to the HABIT model. However, compared to the RANGE model, it failed to reproduce the observed behavior of the experiments where feedback was provided in the transfer phase. This naturally results from the fact that the model assumes diminishing marginal utility as being a static property of the outcomes and therefore cannot account for experience-dependent correction of context-dependent biases. However, also in this case, a priori considerations could have ruled out the UTILITY interpretation. Our experiment involves stakes small enough to make diminishing marginal utility not reasonable. Rabin provides a full treatment of this issue and shows that the explaining risk aversion for small stakes (as those used in the laboratory) using diminishing marginal utility leads to extremely unlikely predictions, such as

turning down gambles with infinite positive expected values (15). Indeed, if anything, following the intuition of Markowitz (31), most realistic models of the utility function suppose risk neutrality (or risk seeking) for small gains.

Our results contribute to the old and still ongoing debate about whether the brain computes option-oriented values, independently from the decision process itself (2, 32). On one side of the spectrum, decision theories such as expected utility theory and prospect theory, postulate that a value is attached to each option independently of the other options simultaneously available (32). On the other side of the spectrum, other theories, such as regret theory, postulate that the value of an option is primarily determined by the comparison with other available options (33). A similar gradient exists in the reinforcement learning framework, between methods such as the Q-learning, on one side, and direct policy learning without value computations, on the other side (34). Recent studies in humans, coupling imaging to behavioral modeling, provided some support for direct policy learning in humans, by showing that, in complete feedback tasks, participants' learning was driven by a teaching signal, essentially determined by the comparison between the obtained and the forgone outcomes (essentially a regret/relief signal) (7, 35). Beyond behavioral model comparison, analysis of neural activity in the ventral striatum (a brain system traditionally thought to encode option-specific prediction errors (36)) was also consistent with direct policy learning. However, while our findings clearly falsify the Q-learning's assumption that option values are learned on a context-independent (or objective) scale, model simulations also reject the other extreme view of direct policy learning (see the Supplementary Materials). Our results are rather consistent with a hybrid scenario where option-specific values are initially encoded on an objective scale and are progressively normalized to eventually represent the context-specific rank of each option. This view is also consistent with previous results using tasks including loss-related options that clearly showed that option valence was taken into account in transfer learning performance (6, 8). Of note, the notion of "valence" (negative versus positive) is unknown to direct policy learning methods. However, several studies using similar paradigms clearly show that other behavioral measures, such as reaction times and confidence, are strongly affected by the valence of the learning context, thus providing additional evidence against pure direct policy learning methods (13, 37). Last, consistent with our intermediate view, other imaging studies found value-related representations more consistent with a partial normalization process (38, 39).

Last, we note that our computational analysis is at the algorithmic and not at the implementational level (40). In other terms, the RANGE model is a model of the mathematical operations that are performed to achieve a computational goal (i.e., to normalize outcomes to bound subjective option values between 0 and 1). To do so, our model learns two context-level variables (R_{MAX} and R_{MIN}), whose values are unbounded (they converge to their objective values). The present treatment is silent on how these context-level variables are represented at the neural level. While it is certain that coding constraints will also apply to these context-level variables (R_{MAX} and R_{MIN}), further modeling and electrophysiological work is needed to address this important issue.

To conclude, we demonstrated that in humans, reinforcement learning values are learnt in a context-dependent manner that is compatible with range adaptation (instantiated as a range normalization process) (41). Specifically, we tested the possibility that this

normalization automatically results from the way outcome information is processed (42), by showing that the lower the task difficulty, the fuller range adaptation. This leads to a paradoxical result: Reducing task difficulty can, in some occasions, decrease choice optimality. This unexpected result can be understood with a perceptual analogy. Going into a dark room forces us to adapt our retinal response to the dark so that when we go back into a light condition, we do not see very well. The longer we are exposed to dim light, the stronger the effect when we go back to normal.

Our findings fit in the debate aimed at deciding whether the computational processes leading to suboptimal decisions have to be considered flaws or features of human cognition (43, 44). Range-adapting reinforcement learning is clearly adaptive in the learning phase. We could hypothesize that the situations in which the process is adaptive are more frequent in real life. In other terms, the performance of the system has to be evaluated as a function of the tasks it has been selected to solve. Coming back to the perceptual analogy, it is true that we may be hit by a bus when we exit a dark room because we do not see well, but on average, the benefit of a sharper perception in a dark room is big enough to compensate for the (rare) event of a bus waiting for us outside the dark room. Ultimately, whether context-dependent reinforcement learning should be considered a flaw or a desirable feature of human cognition should be determined comparing the real-life frequency of the situations where it is adaptive (as in the learning phase) to that where it is maladaptive (as in the transfer phase). However, while our study does not settle this issue, our findings do demonstrate that this process induces, in some circumstances, economically suboptimal choices. Whether or not the same process is responsible for maladaptive economic behavior in real-life situations will be addressed by future studies using more ecological settings and field data (45).

MATERIALS AND METHODS

Participants

For the laboratory experiment, we recruited 40 participants (28 females, aged 24.28 ± 3.05 years) via internet advertising in a local mailing list dedicated to cognitive science-related activities. For the online experiments, we recruited 8×100 participants (414 females, aged 30.06 ± 10.10 years) from the Prolific platform (www.prolific.co). We based the online sample size on a power analysis that was based on the behavioral results of the laboratory experiment. In the $\Delta EV = 1.75$ context, laboratory participants reached a difference between choice rate and chance (0.5) of 0.11 ± 0.30 (mean \pm SD). To obtain the same with a power of 0.95, the MATLAB function "samsizewr.m" indicated a value of 99 participants that we rounded to 100. The research was carried out following the principles and guidelines for experiments including human participants provided in the Declaration of Helsinki (1964, revised in 2013). The INSERM Ethical Review Committee/IRB00003888 approved the study on 13 November 2018, and participants were provided written informed consent before their inclusion. To sustain motivation throughout the experiment, participants were given a bonus depending on the number of points won in the experiment [average money won in pounds: 4.14 ± 0.72 , average performance against chance: 0.65 ± 0.13 , $t(799) = 33.91$, and $P < 0.0001$]. A laboratory-based experiment was originally performed ($n = 40$) to ascertain that online testing would not significantly affect the main conclusions. The results are presented in the Supplementary Materials.

Behavioral tasks

Participants performed an online version of a probabilistic instrumental learning task adapted from previous studies (6). After checking the consent form, participants received written instructions explaining how the task worked and that their final payoff would be affected by their choices in the task. During the instructions the possible outcomes in points (0, 1, and 10 points) were explicitly showed as well as their conversion rate (1 point = 0.005£). The instructions were followed by a short training session of 12 trials aiming at familiarizing the participants with the response modalities. Participants could repeat the training session up to two times and then started the actual experiment.

In our task, options were materialized by abstract stimuli (cues) taken from randomly generated identicons, colored such that the subjective hue and saturation were very similar according to the HSL_{UV} color scheme (www.hsluv.org). On each trial, two cues were presented on both sides of the screen. The side in which a given cue was presented was pseudo-randomized, such that a given cue was presented an equal number of times on the left and the right. Participants were required to select between the two cues by clicking on one cue. The choice window was self-paced. A brief delay after the choice was recorded (500 ms); the outcome was displayed for 1000 ms. There was no fixation screen between trials. The average reaction time was 1.36 ± 0.04 s (median, 1.16), and the average experiment completion time was 325.24 ± 8.39 s (median, 277.30).

As in previous studies, the full task consisted in one learning phase followed by a transfer phase (6–8, 46). During the learning phase, cues appeared in four fixed pairs. Each pair was presented 30 times, leading to a total of 120 trials. Within each pair, the two cues were associated to a zero and a nonzero outcome with reciprocal probabilities (0.75/0.25 and 0.25/0.75). At the end of the trial, the cues disappeared and the selected one was replaced by the outcome (“10,” “1,” or “0”) (Fig. 1A). In experiments E3, E4, E7, and E8, the outcome corresponding to the forgone option (sometimes referred to as the counterfactual outcome) was also displayed (Fig. 1C). Once they had completed the learning phase, participants were displayed with the total points earned and their monetary equivalent.

During the transfer phase after the learning phase, the pairs of cues were rearranged into four new pairs. The probability of obtaining a specific outcome remained the same for each cue (Fig. 1B). Each new pair was presented 30 times, leading to a total of 120 trials. Before the beginning of the transfer phase, participants were explained that they would be presented with the same cues, only that the pairs would not have been necessarily displayed together before. To prevent explicit memorizing strategies, participants were not informed that they would have to perform a transfer phase until the end of the learning phase. After making a choice, the cues disappeared. In experiments E1, E3, E5, and E7, participants were not informed of the outcome of the choice on a trial-by-trial basis, and the next trial began after 500 ms. This was specified in the instruction phase. In experiments E2, E4, E6, and E8, participants were informed about the result of their choices in a trial-by-trial basis, and the outcome was presented for 1000 ms. In all experiments, they were informed about the total points earned at the end of the transfer phase. In addition to the presence/absence of feedback, experiments differed in two other factors. Feedback information could be either partial (experiments E1, E2, E5, and E6) or complete (experiments E3, E4, E7, and E8; meaning, the outcome of the forgone option was also showed). When the transfer phase included

feedback, the information factor was the same as in the learning phase. Trial structure was also manipulated, such that in some experiments (E5, E6, E7, and E8), all trials of a given choice context were clustered (“blocked”), and in the remaining experiments (E1, E2, E3, and E4), they were interleaved, in both the learning phase and the transfer phase (Fig. 1C).

Reanalysis of a previous experiment involving gain and losses

In the present paper, we also include new analyses of previously published experiments (6). The general design of the previous experiments is similar to that used in the present experiments, as they also involved a learning phase and a transfer phase. However, the previous experimental designs differed from the present one in several important aspects. First, in addition to an outcome magnitude manipulation (“10c” versus “1c”, similar to our learning phase), the study also manipulated the valence of the outcomes (gain versus loss), generating to a 2×2 factorial design. In the gain contexts, participants had to maximize gains, while in the loss contexts, they could only minimize losses. As in the other experiments, outcomes were probabilistic (75 or 25%), and an option was associated with only one type of nonzero outcome. Second, the organization of the transfer phase was quite different. Each option was compared with all other possible options. The main dependent variable extracted from the transfer phase is therefore not the correct response rate but simply the choice rate per option (which is proportional to its subjective value). The data were pooled across two experiments featuring partial ($n = 20$) and partial-and-complete feedback trials ($n = 40$). In both experiments, the choice contexts were interleaved. Other differences include the fact that these previous experiments were laboratory-based and featured a slightly different number of trials, different stimuli and timing [see (6) for more details].

Analyses

Behavioral analyses

The main dependent variable was the correct choice rate, i.e., choices directed toward the option with the highest expected value. Statistical effects were assessed using multiple-way repeated measures analyses of variance (ANOVAs) with choice context (labeled in the manuscript by their difference in expected values: ΔEV) as within-participant factor, and feedback information, feedback in the transfer phase and task structure as between-participant factors. Post hoc tests were performed using one-sample and two-sample t tests for respectively within- and between-experiment comparisons. To assess overall performance, additional one sample t tests were performed against chance level (0.5). We report the t statistic, P value, and Cohen’s d to estimate effect size (two-sample t test only). Given the large sample size ($n = 800$), central limit theorem allows us to assume normal distribution of our overall performance data and to apply properties of normal distribution in our statistical analyses, as well as sphericity hypotheses. Concerning ANOVA analyses, we report the uncorrected statistical, as well as Huynh-Feldt correction for repeated measures ANOVA when applicable (47), F statistic, P value, partial eta-squared η_p^2 , and generalized eta-squared η^2 (when Huynh-Feldt correction is applied) to estimate effect size. All statistical analyses were performed using MATLAB (www.mathworks.com) and R (www.r-project.org). For visual purposes, learning curves were smoothed using a moving average filter (span of 5 in MATLAB’s smooth function).

Models

We analyzed our data with variation of simple reinforcement learning models (48, 49). The goal of all models is to estimate in each choice context (or state) the expected reward (Q) of each option and pick the one that maximizes this expected reward Q.

At trial *t*, option values of the current context *s* are updated with the delta rule

$$Q_{t+1}(s, c) = Q_t(s, c) + \alpha_c \delta_{c,t} \tag{5}$$

$$Q_{t+1}(s, u) = Q_t(s, u) + \alpha_u \delta_{u,t} \tag{6}$$

where α_c is the learning rate for the chosen (*c*) option and α_u the learning rate for the unchosen (*u*) option, i.e., the counterfactual learning rate. δ_c and δ_u are prediction error terms calculated as follows

$$\delta_{c,t} = R_{c,t} - Q_t(s, c) \tag{7}$$

$$\delta_{u,t} = R_{u,t} - Q_t(s, u) \tag{8}$$

δ_c is calculated in both partial and complete feedback experiments, and δ_u is calculated in the experiments with complete feedback only.

We modeled participants' choice behavior using a softmax decision rule representing the probability for a participant to choose one option *a* over the other option *b*

$$P_t(s, a) = \frac{1}{1 + e^{(Q_t(s,b) - Q_t(s,a)) / \beta}} \tag{9}$$

where β is the inverse temperature parameter. High temperatures ($\beta \rightarrow 0$) cause the action to be all (nearly) equiprobable. Low temperatures ($\beta \rightarrow +\infty$) cause a greater difference in selection probability for actions that differ in their value estimates (48).

We compared four alternative computational models: the ABSOLUTE model, which encodes outcomes on an absolute scale independently of the choice context in which they are presented; the RANGE model, which tracks the value of the maximum reward in each context and normalizes the actual reward accordingly, rescaling rewards between 0 and 1; the HABIT model, which integrates action weights into the decision process; and the UTILITY model that assumes diminishing marginal utility.

ABSOLUTE model

The outcomes are encoded as the participants see them (i.e., their objective value). In the eight online experiments, they are encoded as their actual value in points: $R_{OBJ,t} \in \{10, 1, 0\}$. In the dataset retrieved from Bavard *et al.* (6), they are encoded as their actual value in euros $R_{OBJ,t} \in \{-1\text{€}, -0.1\text{€}, 0\text{€}, +0.1\text{€}, \text{ and } +1.0\text{€}\}$.

RANGE model

The outcomes (both chosen and unchosen) are encoded on a context-dependent relative scale. On each trial, the relative reward $R_{RAN,t}$ is calculated as follows

$$R_{RAN,t} = \frac{R_{OBJ,t} - R_{MIN,t}(s)}{R_{MAX,t}(s) - R_{MIN,t}(s) + 1} \tag{2}$$

As R_{MIN} is initialized to zero and never changes, in the eight online experiments, this model can be reduced to

$$R_{RAN,t} = \frac{R_{OBJ,t}}{R_{MAX,t}(s) + 1} \tag{10}$$

where *s* is the decision context (i.e., a combination of options) and R_{MAX} and R_{MIN} are context-dependent variables, initialized to 0 and updated at each trial *t* if the outcome is greater (or smaller, respectively) than its current value

$$R_{MAX,t+1}(s) = R_{MAX,t}(s) + \alpha_R (R_{OBJ,t} - R_{MAX,t}(s)) \text{ if } R_{OBJ,t} > R_{MAX,t}(s) \tag{11}$$

$$R_{MIN,t+1}(s) = R_{MIN,t}(s) + \alpha_R (R_{OBJ,t} - R_{MIN,t}(s)) \text{ if } R_{OBJ,t} < R_{MIN,t}(s) \tag{12}$$

Accordingly, outcomes are progressively normalized so that eventually $R_{RAN,t} \in [0, 1]$. The chosen and unchosen option values and prediction errors are updated with the same rules as in the ABSOLUTE model. α_R is an additional free parameter, the contextual—or range—learning rate, that is used to update the range variables. Note that the ABSOLUTE model is nested within the RANGE model ($\alpha_R = 0$).

HABIT model

The outcomes are encoded on an absolute scale, but decisions integrate a habitual component (16, 19). To do so, in addition to the Q values, a habitual (or choice trace) component *H* is tracked and updated (with a dedicated learning rate parameter) that takes into account the selected action (1 for chosen option and 0 for the unchosen option). The choice is performed with a softmax rule based on decision weights *D* that integrate Q values and decision weights *H*

$$D_t(s, c) = (1 - \omega) * Q_t(s, c) + \omega * H_t(s, c) \tag{3}$$

where at each trial *t*, state *s*, and chose option *c*, *D* is the arbiter, *Q* is the goal-directed component (Q values matrix), and *H* is the habitual component. The weight ω is fitted as an additional parameter and governs the relative weights of values and habits (for $\omega = 0$, the model reduces to the ABSOLUTE model).

UTILITY model

The outcomes are encoded as an exponentiation of the absolute reward, leading to a curvature of the value function (29)

$$R_{UTI,t} = (R_{OBJ,t})^v \tag{4}$$

where the exponent *v* is the utility parameter, with $0 < v < 1$ (for $v = 1$ the model reduces to the ABSOLUTE model).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at <http://advances.sciencemag.org/cgi/content/full/7/14/eabe0340/DC1>

REFERENCES AND NOTES

1. K. Louie, P. W. Glimcher, Efficient coding and the neural representation of value. *Ann. N. Y. Acad. Sci.* **1251**, 13–32 (2012).
2. I. Vlaev, N. Chater, N. Stewart, G. D. A. Brown, Does the brain calculate value? *Trends Cogn. Sci.* **15**, 546–554 (2011).
3. K. M. Cox, J. W. Kable, BOLD subjective value signals exhibit robust range adaptation. *J. Neurosci.* **34**, 16533–16543 (2014).
4. S. Nieuwenhuis, D. J. Heslenfeld, N. J. Alting von Geusau, R. B. Mars, C. B. Holroyd, N. Yeung, Activity in human reward-sensitive brain areas is strongly context dependent. *Neuroimage* **25**, 1302–1309 (2005).
5. R. Elliott, Z. Agnew, J. F. W. Deakin, Medial orbitofrontal cortex codes relative rather than absolute value of financial rewards in humans. *Eur. J. Neurosci.* **27**, 2213–2218 (2008).
6. S. Bavard, M. Lebreton, M. Khamassi, G. Coricelli, S. Palminteri, Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences. *Nat. Commun.* **9**, 4503 (2018).
7. T. A. Klein, M. Ullsperger, G. Jocham, Learning relative values in the striatum induces violations of normative decision making. *Nat. Commun.* **8**, 16033 (2017).

8. S. Palminteri, M. Khamassi, M. Joffily, G. Coricelli, Contextual modulation of value signals in reward and punishment learning. *Nat. Commun.* **6**, 8096 (2015).
9. E. Freidin, A. Kacelnik, Rational choice, context dependence, and the value of information in European starlings (*Sturnus vulgaris*). *Science* **334**, 1000–1002 (2011).
10. L. Pompilio, A. Kacelnik, Context-dependent utility overrides absolute memory as a determinant of choice. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 508–512 (2010).
11. A. Rustichini, K. E. Conen, X. Cai, C. Padoa-Schioppa, Optimal coding and neuronal adaptation in economic decisions. *Nat. Commun.* **8**, 1208 (2017).
12. R. Webb, P. W. Glimcher, K. Louie, *The Normalization of Consumer Valuations: Context-Dependent Preferences from Neurobiological Constraints* (Management Science, 2020); <https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2019.3536>.
13. L. Fontanesi, S. Palminteri, M. Lebreton, Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: A meta-analytical approach using diffusion decision modeling. *Cogn. Affect. Behav. Neurosci.* **19**, 490–502 (2019).
14. A. G. E. Collins, M. J. Frank, How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *Eur. J. Neurosci.* **35**, 1024–1035 (2012).
15. M. Rabin, *Diminishing Marginal Utility of Wealth Cannot Explain Risk Aversion* (2000); <https://escholarship.org/uc/item/61d7b4pg>.
16. K. J. Miller, A. Shenhav, E. A. Ludvig, Habits without values. *Psychol. Rev.* **126**, 292–311 (2019).
17. P. Landry, R. Webb, *Pairwise Normalization: A Neuroeconomic Theory of Multi-Attribute Choice* (Social Science Research Network, 2019); <https://papers.ssrn.com/abstract=2963863>.
18. S. Palminteri, V. Wyart, E. Koechlin, The importance of falsification in computational cognitive modeling. *Trends Cogn. Sci.* **21**, 425–433 (2017).
19. K. Katahira, The statistical structures of reinforcement learning with asymmetric value updates. *J. Math. Psychol.* **87**, 31–45 (2018).
20. K. Louie, P. W. Glimcher, R. Webb, Adaptive neural coding: From biological to behavioral decision-making. *Curr. Opin. Behav. Sci.* **5**, 91–99 (2015).
21. T. Dumbalska, V. Li, K. Tsetos, C. Summerfield, A map of decoy influence in human multialternative choice. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 25169–25178 (2020).
22. R. Daviet, R. Webb, *A Double Decoy Experiment to Distinguish Theories of Dominance Effects* (Social Science Research Network, 2019); <https://papers.ssrn.com/abstract=3374514>.
23. S. Gluth, N. Kern, M. Kortmann, C. L. Vitali, Value-based attention but not divisive normalization influences decisions with multiple alternatives. *Nat. Hum. Behav.* **4**, 634–645 (2020).
24. P. B. Goodwin, Habit and hysteresis in mode choice. *Urb. Stud.* **14**, 95–98 (1977).
25. A. Dickinson, L. Weiskrantz, Actions and habits: The development of behavioural autonomy. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **308**, 67–78 (1985).
26. P. Lally, C. H. M. van Jaarsveld, H. W. W. Potts, J. Wardle, How are habits formed: Modelling habit formation in the real world. *Eur. J. Soc. Psychol.* **40**, 998–1009 (2010).
27. E. A. Thrailkill, S. Trask, P. Vidal, J. A. Alcalá, M. E. Bouton, Stimulus control of actions and habits: A role for reinforcer predictability and attention in the development of habitual behavior. *J. Exp. Psychol. Anim. Learn. Cogn.* **44**, 370–384 (2018).
28. R. Akaishi, K. Umeda, A. Nagase, K. Sakai, Autonomous mechanism of internal choice estimate underlies decision inertia. *Neuron* **81**, 195–206 (2014).
29. J. von Neumann, O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton Univ. Press, 1953).
30. D. Bernoulli, Exposition of a new theory on the measurement of risk. *Econometrica* **22**, 23–36 (1954).
31. H. Markowitz, The utility of wealth. *J. Polit. Econ.* **60**, 151–158 (1952).
32. D. Kahneman, A. Tversky, Subjective probability: A judgment of representativeness. *Cogn. Psychol.* **3**, 430–454 (1972).
33. G. Loomes, R. Sugden, Regret theory: An Alternative theory of rational choice under uncertainty. *Econ. J.* **92**, 805–824 (1982).
34. P. Dayan, L. F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (Massachusetts Institute of Technology Press, 2001), p. 460.
35. J. Li, N. D. Daw, Signals in human striatum are appropriate for policy update rather than value prediction. *J. Neurosci.* **31**, 5504–5511 (2011). [cited 2020 Nov 30].
36. S. Palminteri, M. Pessiglione, Opponent brain systems for reward and punishment learning: Causal evidence from drug and lesion studies in humans, in *Decision Neuroscience*, J.-C. Dreher, L. Tremblay, Eds. (San Diego: Academic Press, 2017), chap. 23, pp. 291–303.
37. M. Lebreton, K. Bacily, S. Palminteri, J. B. Engelmann, Contextual influence on confidence judgments in human reinforcement learning. *PLOS Comput. Biol.* **15**, e1006973 (2019).
38. C. J. Burke, M. Baddeley, P. N. Tobler, W. Schultz, Partial adaptation of obtained and observed value signals preserves information about gains and losses. *J. Neurosci.* **36**, 10016–10025 (2016).
39. D. Pischedda, S. Palminteri, G. Coricelli, The effect of counterfactual information on outcome value coding in medial prefrontal and cingulate cortex: From an absolute to a relative neural code. *J. Neurosci.* **40**, 3268–3277 (2020).
40. D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (W. H. Freeman and Co Ltd, 1982).
41. K. E. Conen, C. Padoa-Schioppa, Partial adaptation to the value range in the macaque orbitofrontal cortex. *J. Neurosci.* **39**, 3498–3513 (2019).
42. C. Padoa-Schioppa, A. Rustichini, Rational attention and adaptive coding: A puzzle and a solution. *Am. Econ. Rev.* **104**, 507–513 (2014).
43. G. Gigerenzer, The bias bias in behavioral economics. *Rev. Behav. Econ* **5**, 303–336 (2018).
44. M. G. Haselton, D. Nettle, P. W. Andrews, The evolution of cognitive bias, in *The Handbook of Evolutionary Psychology*, (John Wiley and Sons Ltd, 2015), pp. 724–46; <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470939376.ch25> [cited 27 July 2020].
45. C. F. Camerer, *Prospect Theory in the Wild: Evidence From the Field* (Pasadena, CA, California Institute of Technology, 1998); <https://resolver.caltech.edu/CaltechAUTHORS:20170811-150835361> [cited 30 January 2021].
46. M. J. Frank, L. C. Seeberger, R. C. O'Reilly, By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science* **306**, 1940–1943 (2004).
47. E. R. Girden, *ANOVA: Repeated Measures* (SAGE, 1992).
48. R. S. Sutton, A. G. Barto, *Reinforcement Learning - An Introduction* (MIT Press, 1998).
49. R. A. Rescorla, A. R. Wagner, A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Class. Condition. II, Curr. Res. Theory* **2**, 64–99 (1972).
50. J. Hergueux, N. Jacquemet, Social preferences in the online laboratory: A randomized experiment. *Exp. Econ.* **18**, 251–283 (2015).
51. D. Kahneman, Maps of bounded rationality: Psychology for behavioral economics. *Am. Econ. Rev.* **93**, 1449–1475 (2003).
52. T. Shavit, D. Sonsino, U. Benzion, A comparative study of lotteries-evaluation in class and on the Web. *J. Econ. Psychol.* **22**, 483–491 (2001).
53. E. T. Miller, D. J. Neal, L. J. Roberts, J. S. Baer, S. O. Cressler, J. Metrik, G. A. Marlatt, Test-retest reliability of alcohol measures: Is there a difference between internet-based assessment and traditional methods? *Psychol. Addict. Behav.* **16**, 56–63 (2002).
54. K. Reinecke, K. Z. Gajos, LabintheWild: conducting large-scale online experiments with uncompensated samples, in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)* (Vancouver, BC, Canada, Association for Computing Machinery, 2015), p. 1364–1378; <https://doi.org/10.1145/2675133.2675246>.

Acknowledgments

Funding: S.P. is supported by an ATIP-Avenir grant (R16069JS), the Programme Emergence(s) de la Ville de Paris, the Fondation Fyssen, the Fondation Schlumberger pour l'Éducation et la Recherche, the FrontCog grant (ANR-17-EURE-0017) and the Institut de Recherche en Santé Publique (IRESP, grant number : 201138-00). S.B. is supported by MILDECA (Mission Interministerielle de Lutte contre les Drogues et les Conduites Addictives) and the EHES (Ecole des Hautes Etudes en Sciences Sociales). A.R. thanks the US Army for financial support (contract W911NF2010242). The funding agencies did not influence the content of the manuscript. **Author contributions:** S.B. and S.P. designed the experiments. S.B. ran the experiments. S.B. and S.P. analyzed the data. S.B., A.R., and S.P. interpreted the results. S.B. and S.P. wrote the manuscript. A.R. edited the manuscript. **Competing interests:** The authors declare that they have no financial or other competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials and are available from Github repository (<https://github.com/hrl-team/range>). All custom scripts have been made available from Github repository (<https://github.com/hrl-team/range>). Additional modified scripts can be accessed upon request.

Submitted 27 July 2020

Accepted 12 February 2021

Published 2 April 2021

10.1126/sciadv.abe0340

Citation: S. Bavard, A. Rustichini, S. Palminteri, Two sides of the same coin: Beneficial and detrimental consequences of range adaptation in human reinforcement learning. *Sci. Adv.* **7**, eabe0340 (2021).

Two sides of the same coin: Beneficial and detrimental consequences of range adaptation in human reinforcement learning

Sophie Bavard, Aldo Rustichini and Stefano Palminteri

Sci Adv 7 (14), eabe0340.
DOI: 10.1126/sciadv.abe0340

ARTICLE TOOLS	http://advances.sciencemag.org/content/7/14/eabe0340
SUPPLEMENTARY MATERIALS	http://advances.sciencemag.org/content/suppl/2021/03/29/7.14.eabe0340.DC1
REFERENCES	This article cites 41 articles, 9 of which you can access for free http://advances.sciencemag.org/content/7/14/eabe0340#BIBL
PERMISSIONS	http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the [Terms of Service](#)

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.

Copyright © 2021 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

Supplementary Materials for

Two sides of the same coin: Beneficial and detrimental consequences of range adaptation in human reinforcement learning

Sophie Bavard, Aldo Rustichini, Stefano Palminteri*

*Corresponding author. Email: stefano.palminteri@ens.fr

Published 2 April 2021, *Sci. Adv.* **7**, eabe0340 (2021)
DOI: [10.1126/sciadv.abe0340](https://doi.org/10.1126/sciadv.abe0340)

This PDF file includes:

Supplementary Text
Figs. S1 to S8
References

Supplementary Materials

Comparison between laboratory- and online-based experiments and robustness of our main results to outliers' exclusion

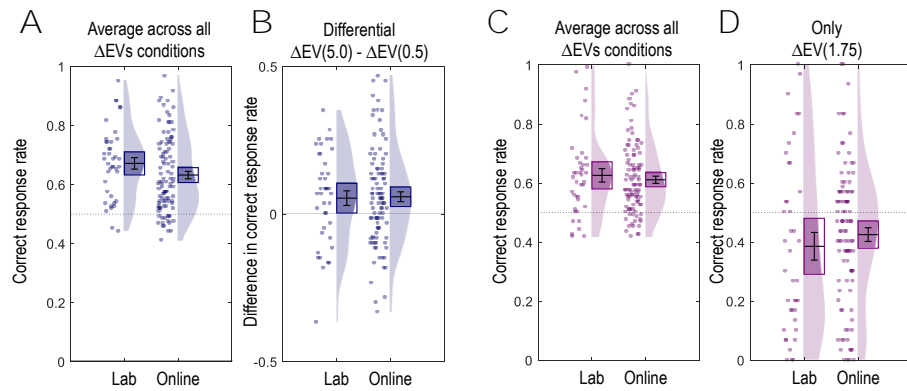
Before moving to online testing, we run a laboratory-based experiment, to ascertain that there was no detectable difference between the two set-ups. We recruited 40 participants (28 females, aged 24.28 ± 3.05 years) via Internet advertising in mailing list dedicated to cognitive science-related activities. The experimental design used in the lab was that of experiment E2 presented in the main text (partial feedback information in both the learning phase and the transfer phase, and trials in an interleaved order; see **Figure 1**).

In order to characterize learning behavior of participants, we analyzed the correct response rate in both phases, i.e., choices directed toward the most favorable option at each trial. To assess successful learning, we first tested participants' correct response rate against chance level. We found it to be above chance level in both the learning phase ($t(39) = 8.88, p < .0001, d = 1.40$, **Supp. Figure 1A**) and the transfer phase ($t(39) = 5.55, p < .0001, d = 0.88$, **Supp. Figure 1C**). We found a significant effect of magnitude in the learning phase ($t(39) = 2.18, p = .036, d = 0.34$, **Supp. Figure 1B**), and the correct choice rate in the $\Delta EV=1.75$ context was significantly below chance level ($t(39) = -2.43, p = .020, d = -0.38$, **Supp. Figure 1D**). Of note, the effect sizes were virtually indistinguishable comparable to those observed in the corresponding online experiment (learning performance $d = 1.04$ vs 1.40 , transfer performance $d = 0.93$ vs 0.88 , magnitude effect $d = 0.35$ vs 0.34 , value inversion $d = -0.32$ vs -0.38).

In addition to checking that the same significant results were detected, to formally assess the similarity between online- and laboratory-based experiments, we explicitly compared their scores. Correct choice rate in the learning phase did not significantly differ between laboratory and online datasets ($t(138) = 1.67, p = .10, d = 0.31$, **Supp. Figure 1A**), neither did the magnitude effect ($t(138) = -0.15, p = .88, d = -0.03$, **Supp. Figure 1B**). Concerning the transfer phase, overall correct choice rate was not significantly different ($t(138) = 0.62, p = .54, d = 0.12$, **Supp. Figure 1C**) and the same result was obtained looking specifically at the $\Delta EV=1.75$ context ($t(138) = -0.84, p = .40, d = -0.16$, **Supp. Figure 1D**). Of note, although the control over the measure of reaction times is arguably limited in online experiments, also this measure did not differ between laboratory- and online-based experiments ($t(138) = -0.50, p = .62, d = -0.09$). This similarity between laboratory- and online-based results supports the usefulness of online-based experiments as a way to target larger, more diversified populations with reduced administrative and financial costs (50). The limitations that can be encountered with online-based experiments - such as lower data quality, faster reaction time, lack of engagement from the participants (51,52) – were not detectable in our data.

However, to further check the robustness of our results, we run analyses of the online data excluding participants presenting unusual task completion time. We approximated participants' total reaction time over the whole task by a normal distribution and removed outliers at a significance level of $p < 0.05$. This led to a removal of only 30 participants (3.75%) for the eight online experiments leading to a final sample of 770 participants. We found that the totality of the statistically significant results described in the **Results** section were observable without these reaction time outliers, thus we decided to include all participants in the statistics reported in the **Results** section. In conclusion, our results successfully replicate in the laboratory and online results are robust to stricter exclusion criteria. Moreover, our results confirm the findings

of recent studies comparing both experimental methods and showing that they produce comparable data quality (53,54).



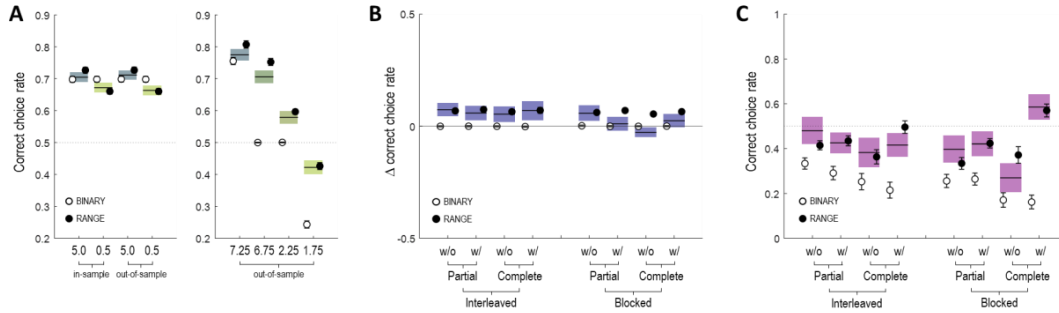
Supp. Figure 1: Comparing laboratory and online experiments. (A) Average correct response rate in the learning phase per experiment. (B) Difference in correct choice rate between the $\Delta EV=5.0$ and the $\Delta EV=0.5$ contexts. (C) Average correct response rate in the transfer phase. (D) Correct choice rate for the $\Delta EV=1.75$ context only.

Additional model comparisons

The computational results presented here follow the same fitting and simulation methods presented in the main text for the main computational models. Also, the general notation is the same.

BINARY model

We analyzed the generative performances of a “full-adaptation” model encoding non-zero outcomes as ones, regardless of their actual magnitude (10pt, 1pt), that we refer to as the BINARY model. At least three behavioral features allow us to reject the BINARY model. Of note, the model is a special case of the UTILITY model for extremely diminishing marginal utility ($v = 0$; $R_{UTI,t} = (R_{OBJ,t})^v$). First, it is not able to capture participants’ behavior in the learning phase by failing to accurately predict the outcome magnitude difference (**Supp. Figure 2A** and **Supp. Figure 2B**); second, the model predicts perfect indifference in the $\Delta EV=6.75$ and the $\Delta EV=2.25$ contexts in the transfer phase, while behavioral results show, respectively, a strong and moderate preference for the high EV options in these contexts; third, the BINARY model predicts an exaggerated rate of suboptimal preferences in the $\Delta EV=1.75$ context in the transfer phase (**Supp. Figure 2A** and **Supp. Figure 2C**). This is true in all 8 experiments and even more striking in E8 where the participants were able to correct their bias.



Supp. Figure 2: Model simulations of the BINARY model. Generative performance of the RANGE model (black dots) compared to a full-adaptation model encoding rewards as 1’s or 0’s (white dots: BINARY model). Black lines represent the empirical averages. Colored squares indicate 95% confidence interval around the empirical averages.

REFERENCE model

We also analyzed the generative performances of a previous context-dependence model (8) that we call here REFERENCE because of its distinctive feature is to apply reference point dependence to outcome encoding:

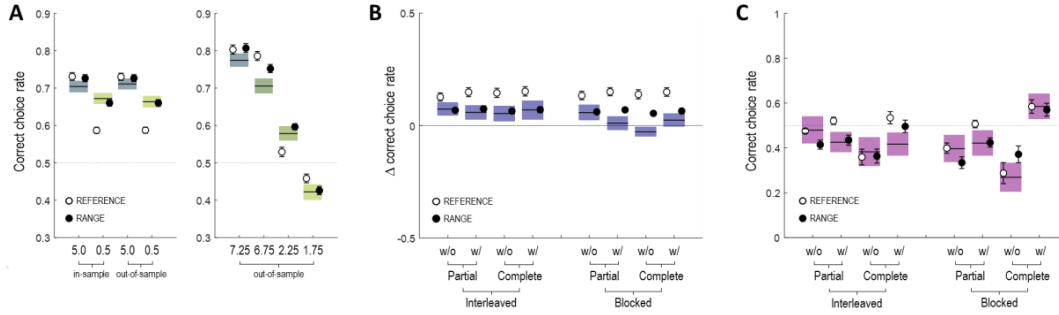
$$Q(s, a) \leftarrow Q(s, a) + \alpha_Q * (R_{OBJ} - V(s) - Q(s, a))$$

Where s is the state (or context: pair of options), $V(s)$ is the state value (or reference point), $Q(s, a)$ is the Q-value (estimated expected value). $V(s)$ is also learnt iteratively, as follows:

$$V(s) \leftarrow V(s) + \alpha_V * \left(\frac{R_{OBJ} + Q(s, u)}{2} - V(s) \right)$$

When the feedback is complete, $Q(s, u)$ (the Q-value of the unchosen option) is replaced by the outcome of the unchosen option. α_V is an additional free parameter for the state value $V(s)$ (which can be considered an off-policy state value).

Concerning the learning phase, model simulation analysis (**Supp. Figure 3**) showed that, while the REFERENCE model matches the performance in the high-magnitude contexts in the learning phase ($\Delta EV=5$), it fails to capture the performance in low-magnitude contexts ($\Delta EV=0.5$). This is expected as the model does not implement range adaptation in any form. Concerning the transfer phase, the REFERENCE model reproduces a pattern that is qualitatively close to the observed results, but still less accurate compared to the RANGE model (out of sample likelihood comparison $LL_{\text{RAN}} = -96.79$ vs $LL_{\text{REF}} = -186.68$, $t(799) = 8.26$, $p < .0001$). To sum up, the REFERENCE model is strongly rejected by the learning phase results (where it essentially behaves like to the ABSOLUTE model) and weakly rejected by the transfer phase results, where it manages to capture the overall pattern, but in a less accurate manner.



Supp. Figure 3: Model simulations of the REFERENCE model. Generative performance of the RANGE model (black dots) compared to the REFERENCE model (white dots). Black lines represent the empirical averages. Colored squares indicate 95% confidence interval around the empirical averages.

GLOBAL model

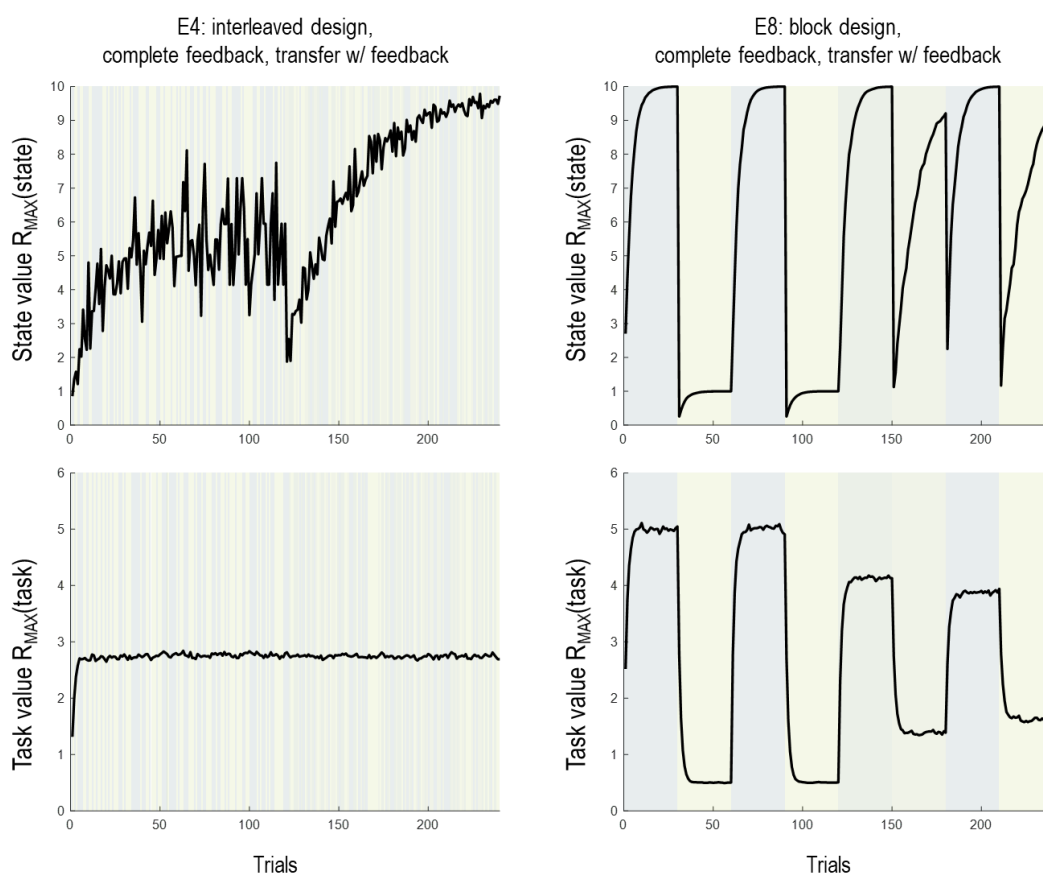
The RANGE model as we implemented it for the analyses presented in main text, does not contain any element to account for the block/interleaved effect. Here we propose a possible computational interpretation to account for the effects of this manipulation (more precisely the fact that contextual effects are exacerbated in block experiments). The key idea of this model is that the notion of ‘context’ can be break down into two components. The ‘local’ context is what we referred to as simply “learning context” in the paper (essentially a pair of cues, or a state ‘s’ in the reinforcement learning framework). In addition to the local context, we also postulate a ‘global’ context that integrate over a time scale larger than a trial (it could be understood as the current average ‘value’ of the task). To instantiate this idea, we built an alternative model (GLOBAL) that includes both “global” (or task-level) and local (or pair of options-level) contextual variables: $R_{\text{MAX}}(\text{task})$ and $R_{\text{MAX}}(\text{state})$. The $R_{\text{MAX}}(\text{state})$ is learnt similarly to the state-value in the REFERENCE model, except that it is not bounded to any particular pair of options:

$$R_{\text{MAX}}(\text{task}) \leftarrow R_{\text{MAX}}(\text{task}) + \alpha_T * \left(\frac{R_{\text{OBJ}} + Q(s, u)}{2} - R_{\text{MAX}}(\text{task}) \right)$$

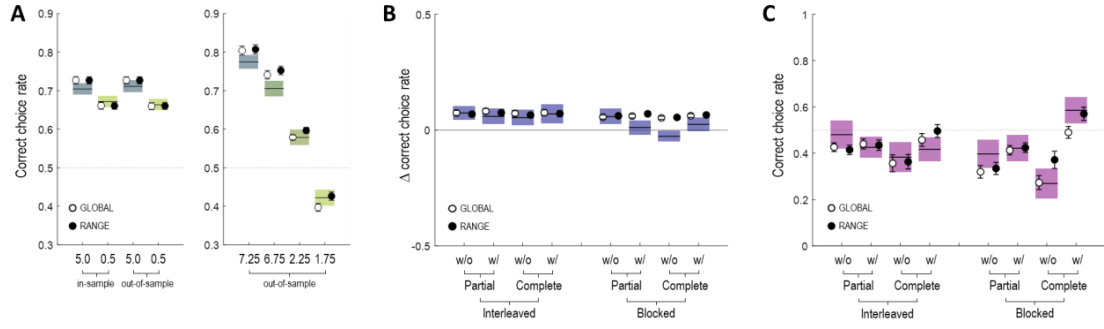
When the feedback is complete, $Q(s, u)$ (the Q-value of the unchosen option) is replaced by the outcome of the unchosen option. α_T is an additional free parameter for the $R_{MAX}(\text{task})$. The range normalization rule (that we write here in its simplified manner that takes into account that $R_{MIN} = 0$ everywhere in our task) in the option value update rule of the GLOBAL model is as follows:

$$Q(s, a) \leftarrow Q(s, a) + \alpha_Q * \left(\frac{R_{OBJ}}{R_{MAX}(\text{state}) + R_{MAX}(\text{task}) + 1} - Q(s, a) \right)$$

This simple model accounts for increased contextual effects in block design, because in the block design, $R_{MAX}(\text{task})$ and the $R_{MAX}(\text{state})$ remain coherent for longer time periods (**Supp. Figure 4**), thus allowing the summation of their effects. As shown in **Supp. Figure 5**, the model seems qualitatively equal than the RANGE model, if not better at matching performance in most of the 8 different versions of the $\Delta EV=1.75$ context.



Supp. Figure 4: State- and task- context values in inter-leaved or blocked designs. The figure illustrates the evolution across the experiment of the hidden variables $R_{MAX}(\text{state})$ and $R_{MAX}(\text{task})$. Simulations concern E4 (interleaved design, complete feedback, transfer with feedback) and E8 (block design, complete feedback, transfer with feedback). Background colors show the choice context (color coded as in Figure 1).



Supp. Figure 5: Model simulations of the GLOBAL model. Generative performance of the RANGE model (black dots) compared to the GLOBAL model (white dots). Black lines represent the empirical averages. Colored squares indicate 95% confidence interval around the empirical averages.

REGRET model

Finally, we analyzed a model assuming that option values are purely encoded by outcome comparison (akin to a relief/regret signal). A similar idea has been put forward by other studies (7,35) where it proved successful in explain ventral striatal neural activity and, to some extent, behavioral data. Of note, this model has the strong handicap that it cannot be straightforwardly extended to the partial feedback case, where the outcome of the unchosen option is not showed. We therefore tested the proposed model in the 4 experiments featuring complete feedback.

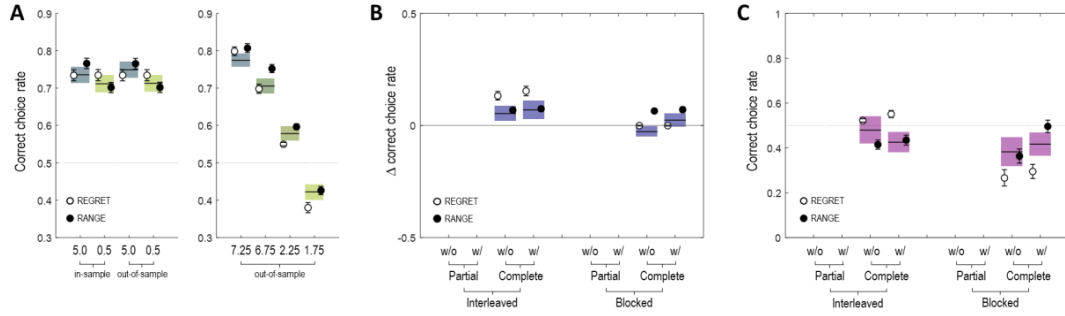
Option values in the REGRET model are updated as follows, with R_C and R_U the outcomes of the chosen option and unchosen option, respectively:

$$R_{\text{REG},t} = \begin{cases} 1 & \text{if } R_C > R_U \\ 0 & \text{if } R_C = R_U \\ -1 & \text{if } R_C < R_U \end{cases}$$

$$Q_{t+1}(s, c) = Q_t(s, c) + \alpha_c * (R_{\text{REG},t} - Q_t(s, c))$$

$$Q_{t+1}(s, u) = Q_t(s, u) + \alpha_u * (R_{\text{REG},t} - Q_t(s, u))$$

As clearly illustrated by the model simulations (**Supp. Figure 6**), the REGRET model does not fit well the behavioral data, especially in the transfer phase, where it overestimates value inversion in the $\Delta EV=1.75$ context. In other terms through a different mechanism, the REGRET model suffers from the same problem the BINARY model: they predict to much option value context dependence.



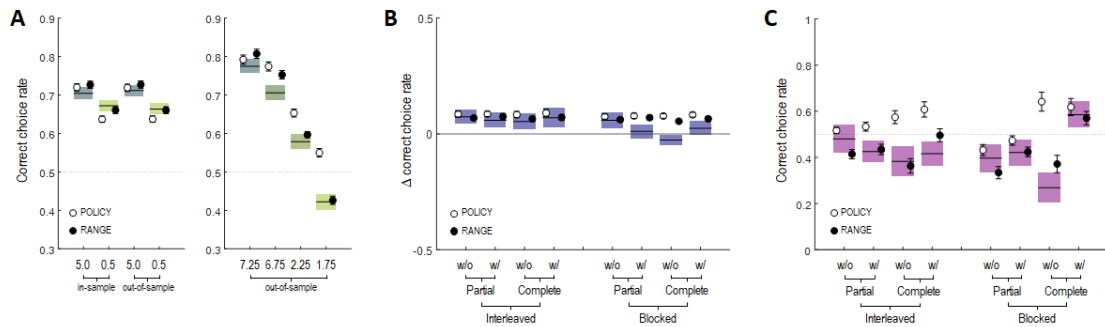
Supp. Figure 6: Model simulations of the REGRET model. Generative performance of the RANGE model (black dots) compared to the REGRET model (white dots). Black lines represent the empirical averages. Colored squares indicate 95% confidence interval around the empirical averages.

POLICY model

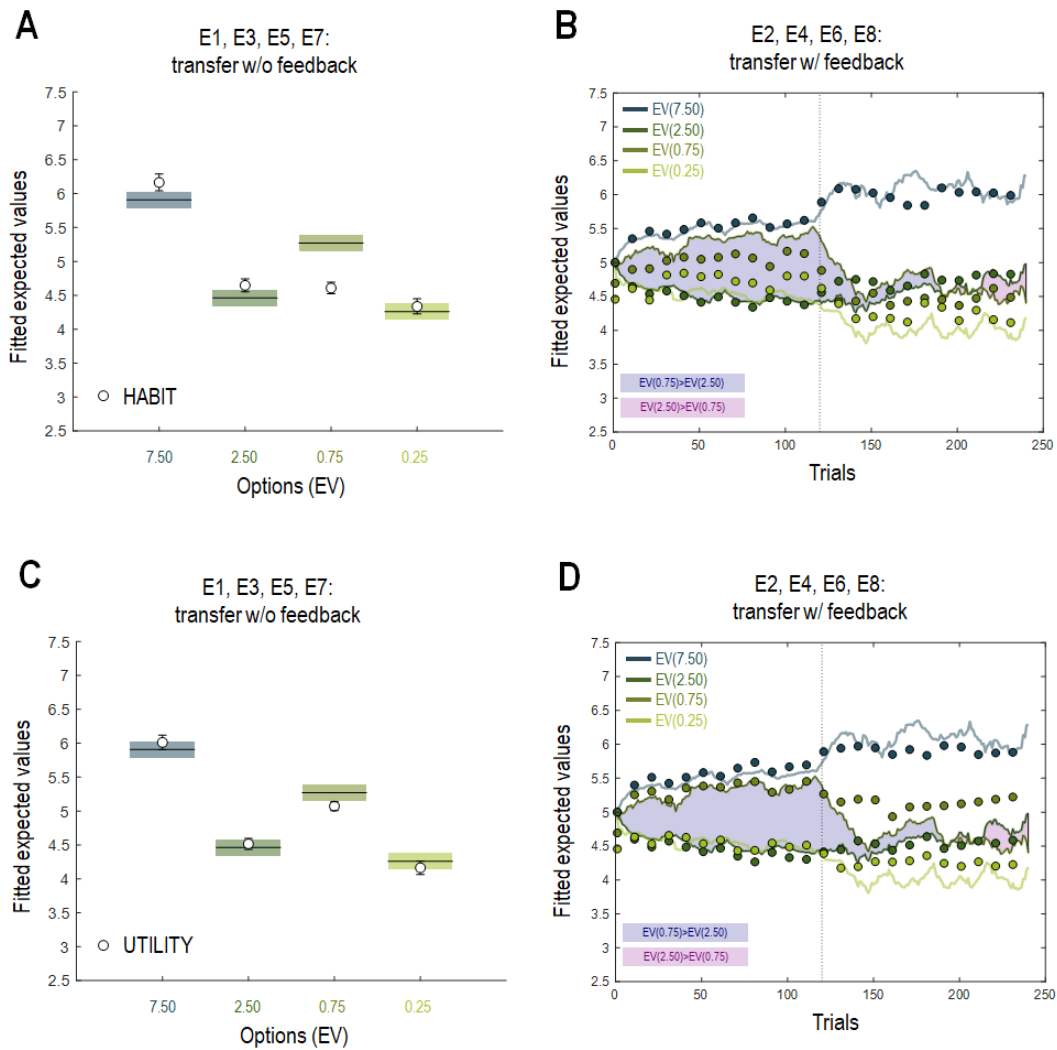
Finally, we considered a model that applies range normalization at the decision step (i.e., in the softmax decision rule), instead of the outcome encoding stage as in the RANGE model. In this model (POLICY) the probability of choosing option a over option b is defined by:

$$P_t(s, a) = \frac{1}{1 + e^{\left(\beta * \frac{Q_t(s,b) - Q_t(s,a)}{1 + \max\{Q_t(s,:)\} - \min\{Q_t(s,:)\}}\right)}}$$

Similarly to the RANGE model, the POLICY model is able to capture the magnitude difference in the learning phase (i.e., the partial range adaptation). In the transfer phase however, the POLICY model fails to predict the value inversion in the $\Delta EV=1.75$ context. This is due to the fact that, despite the normalization process within the softmax function, option values remain encoded in an absolute scale. Whereas in the learning phase the POLICY model predicts a behavior compatible with the RANGE model, in the transfer phase it predicts a behavior consistent with the ABSOLUTE model (Supp. Fig. 6).



Supp. Figure 7: Model simulations of the POLICY model. Generative performance of the RANGE model (black dots) compared to the POLICY model (white dots). Black lines represent the empirical averages. Colored squares indicate 95% confidence interval around the empirical averages.



Supp. Figure 8. Inferred option values from the UTILITY and the HABIT models. (A-C) Average inferred option values for the behavioral data and simulated data for the experiments without trial-by-trial transfer feedback (white dots: HABIT (resp. UTILITY) model). (B-D) Trial-by-trial inferred option values for the behavioral data and simulated data for the experiments with trial-by-trial transfer feedback, where curves indicate trial-by-trial fit of each inferred option value, and colored dots indicate HABIT (resp. UTILITY) model simulations.

REFERENCES AND NOTES

1. K. Louie, P. W. Glimcher, Efficient coding and the neural representation of value. *Ann. N. Y. Acad. Sci.* **1251**, 13–32 (2012).
2. I. Vlaev, N. Chater, N. Stewart, G. D. A. Brown, Does the brain calculate value? *Trends Cogn. Sci.* **15**, 546–554 (2011).
3. K. M. Cox, J. W. Kable, BOLD subjective value signals exhibit robust range adaptation. *J. Neurosci.* **34**, 16533–16543 (2014).
4. S. Nieuwenhuis, D. J. Heslenfeld, N. J. Alting von Geusau, R. B. Mars, C. B. Holroyd, N. Yeung, Activity in human reward-sensitive brain areas is strongly context dependent. *Neuroimage* **25**, 1302–1309 (2005).
5. R. Elliott, Z. Agnew, J. F. W. Deakin, Medial orbitofrontal cortex codes relative rather than absolute value of financial rewards in humans. *Eur. J. Neurosci.* **27**, 2213–2218 (2008).
6. S. Bavard, M. Lebreton, M. Khamassi, G. Coricelli, S. Palminteri, Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences. *Nat. Commun.* **9**, 4503 (2018).
7. T. A. Klein, M. Ullsperger, G. Jocham, Learning relative values in the striatum induces violations of normative decision making. *Nat. Commun.* **8**, 16033 (2017).
8. S. Palminteri, M. Khamassi, M. Joffily, G. Coricelli, Contextual modulation of value signals in reward and punishment learning. *Nat. Commun.* **6**, 8096 (2015).
9. E. Freidin, A. Kacelnik, Rational choice, context dependence, and the value of information in European starlings (*Sturnus vulgaris*). *Science* **334**, 1000–1002 (2011).
10. L. Pompilio, A. Kacelnik, Context-dependent utility overrides absolute memory as a determinant of choice. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 508–512 (2010).
11. A. Rustichini, K. E. Conen, X. Cai, C. Padoa-Schioppa, Optimal coding and neuronal adaptation in economic decisions. *Nat. Commun.* **8**, 1208 (2017).
12. R. Webb, P. W. Glimcher, K. Louie, *The Normalization of Consumer Valuations: Context-Dependent Preferences from Neurobiological Constraints* (Management Science, 2020); <https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2019.3536>.
13. L. Fontanesi, S. Palminteri, M. Lebreton, Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: A meta-analytical approach using diffusion decision modeling. *Cogn. Affect. Behav. Neurosci.* **19**, 490–502 (2019).

14. A. G. E. Collins, M. J. Frank, How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *Eur. J. Neurosci.* **35**, 1024–1035 (2012).
15. M. Rabin, *Diminishing Marginal Utility of Wealth Cannot Explain Risk Aversio* (2000); <https://escholarship.org/uc/item/61d7b4pg>.
16. K. J. Miller, A. Shenhav, E. A. Ludvig, Habits without values. *Psychol. Rev.* **126**, 292–311 (2019).
17. P. Landry, R. Webb, *Pairwise Normalization: A Neuroeconomic Theory of Multi-Attribute Choice* (Social Science Research Network, 2019); <https://papers.ssrn.com/abstract=2963863>.
18. S. Palminteri, V. Wyart, E. Koechlin, The importance of falsification in computational cognitive modeling. *Trends Cogn. Sci.* **21**, 425–433 (2017).
19. K. Katahira, The statistical structures of reinforcement learning with asymmetric value updates. *J. Math. Psychol.* **87**, 31–45 (2018).
20. K. Louie, P. W. Glimcher, R. Webb, Adaptive neural coding: From biological to behavioral decision-making. *Curr. Opin. Behav. Sci.* **5**, 91–99 (2015).
21. T. Dumbalska, V. Li, K. Tsetsos, C. Summerfield, A map of decoy influence in human multialternative choice. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 25169–25178 (2020).
22. R. Daviet, R. Webb, *A Double Decoy Experiment to Distinguish Theories of Dominance Effects* (Social Science Research Network, 2019); <https://papers.ssrn.com/abstract=3374514>.
23. S. Gluth, N. Kern, M. Kortmann, C. L. Vitali, Value-based attention but not divisive normalization influences decisions with multiple alternatives. *Nat. Hum. Behav.* **4**, 634–645 (2020).
24. P. B. Goodwin, Habit and hysteresis in mode choice. *Urb. Stud.* **14**, 95–98 (1977).
25. A. Dickinson, L. Weiskrantz, Actions and habits: The development of behavioural autonomy. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **308**, 67–78 (1985).
26. P. Lally, C. H. M. van Jaarsveld, H. W. W. Potts, J. Wardle, How are habits formed: Modelling habit formation in the real world. *Eur. J. Soc. Psychol.* **40**, 998–1009 (2010).
27. E. A. Thrailkill, S. Trask, P. Vidal, J. A. Alcalá, M. E. Bouton, Stimulus control of actions and habits: A role for reinforcer predictability and attention in the development of habitual behavior. *J. Exp. Psychol. Anim. Learn. Cogn.* **44**, 370–384 (2018).
28. R. Akaishi, K. Umeda, A. Nagase, K. Sakai, Autonomous mechanism of internal choice estimate underlies decision inertia. *Neuron* **81**, 195–206 (2014).

29. J. von Neumann, O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton Univ. Press, 1953).
30. D. Bernoulli, Exposition of a new theory on the measurement of risk. *Econometrica* **22**, 23–36 (1954).
31. H. Markowitz, The utility of wealth. *J. Polit. Econ.* **60**, 151–158 (1952).
32. D. Kahneman, A. Tversky, Subjective probability: A judgment of representativeness. *Cogn. Psychol.* **3**, 430–454 (1972).
33. G. Loomes, R. Sugden, Regret theory: An Alternative theory of rational choice under uncertainty. *Econ. J.* **92**, 805–824 (1982).
34. P. Dayan, L. F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (Massachusetts Institute of Technology Press, 2001), p. 460.
35. J. Li, N. D. Daw, Signals in human striatum are appropriate for policy update rather than value prediction. *J. Neurosci.* **31**, 5504–5511 (2011). [cited 2020 Nov 30].
36. S. Palminteri, M. Pessiglione, Opponent brain systems for reward and punishment learning: Causal evidence from drug and lesion studies in humans, in *Decision Neuroscience*, J.-C. Dreher, L. Tremblay, Eds. (San Diego: Academic Press, 2017), chap. 23, pp. 291–303.
37. M. Lebreton, K. Bacily, S. Palminteri, J. B. Engelmann, Contextual influence on confidence judgments in human reinforcement learning. *PLOS Comput. Biol.* **15**, e1006973 (2019).
38. C. J. Burke, M. Baddeley, P. N. Tobler, W. Schultz, Partial adaptation of obtained and observed value signals preserves information about gains and losses. *J. Neurosci.* **36**, 10016–10025 (2016).
39. D. Pischedda, S. Palminteri, G. Coricelli, The effect of counterfactual information on outcome value coding in medial prefrontal and cingulate cortex: From an absolute to a relative neural code. *J. Neurosci.* **40**, 3268–3277 (2020).
40. D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (W. H. Freeman and Co Ltd, 1982).
41. K. E. Conen, C. Padoa-Schioppa, Partial adaptation to the value range in the macaque orbitofrontal cortex. *J. Neurosci.* **39**, 3498–3513 (2019).
42. C. Padoa-Schioppa, A. Rustichini, Rational attention and adaptive coding: A puzzle and a solution. *Am. Econ. Rev.* **104**, 507–513 (2014).
43. G. Gigerenzer, The bias bias in behavioral economics. *Rev. Behav. Econ* **5**, 303–336 (2018).

44. M. G. Haselton, D. Nettle, P. W. Andrews, The evolution of cognitive bias, in *The Handbook of Evolutionary Psychology*, (John Wiley and Sons Ltd, 2015), pp. 724–46;
<https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470939376.ch25> [cited 27 July 2020]
45. C. F. Camerer, *Prospect Theory in the Wild: Evidence From the Field* (Pasadena, CA, California Institute of Technology, 1998); <https://resolver.caltech.edu/CaltechAUTHORS:20170811-150835361> [cited 30 January 2021]
46. M. J. Frank, L. C. Seeberger, R. C. O'reilly, By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science* **306**, 1940–1943 (2004).
47. E. R. Girden, *ANOVA: Repeated Measures* (SAGE, 1992).
48. R. S. Sutton, A. G. Barto, *Reinforcement Learning - An Introduction* (MIT Press, 1998).
49. R. A. Rescorla, A. R. Wagner, A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Class. Condition. II, Curr. Res. Theory* **2**, 64–99 (1972).
50. J. Hergueux, N. Jacquemet, Social preferences in the online laboratory: A randomized experiment. *Exp. Econ.* **18**, 251–283 (2015).
51. D. Kahneman, Maps of bounded rationality: Psychology for behavioral economics. *Am. Econ. Rev.* **93**, 1449–1475 (2003).
52. T. Shavit, D. Sonsino, U. Benzion, A comparative study of lotteries-evaluation in class and on the Web. *J. Econ. Psychol.* **22**, 483–491 (2001).
53. E. T. Miller, D. J. Neal, L. J. Roberts, J. S. Baer, S. O. Cressler, J. Metrik, G. A. Marlatt, Test-retest reliability of alcohol measures: Is there a difference between internet-based assessment and traditional methods? *Psychol. Addict. Behav.* **16**, 56–63 (2002).
54. K. Reinecke, K. Z. Gajos, LabintheWild: conducting large-scale online experiments with uncompensated samples, in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)* (Vancouver, BC, Canada, Association for Computing Machinery, 2015), p. 1364–1378; <https://doi.org/10.1145/2675133.2675246>

2.2.3 Conclusion

In this study, we demonstrated that reinforcement learning values are learned in a context-dependent manner that is compatible with range adaptation. Manipulation of task difficulty led to a paradoxical result: reducing task difficulty can, in some occasions, decrease choice optimality. Our findings show that context-dependent reinforcement learning induces, in some circumstances, economically suboptimal choices.

Chapter 3

The multiple facets of reinforcement valence in neuropsychiatric diseases

3.1 A meta-analysis

3.1.1 Introduction

Approaching rewards and avoiding punishments are core principles that govern the adaptation of behavior to the environment. Recent neuroscience research suggests that many psychiatric conditions involve behavioral dysfunctions that can be understood in terms of aberrant reinforcement processes, since reward and punishment learning might be underpinned by distinct brain systems. Therefore, one might wonder about the effects of neural perturbation, following drug administration and/or pathological conditions, on reward and punishment learning. For example, in the past decades, wealth of evidence has suggested that patients with psychiatric symptoms such as depression (Henriques et al. 1994, Chen et al. 2015, Rothkirch et al. 2017) and/or anxiety (Grillon et al. 2017, Mkrtchian et al. 2017) might have a hyposensibility to rewards (which should be sought) and hypersensibility to punishments (which should be avoided). On the other side, individuals with substance-related disorders might have an hypersensibility to rewards (Dayan 2009, Keiflin and Janak 2015, Nutt et al. 2015). These findings converge to the hypothesis that pathologies impacting the reward system also have an impact on the *valence bias*, which represents a deviation from the ability to learn equally from rewards and punishments.

To compare reward and punishment learning, typical reinforcement learning tasks are used to dissociate valence-specific and valence-independent processes. The implementation of the

comparison within the same task is necessary to avoid confounds with details of the design and to avoid framing effects. Indeed, individuals might reframe their expectations if they realize that they are in a reward- or punishment-learning task, i.e., they might change their reference point and, for instance, take an absence of reward as a punishment or an absence of punishment as a reward (Seymour and McClure 2008, Vlaev et al. 2011, Rangel and Clithero 2012, Palminteri et al. 2015). In other words, to investigate the valence bias, we want to avoid studies that focus on either only reward or only punishment because of the absence of a referential. We focus precisely on tasks with both reward and punishment learning, for which we can identify a bias, while controlling for a baseline performance. To this aim, we focused on two most influential human reinforcement learning papers, both described in section 1.2.3, which use two classical tasks to compare reward and punishment learning.

In a paper published in 2004 in *Science*, Frank and colleagues designed a task that we refer to as the "Hiragana task", due to the alphabet used for the stimuli. The Hiragana task is designed to reveal in a test session the type of learning (reward seeking versus punishment avoidance) that was operant during the training session. During the training session, participants are presented with fixed pairs of options (typically three pairs), materialized by Hiragana symbols and associated with different, reciprocal probabilities of winning or losing. During the test session, participants are asked to identify the best option, among novel binary combinations, in the absence of feedback. The capacity to correctly identify the best option (choose A) and reject the worst (avoid B) is taken as a measure of the capacity to learn from positive and negative prediction errors (Figure 15).

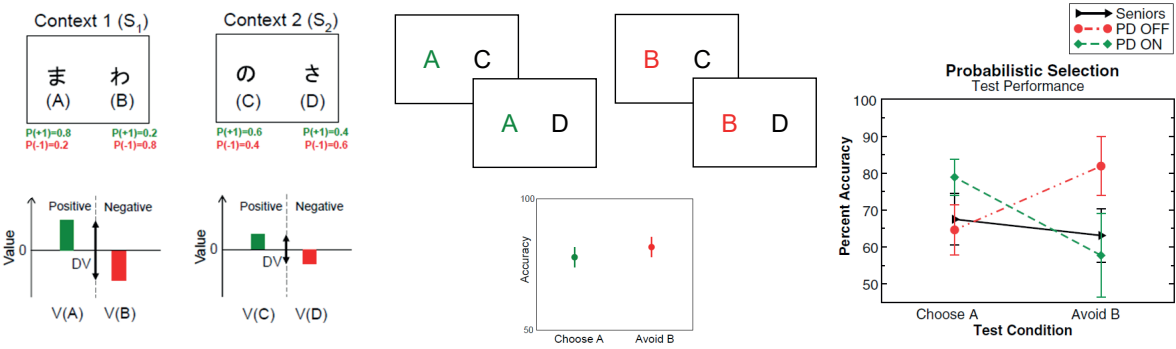


Figure 15. Hiragana task. Decision screens in two possible contexts (pairs of symbols), the probabilistic contingencies associated with each symbol, the two option values (DV is the decision value, i.e., the difference between the two options), and the main performance measure (choice accuracy) expected from a healthy participant. Note that values are the actual values that participants have to learn, before learning they are equal to zero. Figure adapted from Frank et al. 2004 and Palminteri and Pessiglione 2017.

By comparing the choose A / avoid B metrics in three groups of unmedicated PD patients (PD OFF), medicated PD patients (PD ON) and senior controls, Frank and colleagues showed that unmedicated PD patients learned better from punishments than from rewards, while medicated PD patients learned better from rewards than from punishments (Frank et al. 2004). The results further agree with the hypothesis that the depletion of dopamine in unmedicated PD patients leads to a lower tonic activity threshold, and therefore are sensitive to a drop in the activity when a punishment occurs, and insensitive to phasic activity due to a reward, because it doesn't reach a learning threshold. On the contrary, medicated PD patients have a higher tonic threshold and are insensitive to the punishment drop but sensitive to the reward burst (Palminteri and Pessiglione 2017).

In a paper published in 2006 in *Nature*, Pessiglione and colleagues designed a task that we refer to as the "Agathodaimon task", also due to the alphabet used for the stimuli. The Agathodaimon task is designed to compare reward and punishment learning directly during the training session. Participants are also presented with fixed pairs of symbols (typically two pairs), now materialized by Agathodaimon symbols, with the crucial difference that rewards and punishments are never mixed within a pair. Some pairs of options are associated with reciprocal probabilities of winning or getting nothing, and others with reciprocal probabilities of losing or getting nothing. Typically, the rate of correct choice (i.e., choosing the most rewarding or the least punishing option) is extracted on a trial-by-trial basis to assess the capacity to learn from rewards versus punishments (Figure 16).

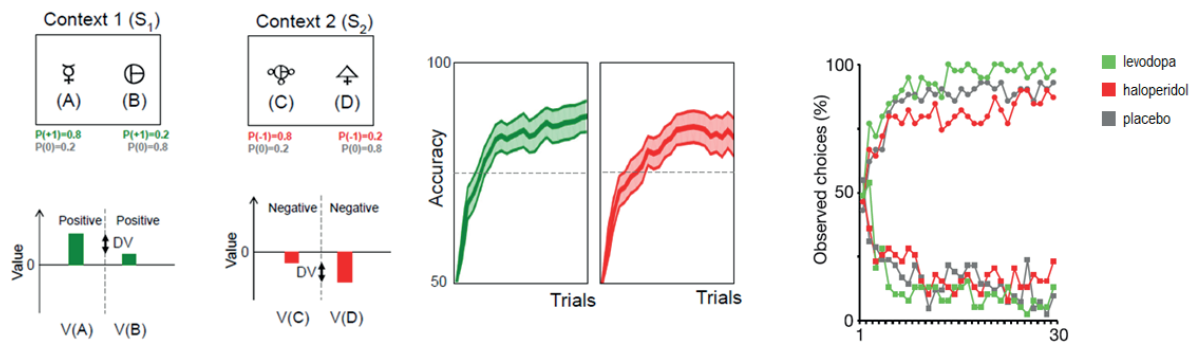


Figure 16. Agathodaimon task. Decision screens in two possible contexts (pairs of symbols), the probabilistic contingencies associated with each symbol, the two option values (DV is the decision value, i.e., the difference between the two options), and the main performance measure (choice accuracy) expected from a healthy participant. Note that values are the actual values that participants have to learn, before learning they are equal to zero. Figure adapted from Pessiglione et al. 2006 and Palminteri and Pessiglione 2017

By comparing the ability to learn from rewards versus punishments in three groups of healthy volunteers, receiving either a dopamine agonist (levodopa) enhancing dopaminergic function, a dopamine antagonist (haloperidol) reducing dopaminergic function, or a placebo, Pessiglione and colleagues show that participants treated with levodopa have a greater propensity to choose the option with the highest expected value in rewarding pairs compared to participants treated with haloperidol. The difference was not significant in punishing pairs, showing evidence of an asymmetry of drug effects between learning from rewards and from punishments (Pessiglione et al. 2006).

Together, these two papers support the hypothesis that dopamine has a specific involvement in reward learning and have been prominent to clinical research. Over the past 15 years, a vast amount of work has contributed to the study of dopamine-related pathologies.

3.1.2 Methods

We are conducting a meta-analysis on clinical papers citing one of these two most influential human reinforcement learning papers. From the electronic database search on Google Scholar, we found 2561 papers citing at least one of the two pioneer papers. After screening, I identified 115 publications using the exact same task and contingencies as the authors (Figure 17).

We are interested in the accuracy in the reward seeking and punishment avoidance conditions and the choose A / avoid B metric. Out of the 115 publications, 24 papers (including the two original papers) provided a table reporting the mean and standard deviation for these metrics, averaged across all relevant trials and calculated separately for each experimental group and control group. After contacting the remaining 91 authors, I have gathered the data for 49 additional papers. Over the 42 remaining publications, I managed to read the metrics on the figures that were provided in 14 papers, using Web Plot Digitizer program (<https://apps.automeris.io/wpd/>). The last 28 papers did not provide any figure, table, or text mention enabling us to infer the metrics. Among the 237 different group measures, we excluded patients receiving placebo from the preliminary analyses. I will now present the preliminary results from the 87 studies for which we have gathered the data, hoping to shed some light on the valence bias in clinical studies. For each study, we extracted the effect size d , which we calculated as follows:

$$d = \frac{M_R - M_P}{\sqrt{\frac{s_R^2 + s_P^2}{2}}} \quad (3.1)$$

where M_R, s_R and M_P, s_P represent the mean accuracy (% of correct choices) and standard

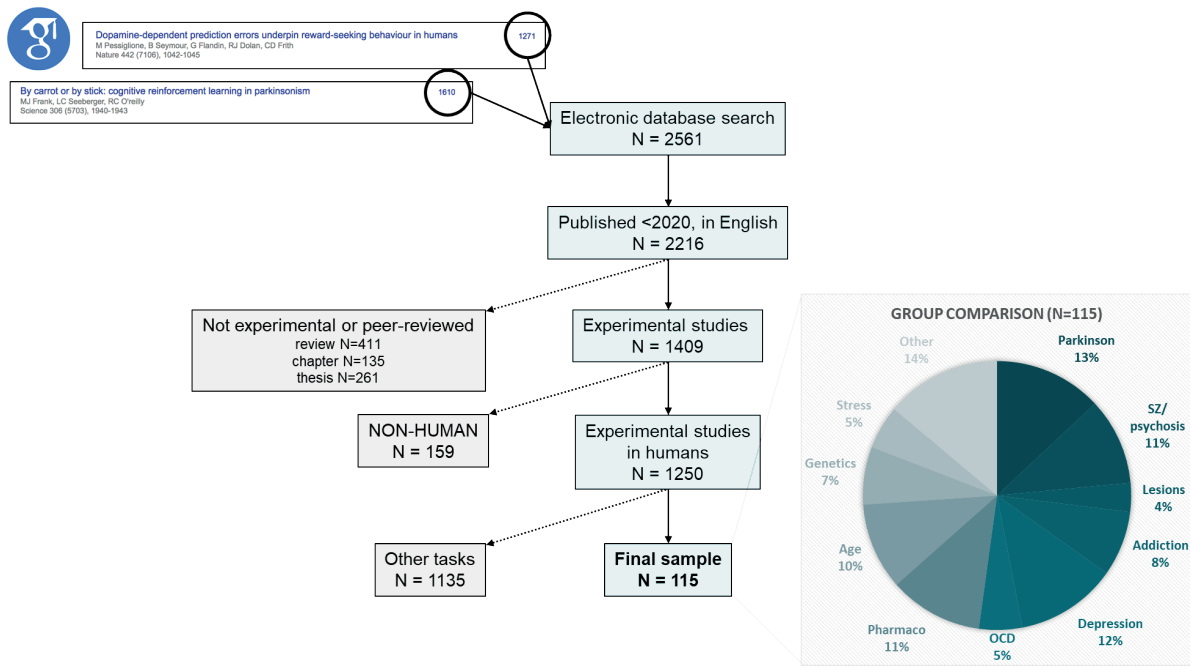


Figure 17. Meta-analysis procedure. From the electronic database search on Google Scholar, I found 2561 publications citing at least one of the two papers [Frank et al. 2004](#) and [Pessiglione et al. 2006](#) (note that the Google Scholar numbers are higher because they do not account for duplicates). Of them, 2216 were published between 2004 and 2019 (included), in English. Of them, 1409 were experimental studies, excluding book chapters, master thesis, PhD thesis, and reviews. Of them, 1250 were performed in humans, not animals. Of them, 115 consisted in a clinical study comparing groups of participants and using the same task as [Frank et al. 2004](#) or [Pessiglione et al. 2006](#).

deviation in the gain or reward contexts and loss or punishment contexts, respectively. Note that, in our case, the effect size is calculated within the same group, so the sample size does not differ between reward and punishment measures.

3.1.3 Preliminary results

When looking at aggregated performance, we found a main effect of valence (reward vs. punishment, $t(217) = 2.39$, $p = .018$) and found the effect size to be significantly different from 0 ($t(217) = 2.63$, $p = .0092$, Figure 18A). This effect seems to be driven by the effect size in patients ($t(104) = 2.53$, $p = .013$), since the effect size in controls does not differ from 0 ($t(112) = 1.11$, $p = .27$, Figure 18B). We found a main effect of group on the correct choice rate, the average accuracy being overall higher in controls than in patients ($t(216) = 3.25$, $p = .0013$, Figure 18B). We found a significant negative Spearman's correlation between age and average performance ($\rho(205) = -0.29$, $p < .0001$, Figure 18C), which was present in both groups (controls: $\rho(102) = -0.20$, $p = .04$, patients: $\rho(101) = -0.29$, $p = .003$), but did not find any significant correlation between age and the absolute value of the effect size ($\rho(205) = 0.13$, $p = .06$), nor

when splitting between controls ($\rho(102) = 0.11$, $p = .26$) and patients ($\rho(101) = 0.16$, $p = .12$).

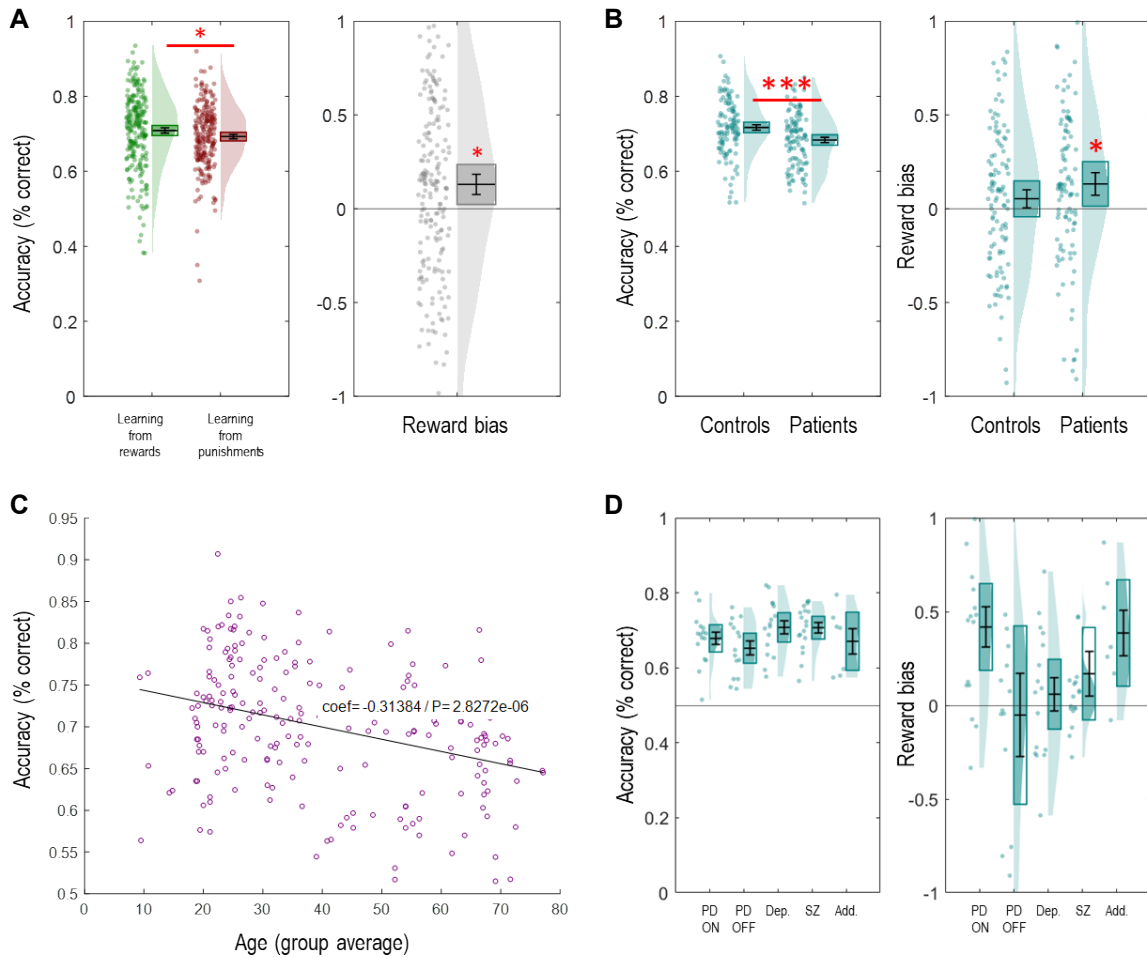


Figure 18. Meta-analysis preliminary results. Each dot represents one group. **(A)** Aggregated performance split over reward- (green) and punishment- (red) learning. **(B)** Aggregated performance split over controls and patients groups. **(C)** Spearman’s correlation between group’s mean age and group’s mean performance. **(D)** Aggregated performance split over the five categories for which we had a sufficient number of studies.

We further analyzed performance and effect size in the 5 groups of patients for which we had gathered the data in the larger number of studies (PD ON: 16 studies; PD OFF: 15 studies; Depression: 17 studies; Schizophrenia and/or Psychosis: 18 studies; Addiction: 7 studies). We found the overall effect size of Addiction studies to be significantly different from zero ($t(6) = 3.21$, $p = .018$, Figure 18D), suggesting a valence bias towards learning from positive rewards, as hypothesized. We found a positive valence bias in the studies involving groups of medicated Parkinson’s disease patients (PD ON, $t(15) = 3.82$, $p = .0017$, Figure 18D), coherent with the results from Frank et al. 2004 and the following research line. However, at first glance, we did not find any other significant bias in the other groups. This might be due to the fact that unmedicated patients with Parkinson’s disease, depression, or schizophrenia/psychosis can

have widely different phenotypes, even when diagnosed with the same pathology.

To assess the valence bias in Parkinson’s disease with more precision, we ran a random-effects model using the R package *metafor*. Results support the significant valence in medicated PD (mean difference $M = 0.09$, 95% CI [0.04, 0.13], $p = .0002$, Figure 19A), indicating better learning from rewards than from punishments. The bias was not significant in unmedicated PD (mean difference $M = -0.01$, 95% CI [-0.10, 0.08], $p = .83$, Figure 19B).

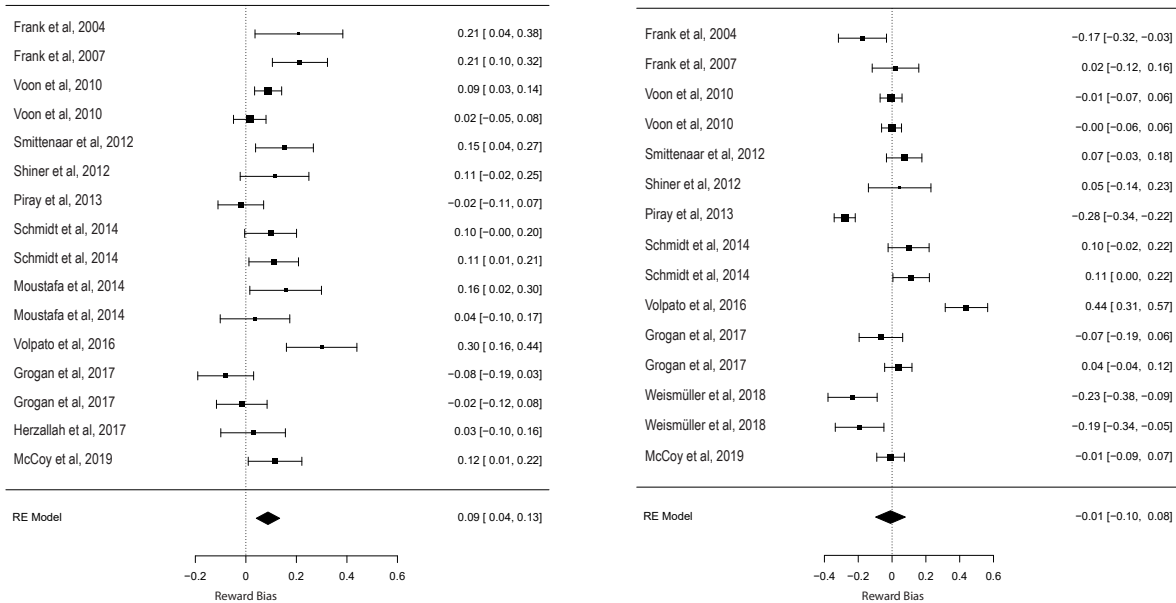


Figure 19. Forest plots for medicated (left) and unmedicated (right) PD patients. The plots show the results of the individual studies together with their 95% confidence intervals. Squares show the average effect size in each group; square size indicates the weight of the study in the meta-analysis, i.e., the sample size. Bottom diamond represents the summary of the random-effect model, with the center corresponding to the estimate and the left/right edges indicating the confidence interval limits.

3.1.4 Conclusion

After screening more than 2500 studies published between 2004 and 2019, we found 115 clinical studies using one of the two tasks from Frank et al. 2004 or Pessiglione et al. 2006. Over the 90 studies from which we managed to gather the data, we found an overall positive valence bias, which seemed to be driven by results in patients, whereas healthy controls learn equally from rewards and punishments. However, over the 20 different pathologies studied in the 115 publications, only a few were in sufficient number to estimate the meta-analytically reliable effect. We found that medicated patients with PD and individuals with substance-related disorders show behavioral evidence for a positive valence bias. The results are inconclusive for unmedicated

patients with PD, depression, or schizophrenia and/or psychosis, as there was no significant bias and a high inter-study variance. Therefore, further work is needed to dig deeper into the results of this meta-analysis.

First, in collaboration with Yulia Worbe, neurologist and professor of neurophysiology, we are aiming at clustering the different pathologies depending on the neural networks involved, to increase the number of studies in each group. In particular, 11% of the studies involved pharmacological groups of healthy controls, as studied in [Pessiglione et al. 2006](#). We are aiming to compare pharmacological studies using dopamine agonists (such as levodopa, cabergoline) or antagonists (such as haloperidol, amisulpride). Second, we will bring our attention to these two tasks, which are widely used when comparing reward- and punishment-learning. Using different reinforcement learning models, implementing the valence bias in several ways, we plan to simulate different forms of valence bias, to determine if they are recoverable in these tasks. We intend to simulate at least three models, implementing:

- context-dependence ([Vlaev et al. 2011](#), [Palminteri et al. 2015](#), [Bavard et al. 2020](#)). The model has a parameter assessing contextualization of values, allowing different value updates in reward and punishment contexts.
- positivity bias ([Sharot 2011](#), [Palminteri et al. 2016](#)). The model has two separate learning rates to learn from positive or negative prediction errors, allowing different learning weights from rewards and punishments.
- loss aversion ([Kahneman and Tversky 1979](#)). The model has a loss aversion parameter, allowing for bigger loss aversion than reward seeking.

Comparing the recoverability of the different models in the two tasks will help us determine which behavioral signature they generate. Finally, we cannot cope for the heterogeneity between studies. For example, a group of patients with bipolar disorder might be tested in specific stages, which might include different patterns of behavior between groups within the same pathological category ([Huys et al. 2014](#)). Therefore, in general, having more information on the psychiatric symptoms or individual conditions can help us make subgroups of patients and perhaps find a specific pattern. This would be in line with a general insight in computational psychiatry, that having discrete diagnostic categories leads to high inter-individual variance within each category ([Gillan and Daw 2016](#)). To account for this general issue, a transnosographic approach allows to inform both categorical and dimensional approaches, which is in direct link with the next project of this PhD.

3.2 A large-scale study

3.2.1 Introduction

This project addresses the fundamental question of addiction in humans and the investigation of its underlying mechanisms. The Diagnostic and Statistical manual of Mental disorders (DSM) defines addiction across several criteria, including the lack of control over substance use (in terms of frequency and duration), the inability to stop using despite the efforts, the considerable time spent consuming despite the consequences, and the irrepressible urge to use. More recently, the latest version of the DSM (DSM-5, [American Psychiatric Association 2013](#)) stresses the term "substance-related disorder" instead of "addiction". The DSM-5 states that addiction strongly correlates with a change in neural circuitry, which can persist even after detoxification. Neural changes associated with addiction particularly affect the dopaminergic circuit, involved in reward and learning. Thus, wealth of recent evidence suggests that the pathology of addiction may be associated with a disorder of reinforcement learning ([Dayan 2009](#), [Huys et al. 2014](#), [Wise and Koob 2014](#), [Keiflin and Janak 2015](#), [Nutt et al. 2015](#)).

In the first part of this PhD, we have developed a satisfactory reinforcement learning model of context-dependence, which accounts for range-adapting coding and rescales outcome values accordingly. This model includes a free parameter, the contextual learning rate α_R , which allows for the relative encoding of values. To account for the valence bias that can be observed when comparing accuracy in reward or punishment contexts, we improved the model to allow different updates when the prediction error is positive or negative. This manipulation has been shown to explain biased behaviors such as optimism ([Sharot 2011](#)) and confirmation bias ([Palminteri et al. 2017a](#)). In the next study, we aimed at correlating our model's parameters with different dimensions of several psychiatric symptoms. Based on a study from Gillan and colleagues published in *Elife* in 2016 ([Gillan et al. 2016](#)), we used a transdiagnostic approach as performed in computational psychiatry. To concentrate our attention on addiction disorders, participants filled in ten self-assessed questionnaires including three substance misuse scales. Five hundred participants performed an online modified version of the previously presented probabilistic selection task, where seeking rewards and avoiding punishments could be dissociated. We hypothesized that addiction symptoms would correlate with an imbalance in learning from rewards or from punishments. In particular, we expected the optimism bias observed in healthy participants ([Lefebvre et al. 2017](#), [Palminteri et al. 2017a](#)) would grow with addiction scores.

Exploratory results did not allow us to link any of our model's parameters to the dimensions we

found with the transdiagnostic factors. However, behavioral results replicated previous findings, and computational analyses allowed for a validation of our model. Further analyses might benefit from the task and the model being tested with another set of questionnaires; to dig deeper into the links between behavior and psychiatric symptoms, we consider developing better modeling that includes a clear analysis of reaction times.

3.2.2 Methods

Participants

We recruited 500 participants (242 females, aged 29.55 ± 10.26 years) from the Prolific platform (www.prolific.co). The research was carried out following the principles and guidelines for experiments including human participants provided in the declaration of Helsinki (1964, revised in 2013). The Inserm Ethical Review Committee / IRB00003888 approved the study on November 13th, 2018 and participants were provided written informed consent prior to their inclusion. To sustain motivation throughout the experiment, participants were given a bonus depending on the number of points won in the experiment (average money won in pounds: 4.52 ± 0.52 , average performance against chance: $M = 0.70 \pm 0.13$, $t(499) = 33.71$; $p < .0001$).

Exclusion criteria: participants were excluded if they displayed a clear side bias, i.e., if they chose the same side more than 95% of the trials ($N=2$). We approximated participants' total reaction time over the whole task by a normal distribution and removed outliers at a significance level of $p < .05$ ($N=25$). In total, 27/500 participants (5.4%) were excluded, leading to a final sample of 473 participants.

Behavioral task

Participants performed an online version of a probabilistic instrumental learning task adapted from previous studies (Frank et al. 2004, Pessiglione et al. 2006, Palminteri et al. 2015, Bavard et al. 2018, 2021). After checking the consent form, participants received written instructions explaining how the task worked and that their final payoff would be affected by their choices in the task. During the instructions, the possible outcomes in points (-1pt, 0pt and +1pt) were explicitly showed as well as their conversion rate (1pt = 2 pence). After the instructions, participants were required to correctly answer a 3-item basic comprehension test regarding the rules of the reinforcement-learning task. If participants failed to answer the questions correctly, they were sent back to the beginning and required to repeat the instructions prior to re-taking the comprehension test. The questions were followed by a short training session of 18 trials

aiming at familiarize the participants with the response modalities. Participants could repeat the training session up to two times and then started the actual experiment if their performance reached a threshold of 60% correct answers in the previous training session. In our task, options were materialized by abstract stimuli (cues) taken from randomly generated identicons, colored such that the subjective hue and saturation were very similar according to the HSLUV color scheme (www.hsluv.org). On each trial, two cues were presented on both sides of the screen. The side in which a given cue was presented was pseudo-randomized, such that a given cue was presented an equal number of times in the left and the right. Participants were required to select between the two cues by clicking on the cue. The choice window was self-paced. A brief delay after the choice was recorded (500 ms), the outcome was displayed for 1000 ms. There was no fixation screen between trials.

The task consisted in one learning phase and a transfer phase, followed by a series of self-assessed questionnaires. During the learning phase, cues appeared in 3 fixed pairs. Each pair was presented 40 times, leading to a total of 120 trials. Within each pair, the two cues were associated to an outcome with reciprocal probabilities (0.75/0.25 and 0.25/0.75). At the end of the trial, the cues disappeared and the selected one was replaced by the outcome ("-1", "0", or "1") (Figure 20). During the transfer phase, the 6 cues from the learning phase were presented in all possible binary 15 combinations (not including pairs formed by the same cue). Each pair of cues was presented eight times, leading to a total of 120 trials. Instructions for the transfer phase were provided orally after the end of the learning phase. Participants were explained that they would be presented with the same cues, but that all pairs would not have been necessarily displayed together before. On each trial, they had to indicate which of the cues was the one with the highest value. In order to prevent explicit memorizing strategies, the outcome was not provided in order not to modify the option values learned during the learning phase.

Self-report psychiatric questionnaires

After the transfer phase, participants completed self-report questionnaires assessing:

- alcohol addiction using the Alcohol Use Disorders Identification Test (AUDIT-10, [Saunders et al. 1993](#))
- cannabis addiction using the Cannabis Abuse Screening Test (CAST-6, [Legleye et al. 2007](#))
- nicotine addiction using the Fagerström Test for Nicotine Dependence (FTND-6, [Heatherton et al. 1991](#))

- anxiety and depression using the Hospital Anxiety and Depression Scale (HADS-14, [Zigmond and Snaith 1983](#))
- hypomania using the Hypomanic Personality Scale (HPS-20, [Meads and Bentall 2008](#))
- social anxiety using the Liebowitz Social Anxiety Scale (LSAS-24, [Liebowitz 1987](#)).
- obsessive-compulsive disorder (OCD) using the Obsessive-Compulsive Inventory – Revised (OCI-R-18, [Foa et al. 2002](#))
- schizotypal traits using the Peters et al. Delusions Inventory (PDI-21, [Peters et al. 1999](#))
- eating disorders using the Reward-based Eating Drive (RED-X5, [Vainik et al. 2019](#))
- sensation seeking using the Sensation Seeking Scale (SSS-13, [Zuckerman et al. 1964](#))

The order of these self-report assessments was fully randomized across participants.

Model space

We analyzed our data with variation of simple associative learning models ([Rescorla and Wagner 1972](#), [Sutton and Barto 1998](#)). The goal of all models is to estimate in each choice context (or state) the expected reward (R) of each option and pick the one that maximizes this expected reward. We modeled participants' choice behavior using a softmax decision rule representing the probability for a participant to choose one option a over the other option b, as in all the studies in this thesis.

We compared three alternative computational models: the ABSOLUTE model, which encodes outcomes in an absolute scale independently of the choice context in which they are presented, the RANGE model which tracks the value of the maximum reward in each context and normalizes the actual reward accordingly, rescaling rewards between 0 and 1, and the A-RANGE model which, in addition to normalizing rewards, allows for different learning from positive and negative prediction errors.

ABSOLUTE model. The outcomes are encoded as the participants see them (i.e., their

objective value). A positive outcome is encoded as its actual positive value (in points):

$$R_{\text{OBJ},t} \in \{-1, 0, 1\} \quad (3.2)$$

RANGE model. The outcomes are encoded on a context-dependent relative scale. On each trial, the relative reward $R_{\text{RAN},t}$ is calculated as follows:

$$R_{\text{RAN},t} = \frac{R_{\text{OBJ},t} - R_{\text{MIN},t}(s)}{R_{\text{MAX},t}(s) - R_{\text{MIN},t}(s)} \quad (3.3)$$

where s is the decision context (i.e., a combination of options), R_{MAX} and $R_{\text{MIN},t}$ are context-dependent variables, initialized to 0 and updated at each trial t if the outcome is greater or smaller than its current value:

$$R_{\text{MAX},t+1}(s) = R_{\text{MAX},t}(s) + \alpha_R(R_{\text{OBJ},t} - R_{\text{MAX},t}(s)) \quad \text{if } R_{\text{OBJ},t} > R_{\text{MAX},t}(s) \quad (3.4)$$

$$R_{\text{MIN},t+1}(s) = R_{\text{MIN},t}(s) + \alpha_R(R_{\text{OBJ},t} - R_{\text{MIN},t}(s)) \quad \text{if } R_{\text{OBJ},t} < R_{\text{MIN},t}(s) \quad (3.5)$$

Accordingly, outcomes are progressively normalized so that eventually $R_{\text{RAN},t} \in [0, 1]$. The chosen option values and prediction errors are updated with the same rules as in the ABSOLUTE model, where α_c is the learning rate for the chosen (c) option and δ_c is the prediction error term:

$$Q_{t+1}(s, c) = Q_t(s, c) + \alpha_c * \delta_{c,t} \quad (3.6)$$

$$\delta_{c,t} = R_{c,t} - Q_t(s, c) \quad (3.7)$$

A-RANGE model. (Asymmetric RANGE) The outcomes are encoded exactly as the RANGE model, but the option value update is performed with an additional free parameter, allowing for asymmetric learning from positive and negative prediction errors:

$$Q_{t+1}(s, c) = \begin{cases} Q_t(s, c) + \alpha_c^+ * \delta_{c,t} & \text{if } \delta_{c,t} > 0 \\ Q_t(s, c) + \alpha_c^- * \delta_{c,t} & \text{if } \delta_{c,t} < 0 \end{cases} \quad (3.8)$$

3.2.3 Preliminary results

Behavioral results

In the learning phase, participants performed above chance level 0.5 (average performance 0.71 ± 0.16 , $t(472) = 28.5$, $p < .0001$, $d = 1.31$). We found a main effect of valence ($F(1.45, 684.98) = 33.08$, $p < .0001$, $\eta_p^2 = .04$, Huynh–Feldt corrected, Figure 20). Interestingly, we found that

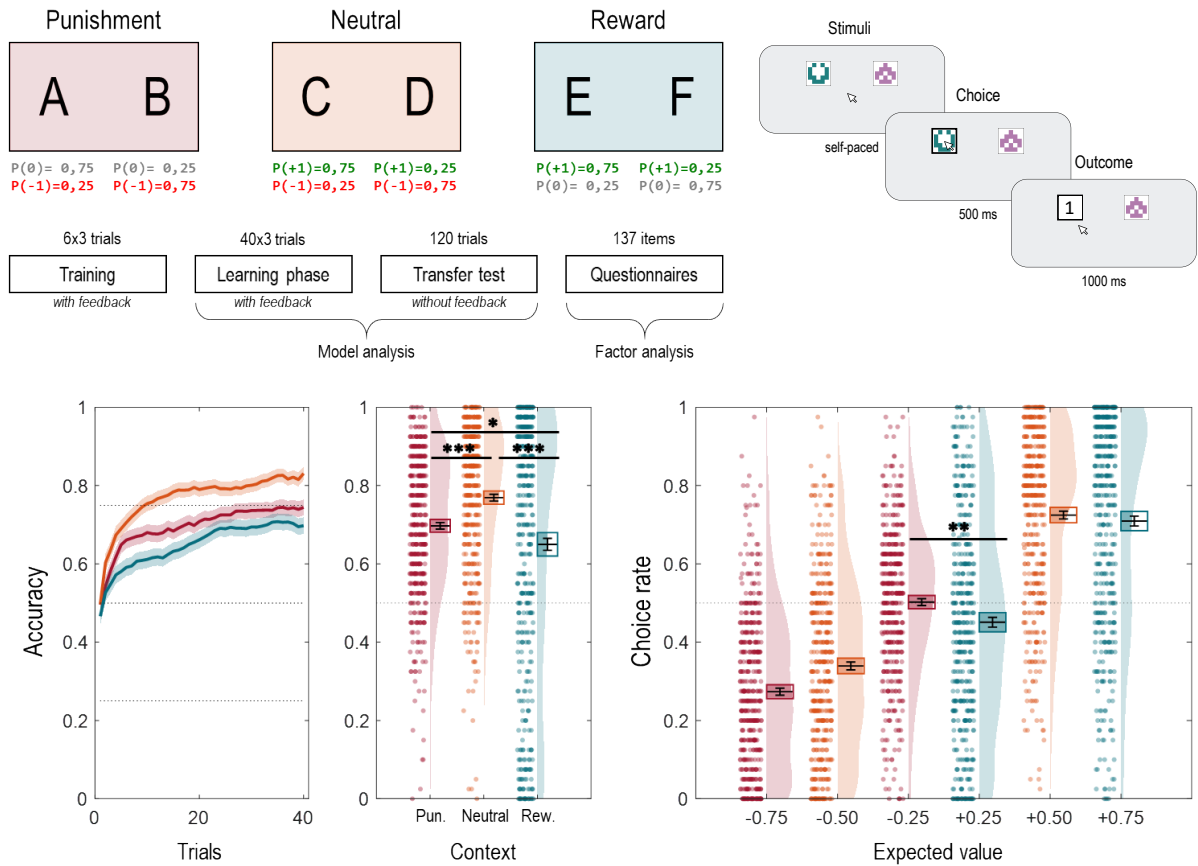


Figure 20. Task design and behavioral results. Top: choice contexts in the learning phase with probabilities and magnitudes, and successive screens of a typical trials (duration is given in milliseconds). Bottom: left, correct choice rate in the learning phase as a function of the choice context; right, choice rate in the transfer test.

participants had a higher performance in the punishment context compared to the reward context ($t(472) = 2.78$, $p = .017$, $d = 0.05$, Bonferroni corrected), due to a pool of participants performing below chance level in the reward context (Figure 20).

In the transfer phase, the correct choice rate was significantly higher than chance, thus providing evidence for significant value transfer and retrieval (average performance 0.69 ± 0.14 , $t(472) = 31.02$, $p < .0001$, $d = 1.43$). As we expected from previous studies (Palminteri et al. 2015, Bavard et al. 2018, 2021), the analysis of the transfer phase revealed that option preference did not linearly follow the objective ranking based on their absolute expected value: the favorable option of the punishment context was chosen more often than the less favorable option of the reward context ($t(472) = 3.03$, $p = .0026$, $d = 0.14$), despite its expected value being smaller (Figure 20). We found a negative Spearman correlation between the average performance in the learning phase and this difference in choice rate between intermediary values ($\rho(471) = -.59$, $p < .0001$, Figure 21), suggesting that a more effective learning will lead to an increased

violation of rational choices when the options are extrapolated from their original context.

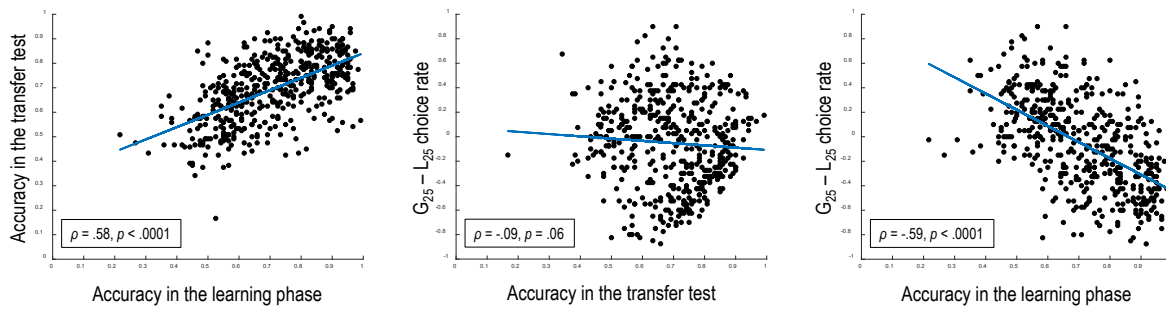


Figure 21. Behavioral sign of range adaptation. Spearman’s correlations between accuracy in the learning phase, accuracy in the transfer test, and transfer test choice rate difference between the intermediary options G_{25} and L_{25} .

Computational modeling

We compared three reinforcement learning models: the ABSOLUTE model which is an adaptation of Q-learning, the RANGE model which normalizes outcomes, and the A-RANGE model which allows for asymmetric learning from positive and negative prediction errors.

In terms of averaged model simulations, the RANGE model seems equivalent to the A-RANGE model (Figure 22A). However, quantitative model comparison suggests that the A-RANGE model is a better fit to the data (BIC_{RAN} vs. BIC_{A-RAN} , $t(472) = 3.80$, $p = .00017$, $d = 0.17$). Moreover, when focusing on the distribution of the simulations in the reward context, the A-RANGE model captures part of the pool of under-performing participants. This is because its implementation, with two distinct learning rates, allows for asymmetric learning: as in previous studies (Palminteri et al. 2017a, Lefebvre et al. 2017), we found the positive learning rate α^+ to be significantly higher than the negative learning rate α^- ($t(472) = 6.65$, $p < .0001$, $d = 0.31$). This suggests that null outcomes in the reward context (i.e., a negative prediction error) do not have the same weight as null outcomes in the punishment context (i.e., a positive prediction error), and that a participant can stick with the wrong option in the reward context whereas it is more unusual in the punishment context. As shown in Figure 22B, the A-RAN model was the closest to capture this behavioral effect, compared to the RAN model which does not account for it.

Factor analysis

The factor analysis was performed using the nFactor package in R, on the 137 items from the 11 self-assessed questionnaires. Factor selection was based on a Scree test (Cattell’s criterion,

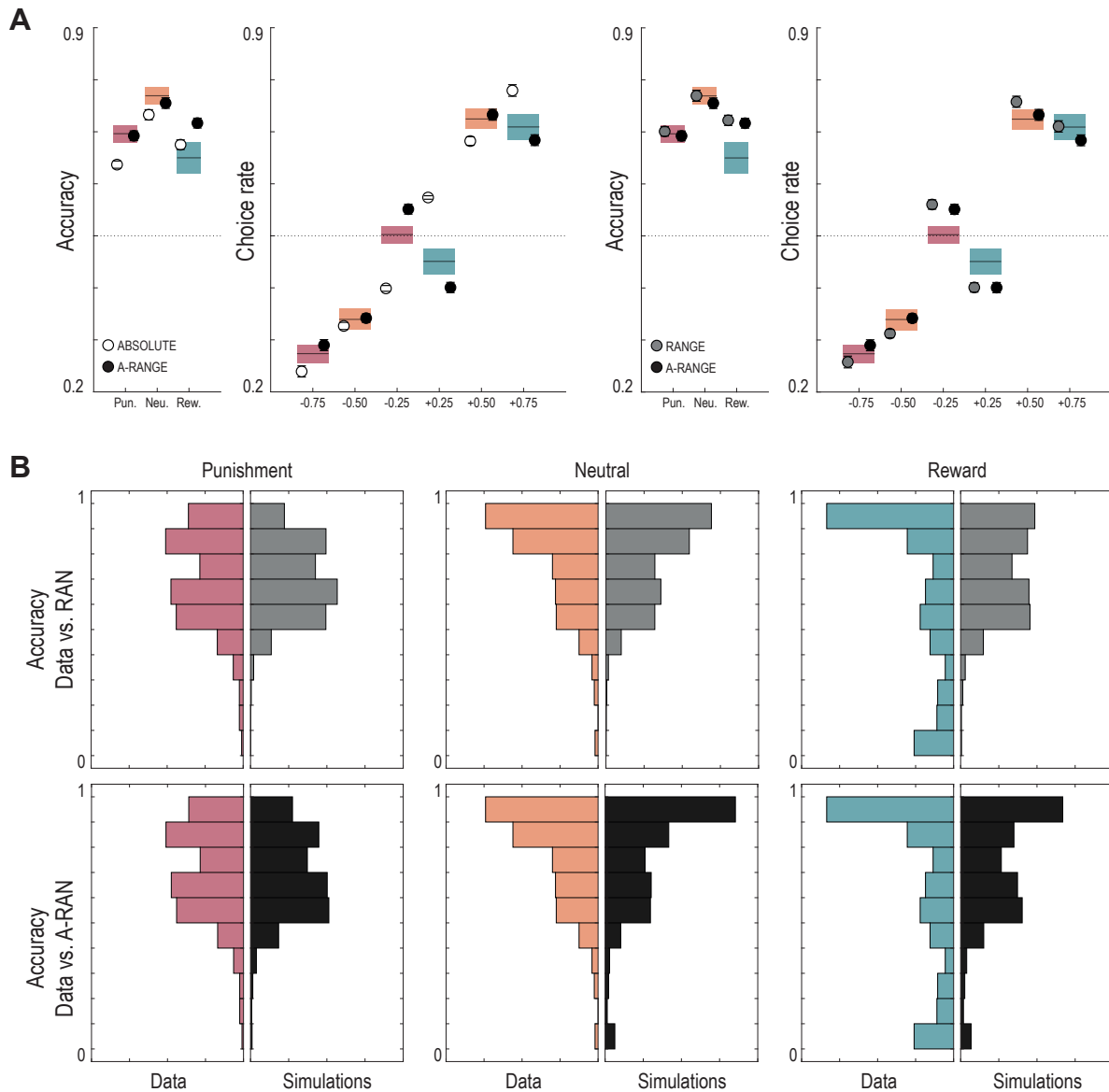


Figure 22. Model simulations. (A) Left: Model simulations of ABSOLUTE (white) and A-RANGE (black) models over the behavioral data (mean and 95% confidence interval) in each context and in the transfer test. Right: Model simulations of RANGE (gray) and A-RANGE (black) models over the behavioral data (mean and 95% confidence interval) in each context and in the transfer test. (B) Distribution of the behavioral data and model simulations, in each learning context.

Cattell 1966) and using the Cattell-Nelson-Gorsuch (CNG) procedure, which retained three factors (Figure 23A) that we labeled "Social Anxiety", "Compulsivity" and "Addiction", based on the 10 strongest individual item loadings (Figure 23B).

Factor 1 was labeled "Social anxiety", as it was dominated by items from the Social Anxiety questionnaire (0.62 ± 0.12), and had a contribution from Anxiety (0.26 ± 0.12) and a low contribution from Depression (0.21 ± 0.12). Interestingly, this factor had low negative contributions from the sensation seeking scale (-0.18 ± 0.08) (Table 1).

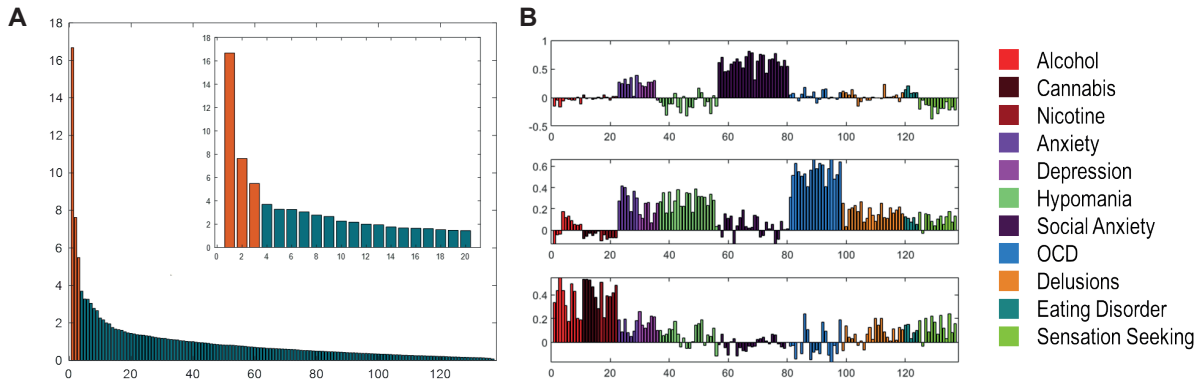


Figure 23. Trans-diagnostic factors. (A) The factor analysis was performed on the 137 questionnaire items and suggested that 3-factor solution best explained these data. Factors were labeled "Social anxiety", "Compulsivity" and "Addiction". (B) Item loadings for each factor are presented, color-codes indicate the questionnaire from which each item was drawn.

Factor 2 was labeled "Compulsivity", the highest average loadings came from the OCD questionnaire (0.54 ± 0.10), followed by Anxiety (0.32 ± 0.08) and Hypomania (0.27 ± 0.09) (Table 1).

Factor 3 was labeled "Addiction", the highest average loadings came from the three substance consumption questionnaires: Cannabis Abuse (0.45 ± 0.10), followed by Nicotine Dependence (0.40 ± 0.11) and Alcohol Use (0.36 ± 0.13) (Table 1).

	Factor 1		Factor 2		Factor 3	
	<i>mean</i>	<i>std</i>	<i>mean</i>	<i>std</i>	<i>mean</i>	<i>std</i>
Alcohol	-0,07	0,06	0,04	0,09	0,36	0,13
Cannabis	0,00	0,03	-0,05	0,03	0,45	0,10
Nicotine	0,00	0,03	-0,08	0,02	0,40	0,11
Anxiety	0,26	0,12	0,32	0,08	0,13	0,06
Depression	0,21	0,12	0,18	0,08	0,17	0,06
Hypomania	-0,11	0,14	0,27	0,09	0,06	0,09
Social anxiety	0,62	0,12	0,03	0,09	-0,02	0,05
OCD	0,03	0,08	0,54	0,10	-0,02	0,12
Delusions	0,02	0,08	0,17	0,06	0,07	0,08
Eating disorder	0,11	0,05	0,10	0,03	0,10	0,05
Sensation seeking	-0,18	0,08	0,10	0,06	0,13	0,08

Table 1. Labeling the factors. Means and standard deviations of loadings for Factor 1 "Social Anxiety", Factor 2 "Compulsivity" and Factor 3 "Addiction" for each questionnaire.

We found significant correlations between the scores of questionnaires assessing symptoms shared by several disorders, such as anxiety and depression ($\rho(456) = 0.52$, $p < .0001$) or social anxiety ($\rho(456) = 0.51$, $p < .0001$). Interestingly, we found a negative correlation between scores of social

anxiety and sensation seeking ($\rho(456) = -0.27, p < .0001$). Coherently with factor loadings, we found a strong positive correlation between Social Anxiety factor scores and questionnaire scores from Social Anxiety ($\rho(456) = 0.90, p < .0001$), as well as a negative correlation with Sensation Seeking ($\rho(456) = -0.42, p < .0001$), Hypomania ($\rho(456) = -0.25, p < .0001$) and Alcohol ($\rho(456) = -0.13, p = .007$). Compulsivity factor scores correlated positively with questionnaires scores from OCD ($\rho(456) = 0.85, p < .0001$), Hypomania ($\rho(456) = 0.66, p < .0001$), and Delusions ($\rho(456) = 0.43, p < .0001$). Addiction factor scores correlated positively with questionnaire scores from Alcohol ($\rho(456) = 0.67, p < .0001$), Cannabis ($\rho(456) = 0.48, p < .0001$) and Nicotine ($\rho(456) = 0.46, p < .0001$), as well as Sensation Seeking ($\rho(456) = 0.38, p < .0001$)(Figure 24).

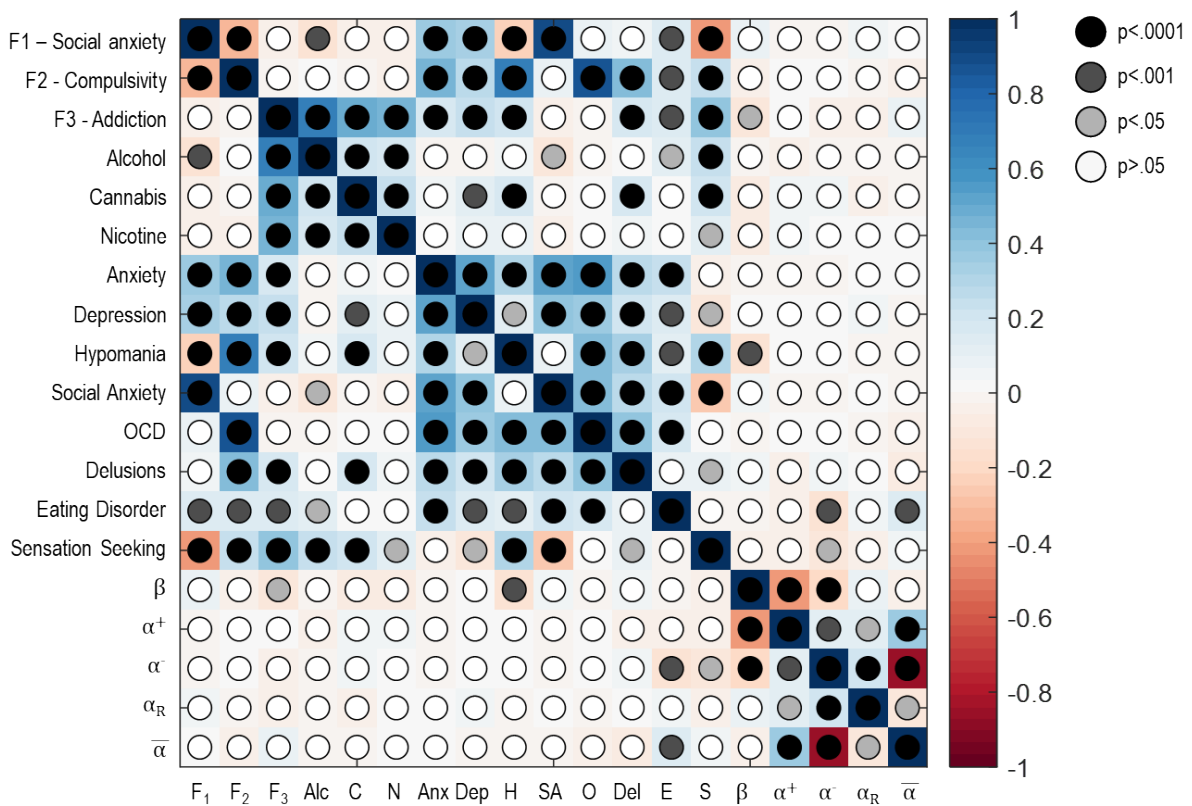


Figure 24. Correlations between factor loadings, questionnaire scores, and model parameters. The square color represent the value of Spearman’s ρ . The circle shade represent the significance for each correlation. $\bar{\alpha}$ is the normalized asymmetric learning rates difference $\frac{\alpha^+ - \alpha^-}{\alpha^+ + \alpha^-}$.

However, we did not find any significant Spearman’s correlation between the factor scores and the context-dependence parameter α_R . The only significant correlation was between the exploration parameter β and factor scores from Addiction ($\rho(456) = -0.10, p = .042$), but was not significant when applying robust regression ($p = .16$). This means that our measure of context-dependent learning in this task, as measured by the contextual learning rate α_R , cannot be linked to the

transdiagnostic dimensions revealed by our factor analysis.

3.2.4 Conclusion

Five hundred participants performed a new version of a reinforcement learning task. Behavioral results allowed us to confirm the goodness-of-fit of our range-adapting model, both in terms of quantitative comparison and model simulations. From a series of self-assessed questionnaires, a factor analysis highlighted 3 factor dimensions, which we labeled "Social Anxiety", "Compulsivity", and "Addiction". We found coherent correlations between factors and questionnaire scores. However, the parameters of our model did not correlate with any of the symptom dimensions. One reason might come from the different questionnaires that we used. For example, some questionnaires, such as Delusions or Sensation Seeking, have a low (i.e., <0.7) Cronbach's alpha in our data, which suggests that the analysis could benefit from other questionnaires to assess these scales. In the same line, Anxiety and Depression are strongly linked, and in our case were assessed in the same questionnaire, which might explain why their score have high correlation with all of the factors. Finally, we might further investigate the task design, which might not have been adapted to these kind of analyses. In a paper published in 2019 in *Plos Computational Biology*, Shahar and colleagues argue that combining choice and reaction time measures improves model estimates (Shahar et al. 2019). Based on the assumption that value discriminability will be reflected in both choice and reaction time, Shahar and colleagues show that results were accounting in a more stable way by a model combining reinforcement learning and drift-diffusion algorithms. Therefore, we plan to develop better modeling, including the analysis of reaction times (Ballard and McClure 2019, Fontanesi et al. 2019). As another perspective, this project being part of a longitudinal study, we also plan to investigate test-retest reliability in both symptom dimensions and computational parameters. To conclude, further work is needed to answer the questions of a potential link between a dysfunction of range-adapting reinforcement learning and psychiatric symptoms.

Chapter 4

Discussion and perspectives

Reinforcement learning is a fundamental cognitive process arising daily from our birth to our death. Our experience gives us the ability of learning to improve our future choices in order to maximize the occurrence of pleasant events (rewards) and to minimize the occurrence of unpleasant events (punishments). The instance of reinforcement learning is observed at several levels of behavior, whether we learn how to use a spoon (motor level), how to reduce the time spent traveling between home and place of work (cognitive level), what is the best method to revise for an exam (educational level), or how to improve a treatment depending on therapeutic results (professional level). As such, an impairment of this process is one of the principal suspects in neurological disorders with behavioral symptoms, such as Parkinson's disease and Tourette's syndrome, as well as psychiatric disorders, such as schizophrenia and addiction. In this framework, a fundamental unanswered question in decision-making and reinforcement learning remains how values are encoded during the learning and decision process. In other words, do we learn values objectively or subjectively? In the past decades, wealth of evidence has shown that human economic behavior often deviates from objective valuation in many circumstances, where the background context, the temporal context, and personal experience play an important role. These description-based sub-optimal behaviors are in contradiction with normative economic decision theory which describes the expected utility of an option as a cardinal function of the outcome value, not affected by the presence and values of other options, offered simultaneously or in the recent past. Starting from prospect theory, whose core assumption is that option values are encoded relative to a reference point, the notion of context-dependence (or relative valuation) has been spread out in behavioral decision-making research involving decisions based on fully described options and prospects. However, research in situations where the values have to be learned by trial-and-error, has comparably neglected the notion of context-dependence.

In an attempt to fill this gap, throughout the work of this PhD, I have used large-scale studies and computational modeling to investigate context-dependent reinforcement learning in human decision making.

In the first part, we developed a satisfactory model to match and explain healthy participant's behavior. Over 2 studies and 10 experiments, we showed that participants' economic choices depend on the surroundings of the different options. Compared to context-independent learning, where options are encoded on an absolute scale, context-dependent learning, where options are encoded on a relative scale, allows for better performances in small magnitude conditions (magnitude bias) and punishment-related conditions (valence bias). However, our data clearly indicates that context-dependent learning leads to irrational choices when the options are extrapolated from their original learning context (transfer phase). Moreover, we confirmed the counter-intuitive prediction that making the task easier led to larger range adaptation: performance was better in the learning phase but even worse in the transfer phase. In addition, range-adapting coding turned out to be economically disadvantageous, supporting the idea that it is the consequence of an automatic, uncontrolled, process. To conclude, the findings in this PhD are in line with behavioral decision-making research involving description-based choice and provide a new insight into the fundamental role played by context in learning from experience. Our results support the idea that values are learned relatively to the value of the alternative options, at the cost of economically sub-optimal decisions. Behavioral and modeling data are consistent with a range adapting form of context-dependence, however one can argue that there might be several possible explanations for these results. Therefore, there is abundant room for further progress in determining the cognitive process involved in context-dependent learning, as well as the underlying neural bases.

In the second part, we turned to impaired reinforcement learning, with a meta-analysis on clinical papers involving one of the two pioneer tasks in the study of valence-specific reinforcement learning. We found that some conditions, such as (medicated) Parkinson's disease and substance-related disorder, are significantly associated with a reward bias, suggesting that patients learn better from rewards than from punishment. Results were inconclusive for several identified conditions, due to a high inter-study variability. This is coherent with a general insight in computational psychiatry, that having discrete diagnostic categories leads to high inter-individual variance within each category. To account for this general issue, a transnosographic approach allows to inform both categorical and dimensional approaches. To investigate the reward bias in the general population using our validated models and the transnosographic approach to

psychiatric dimensions, we designed a large-scale study, including self-assessed questionnaires, based on the two pioneer tasks from the meta-analysis. We found coherent correlations between psychiatric dimensions and questionnaire scores, however the parameters of our model did not correlate with any of the dimensions. We anticipate that better modeling, such as the addition of reaction time analysis, will improve our modeling tools. For now, one might argue about the discrepancy between the reward bias in addiction highlighted in the meta analysis, and the lack of evidence for such a bias in the large-scale experiment. The latter results should be taken into account when considering that the dimension that we labeled "addiction" in the large-scale experiment comes from traits from the general population (assessed by questionnaires measuring alcohol, cannabis or nicotine consumption), contrary to the meta-analysis groups where individuals were either diagnosed with substance-related disorder or regularly consuming drugs such as heroin or cocaine. Hence, further data-driven work is needed to shed some light on the valence bias in impaired reinforcement learning.

4.1 Contrasting adaptive coding and divisive normalization in human reinforcement learning

Context-dependent learning in healthy individuals was well captured by a range-adapting model, which tracks the range of the available options and normalizes the outcomes accordingly. The model originally comes from prospect theory (context-dependence as a reference-point, [Kahneman and Tversky 1979](#)) and recent findings in monkey electrophysiology (context-dependence as a range, [Padoa-Schioppa 2009](#)). While this model is able to capture participants' choices in all of our tasks, further work is needed to investigate other types of context-dependence, such as divisive normalization ([Louie and Glimcher 2012](#), [Louie et al. 2013](#), [Webb et al. 2020b](#)). In fact, our task design always included binary choices between probabilistic options with no volatility. While probabilistic selection tasks are widely used and adapted to our models, one can argue that the different types of context-dependent algorithms might not be differentiated. To fill this gap, we designed a new reinforcement learning task manipulating the range magnitude and the number of options per choice. Option values are drawn from a Gaussian distribution with a mean between 0 and 100 and a fixed variance (Figure 25). The options are arranged in pairs or triplets, so that a model encoding context-dependent learning with range adaptation will have different predictions than a model encoding context-dependent learning with divisive normalization ([Louie and Glimcher 2012](#)). To avoid a sampling bias, feedback is provided for all of the options after choice at each trial.

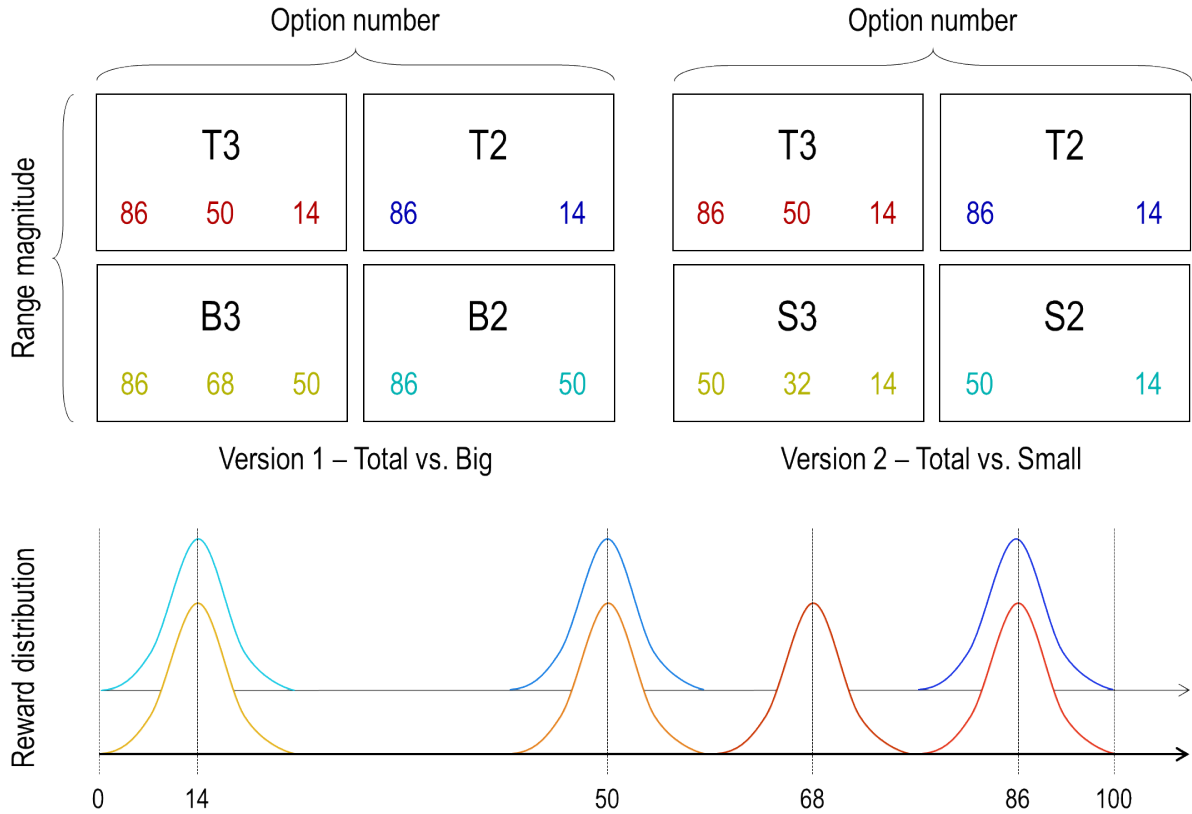


Figure 25. Task design. Top: the task will consist of 2 versions to compare range adaptation and divisive normalization, which both have a 2x2 design manipulating the number of options and the magnitude of the range of the option values. Bottom: example of the distributions from which the outcome will be drawn for each option (version 1 only).

We piloted we versions of the task on $N=2 \times 20$ participants. For the pilot experiments, the options' means were set at 86, 68, 50, 32, 14; the variance was set at 0, and the reward was deterministic (100% chance of getting a reward, Figure 25, top). Preliminary results show that, in the learning phase, participants have similar performance in contexts with 2 or 3 options, contrary to predictions from the divisive normalization model, which predicts that option values will decrease when the number of options increases (since option values are divided by the sum of all the available options). In the transfer phase, it is unclear whether participants' behavior is closest to range adaptation or divisive normalization predictions, even if the choice rates of the best option in each context seem to be equal, which matches a range-adapting model. Interestingly, we found the valuation of the 2 non-favorable options (from contexts with 3 options) to be almost equal in both versions, which does not match any of the models' predictions; therefore, we might consider an additional model to investigate these results. To conclude, preliminary results seem to advantage a range-adapting form of context-dependence, over divisive normalization.

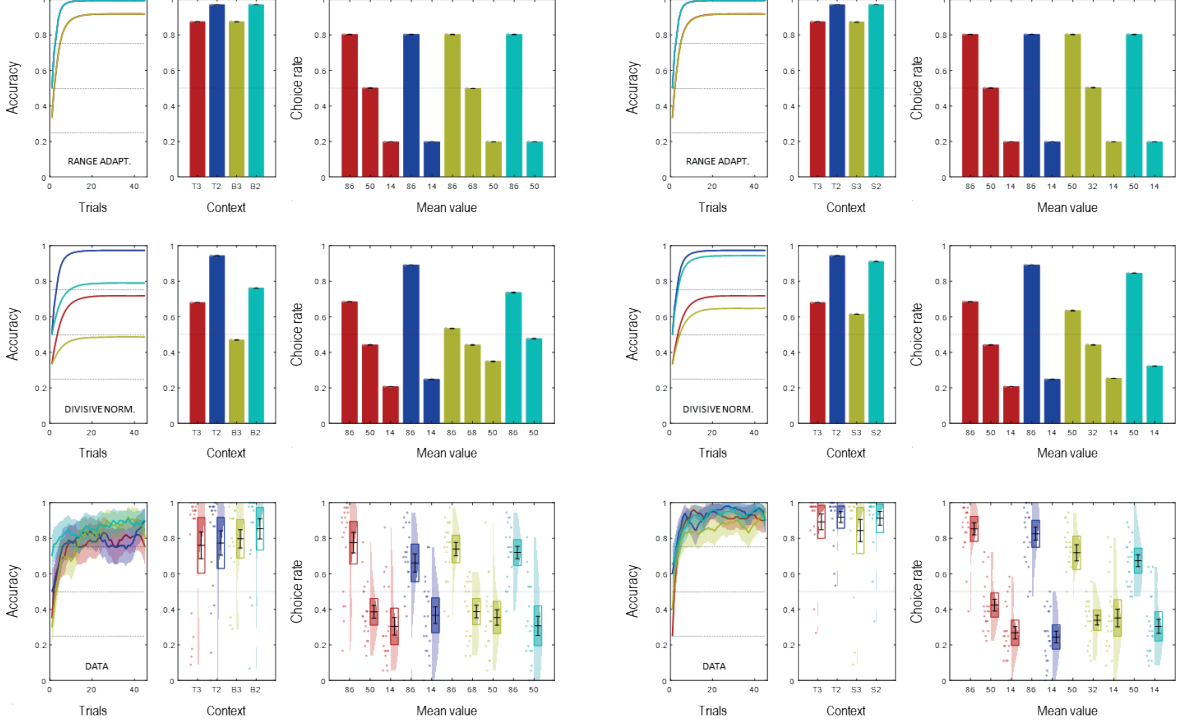


Figure 26. Model predictions and results of the pilot experiment. Model predictions for the range-adaptation model (top) and divisive normalization (middle) for version 1 (left) and version 2 (right) of the experiment. Preliminary results of the pilot experiment (bottom). Model predictions were simulated with fixed parameters.

However, these preliminary results are to be handled with precaution. First, the pilot tasks did not include variance nor probabilities of reward, contrary to model simulations. Second, we used simplified versions of reinforcement learning models (described in sections 2.2 and 1.1.4), adapted to complete feedback information :

$$R_{\text{divisive}} = \frac{R_i}{\sum_{j=1}^3 R_j} \quad R_{\text{range}} = \frac{R_i - R_{\min}}{R_{\max} - R_{\min}}$$

where R_i is the obtained reward for the chosen option i , R_{\max} and R_{\min} are the maximum and minimum outcomes respectively, within each trial. Our RANGE model is not meant to be a model of neural activity, but rather an algorithmic description of how outcome values are normalized. To cope for this limitation, we aim at implementing the complete version of the divisive normalization function presented in section 1.1.4, especially with a parameter assessing the degree of the norm (Webb et al. 2020b). This would not only be a model of how learning processes are implemented at the neural network level, but has been proven to explain context-dependent decision making in several tasks (Louie et al. 2013, 2015). By contrast, in a paper published in 2020 in *Nature Human Behavior*, Gluth and colleagues argue that violations

of IIA reported in the description-based ternary choice task in Louie et al. 2013 (where choice behavior was affected by the addition of one or several distractors of various values) differed from those described in section 1.1.1 (Luce 1959), because the decision relies on a single attribute (Gluth et al. 2020a). Using the same experimental design, the authors did not replicate the results from Louie et al. 2013 and found that divisive normalization might not be an element to take into account when trying to develop models of decision-making (for the benefit of value-based attention), which is in contradiction with the findings of Webb et al. 2020b. Although these results have been recently further discussed (Webb et al. 2020a, Gluth et al. 2020b), the line of research presented in this thesis is more concerned with experience-based decisions, not description-based decisions. In that regard, Gluth and colleagues argue that violations of IIA in experience-based decisions appear to emerge from specific mechanisms during the processing of feedback rather than during the choice process itself (Gluth et al. 2020a). This is in line with previous studies arguing that different forms of IIA violation in experience-based decisions might be due to an interaction between attention and choice: value drives attention, which affects accumulation of evidence (Gluth et al. 2017, 2018, Spektor et al. 2019, Busemeyer et al. 2019).

4.2 Assessing the role of working-memory in range-adapting learning

In the second study, presented in section 2.2, we present the results of 8 versions of a reinforcement learning task manipulating outcome magnitude. Among other variations, half of the experiments had a trial structure in block (i.e., all trials belonging to the same choice contexts are presented in a row), while the others were interleaved (i.e., in a randomized cross-contexts order). We found that, in block experiments, learning performance was higher and contextual effects were exacerbated. However, since the design was between-subjects, each participant performed only one version of the task. Therefore, we discussed in the supplementary materials that the RANGE model presented in the paper would not be able to capture differences between block and interleaved trial structures in a within-subjects design, because the model does not contain a working-memory element. In 2012, Collins and Frank proposed a reinforcement learning model that accounts for working-memory by adding a forgetting parameter: at each trial, the values of the options from not-on-screen contexts are progressively forgotten. If there is a sequence of successive trials from the same context (i.e., the same pair of options), working-memory demand is lower and there is no forgetting (Collins and Frank 2012). However, such

a model is rejected by our data, because of the transfer phase. When trials are in blocks, the values of previously presented contexts are progressively forgotten as the task continues, and at the end of the learning phase, the values of the options from all contexts but one (the last one), will be back to their initial value, which is not what we observe in our behavioral data. Thus, we propose a possible computational interpretation to account for the effects of this manipulation. The key idea of this model is that the notion of *context* can be broken down into two components: the "local" context (which is what we refer to as learning context) and the "global" context (which integrates over a time scale larger than one trial). The model accounts for increased contextual effects in block design, because the local and global context values remain coherent for longer time periods (for simulations, see [Bavard et al. 2021](#), Supp. Figure 4), thus allowing the summation of their effects. We plan to experimentally test this model on a dataset of two additional variants of the experiment. In these variants, the trials in the learning phase were in blocks and the transfer phase interleaved, or the other way around, so that the design was within-subjects. Learning curves are shown and compared to full-block and full-interleaved designs in Figure 27. We anticipate that the improved model, now sensitive to trial structure, will outperform the current RANGE model which does not account for this effect.

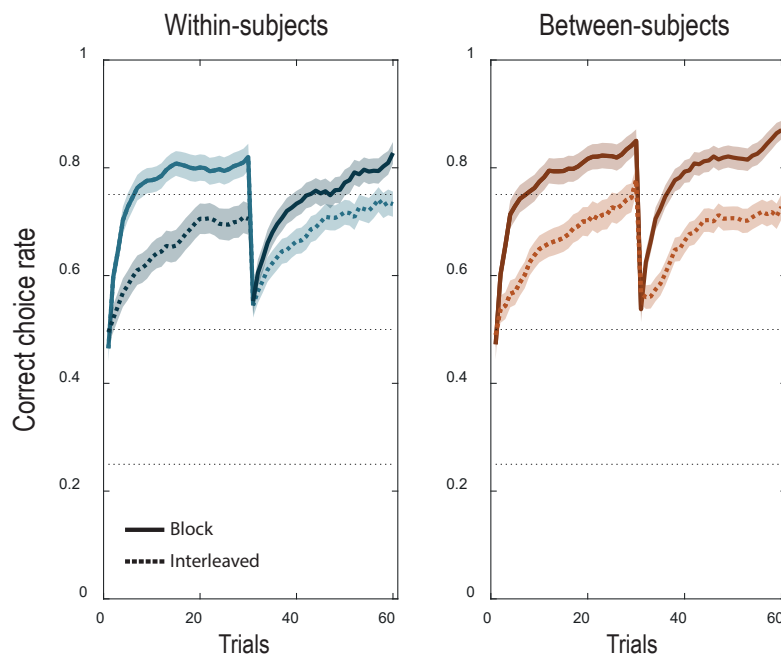


Figure 27. Trial-by-trial performance in within-subjects and between-subjects designs. Learning curves for learning and transfer phase in the two additional experiments (left) and two previous experiments (right). In the within-subjects design, trials were either in blocks in the learning phase and interleaved in the transfer phase (light blue), or the other way around (dark blue). In the between-subjects design, trials were either interleaved (light orange) or in blocks (dark orange) throughout the whole task, corresponding to experiments E4 and E8 from [Bavard et al. 2021](#).

4.3 Cross-cultural study of the impact of contextual information in decision-making

Over all of our experiments, we have replicated results of context-dependent reinforcement learning in human decision making. However, it is important to point out that we did not include demographic measures of our samples of participants. In 2008, J.J. Arnett pointed out that psychological research published in American Psychological Association (APA) journals focuses too narrowly on Americans, when American citizens represent less than 5% of the world's population. The result is an understanding of psychology that is incomplete and does not adequately represent humanity. First, an analysis of articles published in six premier APA journals showed that the contributors, samples, and editorial leadership of the journals are predominantly American. Then, a demographic profile of the human population showed that the majority of the world's population actually lives in conditions vastly different from the conditions of Americans, underlining a lack of basis for assuming psychological processes to be universal and generalizing research findings to the rest of the global population (Arnett 2008). In 2010, Henrich and colleagues reported a systemic bias in conducting psychology studies with participants from "WEIRD" (Western, Educated, Industrialized, Rich and Democratic) societies. Although only 1/8 people worldwide live in regions that fall into the WEIRD classification, the researchers claimed that 60–90% of psychology studies are performed on participants from these areas (Henrich et al. 2010). The article gave examples of results that differ significantly between people from WEIRD and tribal cultures, including the Müller-Lyer illusion (Figure 28). In 2018, Rad and colleagues showed that nearly a decade after Henrich and colleagues's paper, over 80% of the samples used in studies published in the journal *Psychological Science* were from the WEIRD population (Rad et al. 2018).

At our level, a crucial point that remains to be explored regarding our results is whether our optimized model can challenge the "universality" of traditional reinforcement learning models across different cultures. If context-dependence were to improve model fit across cultures, this would portray it as an innate base feature of decision-making. To this aim, in collaboration with Hernan Anllo, doctor of psychology, we are currently testing our task from Bavard et al. 2021 across different cultural samples. Dr Anllo has put together a team including Stefano Palminteri and collaborators from 11 countries (Argentina, France, China, India, Iran, Israel, Japan, Morocco, Russia, UK, US) and has launched preliminary work on this project. Until now, six countries have gathered the experimental data. The task included questionnaires aiming at assessing cross-cultural differences, such as:

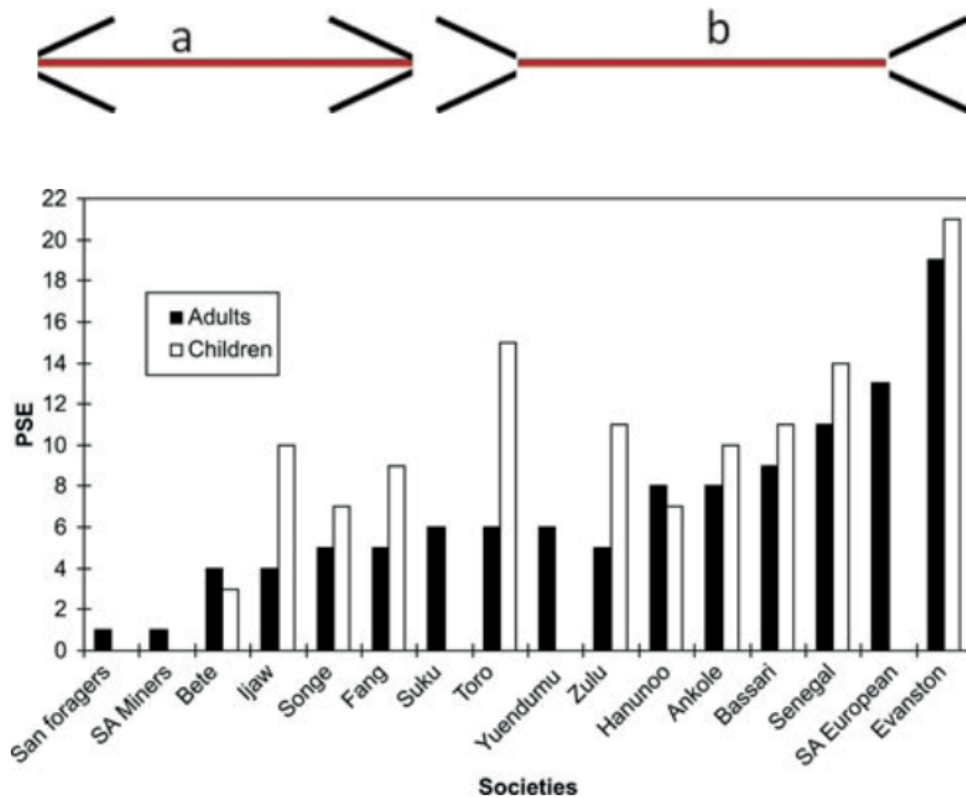


Figure 28. The influence of culture on visual perception. Much like the Ebbinghaus illusion, the Müller-Lyer illusion is an optical illusion consisting of stylized arrows, which tricks viewer into wrongly answering which red line is the shortest, when both red lines are of the same size. In 1963, Segall and colleagues compared susceptibility to four different visual illusions in population samples of several countries. For the Müller-Lyer illusion, the mean fractional misperception of the length of the line segments varied from 1% to 20% across cultures. Figure reproduced from [Segall et al. 1966](#).

- Individualism and Collectivism Scale ([Triandis and Gelfand 1998](#)), a 16-item scale designed to measure four dimensions of collectivism and individualism (vertical collectivism – seeing the self as a part of a collective and being willing to accept hierarchy and inequality within that collective, vertical individualism – seeing the self as fully autonomous, but recognizing that inequality will exist among individuals and accepting this inequality, horizontal collectivism – seeing the self as part of a collective but perceiving all the members of that collective as equal, horizontal individualism – seeing the self as fully autonomous, and believing that equality between individuals is the ideal).
- Centrality of Religiosity ([Huber and Huber 2012](#)), a 15-item scale designed to measure centrality, importance or salience of religious meanings in personality (intellectual dimension – themes of interest, hermeneutical skills, styles of thought and interpretation, and as

bodies of knowledge, ideology dimension – beliefs, unquestioned convictions and patterns of plausibility, public practice dimension – public participation in religious rituals and in communal activities, private practice dimension – patterns of action and a personal style of devotion to the transcendence, experience dimension – patterns of religious perceptions and as a body of religious experiences and feelings).

- Socioeconomic Status (Griskevicius et al. 2013), a 13-item questionnaire that measures perceived socioeconomic status in three dimensions: infancy, adulthood and overall self-rating.

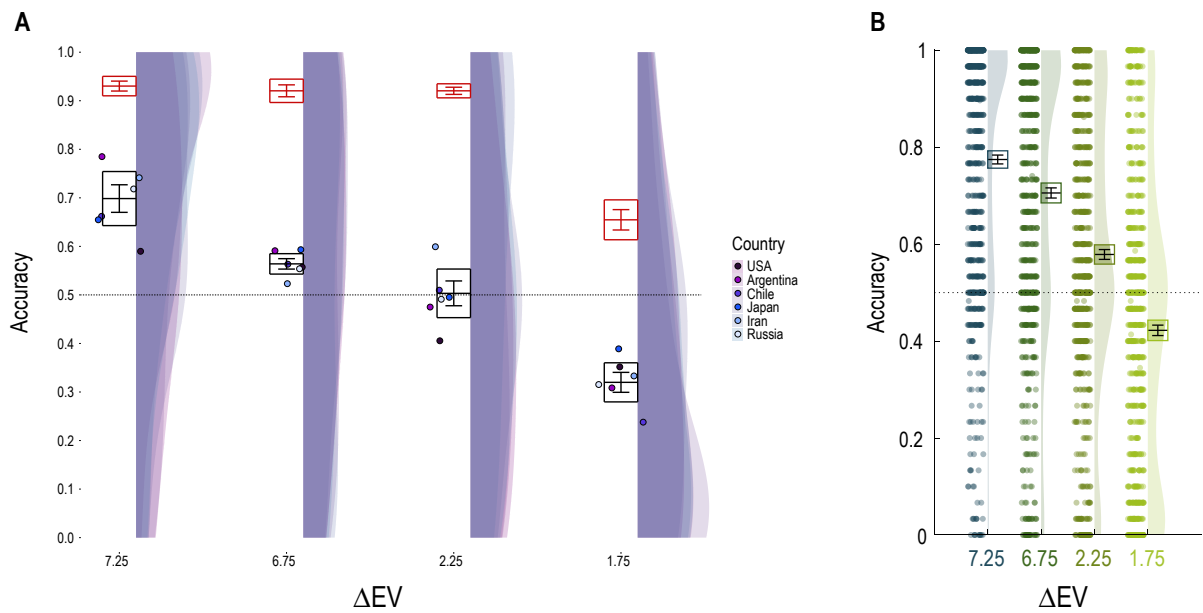


Figure 29. Preliminary results for the cross-cultural investigation of context-dependent reinforcement learning. (A) Each circle represents the average correct response rate for experience-based choices (mean and standard error in black). Mean and standard error of description-based choices are shown in red. (B) Results replicate the findings from Bavard et al. 2021.

Another change to the original task was the addition of a third phase with description-based choices. Participants chose between two options from which they could see both the potential outcome and the probability of reward. This additional phase allows to assess the difference between description-based and experience-based choices. Preliminary results from six countries show that our results from Bavard et al. 2021 replicated in all of them, specifically the sub-optimal choice in the transfer phase. Moreover, this effect disappeared in the equivalent description-based choices where the preference is reversed (Figure 29), showing that participants do not have an objective representation of probabilities and magnitudes when choosing from experience. It also rules out the possibility that the sub-optimal choice was induced by a preference towards the most frequently rewarding option, and not an effect of context-dependent learning.

To conclude, these preliminary results suggest not only that context-dependence in this task arises through learning and not in the decision-making process, but also that context-dependent reinforcement learning is robust across (some) cultures. Further inclusion of participating countries, as well as the analysis of questionnaires assessing cross-cultural measuring, will enlighten us on the robustness of this mechanism over different cultures around the world.

4.4 "There are known knowns..."

I would like to conclude this work by offering a broader view on the research that I have been conducting for the past few years. In 2002, Donald Rumsfeld, US Secretary of State for Defence, stated at a Defence Department briefing:

As we know, there are known knowns; there are things we know that we know. We also know there are known unknowns; that is to say, we know there are some things we do not know. But there are also unknown unknowns - the ones we do not know we don't know.

One might argue that Socrates was using this framework indirectly when he said, according to Plato, "*I only know that I know nothing*". Now, I do not intend to make political or philosophical debates the point of interest here, but this framing of a riddle-like description of knowns and unknowns is very interesting to me, since much scientific research is based on investigating known unknowns. Theories are often built based on previous knowledge, experimental scientists develop a hypothesis to be tested, and design experiments to the aim of testing the null hypothesis. At the outset the researcher does not know whether or not the results will support the null hypothesis. However, it is common for the researcher to believe that the result that will be obtained will be within a range of known possibilities. This leads to incremental improvements, but it cannot take you to a big leap in discovery (occasionally, however, the result is completely unexpected).

From the findings presented here and in line with previous literature, it is reasonable to think that context-dependence plays a crucial role in learning in our daily life. However, generalizing our results to more complex environments is a challenge. What can we infer from experiments conducted in a laboratory, where one can exclude all potential confounding factors to manipulate one and only one variable at a time? In real life, the environment is more complex and multidimensional and full of unknown unknowns, but as we inherently cannot know of their existence, studying the known unknowns is our best approximation, and it is the goal of fundamental science to work forward in that direction.

Once again, the discovery of a previously unknown unknown shows us how little we know and leads to the propagation of a family of known unknowns which can then be tackled by traditional hypothesis forming and testing, occasionally throwing-up another unknown unknown, and so the cycle continues. In the end, the prospect of unknown unknowns is what makes the journey so exciting.

References

- Adams, R. A., Huys, Q. J. M., and Roiser, J. P. (2015). Computational Psychiatry: towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery & Psychiatry*, 87(1):53–63. Publisher: BMJ Publishing Group Ltd Section: Neuropsychiatry. [37](#)
- Allenby, G. M., Rossi, P. E., and McCulloch, R. E. (2005). Hierarchical Bayes Models: A Practitioners Guide. SSRN Scholarly Paper ID 655541, Social Science Research Network, Rochester, NY. [31](#)
- Alpaydm, E. (2004). Introduction to Machine Learning. In *Machine Learning*, volume 56. Journal Abbreviation: Machine Learning. [22](#)
- American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, fifth edition edition. [37](#), [97](#)
- Arnett, J. J. (2008). The neglected 95%: why American psychology needs to become less American. *The American Psychologist*, 63(7):602–614. [115](#)
- Ballard, I. C. and McClure, S. M. (2019). Joint modeling of reaction times and choice improves parameter identifiability in reinforcement learning models. *Journal of Neuroscience Methods*, 317:37–44. [107](#)
- Bateson, M., Healy, S., and Hurly, T. (2002). Irrational choices in hummingbird foraging behaviour. *Animal Behaviour*, 63:587–596. [6](#)
- Bavard, S., Lebreton, M., Khamassi, M., Coricelli, G., and Palminteri, S. (2018). Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences. *Nature Communications*, 9(1):4503. Number: 1 Publisher: Nature Publishing Group. [98](#), [102](#)
- Bavard, S., Rustichini, A., and Palminteri, S. (2020). The construction and deconstruction of sub-optimal preferences through range-adapting reinforcement learning. *bioRxiv*, page 2020.07.28.224642. Publisher: Cold Spring Harbor Laboratory Section: New Results. [96](#)
- Bavard, S., Rustichini, A., and Palminteri, S. (2021). Two sides of the same coin: Beneficial and detrimental consequences of range adaptation in human reinforcement learning. *Science Advances*, 7(14):eabe0340. Publisher: American Association for the Advancement of Science Section: Research Article. [98](#), [102](#), [114](#), [115](#), [117](#)

- Bayes, M. and Price, M. (1763). An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions (1683-1775)*, 53:370–418. Publisher: The Royal Society. [33](#), [135](#)
- Behrens, T. E. J., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9):1214–1221. Number: 9 Publisher: Nature Publishing Group. [31](#)
- Bellman, R. (1958). Dynamic programming and stochastic control processes. *Information and Control*, 1(3):228–239. [22](#)
- Benner, M. J. and Tushman, M. L. (2003). Exploitation, Exploration, and Process Management: The Productivity Dilemma Revisited. *The Academy of Management Review*, 28(2):238–256. Publisher: Academy of Management. [23](#)
- Bernoulli, D. (1738). *Specimen Theoriae Novae de Mensura Sortis*. [5](#), [6](#), [36](#)
- Bernoulli, D. (1954). Exposition of a New Theory on the Measurement of Risk. *Econometrica*, 22(1):23–36. [5](#)
- Berridge, K. C. and Kringelbach, M. L. (2008). Affective neuroscience of pleasure: reward in humans and animals. *Psychopharmacology*, 199(3):457–480. [18](#), [19](#)
- Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific, Belmont, Mass. [23](#)
- Bhat, H. and Kumar, N. (2010). On the Derivation of the Bayesian Information Criterion. [34](#)
- Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer-Verlag, New York. [31](#), [34](#)
- Bouton, M. E. (2007). *Learning and behavior: A contemporary synthesis*. Learning and behavior: A contemporary synthesis. Sinauer Associates, Sunderland, MA, US. Pages: xiii, 482. [26](#)
- Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, 71(356):791–799. [31](#)
- Bridle, J. S. (1990). Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition. In Soulié, F. F. and Héroult, J., editors, *Neurocomputing*, NATO ASI Series, pages 227–236, Berlin, Heidelberg. Springer. [30](#)
- Brodersen, K. H., Schofield, T. M., Leff, A. P., Ong, C. S., Lomakina, E. I., Buhmann, J. M., and Stephan, K. E. (2011). Generative Embedding for Model-Based Classification of fMRI Data. *PLOS Computational Biology*, 7(6):e1002079. Publisher: Public Library of Science. [37](#)
- Buduma, N. and Locascio, N. (2017). *Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms*. O’Reilly Media, 1er édition edition. [27](#)

- Buonomano, D. V., Bramen, J., and Khodadadifar, M. (2009). Influence of the interstimulus interval on temporal processing and learning: testing the state-dependent network model. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 364(1525):1865–1873. 26
- Busemeyer, J. R., Gluth, S., Rieskamp, J., and Turner, B. M. (2019). Cognitive and Neural Bases of Multi-Attribute, Multi-Alternative, Value-based Decisions. *Trends in Cognitive Sciences*, 23(3):251–263. 113
- Camerer, C., Loewenstein, G., and Rabin, M. (2004). *Advances in Behavioral Economics*. 6
- Carandini, M. and Heeger, D. J. (2011). Normalization as a canonical neural computation. *Nature Reviews. Neuroscience*, 13(1):51–62. 36
- Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, 1(2):245–276. Publisher: Routledge _eprint: https://doi.org/10.1207/s15327906mbr0102_10. 104
- Chen, C., Takahashi, T., Nakagawa, S., Inoue, T., and Kusumi, I. (2015). Reinforcement learning in depression: A review of computational research. *Neuroscience & Biobehavioral Reviews*, 55:247–267. 89
- Chen, M., Lakshminarayanan, V., and Santos, L. (2006). How Basic Are Behavioral Biases? Evidence from Capuchin Monkey Trading Behavior. *Journal of Political Economy*, 114(3):517–537. Publisher: The University of Chicago Press. 6
- Claeskens, G. and Hjort, N. (2007). *Model Selection And Model Averaging*. Journal Abbreviation: Cambridge Publication Title: Cambridge. 34
- Cohen, J. D., McClure, S. M., and Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 362(1481):933–942. 23
- Collins, A. G. E. and Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *The European Journal of Neuroscience*, 35(7):1024–1035. 113
- Conen, K. E. and Padoa-Schioppa, C. (2019). Partial Adaptation to the Value Range in the Macaque Orbitofrontal Cortex. *The Journal of Neuroscience*. 12
- Cross-Disorder Group of the Psychiatric Genomics Consortium (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet (London, England)*, 381(9875):1371–1379. 37
- D’Ardenne, K., McClure, S. M., Nystrom, L. E., and Cohen, J. D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science (New York, N.Y.)*, 319(5867):1264–1267. 20

- Davis, M. (1970). Effects of interstimulus interval length and variability on startle-response habituation in the rat. *Journal of Comparative and Physiological Psychology*, 72(2):177–192. [26](#)
- Daw, N. (2011). Trial-by-trial data analysis using computational models. *Affect, Learning and Decision Making, Attention and Performance XXIII*, 23. [32](#), [34](#)
- Daw, N. D. and Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, 16(2):199–204. [24](#)
- Dayan, P. (2009). Dopamine, reinforcement learning, and addiction. *Pharmacopsychiatry*, 42 Suppl 1:S56–65. [89](#), [97](#)
- Dayan, P. and Abbott, L. F. (2005). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press. [22](#)
- Druckman, J. N. (2001). Using Credible Advice to Overcome Framing Effects. *The Journal of Law, Economics, and Organization*, 17(1):62–82. [3](#)
- Enzi, B., Edel, M.-A., Lissek, S., Peters, S., Hoffmann, R., Nicolas, V., Tegenthoff, M., Juckel, G., and Saft, C. (2012). Altered ventral striatal activation during reward and punishment processing in premanifest Huntington’s disease: A functional magnetic resonance study. *Experimental Neurology*, 235(1):256–264. [142](#)
- Eshel, N., Bukwich, M., Rao, V., Hemmelder, V., Tian, J., and Uchida, N. (2015). Arithmetic and local circuitry underlying dopamine prediction errors. *Nature*, 525(7568):243–246. Number: 7568 Publisher: Nature Publishing Group. [24](#)
- Fechner, G. T. (1860). *Elemente der psychophysik*. Leipzig : Breitkopf und Härtel. [36](#)
- Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., and Salkovskis, P. M. (2002). The Obsessive-Compulsive Inventory: development and validation of a short version. *Psychological Assessment*, 14(4):485–496. [100](#)
- Fontanesi, L., Palminteri, S., and Lebreton, M. (2019). Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: a meta-analytical approach using diffusion decision modeling. *Cognitive, Affective & Behavioral Neuroscience*, 19(3):490–502. [107](#)
- Frank, M. J., Seeberger, L. C., and O’reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science (New York, N. Y.)*, 306(5703):1940–1943. [21](#), [37](#), [38](#), [90](#), [91](#), [93](#), [94](#), [95](#), [98](#), [143](#)
- Gächter, S., Orzen, H., Renner, E., and Starmer, C. (2009). Are experimental economists prone to framing effects? A natural field experiment. *Journal of Economic Behavior & Organization*, 70(3):443–446. Publisher: Elsevier. [3](#)

- Ghemawat, P. and Costa, J. E. I. R. (1993). The organizational tension between static and dynamic efficiency. *Strategic Management Journal*, 14(S2):59–73. [23](#)
- Gillan, C. M. and Daw, N. D. (2016). Taking Psychiatry Research Online. *Neuron*, 91(1):19–23. [38](#), [96](#)
- Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., and Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife*, 5:e11305. Publisher: eLife Sciences Publications, Ltd. [38](#), [97](#)
- Gluth, S., Hotaling, J. M., and Rieskamp, J. (2017). The Attraction Effect Modulates Reward Prediction Errors and Intertemporal Choices. *Journal of Neuroscience*, 37(2):371–382. Publisher: Society for Neuroscience Section: Research Articles. [113](#)
- Gluth, S., Kern, N., Kortmann, M., and Vitali, C. L. (2020a). Value-based attention but not divisive normalization influence decisions with multiple alternatives. *Nature Human Behaviour*. [113](#)
- Gluth, S., Kern, N., and Vitali, C. L. (2020b). Reply to: Divisive normalization does influence decisions with multiple alternatives. *Nature human behaviour*. [113](#)
- Gluth, S., Spektor, M. S., and Rieskamp, J. (2018). Value-based attentional capture affects multi-alternative decision making. *eLife*. [113](#)
- Gold, J. M., Waltz, J. A., Matveeva, T. M., Kasanova, Z., Strauss, G. P., Herbener, E. S., Collins, A. G. E., and Frank, M. J. (2012). Negative symptoms and the failure to represent the expected reward value of actions: behavioral and computational modeling evidence. *Archives of General Psychiatry*, 69(2):129–138. [37](#)
- Grillon, C., Robinson, O. J., O’Connell, K., Davis, A., Alvarez, G., Pine, D. S., and Ernst, M. (2017). Clinical anxiety promotes excessive response inhibition. *Psychological Medicine*, 47(3):484–494. [89](#)
- Griskevicius, V., Ackerman, J. M., Cantú, S. M., Delton, A. W., Robertson, T. E., Simpson, J. A., Thompson, M. E., and Tybur, J. M. (2013). When the Economy Falts, Do People Spend or Save? Responses to Resource Scarcity Depend on Childhood Environments. *Psychological Science*, 24(2):197–205. Publisher: SAGE Publications Inc. [117](#)
- Healy, A. F., Havas, D. A., and Parker, J. T. (2000). Comparing serial position effects in semantic and episodic memory using reconstruction of order tasks. *Journal of Memory and Language*, 42(2):147–167. Place: Netherlands Publisher: Elsevier Science. [26](#)
- Heath, T. B. and Chatterjee, S. (1995). Asymmetric Decoy Effects on Lower-Quality versus Higher-Quality Brands: Meta-analytic and Experimental Evidence. *Journal of Consumer Research*, 22(3):268–84. Publisher: Oxford University Press. [4](#)
- Heatherton, T. F., Kozlowski, L. T., Frecker, R. C., and Fagerström, K. O. (1991). The Fagerström Test for Nicotine Dependence: a revision of the Fagerström Tolerance Questionnaire. *British Journal of Addiction*, 86(9):1119–1127. [99](#)

- Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9(2):181–197. [9](#)
- Henrich, J., Heine, S., and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*. [115](#)
- Henriques, J. B., Glowacki, J. M., and Davidson, R. J. (1994). Reward fails to alter response bias in depression. *Journal of Abnormal Psychology*, 103(3):460–466. [89](#)
- Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, 4(3):267–272. [17](#)
- Hollerman, J. R. and Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, 1(4):304–309. [24](#)
- Howard, R. A. (1960). *Dynamic Programming and Markov Process*. MIT Press, Cambridge. [23](#)
- Huber, J., Payne, J. W., and Puto, C. (1982). Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis. *Journal of Consumer Research*, 9(1):90–98. Publisher: Oxford University Press. [4](#)
- Huber, S. and Huber, O. W. (2012). The Centrality of Religiosity Scale (CRS). [116](#)
- Huys, Q. J. M., Tobler, P. N., Hasler, G., and Flagel, S. B. (2014). The role of learning-related dopamine signals in addiction vulnerability. *Progress in Brain Research*, 211:31–77. [96](#), [97](#)
- Ikemoto, S. (2010). Brain reward circuitry beyond the mesolimbic dopamine system: a neurobiological theory. *Neuroscience and Biobehavioral Reviews*, 35(2):129–150. [20](#)
- Jackson, K. and Hackenberg, T. D. (1996). Token reinforcement, choice, and self-control in pigeons. *Journal of the Experimental Analysis of Behavior*, 66(1):29–49. Place: US Publisher: Journal of the Experimental Analysis of Behavior. [17](#)
- Kable, J. W. and Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, 10(12):1625–1633. [8](#)
- Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: a survey. *Journal of Artificial Intelligence Research*, 4(1):237–285. [23](#)
- Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–291. Publisher: [Wiley, Econometric Society]. [7](#), [36](#), [96](#), [110](#)
- Kahneman, D. and Tversky, A. (2000). *Choices, values, and frames*. Choices, values, and frames. Cambridge University Press, New York, NY, US. Pages: xx, 840. [6](#)
- Kamin, L. (1967). Predictability, surprise, attention, and conditioning. [26](#)

- Kandel, E. R., Schwartz, J. H., and Jessell, T. M. (2000). *Principles of neural science*. McGraw-Hill, Health Professions Division, New York. OCLC: 42073108. [19](#)
- Keiflin, R. and Janak, P. H. (2015). Dopamine prediction errors in reward learning and addiction: from theory to neural circuitry. *Neuron*, 88(2):247–263. [89](#), [97](#)
- Kim, H., Shimojo, S., and O’Doherty, J. P. (2006). Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. *PLoS biology*, 4(8):e233. [36](#), [143](#)
- Klein, T. A., Ullsperger, M., and Jocham, G. (2017). Learning relative values in the striatum induces violations of normative decision making. *Nature Communications*, 8(1):16033. Number: 1 Publisher: Nature Publishing Group. [36](#)
- Knutson, B., Westdorp, A., Kaiser, E., and Hommer, D. (2000). fMRI visualization of brain activity during a monetary incentive delay task. *NeuroImage*, 12(1):20–27. [142](#)
- Kobayashi, S., Pinto de Carvalho, O., and Schultz, W. (2010). Adaptation of Reward Sensitivity in Orbitofrontal Neurons. *The Journal of Neuroscience*, 30(2):534–544. [12](#)
- Kőszegi, B. and Rabin, M. (2006). A Model of Reference-Dependent Preferences*. *The Quarterly Journal of Economics*, 121(4):1133–1165. [36](#)
- Kumar, P., Waiter, G., Ahearn, T., Milders, M., Reid, I., and Steele, J. D. (2008). Abnormal temporal difference reward-learning signals in major depression. *Brain: A Journal of Neurology*, 131(Pt 8):2084–2093. [37](#)
- Laplace, P. S. (1820). *Théorie analytique des probabilités*. Publisher: Courcier(Paris). [34](#), [135](#)
- Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., and Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, 1(4):1–9. Number: 4 Publisher: Nature Publishing Group. [97](#), [103](#)
- Legleye, S., Karila, L., Beck, F., and Reynaud, M. (2007). Validation of the CAST, a general population Cannabis Abuse Screening Test. *Journal of Substance Use*, 12(4):233–242. [99](#)
- Liebowitz, M. R. (1987). Social phobia. *Modern Problems of Pharmacopsychiatry*, 22:141–173. [100](#)
- Louie, K. and Glimcher, P. W. (2012). Efficient coding and the neural representation of value. *Annals of the New York Academy of Sciences*, 1251:13–32. [8](#), [110](#)
- Louie, K., Glimcher, P. W., and Webb, R. (2015). Adaptive neural coding: from biological to behavioral decision-making. *Current Opinion in Behavioral Sciences*, 5:91–99. [112](#)
- Louie, K., Grattan, L. E., and Glimcher, P. W. (2011). Reward Value-Based Gain Control: Divisive Normalization in Parietal Cortex. *Journal of Neuroscience*, 31(29):10627–10639. Publisher: Society for Neuroscience Section: Articles. [9](#), [10](#)

- Louie, K., Khaw, M. W., and Glimcher, P. W. (2013). Normalization is a general neural mechanism for context-dependent decision making. *Proceedings of the National Academy of Sciences*, 110(15):6139–6144. Publisher: National Academy of Sciences Section: Biological Sciences. [11](#), [110](#), [112](#), [113](#)
- Luce, R. D. (1959). *Individual choice behavior*. Individual choice behavior. John Wiley, Oxford, England. Pages: xii, 153. [2](#), [30](#), [113](#)
- Maia, T. V. and Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, 14(2):154–162. [38](#)
- Marsh, B. and Kacelnik, A. (2002). Framing effects and risky decisions in starlings. *Proceedings of the National Academy of Sciences*, 99(5):3352–3355. Publisher: National Academy of Sciences Section: Social Sciences. [6](#)
- McClure, S. M. and D’Ardenne, K. (2009). Computational neuroimaging. Monitoring reward learning with blood flow. *Handbook of Reward and Decision Making*, pages 229–247. Publisher: Elsevier Inc. [24](#)
- Meads, D. M. and Bentall, R. P. (2008). Rasch analysis and item reduction of the hypomanic personality scale. *Personality and Individual Differences*, 44(8):1772–1783. [100](#)
- Miller, R. R., Barnet, R. C., and Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, 117(3):363–386. [26](#)
- Mkrtchian, A., Aylward, J., Dayan, P., Roiser, J. P., and Robinson, O. J. (2017). Modeling Avoidance in Mood and Anxiety Disorders Using Reinforcement Learning. *Biological Psychiatry*, 82(7):532–539. [89](#)
- Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of Machine Learning, second edition*. The MIT Press, Cambridge, Massachusetts, second edition edition. [23](#)
- Montague, P. R., Dolan, R. J., Friston, K. J., and Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, 16(1):72–80. [37](#)
- Montmort, P. R. d. (1713). *Essai d’analyse sur les jeux de hasard*. Quillan. Google-Books-ID: 9K1ntwAACAAJ. [5](#)
- Moutoussis, M., Bentall, R. P., El-Deredy, W., and Dayan, P. (2011). Bayesian modelling of Jumping-to-Conclusions bias in delusional patients. *Cognitive Neuropsychiatry*, 16(5):422–447. [37](#)
- Nestler, E. J., Hyman, S. E., and Malenka, R. C. (2009). *Molecular neuropharmacology: a foundation for clinical neuroscience*. McGraw-Hill Medical, New York. OCLC: 273018757. [20](#)
- Niv, Y. and Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences*, 12(7):265–272. [26](#), [27](#)
- Nutt, D. J., Lingford-Hughes, A., Erritzoe, D., and Stokes, P. R. A. (2015). The dopamine theory of addiction: 40 years of highs and lows. *Nature Reviews. Neuroscience*, 16(5):305–312. [89](#), [97](#)

- O’Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science (New York, N.Y.)*, 304(5669):452–454. [20](#), [28](#)
- O’Doherty, J. P. (2014). The problem with value. *Neuroscience and Biobehavioral Reviews*, 43:259–268. [18](#)
- O’Doherty, J. P., Dayan, P., Friston, K., Critchley, H., and Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, 38(2):329–337. [24](#), [28](#)
- Padoa-Schioppa, C. (2009). Range-Adapting Representation of Economic Value in the Orbitofrontal Cortex. *Journal of Neuroscience*, 29(44):14004–14014. Publisher: Society for Neuroscience Section: Articles. [11](#), [13](#), [110](#)
- Padoa-Schioppa, C. (2013). Neuronal origins of choice variability in economic decisions. *Neuron*, 80(5):1322–1336. [11](#)
- Padoa-Schioppa, C. and Assad, J. A. (2008). The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. *Nature Neuroscience*, 11(1):95–102. [11](#)
- Palminteri, S., Justo, D., Jauffret, C., Pavlicek, B., Dauta, A., Delmaire, C., Czernecki, V., Karachi, C., Capelle, L., Durr, A., and Pessiglione, M. (2012). Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning. *Neuron*, 76(5):998–1009. [138](#), [139](#), [142](#), [143](#)
- Palminteri, S., Khamassi, M., Joffily, M., and Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, 6(1):8096. Number: 1 Publisher: Nature Publishing Group. [36](#), [38](#), [90](#), [96](#), [98](#), [102](#), [138](#), [143](#)
- Palminteri, S., Kilford, E. J., Coricelli, G., and Blakemore, S.-J. (2016). The Computational Development of Reinforcement Learning during Adolescence. *PLoS Computational Biology*, 12(6):e1004953. Publisher: Public Library of Science. [36](#), [96](#)
- Palminteri, S., Lebreton, M., Worbe, Y., Grabli, D., Hartmann, A., and Pessiglione, M. (2009). Pharmacological modulation of subliminal learning in Parkinson’s and Tourette’s syndromes. *Proceedings of the National Academy of Sciences*, 106(45):19179–19184. Publisher: National Academy of Sciences Section: Biological Sciences. [143](#)
- Palminteri, S., Lefebvre, G., Kilford, E. J., and Blakemore, S.-J. (2017a). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS Computational Biology*, 13(8):e1005684. Publisher: Public Library of Science. [38](#), [97](#), [103](#)
- Palminteri, S. and Pessiglione, M. (2017). Chapter 23 - Opponent Brain Systems for Reward and Punishment Learning: Causal Evidence From Drug and Lesion Studies in Humans. In Dreher, J.-C. and Tremblay, L., editors, *Decision Neuroscience*, pages 291–303. Academic Press, San Diego. [90](#), [91](#), [143](#)

- Palminteri, S., Wyart, V., and Koechlin, E. (2017b). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, 21(6):425–433. [31](#), [34](#), [35](#)
- Pavlov, I. P. (1927). *Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex*. Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. Oxford Univ. Press, Oxford, England. Pages: xv, 430. [13](#), [14](#)
- Pessiglione, M. and Delgado, M. R. (2015). The good, the bad and the brain: Neural correlates of appetitive and aversive values underlying decision making. *Current Opinion in Behavioral Sciences*, 5:78–84. [143](#)
- Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., and Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442(7106):1042–1045. [37](#), [38](#), [91](#), [92](#), [93](#), [95](#), [96](#), [98](#), [143](#)
- Peters, E. R., Joseph, S. A., and Garety, P. A. (1999). Measurement of delusional ideation in the normal population: Introducing the PDI (Peters et al. Delusions Inventory). *Schizophrenia Bulletin*, 25(3):553–576. [100](#)
- Plous, S. (1993). *The psychology of judgment and decision making*. The psychology of judgment and decision making. Mcgraw-Hill Book Company, New York, NY, England. Pages: xvi, 302. [3](#)
- Rad, M. S., Martingano, A. J., and Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences of the United States of America*, 115(45):11401–11405. [115](#)
- Rangel, A. and Clithero, J. A. (2012). Value normalization in decision making: Theory and evidence. *Current Opinion in Neurobiology*, 22(6):970–981. Place: Netherlands Publisher: Elsevier Science. [12](#), [90](#)
- Rangel, A. and Hare, T. (2010). Neural computations associated with goal-directed choice. *Current Opinion in Neurobiology*, 20(2):262–270. [8](#)
- Rescorla, R. A. (1988). Pavlovian conditioning: It’s not what you think it is. *American Psychologist*, 43(3):151–160. Place: US Publisher: American Psychological Association. [22](#)
- Rescorla, R. A. and Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, 2:64–99. [20](#), [25](#), [100](#)
- Rieskamp, J., Busemeyer, J., and Mellers, B. (2006). Extending the Bounds of Rationality: Evidence and Theories of Preferential Choice. *Journal of Economic Literature*, pages 631–661. [2](#)
- Rigoli, F., Chew, B., Dayan, P., and Dolan, R. J. (2016). Multiple value signals in dopaminergic midbrain and their role in avoidance contexts. *Neuroimage*, 135:197–203. [143](#)

- Rigoli, F., Chew, B., Dayan, P., and Dolan, R. J. (2018). Learning Contextual Reward Expectations for Value Adaptation. *Journal of Cognitive Neuroscience*, 30(1):50–69. [36](#)
- Rothkirch, M., Tonn, J., Köhler, S., and Sterzer, P. (2017). Neural mechanisms of reinforcement learning in unmedicated patients with major depressive disorder. *Brain: A Journal of Neurology*, 140(4):1147–1157. [89](#)
- Russell, S. J., Russell, S. J., Norvig, P., and Davis, E. (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall. [23](#)
- Rustichini, A., Conen, K. E., Cai, X., and Padoa-Schioppa, C. (2017). Optimal coding and neuronal adaptation in economic decisions. *Nature Communications*, 8(1):1208. Number: 1 Publisher: Nature Publishing Group. [11](#)
- Saunders, J. B., Aasland, O. G., Babor, T. F., de la Fuente, J. R., and Grant, M. (1993). Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on Early Detection of Persons with Harmful Alcohol Consumption–II. *Addiction (Abingdon, England)*, 88(6):791–804. [99](#)
- Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80(1):1–27. [24](#)
- Schultz, W., Dayan, P., and Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, 275(5306):1593–1599. Publisher: American Association for the Advancement of Science Section: Articles. [20](#), [21](#), [26](#)
- Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464. Publisher: Institute of Mathematical Statistics. [34](#)
- Segall, M. H., Campbell, D. T., and Herskovits, M. J. (1966). *The influence of culture on visual perception*. The influence of culture on visual perception. Bobbs-Merrill, Oxford, England. Pages: xvii, 268. [116](#)
- Sescousse, G., Caldú, X., Segura, B., and Dreher, J.-C. (2013). Processing of primary and secondary rewards: a quantitative meta-analysis and review of human functional neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, 37(4):681–696. [18](#)
- Seymour, B. and McClure, S. M. (2008). Anchors, scales and the relative coding of value in the brain. *Current Opinion in Neurobiology*, 18(2):173–178. [90](#)
- Seymour, B., O’Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., Friston, K. J., and Frackowiak, R. S. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, 429(6992):664–667. Number: 6992 Publisher: Nature Publishing Group. [143](#)
- Shafir, S., Waite, T., and Smith, B. (2002). Context-dependent Violations of Rational Choice in Honeybees (*Apis Mellifera*) and Gray Jays (*Perisoreus Canadensis*). *Behavioral Ecology and Sociobiology*, 51:180–187. [6](#)

- Shahar, N., Hauser, T. U., Moutoussis, M., Moran, R., Keramati, M., Consortium, N., and Dolan, R. J. (2019). Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLoS Computational Biology*, 15(2):e1006803. Publisher: Public Library of Science. [107](#)
- Sharot, T. (2011). The optimism bias. *Current Biology*, 21(23):R941–R945. [96](#), [97](#)
- Shiner, T., Seymour, B., Wunderlich, K., Hill, C., Bhatia, K. P., Dayan, P., and Dolan, R. J. (2012). Dopamine and performance in a reinforcement learning task: evidence from Parkinson’s disease. *Brain: A Journal of Neurology*, 135(Pt 6):1871–1883. [143](#)
- Siegel, S. and Allan, L. G. (1996). The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin & Review*, 3(3):314–321. Place: US Publisher: Psychonomic Society. [26](#)
- Simonson, I. and Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research*, 29(3):281–295. Place: US Publisher: American Marketing Association. [5](#), [6](#)
- Skinner, B. F. (1938). *The behavior of organisms, an experimental analysis*. D. Appleton-Century Company, Incorporated, New York, London. OCLC: 553295. [14](#), [17](#)
- Soltani, A., Chaisangmongkon, W., and Wang, X. J. (2017). Chapter 13 - Neural Circuit Mechanisms of Value-Based Decision-Making and Reinforcement Learning. In Dreher, J.-C. and Tremblay, L., editors, *Decision Neuroscience*, pages 163–176. Academic Press, San Diego. [12](#)
- Spektor, M. S., Gluth, S., Fontanesi, L., and Rieskamp, J. (2019). How similarity between choice options affects decisions from experience: The accentuation-of-differences model. *Psychological Review*, 126(1):52–88. [113](#)
- Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44. [26](#)
- Sutton, R. S. and Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In *Learning and computational neuroscience: Foundations of adaptive networks*, pages 497–537. The MIT Press, Cambridge, MA, US. [26](#)
- Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning - An Introduction*. Mit Press. [22](#), [23](#), [24](#), [26](#), [28](#), [100](#)
- Thaler, R. H. (1992). *The winner’s curse: Paradoxes and anomalies of economic life*. The winner’s curse: Paradoxes and anomalies of economic life. Free Press, New York, NY, US. Pages: ix, 230. [6](#)
- Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. Animal intelligence: Experimental studies. Macmillan Press, Lewiston, NY, US. Pages: viii, 297. [15](#)

- Thorndike, E. L. E. L. (1898). *Animal intelligence : an experimental study of the associative processes in animals*. New York : Macmillan. [13](#)
- Tom, S. M., Fox, C. R., Trepel, C., and Poldrack, R. A. (2007). The Neural Basis of Loss Aversion in Decision-Making Under Risk. *Science*, 315(5811):515–518. Publisher: American Association for the Advancement of Science Section: Report. [31](#)
- Triandis, H. C. and Gelfand, M. J. (1998). Converging measurement of horizontal and vertical individualism and collectivism. *Journal of Personality and Social Psychology*, 74(1):118–128. Place: US Publisher: American Psychological Association. [116](#)
- Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458. Publisher: American Association for the Advancement of Science Section: Articles. [3](#), [6](#)
- Tversky, A. and Kahneman, D. (1986). Rational Choice and the Framing of Decisions. *The Journal of Business*, 59(4):S251–S278. Publisher: University of Chicago Press. [3](#)
- Vaidya, A. R., Pujara, M. S., Petrides, M., Murray, E. A., and Fellows, L. K. (2019). Lesion Studies in Contemporary Neuroscience. *Trends in Cognitive Sciences*, 23(8):653–671. [21](#)
- Vainik, U., Eun Han, J., Epel, E. S., Janet Tomiyama, A., Dagher, A., and Mason, A. E. (2019). Rapid Assessment of Reward-Related Eating: The RED-X5. *Obesity (Silver Spring, Md.)*, 27(2):325–331. [100](#)
- Van Os, J., Gilvarry, C., Bale, R., Van Horn, E., Tattan, T., White, I., and Murray, R. (1999). A comparison of the utility of dimensional and categorical representations of psychosis. UK700 Group. *Psychological Medicine*, 29(3):595–606. [37](#)
- Vlaev, I., Chater, N., Stewart, N., and Brown, G. D. A. (2011). Does the brain calculate value? *Trends in Cognitive Sciences*, 15(11):546–554. [90](#), [96](#)
- Von Neumann, J. and Morgenstern, O. (1947). *Theory of games and economic behavior*, 2nd rev. ed. Theory of games and economic behavior, 2nd rev. ed. Princeton University Press, Princeton, NJ, US. Pages: xviii, 641. [6](#)
- Voon, V., Pessiglione, M., Brezing, C., Gallea, C., Fernandez, H. H., Dolan, R. J., and Hallett, M. (2010). Mechanisms underlying dopamine-mediated reward bias in compulsive behaviors. *Neuron*, 65(1):135–142. [143](#)
- Watkins, C. (1989). Learning From Delayed Rewards. [28](#)
- Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3):279–292. [28](#)
- Webb, R., Glimcher, P., and Louie, K. (2020a). Divisive normalization does influence decisions with multiple alternatives. *Nature human behaviour*. [113](#)

- Webb, R., Glimcher, P. W., and Louie, K. (2020b). The Normalization of Consumer Valuations: Context-Dependent Preferences From Neurobiological Constraints. *Management Science*. Publisher: INFORMS. [9](#), [10](#), [11](#), [110](#), [112](#), [113](#)
- Werbos, P. J. (1992). Neurocontrol and fuzzy logic: Connections and designs. *International Journal of Approximate Reasoning*, 6(2):185–219. [23](#)
- Wiering, M. and Otterlo, M. v., editors (2012). *Reinforcement Learning: State-of-the-Art*. Adaptation, Learning, and Optimization. Springer-Verlag, Berlin Heidelberg. [23](#)
- Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., and Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. *Journal of Experimental Psychology. General*, 143(6):2074–2081. [23](#)
- Wise, R. A. and Koob, G. F. (2014). The Development and Maintenance of Drug Addiction. *Neuropsychopharmacology*, 39(2):254–262. [97](#)
- Zigmond, A. S. and Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica*, 67(6):361–370. [100](#)
- Zimmermann, J., Glimcher, P. W., and Louie, K. (2018). Multiple timescales of normalized value coding underlie adaptive choice behavior. *Nature Communications*, 9(1):3206. Number: 1 Publisher: Nature Publishing Group. [12](#)
- Zuckerman, M., Kolin, E. A., Price, L., and Zoob, I. (1964). Development of a sensation-seeking scale. *Journal of Consulting Psychology*, 28(6):477–482. [100](#)

Appendices

Appendix A

Estimation of the model evidence

One approach to model selection is to pick the candidate model with the highest probability given the data, regardless of the set of parameters. To determine how well a model fits the data, one might consider its *posterior probability*, which is the conditional probability of the model M being "true" after observing some data set D , $P(M | D)$. According to Bayes' theorem, or Bayes' rule ([Bayes and Price 1763](#)):

$$P(M | D) = P(D | M) \cdot \frac{P(M)}{P(D)} \quad (\text{A.1})$$

We know that D is fixed and we wish to consider the impact of D having been observed on our belief in M . Therefore, $P(D)$ is also fixed and we can write:

$$P(M | D) \propto P(D | M) \cdot P(M) \quad (\text{A.2})$$

The key quantity, $P(D | M)$ is called the *model evidence* and represents the probability of model M generating data D . In the literature, model evidence is also referred to as evidence, marginal likelihood, or integrated likelihood. This is because, to evaluate the model evidence, one might integrate it over all the possible sets of parameters of model M :

$$P(D | M) = \int_{\theta} P(D, \theta | M) d\theta \quad (\text{A.3})$$

Then, the quantity is evaluated using Laplace approximation ([Laplace 1820](#)), which combines Taylor expansion and the Gaussian integral. For a multivariable function f , the Taylor expansion to the second order around z_0 is given by:

$$f(z) \approx f(z_0) + \nabla f(z_0)(z - z_0) + \frac{1}{2}(z - z_0)^\top \nabla^2 f(z_0)(z - z_0) \quad (\text{A.4})$$

Note that if z_0 is the maximum, $\nabla f(z_0) = 0$ and $\nabla^2 f(z_0) < 0$. Let us note $H = -\nabla^2 f(z_0)$ the Hessian matrix, which is negative-definite. We estimate the conditional probability $P(D, \theta | M)$ via a multivariable case of Taylor expansion to the second order around the maximum, i.e., the optimal set of parameters, θ^* , of size d :

$$\begin{aligned} P(D | M) &= \int_{\theta} P(D, \theta | M) d\theta \\ &= \int_{\theta} \exp \log P(D, \theta | M) d\theta \\ &\approx \int_{\theta} \exp \left(\log P(D, \theta^* | M) + \nabla \log P(D, \theta^* | M)(\theta - \theta^*) \right. \\ &\quad \left. + \frac{1}{2}(\theta - \theta^*)^\top \nabla^2 \log P(D, \theta^* | M)(\theta - \theta^*) \right) d\theta \\ &\approx \int_{\theta} P(D, \theta^* | M) \cdot \exp \left(\frac{1}{2}(\theta - \theta^*)^\top \nabla^2 \log P(D, \theta^* | M)(\theta - \theta^*) \right) d\theta \\ &\approx P(D, \theta^* | M) \cdot \int_{\theta} \exp \left(\frac{1}{2}(\theta - \theta^*)^\top \cdot -H \cdot (\theta - \theta^*) \right) d\theta \\ &\approx P(D, \theta^* | M) \cdot \int_{\theta} \exp \left(-\frac{1}{2}(\theta - \theta^*)^\top \cdot H \cdot (\theta - \theta^*) \right) d\theta \end{aligned}$$

The multivariate Gaussian integral over \mathbb{R}^d has closed form solution:

$$\int_{z \in \mathbb{R}^d} \exp \left(-\frac{1}{2} z^\top H z \right) dz = \frac{(2\pi)^{\frac{d}{2}}}{|H|^{\frac{1}{2}}} \quad (\text{A.5})$$

where H is a symmetric positive-definite matrix and $|H|$ its determinant. We now have:

$$\begin{aligned} P(D | M) &\approx P(D, \theta^* | M) \cdot \int_{\theta} \exp \left(-\frac{1}{2}(\theta - \theta^*)^\top \cdot H \cdot (\theta - \theta^*) \right) d\theta \\ &\approx P(D, \theta^* | M) \cdot \frac{(2\pi)^{\frac{d}{2}}}{|H|^{\frac{1}{2}}} \\ &\approx P(D | \theta^*, M) \cdot P(\theta^* | M) \cdot \frac{(2\pi)^{\frac{d}{2}}}{|H|^{\frac{1}{2}}} \end{aligned}$$

To facilitate the estimation, we apply the logarithm again:

$$\log P(D | M) \approx \log P(D | \theta^*, M) + \log P(\theta^* | M) + \frac{d}{2} \log 2\pi - \frac{1}{2} \log |H| \quad (\text{A.6})$$

In this case, the difficult component to evaluate might be the Hessian H . To simplify the Laplace approximation even further, let us assume that the number of observation n grows to infinity. Since $\log P(D | \theta^*, M)$ grows with n when n is large, the term will dominate the rest and we can drop the terms which do not depend on n . According to the weak law of large numbers, the matrix H grows as nH_0 for some constant matrix H_0 , so:

$$-\frac{1}{2} \log |H| \approx -\frac{1}{2} \log |nH_0| = -\frac{d}{2} \log n - \frac{1}{2} \log |H_0|$$

By dropping all terms which are independent of n , we have:

$$\log P(D | M) \approx \log P(D | \theta^*, M) - \frac{d}{2} \log n \tag{A.7}$$

↑

Appendix B

Additional results

I will now briefly present some additional clinical results investigating the neural bases of learning from rewards, punishments and counterfactuals. Using the same task contrasting monetary gains and losses, Palminteri and colleagues investigated the role of cortical and subcortical candidate regions (namely the anterior insula and dorsal striatum) in behavioral impairments in patients presenting brain tumor and Huntington’s disease ([Palminteri et al. 2012](#)); both groups exhibited selective impairment of punishment-based learning. During this PhD, I took part in three projects which used the block design version of this task (i.e., all trials belonging to the same choice contexts are presented in a row), in patients with Huntington’s disease, Parkinson’s disease, and brain lesions.

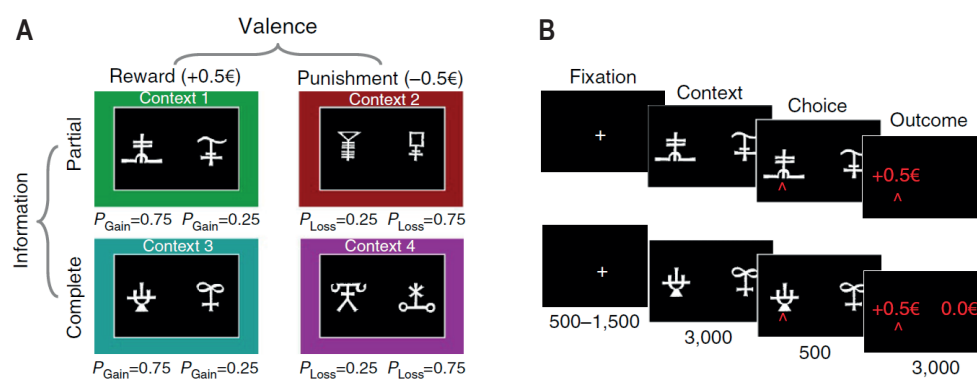


Figure 30. Task design for the clinical experiment. The task was similar to [Palminteri et al. 2015](#) except the trials were in blocks (i.e., all trials belonging to the same choice contexts are presented in a row). (A) Choice contexts manipulating outcome valence and feedback information in a 2x2 orthogonal design. (B) Successive screens for one trial in the partial (top) and complete (bottom) contexts. Figure adapted from [Palminteri et al. 2015](#).

The new task design also allows to compare learning in different informational contexts (partial

and complete feedback, Figure 30). After the learning phase, participants performed a transfer phase where all possible combinations were presented, in order to assess participants' preference for each option. We were aiming at potentially supporting previous results from Palminteri et al. 2012, namely that patients with Huntington's disease or insular lesions should learn less well from punishments than from rewards, as well as at assessing the differences in learning from partial vs. complete feedback contexts.

B.1 Huntington's disease

Twenty-nine Huntington's disease gene carriers far from symptom onset (>10 years), and 30 healthy controls performed the reinforcement learning task, where we manipulated feedback information (partial vs. complete) and outcome valence (reward vs. punishment). The experiment was performed as part of a longitudinal study in collaboration with Alexandra Durr, professor of neurogenetics.

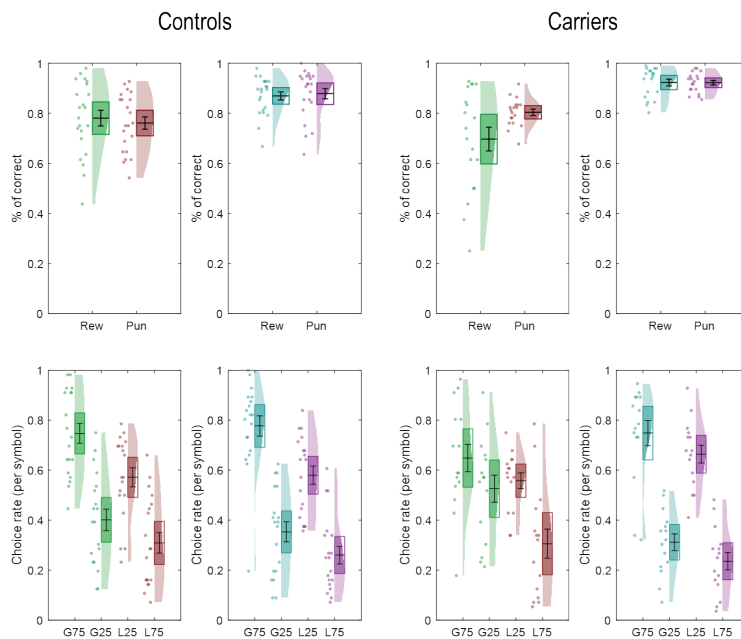


Figure 31. Behavioral results for the study on Huntington's disease. Top: correct choice rate in the learning phase. Bottom: choice rate for the transfer phase. Color coded as in Figure 30.

In the learning phase, we found a main effect of feedback information ($F(1, 37) = 72.10, p < .0001$) meaning that participants had a better performance in complete feedback information contexts, compared to partial feedback. We found an interaction between feedback information and group ($F(1, 37) = 4.78, p = .035$). Post-hoc tests suggest that the performance improvement between complete and partial is higher in carriers than in controls ($t(39) = 2.20, p = .034$) (Figure 31, top).

In the transfer phase, there was no significant main effect, nor interaction, involving the group factor. However, we replicated previous transfer phase results and found a significant effect of *favorableness* (i.e., was the option the most favorable option in its learning context) ($F(1, 34) = 123.88, p < .0001$), suggesting that an option previously more favorable is more likely to be chosen out of context; we found a main effect of valence ($F(1, 34) = 15.49, p = .00039$), suggesting that an option associated with punishment is less likely to be chosen; we found an interaction between favorableness and feedback information ($F(1, 34) = 9.99, p = .0033$), suggesting that the favorableness effect is even greater for options learned in complete feedback information contexts (Figure 31, bottom).

B.2 Parkinson’s disease

Thirty-five patients with Parkinson’s disease performed the reinforcement learning task, where we manipulated feedback information (partial vs. complete) and outcome valence (reward vs. punishment). All patients were medicated, 15 were diagnosed with a dopaminergic dysregulation syndrome (characterized by self-control problems, such as addiction to medication, gambling, or sexual behavior), and 20 were regulated. No control group has performed the task yet. The experiment is performed as part of a longitudinal study in collaboration with Yulia Worbe, neurologist and professor of neurophysiology.

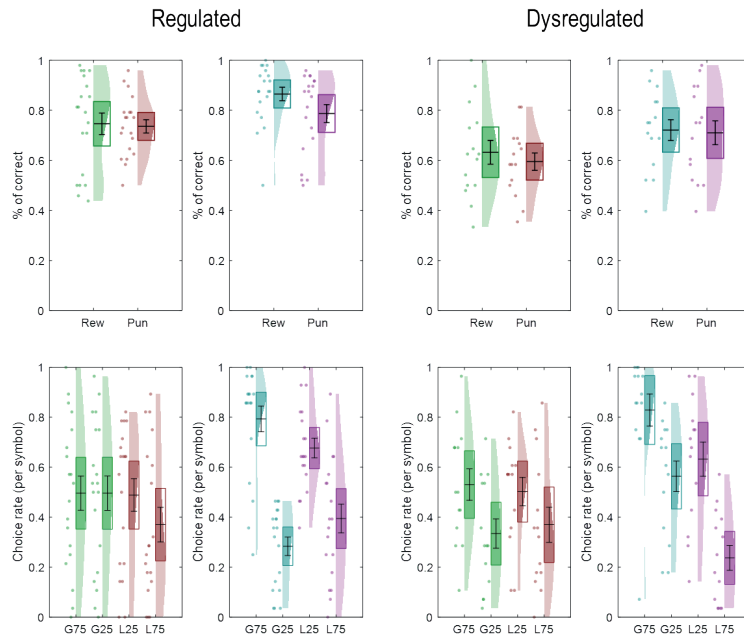


Figure 32. Behavioral results for the study on Parkinson’s disease. Top: correct choice rate in the learning phase. Bottom: choice rate for the transfer phase. Color coded as in Figure 30.

In the learning phase, we found a main effect of feedback information ($F(1, 33) = 25.96$,

$p < .0001$) meaning that patients had a better performance in complete feedback information contexts, compared to partial feedback. We found a main effect of group ($F(1, 33) = 9.05$, $p = .0050$); post-hoc tests suggest that regulated patients performed better than dysregulated patients ($t(33) = 3.01$, $p = .0050$). We found no other significant effect nor interaction (Figure 32, top).

In the transfer phase, there was no significant main effect, nor interaction, involving the group factor. We found a significant effect of favorableness ($F(1, 31) = 57.93$, $p < .0001$), information ($F(1, 31) = 13.04$, $p = .0011$), and a marginal effect of valence ($F(1, 31) = 3.45$, $p = .073$). We found an interaction between favorableness and feedback information ($F(1, 31) = 15.79$, $p = .00039$), suggesting that the favorableness effect is even greater for options learned in complete feedback information contexts (Figure 32, bottom).

B.3 Brain lesions

Sixteen patients with insular lesions, 16 patients with frontal lesions, and 20 healthy controls performed the reinforcement learning task, where we manipulated feedback information (partial vs. complete) and outcome valence (reward vs. punishment). The experiment is performed as part of a study in collaboration with Vasilisa Skvortsova, Anush Ghambaryan, and collaborators.

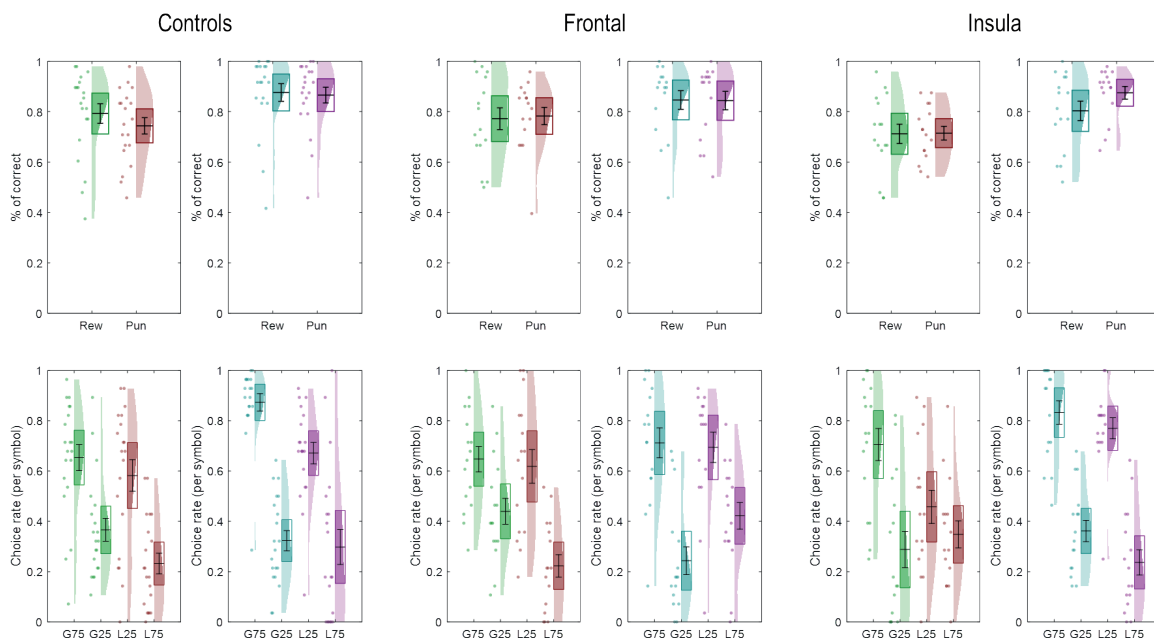


Figure 33. Behavioral results for the study on frontal and insular brain lesions. Top: correct choice rate in the learning phase. Bottom: choice rate for the transfer phase. Color coded as in Figure 30.

In the learning phase, we found a main effect of feedback information ($F(1, 49) = 47.34$, $p < .0001$) meaning that patients had a better performance in complete feedback information contexts, compared to partial feedback. We found no other significant main effect nor interaction (Figure 33, top).

In the transfer phase, we found a significant effect of favorableness ($F(1, 48) = 228.11$, $p < .0001$), information ($F(1, 48) = 15.49$, $p = .00039$), and valence ($F(1, 48) = 8.63$, $p = .0051$). We found an interaction between favorableness and feedback information ($F(1, 48) = 8.05$, $p = .0066$), and a significant interaction between valence, information, favorableness, and group ($F(2, 48) = 8.04$, $p = .00098$). Although challenging to interpret, this interaction suggests that the insular lesions group chose less often the most favorable option from the punishment, partial feedback context (L25), when compared to other groups. This is the only significant result providing evidence for impaired punishment learning in the group of patients with insular lesions (Figure 33, bottom).

B.4 Conclusion

To conclude, our results did not replicate the findings from [Palminteri et al. 2012](#). For the Huntington's disease group, our results have to be interpreted considering that in the original study, the groups consisted in presymptomatic and symptomatic patients, whereas our group of patients represents gene carriers far from symptomatic onset. Therefore, we can conclude that gene mutation far from onset leads no detectable change in learning from punishments. To my knowledge, only few studies have directly investigated punishment-based avoidance learning in Huntington's disease. Nevertheless, our results are in line with another study comparing Huntington's disease gene carriers near to (< 5 years) and far from (> 5 years) motor symptom onset ([Enzi et al. 2012](#)). Using a different task than ours, the Monetary Incentive Delay (MID) task ([Knutson et al. 2000](#)) during fMRI data acquisition, Enzi and colleagues found that healthy controls and the "HD-far" group exhibited similar patterns of brain activations when discriminating between rewards and punishments, whereas the "HD-near" group showed impairments in punishments conditions. The number of successful reward and punishment trials did not differ significantly between the three groups, however HD-near patients showed longer reaction times concerning both trial types than healthy controls, reflecting a decline in motor performance associated with disease progression ([Enzi et al. 2012](#)). Of note, we did not find any group effect of reaction times in our data.

As discussed in the previous chapters of my thesis, dopamine enhancers given to healthy subjects or to patients with Parkinson's disease improve reward learning but leave unaffected, or some-

times even impair, punishment learning (Frank et al. 2004, Pessiglione et al. 2006, Palminteri et al. 2009, Shiner et al. 2012, Pessiglione and Delgado 2015). In a paper published in 2010 in *Neuron*, Voon and colleagues compared two groups of patients with Parkinson’s disease, with or without dopaminergic dysregulation syndrome, and found that dysregulated patients showed better performance in learning from rewards than from punishments, and learned faster than regulated patients in this context (Voon et al. 2010). Based on this previous literature, by using a task that orthogonally manipulates reward and punishment learning, we were expecting medicated patients with Parkinson’s disease to learn better from rewards than from punishments. However, our results do not replicate this effect, neither in the regulated group nor the dysregulated group (we note that we do not have behavioral data for a control group yet).

Regarding the study including patients with brain damage, the role of the anterior insula in punishment-based avoidance learning has been investigated using fMRI (Seymour et al. 2004, Kim et al. 2006, Palminteri et al. 2015, Rigoli et al. 2016) and patients with brain tumors or lesions (Palminteri et al. 2012). Results from studies involving patients with brain damage showed that insular lesions specifically impair punishment learning (Palminteri and Pessiglione 2017). From these findings, we expected patients with lesions located in the insular cortex to learn less from punishments than from rewards, when compared with patients with lesions in the frontal cortex and healthy controls. However, our results, at least in the learning phase, do not indicate a significant difference between reward and punishment learning in patients with insular lesions, nor when comparing with patients with frontal lesions and healthy controls. Regarding these results, as well as the absence of valence bias in the current study with Parkinson’s disease patients, one might argue that, by making the task easier with a block design, we might have made the learning phase of the task *too* easy, which might have blunted the significant results observed in previous studies (Palminteri et al. 2009, 2012). This might explain why no significant valence bias was observed in the Parkinson’s groups nor in the lesions groups, whereas some significant differences were assessed in the transfer phase.

B.5 Major depressive disorder

In this additional section, I will now present a draft of a clinical study assessing value-based decision making impairment in major depressive disorder, co-first authored by Henri Vandendriessche and Amel Demmou. To test whether reward sensitivity deficits are dependent on the overall value of the decision problem, we used a reinforcement learning task that includes two different contexts: one "rich" context where both options are associated with an overall positive

expected value, and a "poor" context where options are associated with overall negative expected value. The task was performed by 30 patients undergoing a major depressive episode and 26 age-, gender- and socioeconomically-matched controls.

We found that contrary to healthy participants, patients showed reduced learning in the "poor" context when compared with the "rich" context. Analysis of the transfer phase showed that the context-dependent deficit in patients transferred when the options were extrapolated from their original context. Together, these results suggest that the detrimental effect of major depressive episodes is a learning, rather than a decision, impairment.

↑

Context-dependent reinforcement learning impairment in depression

Henri Vandendriessche*, Amel Demmou*, Julien Yadak, Sophie Bavard, Thomas Mauras[°], Stefano Palminteri[°]

*co-first author

[°]co-last author

Abstract: (250w)

Backgrounds:

Value-based decision-making impairment in depression is a complex phenomenon: while some studies did find evidence of blunted reward learning and reward-related signals in the brain, others indicate no effect. Here we test whether such reward sensitivity deficits are dependent on the overall value of the decision problem.

Methods:

We used a classical two-armed bandit task that includes two different contexts: one ‘rich’ context where both options were associated with an overall positive expected value and a ‘poor’ context where options were associated with overall negative expected value. We tested N=30 patients undergoing a major depressive episode and N=26 age, gender and socio-economically matched controls. To assess whether context-induced reinforcement deficit in patients was due to a decision or a value-update process, we analysed performance in a transfer test, performed immediately after the learning test, where we asked to indicate the most rewarding option in all possible combinations.

Results

Healthy subjects showed similar learning performance in both the ‘rich’ and the ‘poor’ context, while patients showed reduced learning in the ‘poor’ context. Analysis of the transfer test showed that the context-dependent deficit in patients replicated when the options were extrapolated from their original context. This suggests that the effect of depression is a learning, rather than a decision, impairment.

Conclusions

Our results illustrate that reinforcement learning deficits in depression are complex and depend on the value of the context. We show that depressive patients have a specific trouble in environment with an overall negative state value. Relevance for clinic.

Key Words:

Depression, reward processing, reinforcement learning, context

Core Text (4500w)

Introduction:

Depression is a common debilitating disease that is a worldwide leading cause of morbidity and mortality. According to the latest estimates from World Health Organization, more than 300 million people are now living with depression. Low mood and anhedonia are core symptoms of major depressive disorder. Those two symptoms are key criteria to the diagnostic of Major Depressive Disorder in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) (American Psychiatric Association, 2013). Anhedonia is broadly defined as a decreased ability to experience pleasure from positive stimuli. More specifically it is described as a reduced motivation to engage in daily life activities (motivational anhedonia) and reduced enjoyment of usually enjoyable activities (consummator anhedonia).

Depression is a complex and heterogeneous disorder implying instinctual, emotional and cognitive dysfunctions. Although its underlying mechanisms remain unclear, both neurobiological and neurofunctional processes seem to be at work. It has been proposed that reduced reward processing, both in terms of incentive motivation and reinforcement learning, plays a key role in the clinical manifestation of depression (Chen et al., 2015; Eshel & Roiser, 2010; Huys et al., 2013; Whitton et al., 2016). This hypothesis implies that depressive subjects should display reduced reward sensitivity both at behavioral and neural level in value-based learning .

Following up on this hypothesis, numerous studies tried to identify and characterize such reinforcement learning deficits, however the results have been mixed so far. Indeed, if some studies did find evidences of blunted reward learning and reward-related signals in the brain, others indicate limited or no effect (Hägele et al., 2015; Chung et al., 2017; Rutledge et al., 2017; Shah, O'carroll, Rogers, Moffoot, & Ebmeier, 1999; Huys et al., 2013; Rothkirch et al., 2017). Outside the learning domain, others recent studies showed no disrupted valuation during decision-making under risk (Moutoussis 2018; Chung et al., 2017). It is also worth noting that many of previous studies identifying value-related deficits in depression, only included one valence domain (only rewards or only punishment) and did not directly contrasted between rewards and punishments or separated the two valence domains in different experimental sessions (Elliott et al., 1997; Steele et al., 2007; Kumar et al., 2008; Gradin et al., 2011; Forbes and Dahl, 2012 ; Vrieze et al., 2013 ; Zhang et al., 2013; Pizzagalli, 2014).

Here we hypothesized that this absence of concordant results may be in part explained by the fact that reinforcement learning impairment in depression is dependent on the context value of the decision problem. To test this hypothesis, we modified a standard reinforcement learning task including a learning phase and a post-learning transfer test (with no feedback in order to probe the subjective values of the options without modifying it). The learning phase included two different contexts: one defined as "rich" (in which the two options have an overall positive expected value) and the other as "poor" (two options with a negative expected value). We assessed performance in the learning test and a function of the context and the patients group, and we found a significant interaction, where depressive patients were specifically impaired in the 'poor' environment. We also analyzed the transfer test performance, where patients found more easily the richest option than the poorest one, which confirmed this depression-induced deficit.

Materials and Methods

Subjects and inclusion criteria

The subjects were recruited in clinical centers. Inclusion criteria were a diagnostic of major unipolar depression diagnosed by a psychiatrist and an age between 18 and 65 years old. A clear, oral and written information was also delivered to all participants. All procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. Exclusion criteria were the presence of psychotic symptoms or a diagnostic of chronic psychosis, severe personality disorder, neurological or any somatic disease that might cause cognitive alterations, neuroleptic treatment, electro-convulsive therapy in the past 12 months and active toxic use. Antidepressant, benzodiazepine and antihistaminic treatments were allowed. Psychiatric co-morbidities were established by a clinician in a usual psychiatric assessment and a semi-structured interview based on the Mini International Neuropsychiatric Interview (MINI) (Sheehan et al 1998). In addition to these criteria, the control subjects had to be free of any past or present depressive episode or psychiatric treatment. In total, we tested **N=30** patients undergoing a major depressive episode (MDE) and **N=26** age, gender and socioeconomically matched controls.

Behavioral testing

Voluntary patients were welcomed to the crisis center and seated in an office away from the center's activity where they were given information about the aim and the procedure of the study. The study was orally described as an evaluation of cognitive functions through a computer « game ». The diagnostic of major depressive episode and the presence of psychiatric co-morbidities were assessed with the MINI screener completed in a semi-structured interview with a psychiatrist by the MINI. The subjects were then asked to complete several questionnaires assessing their level of optimism (Life Orientation Test- Revised (LOT-R), an optimism analogue scale (created for this study to contrast usual and current level of optimism) and the severity of depression (Beck Depression Inventory – II) (see Supplementary materials for more details).

The participants were told they are going to play a simple computer game which goal is to win as many points as possible. Written instruction were provided and orally reformulated.

The task was a probabilistic reinforcement-learning game in which two stable pairs of abstract symbols (or choice contexts) appeared alternatively on a black screen. The subjects were told that one of the two symbols was more rewarding than the other and encouraged to find out which one. The reward probability attached to each symbol was never specifically given and the subjects had to learn it through trial and error. Each symbol was associated to a fixed reward probability. Reward probabilities were distributed across symbols as follows: 10% -40% (“poor” context), 60% - 90% (“rich context”). The reward probabilities were decided in order to have the same choice difficulty across choice contexts. When the symbols appeared on the screen, the subject had to choose between the two symbols by pushing a right or a left key on a keyboard. In rewarded/punished trials a green/red smiley/sad face appeared on screen . In order to be sure that the subjects payed attention to the feedback, they had to push the up key after a win and the down key after a loss to move to the next trial.

The two learning sessions of **100** trials each (involving different set of stimuli - 8 different symbols in total) were followed by a transfer-test of **112** trials in which the 8 different symbols were presented by pairs in all binary combinations four times (including pairing that had never been displayed together in the previous task). The subjects had to choose which symbol deemed the more rewarding, however, in the transfer test, no feedback was provided in order to not interfere with subject's final estimates of option values. The subjects were told to use instinct when doubting. The aim of the transfer-test was to evaluate the subject's capacity to remember and extrapolate the symbol's value out of its initial context (generalization).

Dependent variables

The main behavioral dependent variable in this study is the correct choice rate. A correct choice is defined, both in the learning and in the transfer test, as a choice toward the reward maximizing symbol. In the learning test, the correct symbols were the 40% p(reward) (in the “poor” environment) and the 90% p(reward) in the “rich” environment. In the transfer test the correct symbols were defined in a trial-by-trial basis and dependent on the particular combination presented (note that in some trials a correct symbol could not be defined, as the comparison involved two symbols with the same value). The learning curves (**Figure 2**) were generated applying smoothing window of five trials. Statistical analyses were performed on unsmoothed data. As exploratory dependent variable we also extract the reaction time and the outcome observation time (see **Supplementary materials**), as a function of the choice context and group (patients vs. controls). Psychometric personality scales were also considered and compared across groups.

Statistical analyses

To assess the effect of choice context and clinical group in the learning test, we submitted our dependent variables (correct choice rate, reaction time and outcome observation time) to a General Linear Model (GLM). At the individual level, the trial-by-trial variable was modeled as:

$$Y_{i,j} = \beta_{0,j} + \beta_{1,j} * X1$$

Where $j \in [1 ;56]$ was the number of subjects, $i \in [1 ;200]$ was the trial number, $X1$ was a binary vector of the choice context (rich=1; poor=-1), β_1 was the regression coefficient associated to the choice context and β_0 was the intercept. We also run control GLMs, where we added additional predictors ($X2$ = trial number; $X3$ = trial-by-choice context interaction; see **Supplementary materials**).

The statistical analyses of the transfer test correct choice rate was restricted to comparisons involving the best possible (reward probability=0.9) or the worst possible (reward probability = 0.1) options and we did not include comparisons involving *both* the best and the worst option, *neither* the best and worst option, or two options with the same expected value. The resulting N=64 trials were analyzed also with a GLM approach, where $X1$ was a vector indicating presence (=1) or absence (= -1) of the best possible option.

The between group were assessed by comparing the resulting regression coefficients (β_0 and β_1) using two-sample t-test. Note that, since the correct choice rate was coded as incorrect = -1 and correct = +1 (correct), an intercept greater than zero indicate above chance performance.

Results:

Demographics.

Patients and controls were matched in age ($t(51)=-1.1$, $p=0.28$), gender ($t(3)=1.71$, $p=0.63$) and years of education ($t(54)=-1.59$, $p=0.12$). Concerning the optimist personality measures, depressive patients were found to be less optimistic in all scales (LOT-R: $t(47)=-7.42$, $p=1.76e-09$; usual optimism: $t(51)=-2.29$, $p=0.03$; current optimism: $t(50)=-10.34$, $p=4.19e-14$). Furthermore, the comparison between usual vs. current optimism in patients and controls, revealed that only patients were significantly less optimistic than usual at the moment of the test (patients: $t(29)=8.26$,

$p=4.21e-09$; controls $t(25)=-1.53$, $p=0.14$), consistent with the fact that they were undergoing an major depressive episode. All patients were taking at least one psychotropic medication at the moment of test. Their average BDI at the moment of test was: **29.37** and they had, in average, **1.8** previous MDE in the past.

Learning test results

Global inspection of the learning curves (**Figure 2A**) suggests that, overall participants were able to learn to respond correctly. Indeed, all the learning curves are above chance whatever the group or the environment. A more detailed inspection reveals that controls' learning curves were unaffected by the choice environment ('rich' vs. 'poor'), while patients' learning curves were different depending on the choice environment (with a lower correct response rate in the 'poor environment').

A between-group comparison of the baseline correct response rate (as proxied by the intercept of our GLM) in the learning phase (**Figure 3A**) indicated that were significantly above chance in both groups (controls: $t(25)=5.44$, $p=1.19e-5$ and patients: $t(29)=5.35$, $p=9.69e-6$) and not different between the two groups ($t(54)=-0.23$, $p=0.82$). This confirms that there is no difference in term of baseline performance between controls and patients. On the other hand the effect of the choice environment value was significantly different between patients and controls ($t(54)=-2.46$, $p=0.017$). Looking at the effect of environment on the performance we see that controls performed equally in both environments when patients performed better in rich environment than in poor one. More precisely, the effect of valence was significantly different from zero only in the patients group ($t(29)=3.32$, $p=0.002$)

These results show a context-specific impairment in the patients group, which is absent in the controls who do not seem affected at all by the value of the environment. Looking at the learning phase only, we cannot establish if this impairment stems from a learning or a decision-making deficit. To tease apart these interpretations we turned to the analysis of the transfer test performance.

Transfer test analysis

Similarly, the transfer test results (**Figure 2B**) indicates that subjects were able to retrieve the value of the stimuli. Accordingly, option 'A' was chosen much more frequently compared to option 'D' in both groups. Crucially, and in accordance with the learning phase results, the difference between the 'C' and the 'D' option was smaller in the patients' group.

Baseline correct response rate once again showed no statistical difference between the two groups ($t(54)=0.44$, $p=0.67$). However the ability to choose A and avoid D revealed a clear difference between the two groups ($t(54)=-3.04$, $p=0.0036$). Controls showed no preference and were able to choose A as frequently as they were to avoid D whereas patients were strikingly better at choosing A than avoiding D ($t(29)=4.7$, $p=5.38e-05$).

These results are consistent with the learning test results. The context-specific deficit in patients that we found in the learning test was also present in the transfer phase where all the different options were extracted from their initial context and displayed with other options. Therefore, it allows us to conclude that the deficit is not only a decision-making deficit but also a learning deficit that is probably induced by a negative affective bias triggered by negative feedbacks in the poor environment (Roiser et al., 2012).

Discussion

In the present study, we assessed reinforcement learning with a behavioral paradigm involving two different reward environments - one 'rich' with a positive overall expected value and one 'poor' with a negative overall expected value - in patients undergoing a major depressive episode and age and education matched healthy controls.

As expected, healthy subjects learned equally well in both environments. On the other hand, depressed patients displayed reduced learning rate in the 'poor' environment. This context-dependent learning asymmetry was found in both the learning phase and in a transfer test, where subjects were asked to retrieve and generalize the values learned during the learning sessions.

This suggests that this depression-related learning asymmetry does not stem from the learning process per se (*primary* learning deficit) and not from a decision process (*secondary* learning deficit). We can hypothesize that the context-value induced deficit observed at the learning phase can be caused by negative affective biases when confronted to aversive feedback as a loss of a point in our case. Confrontation with negative affective stimuli seems to affect the updating process of the state value in environment with an overall negative state value. On the other side, the confrontation with positive affective stimuli does not affect their performances at all.

On the other hand, another interesting point in literature might lead to a different hypothesis to explain the present results. Indeed some studies found an impairment in processing of positive feedback for depressed subjects, based on a diminution of positive prediction error signal in this population (Knutson 2008; Kumar et al 2008; Gradin et al. 2011; Ubl et al. 2015; Whitton et al., 2016). This signal would code, at a neurobiological and neurofunctional level, the surprise effect caused by a better than expected information. It would be involved in the process of learning through positive feedback. We could here hypothesize that in the poor environment, positive feedback, scarcer, is therefore more salient and more determinant to the learning process. A decreased sensibility to positive feedback could in that case explain the weakest learning performance of depressed subjects in this environment.

We can also infer from the present results that the learning deficit for patients is not a pure decision problem, as it is not observed in the rich environment. It would reflect a dysfunction during the learning process, dependent of context, rather than a learning deficit per se. Some previous results seem to suggest that this dysfunction would not be at a perceptual level because valuation in major depression is intact in a non-learning environment (Chung et al., 2017). The fact that the deficit was still present in the transfer test, away from feedback, seem to imply that this dysfunction is not just a short-term effect on valuation due to negative affective stimuli. It would rather involve complex mechanisms, embedded in the learning process, and triggered by negative affective stimuli.

Place in the literature

It's a significant step forward to better understand major depressive disorder and its cognitive implications on patients' behavior. These results should help disentangling the conflicting results in the literature on blunted reward learning in patients suffering from major depressive disorder.

On a behavioral aspect, the good performance of patients in the rich environment is not very common in the literature but can probably be explain by the interleaved design of the task.

Switching from poor to rich trials may boost patient's confidence and motivate them to perform better. It can also explain the absence of reduced positive affect observed in certain studies (Knuston et al. 2008)

Another interesting result, also present in the literature that is visible on both controls and patients is the learning asymmetry present in the transfer test (figure 2b). Symbol B should be higher than symbol C. Every symbol's value is learned in relation and comparison to its "partner" symbol within the initial pair. The participants' inability to differentiate B from C seem to reflect their inability to determine the absolute value of symbols.

Consequences for clinical practice, research and understanding of the symptoms

The consequences of this result deserve to be more thoroughly explored especially by psychiatrists in charge of patients. The fact that patient's performance do not differ from controls in the rich environment is very encouraging and should be exploited as in some psychotherapeutic practices, notably cognitive-behavioral, where the patient is placed in a spiral of success. Splitting burdensome activities in smaller and simpler tasks achievable more easily should provide more positive affective stimuli. It is a question of prioritizing the tasks and prescribing them in a graduated way so as to meet only successes.

Limitation and perspectives

One of the limitations of our study is that patients were medicated at the time of the experiment. It is possible that antidepressants had an effect on patients and therefore on their cognitive mechanism. Even though studies have found effects on performance on medicated and unmedicated patients (Steele et al., 2007, Douglas et al., 2009) it is always very difficult to control for this effect especially when certain patients take medications for other comorbidities.

The overall good performances of patients and more specifically in the rich environment could be explained as explained earlier by interleaved design of the experiment but also by the fact that patients in general are more focused and more involved than controls in this type of study. The result of this test is much more meaningful for them than it is for controls that are not really impacted by the outcomes of the experiment.

In the literature it has been shown numerous times that controls perform equally when they have to choose a reward or avoid a punishment and it's also frequent that patients with mental or neurological disorders other than major depression disorder show an imbalance behavior when implicated in a task with a reward selection and avoiding a punishment (Frank et al., 2004). Studying several aspects of reward processing that correspond to different neurobiological circuits and exploring dysregulation across different psychiatric disorders could be a very efficient way to unfold abnormalities in reward-related decision making. It could be very interesting to apply that task to other psychiatric disorders in order to identify neurobiological signatures and develop more targeted and promising treatments. (Insel et al., 2010; Whitton et al., 2015)

Besides the questions raised on care and support of patients undergoing a MDD, this study makes us discuss polysemic clinical concepts such as anhedonia, which appears to be relative to the context. A natural follow up to that study would be the development of a reinforcement-learning model to understand more deeply this context dependent learning deficit. Another option would be to

replicate these results in an fMRI to characterize this deficit from an anatomical and functional point of view.

F- References and Notes:

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Pub.

Blanchard, J. L., Horan, W. P., & Brown, S. A. (2001). Diagnostic differences in social anhedonia: A longitudinal study of schizophrenia and major depressive disorder. *Journal of Abnormal Psychology, 110*(3), 363–371. <https://doi.org/10.1037/0021-843X.110.3.363>

Chen, C., Takahashi, T., Nakagawa, S., Inoue, T., & Kusumi, I. (2015). Reinforcement learning in depression: A review of computational research. *Neuroscience & Biobehavioral Reviews, 55*, 247–267. <https://doi.org/10.1016/j.neubiorev.2015.05.005>

Chung, D., Kadlec, K., Aimone, J. A., McCurry, K., King-Casas, B., & Chiu, P. H. (2017). Valuation in major depression is intact and stable in a non-learning environment. *Scientific Reports, 7*, 44374. <https://doi.org/10.1038/srep44374>

Cooper, J. A., Arulpragasam, A. R., & Treadway, M. T. (2018). Anhedonia in depression: Biological mechanisms and computational models. *Current Opinion in Behavioral Sciences, 22*, 128–135. <https://doi.org/10.1016/j.cobeha.2018.01.024>

Douglas, K. M., Porter, R. J., Frampton, C. M., Gallagher, P., Young, A. H. (2009). Abnormal response to failure in unmedicated major depression. *Journal of Affective Disorders*, Volume 119, Issues 1–3, Pages 92-99, ISSN 0165-0327, <https://doi.org/10.1016/j.jad.2009.02.018>.

Eshel, N., & Roiser, J. P. (2010). Reward and Punishment Processing in Depression. *Biological Psychiatry, 68*(2), 118–124. <https://doi.org/10.1016/j.biopsych.2010.01.027>

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. *Science, 306*(5703), 1940–1943. <https://doi.org/10.1126/science.1102941>

Hägele, C., Schlagenhauf, F., Rapp, M., Sterzer, P., Beck, A., Bermpohl, F., ... Heinz, A. (2015). Dimensional psychiatry: Reward dysfunction and depressive mood across psychiatric disorders. *Psychopharmacology, 232*(2), 331–341. <https://doi.org/10.1007/s00213-014-3662-7>

Harmer Catherine J., & Cowen Philip J. (2013). ‘It’s the way that you look at it’—A cognitive neuropsychological account of SSRI action in depression. *Philosophical Transactions of the Royal Society B: Biological Sciences, 368*(1615), 20120407. <https://doi.org/10.1098/rstb.2012.0407>

Huys, Q. J. M., Daw, N. D., & Dayan, P. (2015). Depression: A Decision-Theoretic Analysis. *Annual Review of Neuroscience, 38*(1), 1–23. <https://doi.org/10.1146/annurev-neuro-071714-033928>

Huys, Q. J., Pizzagalli, D. A., Bogdan, R., & Dayan, P. (2013). Mapping anhedonia onto reinforcement learning: A behavioural meta-analysis. *Biology of Mood & Anxiety Disorders, 3*(1),

12. <https://doi.org/10.1186/2045-5380-3-12>

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., ... Wang, P. (2010). Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. *American Journal of Psychiatry*, *167*(7), 748–751. <https://doi.org/10.1176/appi.ajp.2010.09091379>

Ironside, M., Amemori, K., McGrath, C. L., Pedersen, M. L., Kang, M. S., Amemori, S., ... Pizzagalli, D. A. (2019). Approach-avoidance conflict in major depression: Congruent neural findings in human and non-human primates. *Biological Psychiatry*. <https://doi.org/10.1016/j.biopsych.2019.08.022>

Kennedy, S. H. (2008). Core symptoms of major depressive disorder: Relevance to diagnosis and treatment. *Dialogues in Clinical Neuroscience*, *10*(3), 271–277.

Knutson, B., Bhanji, J. P., Cooney, R. E., Atlas, L. Y., & Gotlib, I. H. (2008). Neural Responses to Monetary Incentives in Major Depression. *Biological Psychiatry*, *63*(7), 686–692. <https://doi.org/10.1016/j.biopsych.2007.07.023>

Moutoussis, M., Rutledge, R. B., Prabhu, G., Hrynkiewicz, L., Lam, J., Ousdal, O.-T., ... Dolan, R. J. (2018). Neural activity and fundamental learning, motivated by monetary loss and reward, are intact in mild to moderate major depressive disorder. *PLOS ONE*, *13*(8), e0201451. <https://doi.org/10.1371/journal.pone.0201451>

Must, A., Horvath, S., Nemeth, V. L., & Janka, Z. (2013). The Iowa Gambling Task in depression – what have we learned about sub-optimal decision-making strategies? *Frontiers in Psychology*, *4*. <https://doi.org/10.3389/fpsyg.2013.00732>

Pizzagalli, D. A., Holmes, A. J., Dillon, D. G., Goetz, E. L., Birk, J. L., Bogdan, R., ... Fava, M. (2009). Reduced Caudate and Nucleus Accumbens Response to Rewards in Unmedicated Individuals With Major Depressive Disorder. *American Journal of Psychiatry*, *166*(6), 702–710. <https://doi.org/10.1176/appi.ajp.2008.08081201>

Pizzagalli, D. A., Jahn, A. L., & O’Shea, J. P. (2005). Toward an objective characterization of an anhedonic phenotype: A signal-detection approach. *Biological Psychiatry*, *57*(4), 319–327. <https://doi.org/10.1016/j.biopsych.2004.11.026>

Roiser, J. P., Elliott, R., & Sahakian, B. J. (2012). Cognitive Mechanisms of Treatment in Depression. *Neuropsychopharmacology*, *37*(1), 117–136. <https://doi.org/10.1038/npp.2011.183>

Rothkirch, M., Tonn, J., Köhler, S., & Sterzer, P. (2017). Neural mechanisms of reinforcement learning in unmedicated patients with major depressive disorder. *Brain*, *140*(4), 1147–1157. <https://doi.org/10.1093/brain/awx025>

Rutledge, R. B., Moutoussis, M., Smittenaar, P., Zeidman, P., Taylor, T., Hrynkiewicz, L., ... Dolan, R. J. (2017). Association of Neural and Emotional Impacts of Reward Prediction Errors With Major Depression. *JAMA Psychiatry*, *74*(8), 790–797. <https://doi.org/10.1001/jamapsychiatry.2017.1713>

Safra, L., Chevallier, C., & Palminteri, S. (2019). Depressive symptoms are associated with blunted reward learning in social contexts. *PLOS Computational Biology*, *15*(7), e1007224. <https://doi.org/10.1371/journal.pcbi.1007224>

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., & Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of clinical psychiatry*, *59 Suppl 20*, 22–57

Seligman ME (1972): Learned helplessness. *Annu Rev Med* 23:407– 412

Shah, P. J., O'carroll, R. E., Rogers, A., Moffoot, A. P. R., & Ebmeier, K. P. (1999). Abnormal response to negative feedback in depression. *Psychological Medicine*, *29*(1), 63–72. <https://doi.org/10.1017/S0033291798007880>

Steele, J. D., Kumar, P., & Ebmeier, K. P. (2007). Blunted response to feedback information in depressive illness. *Brain*, *130*(9), 2367–2374. <https://doi.org/10.1093/brain/awm150>

Talmi, D., Lohnas, L. J., & Daw, N. D. (2019). A retrieved context model of the emotional modulation of memory. *Psychological Review*. <https://doi.org/10.1037/rev0000132>

Treadway, M. T., & Zald, D. H. (2011). Reconsidering anhedonia in depression: Lessons from translational neuroscience. *Neuroscience & Biobehavioral Reviews*, *35*(3), 537–555. <https://doi.org/10.1016/j.neubiorev.2010.06.006>

Wacker, J., Dillon, D. G., & Pizzagalli, D. A. (2009). The role of the nucleus accumbens and rostral anterior cingulate cortex in anhedonia: Integration of resting EEG, fMRI, and volumetric techniques. *NeuroImage*, *46*(1), 327–337. <https://doi.org/10.1016/j.neuroimage.2009.01.058>

Whitton, A. E., Treadway, M. T., & Pizzagalli, D. A. (2015). Reward processing dysfunction in major depression, bipolar disorder and schizophrenia. *Current Opinion in Psychiatry*, *28*(1), 7–12. <https://doi.org/10.1097/YCO.0000000000000122>

Supplementary materials

Scales and diagnostic questionnaires

the Life Orientation Test- Revised (LOT-R) : self-questionnaire establishing the level of optimism. The test is composed of 10 affirmations including 4 decoys. The other 6 affirmations concern expectations for the future. The subject has to cote between one and five points to what extent he agrees with the affirmation.

Optimism Analogue Scale, created for this study. The subject has to evaluate his optimism by placing a cross on a 10 cm line. The origin and the end of the line correspond to the worst and the best state of optimism imaginable respectively. Current and usual levels of optimism are estimated.

Beck Depression Inventory – II (BDI-II): self-questionnaire estimating symptoms of depression during the past two weeks. This scale is composed with 21 groups of 4 graduated affirmations among which the subject has to choose the one better corresponding to his state. Each group of affirmation explores a specific dimension of depressive syndrome (sadness, guilt, feeling of failure...).

Scores are interpreted as follows:

0-13 : normal mood variations

14-19 : mild depression

20-28 : moderate depression

>29 : severe depression

MINI-Screener. Quick self-questionnaire screening psychiatric diagnostics with the filter questions (first, necessary and eliminatory questions) of each section of the MINI.

When the answer to one of the MINI-screener items is positive, it is completed with the entire corresponding section of the MINI in a semi-structured interview.

Mini International Neuropsychiatric Interview (MINI): diagnostic questionnaire using DSM-IV criteria to assess several axis I psychiatric diagnostics: current or past major depressive episode, melancholia, suicidal risk, hypomania, mania, social phobia, panic disorder, agoraphobia, obsessive-compulsive disorder, addiction or toxic use (alcohol or other substances), psychosis, anorexia, bulimia, post-traumatic stress disorder.

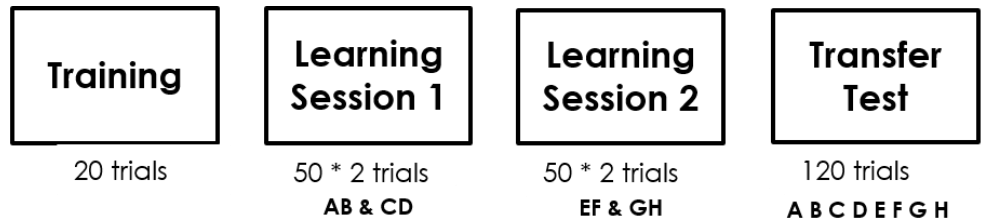
In addition to the LOT-R, the analogue optimism scale and the MINI-screener, controls completed :

The Big Five Inventory (BFI) : Self-questionnaire evaluating five dimension of personality : extraversion, agreeableness, conscientiousness, neuroticism, openness.

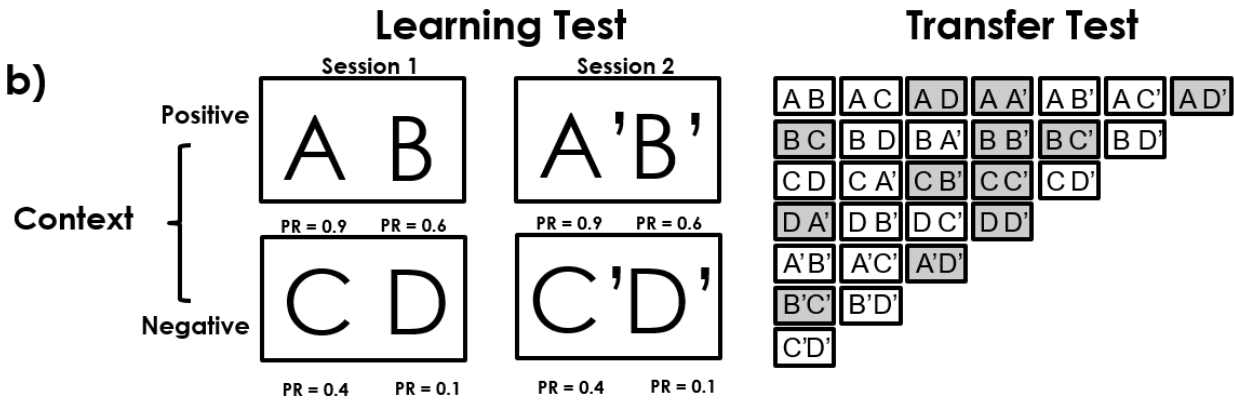
Figure

1:

a)



b)



c)

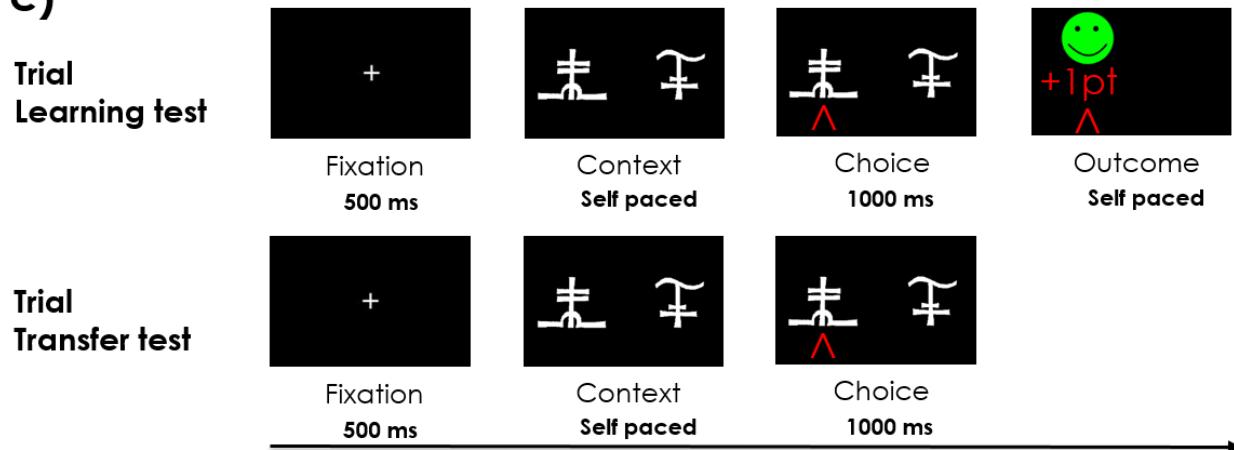


Figure 1: Experimental design:

a) Experimental design: the experiment course is composed of a short training with neutrals stimuli (letters) which is followed by two learning sessions with 4 different stimuli each. The last session is the transfer test where all stimuli from the learning sessions are shuffled and presented pair-wise.

b). One learning session is composed of 2 different contexts: a rich one with an overall positive expected value (one symbol with a 0.9 gain probability and the second symbol with a 0.6 gain probability) and a poor context (one symbol at 0.4 and the second one with 0.1 gain probability). The two contexts are interleaved during the learning phase with a limit of repetition. Participants are told to find the most rewarding symbol in every trial. On the transfer phase all 8 symbols from the learning phase (2 symbols x 2 contexts x 2 learning sessions) are presented in every possible combination no matter what context they belong to.

c). Trials are following the same course in the test phase and in the transfer-test phase except that the transfer-test doesn't have any outcome.

Figure 2

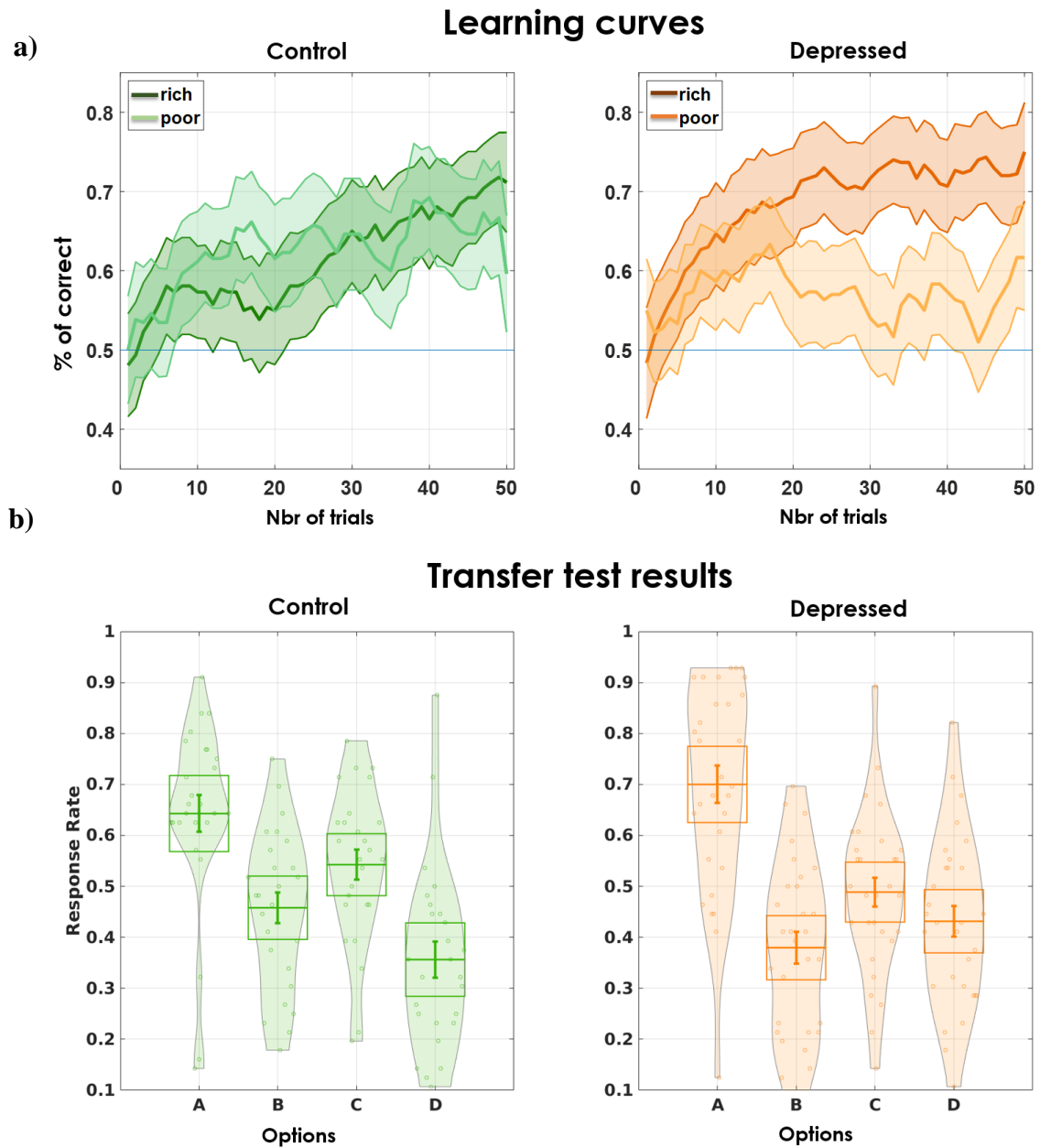


Figure 2:

a) Learning test:

Learning curves in percentage of correct response across trials. The darker curve represents the rich environment and the lighter curves represent the poor environment. Curves are pooled for every session of every participant and smoothed with 5 points. The standard error of the mean is displayed in transparent color around the curve.

b) Transfer phase

Response rate for every symbol during the transfer test. All eight symbols of the two learning phase sessions are presented together (A and A' are presented together as well as B and B' etc.). Every symbol appears 14 times and each two-symbol comparison appears 2 times balanced left and right. Gray dots are the value for each participant.

Figure 3:

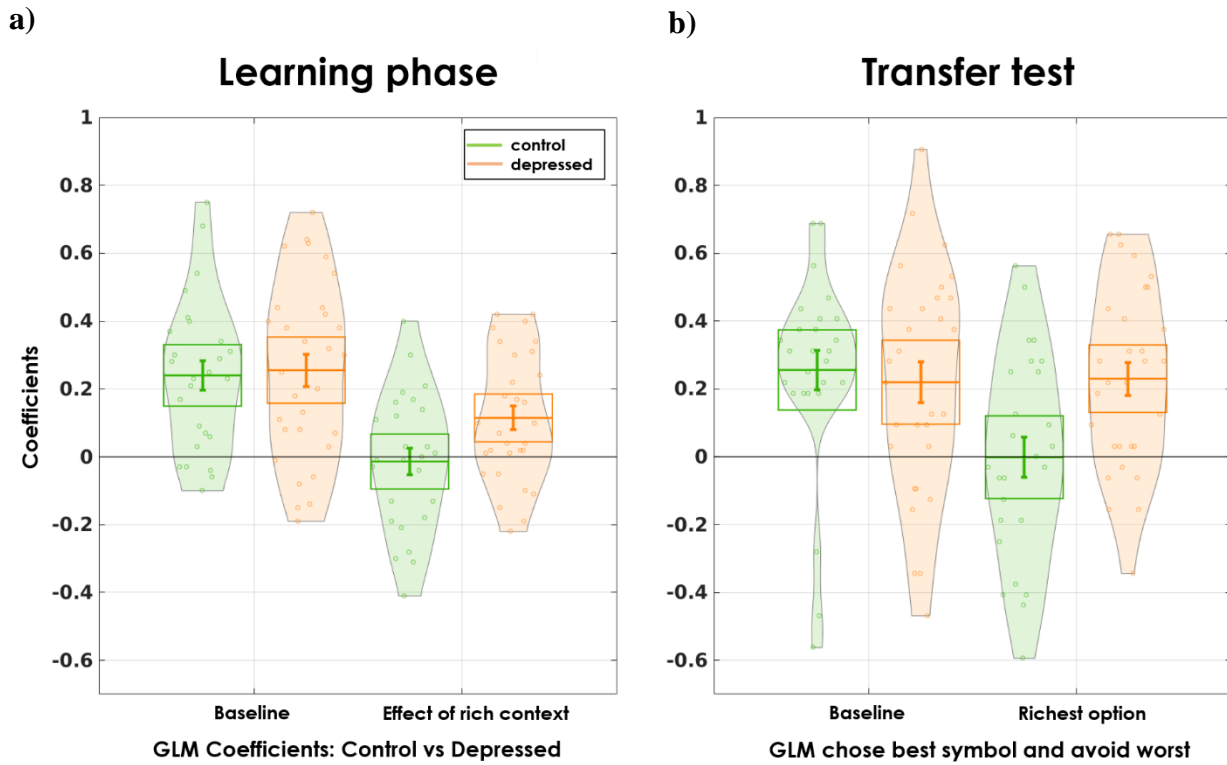


Figure 3:

a). General linear model of the learning phase.

$$Y_{i,j} = \beta_0 + \beta_1 * X1$$

Where X1 represents the environment (rich=1, poor=-1). β_0 quantify for each subject the baseline performance. β_1 represents the environment effect. If positive, subject learns better from rich environment. Grey dots are the individual performance for every subject.

b). General linear model of the post-learning phase.

The data used to generate this figure takes in account symbol comparison without the most obvious comparison (A vs D) and the less obvious comparison (B vs C) (cf figure 1, transparent diapos). Symbol B and C were equally confused by all the participants and symbol A and D comparison was obvious to every participants whatever the population.

$$Y_{i,j} = \beta_0 + \beta_1 * X1$$

Where X1 code for the ability to choose the best symbol over avoiding the worst (Choose A = 1, avoid D = -1). β_0 quantify for each subject the baseline performance, β_1 represent the tendency to choose A and avoid D. The value 0 represents equal performance in choosing the best symbol and avoiding the worst. Grey dots are the individual performance for every subject.

Table 1

Group	Patients	Controls	Difference (P)
Age (mean±sen)	36.5 ± 2.80	40.35 ± 2.09	Df= 51.71 , P= 0.28
Gender (%female)	30 (53.33)	26 (61.53)	Df=3, P = 0.63
Education (years after BAC)	1.97 ± 0.24	2.42 ± 0.21	Df = 54 , P = 0.12
Usual Optimism	5.98 ± 0.42	7.16 ± 0.30	Df= 51.33 , P= 0.03
Current Optimism	2.38 ± 0.40	7.46 ± 0.29	Df = 50.82 , P= 4.19e-14
LOTR	9.1 ± 0.79	16 ± 0.49	Df = 47.46 , P= 1.76e-09
BDI	29.37 ± 0.22	-	-
MDE	1.8 ± 0.38	-	-

Table 1:

Descriptive statistics for age, gender, education, usual optimism, current optimism, life orientation, depression scores and number of major depressive episodes. For each sample, the mean of each variable is presented with its standard error of the mean.

Appendix C

Assessing inter-individual differences with task-related functional neuroimaging

Assessing inter-individual differences with task-related functional neuroimaging

Maël Lebreton^{1,2,3,4*}, Sophie Bavard^{5,6,7}, Jean Daunizeau^{8,9} and Stefano Palminteri^{5,6,7}

Explaining and predicting individual behavioural differences induced by clinical and social factors constitutes one of the most promising applications of neuroimaging. In this Perspective, we discuss the theoretical and statistical foundations of the analyses of inter-individual differences in task-related functional neuroimaging. Leveraging a five-year literature review (July 2013–2018), we show that researchers often assess how activations elicited by a variable of interest differ between individuals. We argue that the rationale for such analyses, typically grounded in resource theory, offers an over-large analytical and interpretational flexibility that undermines their validity. We also recall how, in the established framework of the general linear model, inter-individual differences in behaviour can act as hidden moderators and spuriously induce differences in activations. We conclude with a set of recommendations and directions, which we hope will contribute to improving the statistical validity and the neurobiological interpretability of inter-individual difference analyses in task-related functional neuroimaging.

Researchers in psychology have long ago acknowledged the importance of building and testing theories that account for both the typical behaviour observed in a representative sample of the population and the observed differences between people^{1–3}. Since the mid-twentieth century, scientific psychology has benefited from important complementary insights from experimental psychology, which studies variance among treatments, and from correlational psychology, which studies variance among participants. Similarly nowadays, understanding the average typical brain and understanding the differences between individuals constitute the two complementary goals of cognitive neuroscience^{4,5}. Inter-individual differences in neural activities can be a source of statistical noise when considering the typical brain, but may also represent the very object of interest^{6–9} and can help provide an accurate and representative picture of brain function¹⁰.

Across the whole spectrum of neuroscience subfields, understanding how differences in neural activity across individuals produce differences in behavioural responses appears necessary, not only to test key predictions of neurobiological theories, but also to realize the potential of neuroimaging applications. For instance, developmental neuroscience and neuroscience of ageing rely, by nature, on the comparison of different individuals characterized by different ages or life histories¹¹. Likewise, some neurobiological concepts, like cognitive reserve, are entirely designed to explain differences in symptoms between individuals faced with the same neural pathology^{12,13}. Inter-individual differences are also important in neuroscience subfields investigating cognitive processes such as learning¹⁴ or executive control⁶, where the neural data could shed light on why some individuals perform better than others. Regarding applications, clinical diagnostics in psychiatry are expected to greatly benefit from the joint analysis of individual behaviour and brain activity, as such complementary techniques will allow doctors

to better dissociate between neurotypical and affected cases^{15–17}. The most promising socioeconomic applications of neuroimaging, such as the characterization of individual preferences and cognitive abilities, also critically depend on our ability to understand how inter-individual differences in brain functions relate to inter-individual differences in behaviour^{18–22}.

One appealing strategy to investigate how inter-individual differences in brain functions relate to inter-individual differences in behaviour involves task-related functional MRI (fMRI). Task-related fMRI is claimed to be able to target the mechanisms underpinning cognitive processes, because—unlike other biomarkers, such as genetics, neuroanatomy, or measures estimated from resting-state functional imaging^{8,23}—it allows measuring the neural activity directly elicited by the cognitive processes of interest^{16,24}. This is particularly true when fMRI is combined with computational modelling, an approach called model-based fMRI, as mechanistic measures of cognitive function are explicitly incorporated in the analysis framework in the form of a computational variables^{16,25–28}.

In the following section, we develop a concrete example of inter-individual difference analyses in task-related fMRI. This example is inspired by the human reinforcement-learning literature, as it is one of the most typical examples of model-based fMRI^{25,29,30}. We then use this example to expose and discuss important assumptions and requirements underlying the standard inter-individual brain-behaviour differences (IBBD) analytical strategy.

An IBBB analysis example from human reinforcement-learning

Reinforcement learning, i.e., learning by trial and error, is thought to be a fundamental cognitive building block and is used to achieve behavioural goals ranging from tuning motor actions to making decisions in social contexts^{31,32}. Reinforcement learning is one of the

¹Neurology and Imaging of Cognition (LabNIC), Department of Basic Neuroscience, University of Geneva, Geneva, Switzerland. ²Swiss Center for Affective Science, Campus Biotech, Geneva, Switzerland. ³Amsterdam Brain and Cognition (ABC), Universiteit van Amsterdam, Amsterdam, The Netherlands.

⁴CREED, Amsterdam School of Economics (ASE), Universiteit van Amsterdam, Amsterdam, The Netherlands. ⁵Laboratoire de Neurosciences Cognitives et Computationnelles, Institut National de la Santé et de la Recherche Médicale, Paris, France. ⁶Département d'Etudes Cognitives, Ecole Normale Supérieure, Paris, France. ⁷Human Reinforcement Learning team, Université de Recherche Paris Sciences et Lettres, Paris, France. ⁸Motivation, Brain & Behavior team, Institut du Cerveau et de la Moelle épinière (ICM), Paris, France. ⁹CNRS UMR 7225, INSERM U 1127, Pierre & Marie Curie University, Paris, France.

*e-mail: mael.lebreton@unige.ch

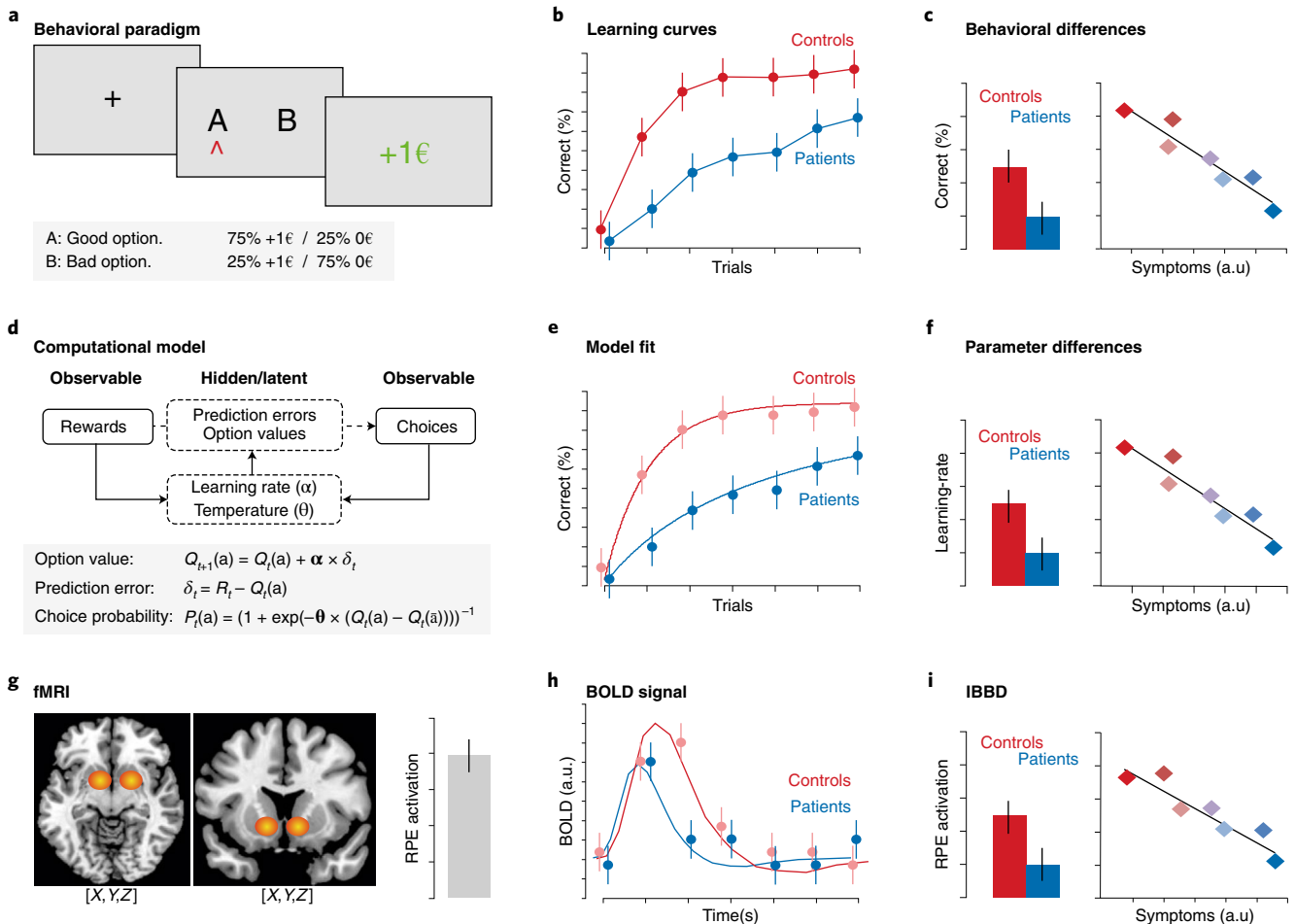


Fig. 1 | A case study: explaining inter-individual differences in learning with model-based fMRI. **a**, Behavioural paradigm. This case-study builds from a learning task, similar to the one used by Pessiglione and colleagues⁴¹. In the learning task, participants have to repeatedly choose between two symbols, probabilistically paired with monetary outcome. The goal is to learn to choose the correct option, i.e., the one that yields a higher reward probability. **b**, In clinical settings, the most common reported result is a deficit in learning of a patient sample, characterized by a slower increase in the rate of correct choice over time. **c**, This is often summarized as a lower average performance in the patient group, or as a negative correlation between symptom severity and average performance. **d**, Computational model. Models of trial-and-error learning typically rely on simple delta-rule algorithms: the values of symbols are updated in proportion to RPE (reward obtained - reward expected), weighted by a learning-rate parameter (α). The model assumes that participants compute latent variables: option values and prediction errors. **e**, The model-generated choices typically capture the difference in learning between the patient and control group. **f**, This is generally paralleled by a difference in estimated parameters (for example, learning rate) between the groups and is sometimes illustrated in a continuous way, by a correlation between symptom severity and the parameter of interest. **g**, fMRI. In reinforcement-learning, one of the most robust finding is that BOLD activity in VS correlates with RPE at the population level. **h**, **i**, To explain the learning deficits in the patient group, a common practice is to compare activations—BOLD signal (**h**) or unstandardized regression coefficients of the prediction errors (**i**)—between controls and patients or to correlate these activations with symptom severity. Those results are taken as evidence that the neurocognitive process of interest (here, the encoding of prediction error in VS) is impaired in patients suffering from the considered pathology.

rare cognitive processes for which we possess satisfactory models at the computational, algorithmic and implementational levels³³: these models account for a wide range of behavioural and neural data, transcending animal models and recording techniques³⁴. A popular reinforcement-learning paradigm is the two-armed bandit task, in which participants are repeatedly faced with pairs of abstract symbols, each probabilistically associated with a monetary outcome³⁵. Their goal is to use trial-by-trial feedback to learn the association between symbols and reward so that they can earn the most money (Fig. 1a). The participants' learning process, as measured by the progression of the frequency of correct choices (Fig. 1b), is generally satisfactorily accounted for by the Rescorla–Wagner model and its variants^{29,36,37}. These models use a simple recursive error-correcting (delta-rule) mechanism, updating the chosen stimulus' expected value with a prediction error (the received outcome minus

the expected value) weighted by a parameter called learning rate^{36,37} (Fig. 1d). Reward prediction errors (RPE) are one of the model's latent variables, i.e., a variable that is not directly observable in individual behaviour, but that is assumed by the model to explain the observable behaviour^{25–27}. One of the most robust finding in cognitive neuroscience is that blood-oxygen-level dependent (BOLD) signal in the ventral striatum (VS) correlates with reward prediction errors (Fig. 1g)^{25,38}.

While initial studies investigated the neural mechanisms of reinforcement learning in the general population^{39–41}, similar tasks have been increasingly used in clinical settings with the aim of explaining symptoms associated with some neuropsychiatric pathologies, using an IBBD approach^{34,42,43}. This model-based IBBD strategy—among others—is embraced by the emerging field of computational psychiatry. Through the combination of task-related fMRI,

computational modelling and IBB analysis, computational psychiatry holds the promise of better characterizing the neural bases of pathological behaviour, thus improving diagnostic and therapeutic tailoring^{42,44,45}. A typical study unfolds as follows: first, a behavioural difference between affected participants and neurotypical controls is revealed, evidencing the distortion of reinforcement-learning mechanisms in the pathology (Fig. 1c). Second, computational modelling is used to show that this difference is generated by a difference in learning rates between affected participants and controls (Fig. 1f). Finally, activations correlating with a learning-related variable in a specific brain region of interest (ROI) (for example, RPE activations in the VS) are shown to be significantly smaller in the affected group than in the control group (Fig. 1i). In general, this whole pattern of results is associated with two main claims: first, the presence of a significant inter-individual correlation is taken as additional statistical evidence supporting the correlation between BOLD signal in the ROI and the variable of interest. Second, the deficit in activation in the affected group is taken as a causal explanation for the behavioural deficit (for example, learning performance). Alternatively, binary classifications (affected participants vs neurotypical controls) are often replaced—or complemented—by an assessment of a continuous variable such as symptom severity or model parameter, in accordance with the dimensional approach to psychiatric disorders (Fig. 1i)⁴⁶. Importantly, despite the focus of the current Perspective on a clinical example, the same conclusions apply to any measure of inter-individual heterogeneity, ranging from task-related performance metrics to political attitudes, for example.

Inter-individual brain–behaviour analyses similar to this example are very common in clinical and non-clinical neuroscience literatures. To provide quantitative support to this claim, we performed a systematic literature review, looking for studies of human reinforcement learning using functional neuroimaging published in leading journals in the period 2013–2018 (Box 1). Crucially, we found IBB analyses in more than 70% of the 207 reviewed studies, thus confirming the typicality of these approaches. In the following paragraphs, we first review and question important theoretical and statistical assumptions underlying the study of IBB. Then, we specifically focus on two related questions: how the differences in behaviour influence the neural and imaging measures and how this can generate spurious results and interpretational problems.

The rationale behind typical IBB analyses

The rationale behind IBB analyses is rooted in resource theory⁴⁷, which has a long psychological history^{48–50}. This theory proposes that “[behavioural] performance is determined by the amount of resources invested and by their efficiency”⁴⁹.

Factors such as motivation or task demand levels have been proposed to modulate the performance–resource function, by impacting either the amount of resources allocated to the task or the efficiency of a resource unit for producing the output needed to accomplish the task⁵¹. This resource theory has been almost literally translated to functional imaging, where resource amounts to BOLD activation (or cerebral blood flow) in a brain ROI^{47,52}. From there, it is commonly assumed that individuals exhibit behavioural differences, either because the ROI is more activated or because activations in the ROI are more efficient^{6,53}. For example, assuming that the RPE is linked to BOLD activity in the VS, the way IBB results are typically interpreted, depends on the directionality of the effects. When good learners exhibit higher RPE activity in the VS, they are thought to have mobilized greater amounts of BOLD, which improves learning performances (Fig. 2a). This proportional coding narrative is the interpretation outlined in the initial example (Fig. 1). On the other hand, when good and poor learners mobilize similar amounts of BOLD activity in the VS, activations in the good learners are thought to be more efficient, leading to better error-correction for a same amount of neural resources (Fig. 2b). Note

that, in this case, this efficiency narrative practically corresponds to an inter-individual range-adaptation coding principle. Range adaptation is a pervasive assumption in the field of decision neuroscience, where it provides a simple computational explanation for the phenomenon of adaptive coding^{54–57}.

It is clear from this example that, although intuitive in its formulation, resource theory does not have strong theoretical constraints, which makes it able to accommodate almost any pattern of results ad hoc⁵¹. There is little reason, a priori, to determine whether proportional or range-adaptation coding is applicable to the experimental paradigm at hand. In addition, the translation of resource theory—originally developed in psychology—to functional imaging seems to rely on very little rationale or experimental evidence: currently, numerous studies appear to assume that BOLD levels provide a reasonable proxy for resource consumption, despite the absence of serious neurophysiological basis to this assumption⁴⁷. Overall, these theoretical weaknesses point to the risk that current and past attempts to explain inter-individual differences in behaviour with IBB analysis are contaminated by ad hoc interpretation of significant correlations, rather than reliable a priori hypothesis-testing^{58–61}.

The typical IBB analysis strategy

Most analyses of inter-individual differences in task-related fMRI typically follow the initial example, and consist in three steps (see also Supplementary Notes 1 and 2). The first step consists in estimating measures of brain activations elicited by the behavioural or computational variable of interest (thereafter simply referred to as the ‘explanatory variable’; for example, the RPE), at the individual level. The second step consists in identifying brain regions with activity significantly correlated with this explanatory variable at the population level, using random-effects analyses. Finally, IBB analyses proceed by testing statistical associations between individual ‘activations’ extracted from these ROIs and individual heterogeneity factors (for example, pathological diagnosis or learning rate).

In the context of resource theory, IBB analyses require individual measures of activations quantifying absolute levels of BOLD activation: in the initial example, one wants to estimate in every individual how much the BOLD signal increases in the VS when the RPE increases by one unit and then compare this quantity between individuals. In functional neuroimaging, numerous measures of individual brain activation elicited by a behavioural measure are available, but only a subset is sensitive to absolute levels of BOLD activations (Supplementary Note 3). This specific subset of measures, which best correspond to the resource theory underlying IBB analyses, typically derives from unstandardized first-level coefficient of regressions (hereafter referred to as ‘betas’).

General issues of IBB

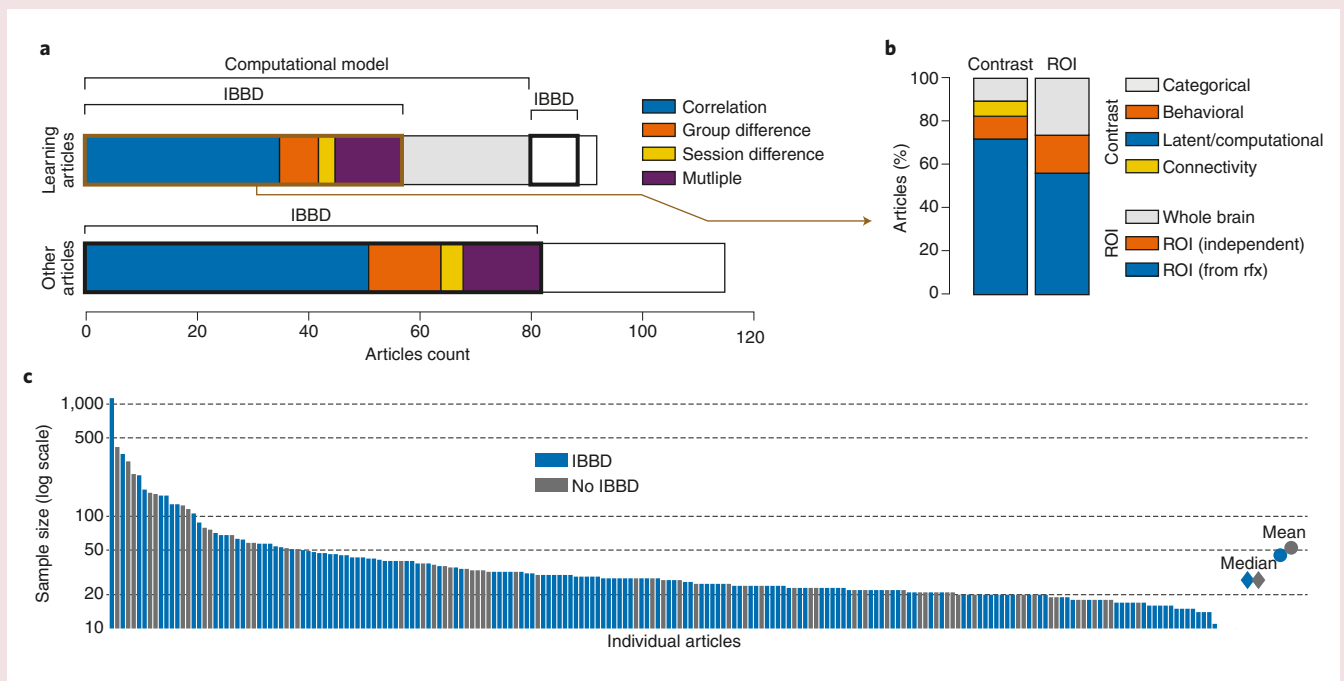
It is worth noting that, even when the standard population-based analyses are well powered and well executed, the statistical requirements for correlating heterogeneity factors and task-related fMRI betas from a given study are not necessarily met. For instance, IBB analyses have been applied to a large body of heterogeneity factors with little concern for internal consistency, i.e., for the extent to which those factors actually provide a robust estimate of individual dimensions⁶². This is an issue because standard tasks in psychology are designed to elicit robust population effects rather than robust inter-individual differences⁶³, and commonly used individual factors such as risk preference seem to lack consistency across different methods of elicitation⁶⁴.

IBB analyses should also be appropriately powered^{65,66} and based on credible and reliable effect size^{61,67}. Yet despite the abundance of studies on the intra-subject reliability of BOLD-signal estimation^{68–70}, little is known about how this translates to inter-subject reliability and effect size, as assessed by popular measures of activation. There are numerous technological and

Box 1 | Assessing IBBD practices in the human reinforcement-learning literature

To quantify the extent of the IBBD issue raised in this Perspective, we conducted a literature review of years of neuroimaging studies investigating human reinforcement-learning processes (July 2013–July 2018). We used the query terms {reinforcement-learning OR reward-learning OR value learning} AND {fMRI}, and focused on the following journals: *Nature Neuroscience*, *Neuron*, *PLoS Biology*, *PNAS*, *Nature Communications*, *eLife*, *Journal of Neuroscience*, *Brain*, *Biological Psychiatry* and *Molecular Psychiatry*. We excluded studies that used animal models and studies that did not use a task-based, event-related fMRI framework (for example, morphometry, resting state or neurofeedback studies). This resulted in the inclusion of 207 studies, which we further split into two groups, depending on whether studies actually used an instrumental-learning paradigm ($N = 92$) or did not use such a paradigm ($N = 115$; typically, other decision-making tasks somehow related to reward processing). We then evaluated whether and how those studies conducted IBBD analyses. Overall, we found that the majority of studies (71% of non-learning and in 72% of learning studies) engaged in IBBD analyses, regardless of whether they focus on instrumental learning or on other types of decision-making processes (Box Fig. a). Yet this prevalence in the reporting of IBBD results was not matched by a consensus in the implementation of the analyses. There was no consensus on the activation measure used (standardized or not beta/regression coefficient, z -score, t value,

etc.). There was also no consensus on the contrast type and the anatomical localization inference to be used in IBBD (Box Fig. b). We found four main types of contrasts: (i) categorical contrast between different event types (grey), (ii) categorical or parametric contrasts derived from individual behaviour, i.e., choices, choice correctness, ratings (orange), (iii) parametric contrasts derived from a model latent variable (blue) and (iv) contrast deriving from psychophysiological interaction analyses or other connectivity measures (yellow). While the first type may not be subject to the issue raised in this Perspective, the three others (89%) may be subject to some analytical or interpretational concerns. Regarding the anatomical localization strategy, we found ROI-based approaches as being preponderant (~70%), with only a minority of studies using independent ROIs. Another issue that arose during the literature review concerns the descriptions of the activation measures, which are often quite uninformative about what mathematical quantity they represent: most common terms simply refer to ‘betas’ or ‘[regression] coefficients’. Likewise, the processing of the behavioural or latent variable is hard to track (standardized across participants or not). Finally, the fact that there is no detectable difference in sample size between the studies including or not IBBD analyses (Box Fig. c) suggests that a large fraction of IBBD analyses may be underpowered and probably opportunistic (i.e., done in complement to planned random-effect (rfx) analyses).



Results of the literature review. a, IBBD prevalence. The horizontal stacked bars display a characterization of IBBD analyses in reinforcement-learning (top) and non-reinforcement-learning (bottom) studies. Studies were included as using IBBD if they report an inter-individual correlation between brain activity and a heterogeneity criterion (blue) or a group difference (orange). We also tagged studies that report between-session analyses (yellow), as they are subject to concerns similar to those regarding IBBD analysis, and studies reporting several between-session analyses (purple). In the pool of reinforcement-learning studies, we used the same coding scheme, and additionally report whether studies make use of computational models. **b**, IBBD practices in the human reinforcement-learning literature. For this second analysis, we focused on human reinforcement-learning studies that report IBBD and make use of computational models. We evaluated the IBBD practices among those studies with respect to the type of neuroimaging contrast used to model subject-level activation in the IBBD analysis (left) and the type of anatomical inference (right). **c**, Sample size (in log-scale). We sorted the 207 studies by sample size and indicated whether they used IBBD (blue) or not (grey). On the right end of the histogram, diamonds and dots, respectively, represent median and mean sample sizes (median_{IBBD} = 27, median_{NO-IBBD} = 27; mean_{IBBD} = 45.0, mean_{NO-IBBD} = 52.4).

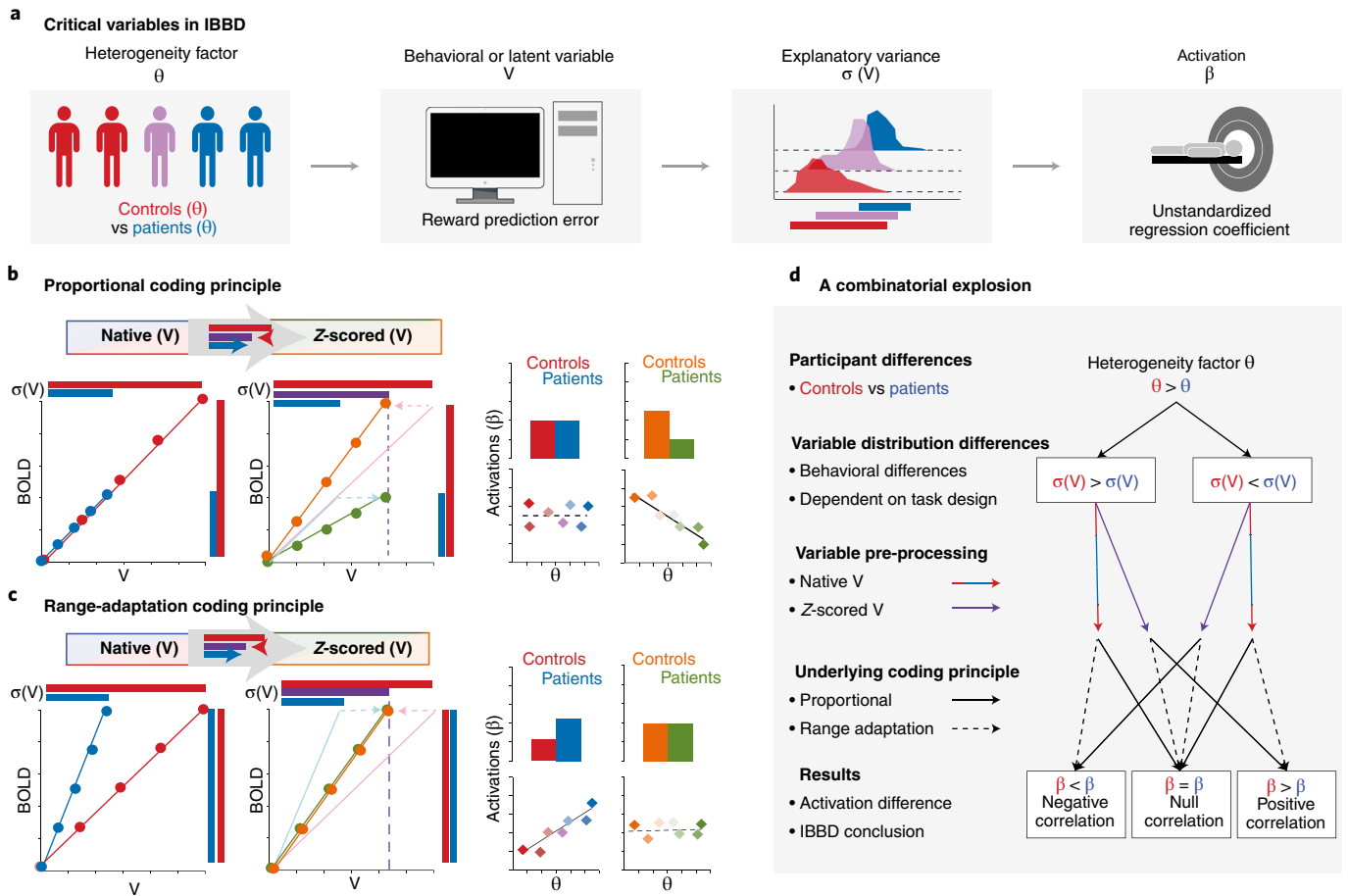


Fig. 2 | From differences in behaviour to IBBD. **a**, General IBBD analysis framework and important variables. **b,c**, Inter-individual coding principles and consequences of inter-individual differences. Left: for the two proposed coding principles, the link between the explanatory variable (a behavioural or latent variable such as RPE) and BOLD signal in two individuals (or groups of individuals) whose explanatory variances differ. The explanatory variable can be considered in its native scaling (blue or red) or after standardization (orange or green). Right: translating these links in terms of group level activations. Top: the resulting IBBD in RPE activations between the two groups as a categorical difference. Bottom: the same result in a continuous (correlational) framing. **b**, The proportional coding principle. In this case, an increase in BOLD signal is linked to an increase in the behavioural measure at both the intra- and inter-individual levels. As a consequence, the activation measure (the slope) is identical in individuals who have dissimilar explanatory variance (left). Standardizing (i.e., z-scoring) the behavioural measure (right) induces a difference in the estimated activations, creating apparent inter-individual differences (right). **c**, The range-adaptation coding principle. In this case, an increase in BOLD signal is linked to an increase in the behavioural measure at the intra-individual level, but all individuals exhibit the same range of BOLD activation (the slope). As a consequence, individuals with a smaller explanatory variance (blue) exhibit higher activations (left). Standardizing (i.e., z-scoring) the behavioural measure (right) erases this difference. **d**, A combinatorial explosion. This graphic depicts how different steps of an IBBD analyses (experimental choices such as task design (see also Supplementary Note 4); analytical choices such as variable pre-processing; physiological constraints such as underlying coding principles) combine to ultimately lead to different IBBD results and conclusions.

neurophysiological factors that could undermine our ability to accurately assess individual differences in absolute levels of BOLD activations—see notably Table 1 in ref. 4, which lists important sources of variance in functional-anatomic imaging, and see ref. 5 for a review. Studies even report that some fMRI activations might be artefactual⁷¹ and that sources of within-subject variability versus between-subject variability might be distinct^{72,73}. Despite these indications that IBBD analyses might have lower signal-to-noise ratio than classical random-effect analyses, it seems that most IBBD analyses are typically conducted as an opportunistic complement to within-subject analyses and hence leverage relatively small sample sizes⁵³. Confirming this interpretation, and contrary to the recommendation that inter-individual differences studies should be supported with higher statistical power⁶⁵, our literature review shows no difference in sample size between studies that include IBBD and studies that do not (Box 1).

From differences in behaviour to IBBD

Having outlined and questioned the main assumptions behind IBBD analysis, we now turn to the central issue of this perspective: a commonly overlooked property of the individual activation measures (unstandardized betas) is that they are inversely proportional to the individual variance of the explanatory variable (hereafter simply referred to as ‘explanatory variance’). Critically, under the two neurobiologically plausible coding principles inspired from resource theory, this property generates statistical dependencies between the activation measure (unstandardized betas) and this explanatory variance. The directions of these dependencies depend on how the explanatory variable was pre-processed in the neuroimaging analysis pipeline. The two current options are either to use the native variable or to proceed with a within-subject standardization (z-scoring) of this variable such that the z-scored explanatory variable has a mean of zero and a standard deviation of one for all participants.

Under the proportional coding principle, it can be easily shown (Supplementary Note 2) that, while individual activations (betas) are uncorrelated with the explanatory variance when activations are estimated with the native explanatory variable, they are positively correlated with it when activations are estimated with the *z*-scored explanatory variable (Fig. 2b). Under the range-adaptation coding principle, however, activations are negatively correlated with the explanatory variance when activations are estimated with the native explanatory variable and independent from it when activations are estimated with the *z*-scored explanatory variable (Fig. 2c). Of course, one never directly uses the explanatory variance to devise groups of individuals or as an explanatory variable in inter-individual correlations with activations, because this variance rarely represents the trait or the behavioural pattern of interest. However, individual differences in variance are very often a by-product of other behavioural differences: for instance, in the learning example, an initial difference in performance naturally translates to a difference in learning rates, which typically generate differences in mean and variance of RPE (Supplementary Note 5).

Some counter-intuitive aspects of IBBD analyses are worth highlighting: using native explanatory variable can lead to false negative interpretations in the proportional case and to false positive interpretations in the range-adaptation case. Returning to the learning example, when higher levels of BOLD signals in the VS of some participants are actually responsible for higher learning rate (proportional coding), IBBD analyses with native explanatory variables would come out non-significant. Reciprocally, when similar levels of BOLD signals in the VS of all participants are associated with different learning performances (for example, because of range-adaptation coding or because individual differences in performance are caused by differential activations in another brain region), IBBD analyses would result in higher activation in the slow-learners group.

Interpretational issues in IBBD

Overall, we believe that the systematically overlooked dependences between activation measures and explanatory variance have important consequences on IBBD analyses and their interpretations. Specifically, IBBD correlations may not constitute additional independent statistical evidence for the implication of the ROI in the generation of the behaviour; they can simply derive from individual differences in the variance of the explanatory variable used to estimate brain activations. When (i) an ROI is shown (or known) to correlate with the explanatory variable at the population level and (ii) some inter-individual differences in the explained variance correlate with the heterogeneity factor of interest, significant IBBD results should be interpreted with caution. This is because they may in fact be artefactual consequences of one's (otherwise valid) methodological approach to testing the significance of population averages. In other words, standard group-level and IBBD results may be two sides of the same coin, rephrasing the same piece of evidence twice.

In addition, IBBD analyses may not assess individual differences associated with average performance (a proxy of efficiency or motivation) as straightforwardly as it is frequently assumed⁴⁷. Rather, significant IBBD results might merely reflect individual differences in explanatory variance. Therefore, testing hypotheses concerning IBBD and interpreting the consequent results should account for how individual performance (efficiency or motivation, as the case may be) correlates with this explanatory variance. This can be influenced by many factors, including the task difficulty and structure, as well as modelling options (Fig. 2d and Supplementary Note 4). Taking these dependences into account could help to understand puzzling results in model-based fMRI (Box 2).

Finally, it seems that currently it is extremely difficult to derive precise IBBD predictions. Indeed, the statistical dependencies between the individual behavioural variance and the individual activations depend on the underlying neurophysiological coding

principles linking the BOLD signal and the variable (namely, proportional vs range-adaptation), which are largely undocumented. As a consequence, almost any significant statistical pattern of inter-individual seems to be explainable ad hoc under certain assumptions (Fig. 2d)—a criticism that was also raised toward resource theory in general⁵¹. In our view, this largely impairs current efforts to derive robust and replicable inter-individual findings in task-related neuroimaging.

A generalization of IBBD issues

Although the 'case study' used to illustrate the theoretical and statistical issues at stake might seem overly specific (reinforcement learning, model-based fMRI, explanatory variables derived from individual behaviour), we believe that the issues raised have more general and broader implications. Most importantly, the lack of a clear specification of how the resource theory applies to fMRI is not restricted to model-based and parametric designs, but actually generalizes to almost all designs, including the class of simpler, categorical designs (i.e., where activations are estimated from experimental conditions, not from behaviourally derived variables). The excessive theoretical flexibility underlying IBBD analyses, which opens the door to ad hoc interpretations of (potentially spurious) correlations, should raise concerns about the validity of statistical claims about IBBD in a wide range of experimental designs allowed by fMRI^{58–61}.

The issues arising from comparing activations in the presence of behavioural differences are also not restricted to the investigation of between-subject differences: they naturally extend to within-subject, between-sessions designs—for example, when behaviour and BOLD activities are recorded in the same individuals but in different sessions. If the explanatory variance is susceptible to being modulated between different sessions, assessing inter-session differences of brain activity is subject to all the aforementioned issues. This cautionary message applies, for example, to typical experimental designs investigating the effects of a pharmacological manipulation⁷⁴, a stimulation protocol (for example, transcranial magnetic stimulation^{75,76}), or general 'contextual' effect on a behaviour-related activation.

Recommendations and avenues for future research

In the present Perspective, we raise awareness about possible pitfalls of the analyses routinely performed to assess how individual differences in brain functions correlate with differences in behaviour (IBBD) and their interpretations. Some of those concerns (for example, regarding statistical power or internal validity) are not specific to functional imaging^{65,66} and might be addressed by the ongoing cultural changes in the field, such as the rise of transparent and reproducible neuroimaging research practices^{61,77} or the collection of larger datasets including task-related fMRI paradigms^{78–80}. Amidst those general concerns, we outlined problems specific to IBBD analyses.

We feel it is mandatory to re-evaluate IBBD theoretical and analytical underpinnings and their potential confounds. A first important area of focus is the statistical impact of individual differences in the behavioural explanatory variables, which can be sources of non-independence issues and spurious results. The presence of this potential confound in previous reports should raise caution about the interpretation of both the directionality and the statistical significance of published IBBD results. As an immediate step to alleviate or to assess the impact of these potential issues in future studies, we recommend systematically documenting dependencies between heterogeneity factors and explanatory variance (i.e., the variance of the behavioural or model-derived variables used to estimate activations) and specifying which measure of activations are used in IBBD analyses (Supplementary Notes 1–3) in order for results to be evaluable, interpretable and reproducible. As illustrated in this perspective, these data could quite straightforwardly help to make sense of seemingly highly contradictory IBBD findings. Before even engaging

Box 2 | Explaining puzzling results of model-based fMRI

In this Box, we illustrate how the IBBD issues outlined in this Perspective may explain self-contradictory practices in model-based fMRI. Model-based fMRI typically uses as dependent variables latent variables that are inferred from observable behaviour, as follows: a computational model is fitted in order to obtain the free-parameters' values that maximize the likelihood of observing the data given the model²⁵. Notably, the free-parameters can be either considered as fixed effects (i.e., shared across individuals) or random effects (i.e., each subject's parameters are drawn from a common population distribution)²⁹.

Counterintuitively, while treating model-free parameters as random effects often provides the best account of individuals' behaviour as assessed by rigorous model-comparisons, a common practice is to treat them as fixed effects—i.e., using the population-level parameters—to generate the latent variables to be fed into the fMRI analysis^{39,40,96–101}. This is often justified by arguing that parameter estimates at the individual level are 'noisy' and estimating them from collapsing all participants is an efficient way to regularize them. However, if individual parameters still provide a better account of the population behavioural data according to rigorous, complexity penalizing, model-comparison procedures, then the variance modelled in the individual parameters actually captures a true inter-individual variability. Therefore, using population-level parameters does not seem justified. As a matter of fact, the advantage of using the population-level parameters could be explained in the light of the IBBD issues highlighted in this Perspective. As depicted in Supplementary Note 5, the value of the learning rate α affects the variance of the latent variables (option

values and prediction errors). Accordingly, using population-level parameters constrains the individual explanatory variance to a similar value, provided that individuals are given a similar input. Under the range-adaptation coding principle, this ensures that individual activations take similar values, hence substantially increase the statistical power of second-level random effects analyses. However, under the range-adaptation coding principle, a more appropriate way to model brain activation would involve using individual model parameters and z-scoring latent variables.

Note that using population-level parameters for the neuroimaging analysis can also spuriously create IBBD patterns, notably under the range-adaptation coding principle: using population-level parameters de facto underestimates the variability of the latent variable in some individuals (those whose individual parameters would have generated a larger explanatory variance). In those individuals, the (population) latent variable (explanatory variable) has to be magnified in order to match the individual's BOLD time-series by inflating estimated activations. The same reasoning can show that using population-level parameters deflates estimated activations in other individuals (those whose individual-parameters would have generated a smaller explanatory variance). In the end, these statistical associations provide a basis for monotonical dependencies between individual parameters values and activations, creating spurious inter-individual brain-behaviour correlation.

This Box further illustrates that common practices in model-based fMRI would benefit from a better understanding of the neurobiological and statistical bases of inter-individual differences.

in IBBD analyses, several steps could be taken, in theory, to evaluate whether the inter-individual difference observed in traits or behaviour and the variance captured in model-based fMRI activations can be related in a meaningful way. For instance, one would expect that subject-specific (latent) variables would better account for the BOLD signal in an ROI than the same variable estimated from another individual. These sanity checks might, however, be hampered by the low signal-to-noise ratio of fMRI, which limits its ability to capture these subtle differences⁸¹.

Our Perspective also highlights the urgent need for statistical tools to clarify the underlying coding principle, to constrain analyses and interpretations and to reduce unnecessary degrees of freedom in analytical pipelines—see, for example, recent developments leveraging analytic tools from psychometric theory⁸². We speculate that a promising avenue for improving IBBD assessment is to depart from the simple reliance on statistical comparisons between individual parameters and/or activations and to turn to more comprehensive neurocomputational approaches, paired with model comparison. Such model-based approaches could notably be tailored to address two specific issues of current IBBD approaches. First, neurocomputational models could, in theory, explicitly incorporate constraints imposed by inter-individual coding principles (i.e., some variant of range adaptation and/or proportional coding). These models could be jointly fitted to behaviour and fMRI concurrently and then compared in their ability to account for the data. Second, current IBBD analyses assume that all participants use the same strategy, implemented as a single computational model, so that inter-individual differences in behaviour are reasonably captured by inter-individual differences in model parameters. If different participants use different strategies, implemented as different computational models, this may confound inter-individual variations in the relationship between brain activity and the computational variables estimated with the single model. Fitting and comparing different

computational models (potentially, jointly to both behavioural and fMRI data) in different participants could, again, be an efficient way of capturing the true essence of inter-individual differences⁴⁵. Eventually, dual fitting approaches could yield new interpretational issues. For example, BOLD signal may be very well explained, but not the behaviour. An important prerequisite would be to ensure that the dual fitting approach actually captures a 'reasonable' amount of inter-individual variance in both BOLD and behavioural data.

A different, recent and increasingly common strategy to investigate inter-individual differences in cognitive neuroscience leverages data-driven approaches, such as unsupervised classification tools, to identify subtypes of individuals⁴⁴. Although this approach has a lot of potential and promises, it is still limited by the quality of the features used by the classifier: if one wishes to classify individuals based on computational model parameters, or on model-based fMRI activations, most of the issues raised in this Perspective would still limit the interpretability of the results.

In parallel with the improvement of IBBD statistical tools, an important area of focus is the development of better theories of IBBD. The dominant, naive translation of resource theory to BOLD signal should be superseded, as it seems to lack solid neurophysiological support⁴⁷. Better, comprehensive IBBD theories should probably depart from a static structure–function mapping and treat brain regions as information-processing nodes, embedded in functional networks and characterized by specific inputs, outputs and canonical computations^{83–85}. To feed these theoretical developments, more basic research regarding the biophysical models and neurophysiological bases underlying inter-individual differences in neuroimaging is needed. Joint fMRI and neural recordings in animals have been an outstanding source of information about the neurophysiological basis of the BOLD signal^{86,87}, yet have rarely addressed inter-individual questions so far. In humans, two diametrically opposed and complementary approaches could

eventually contribute to improve our understanding and modelling of the sources of inter-individual variance: the investigation of highly sampled fMRI datasets, designed to improve the anatomical and functional characterization of individual participants^{72,73,88,89}; and the exploration of large, longitudinal, population-based neuroimaging datasets, designed to evaluate inter-individual variability in brain structure and function across individuals, environments and developmental stages^{79,80,90,91}.

Although these developments appear necessary to unlock the potential of task-related fMRI to explain and understand inter-individual differences in behaviour, other relative agnostic uses of the inter-individual variance in fMRI might still be able to deliver outcomes of important societal value. For instance, combined with multimodal imaging^{67,92,93} and genetics^{94,95}, task-related fMRI can be an important component of neuromarkers^{8,9}, irrespective of its ability to truly—i.e., mechanistically—explain what drives inter-individual differences in traits or behaviour.

Received: 20 February 2018; Accepted: 9 July 2019;

Published online: 26 August 2019

References

- Cronbach, L. J. *Am. Psychol.* **12**, 671–684 (1957).
- Underwood, B. J. *Am. Psychol.* **30**, 128–134 (1975).
- Vogel, E. K. & Awh, E. *Curr. Dir. Psychol. Sci.* **17**, 171–176 (2008).
- Van Horn, J. D., Grafton, S. T. & Miller, M. B. *Brain Imaging Behav.* **2**, 327–334 (2008).
- Dubois, J. & Adolphs, R. *Trends Cogn. Sci.* **20**, 425–443 (2016).
- Braver, T. S., Cole, M. W. & Yarkoni, T. *Curr. Opin. Neurobiol.* **20**, 242–250 (2010).
- McGonigle, D. J. *Neuroimage* **62**, 1116–1120 (2012).
- Gabrieli, J. D. E., Ghosh, S. S. & Whitfield-Gabrieli, S. *Neuron* **85**, 11–26 (2015).
- Seghier, M. L. & Price, C. J. *Trends Cogn. Sci.* **22**, 517–530 (2018).
- Falk, E. B. et al. *Proc. Natl. Acad. Sci. USA* **110**, 17615–17622 (2013).
- Luna, B., Padmanabhan, A. & O’Hearn, K. *Brain Cogn.* **72**, 101–113 (2010).
- Barulli, D. & Stern, Y. *Trends Cogn. Sci.* **17**, 502–509 (2013).
- Gregory, S., Long, J. D., Tabrizi, S. J. & Rees, G. *Curr. Opin. Neurol.* **30**, 380–387 (2017).
- Commins, S. *Rev. Neurosci.* **29**, 183–197 (2017).
- Matthews, P. M., Honey, G. D. & Bullmore, E. T. *Nat. Rev. Neurosci.* **7**, 732–744 (2006).
- Kishida, K. T., King-Casas, B. & Montague, P. R. *Neuron* **67**, 543–554 (2010).
- Dagher, A. *Neuroimage* **151**, 128–129 (2017).
- Camerer, C. F. *Econ. J.* **117**, C26–C42 (2007).
- Fehr, E. & Camerer, C. F. *Trends Cogn. Sci.* **11**, 419–427 (2007).
- Rustichini, A. *Curr. Opin. Neurobiol.* **19**, 672–677 (2009).
- Kable, J. W. & Levy, I. *Curr. Opin. Behav. Sci.* **5**, 100–107 (2015).
- Katsnelson, A. *Proc. Natl. Acad. Sci. USA* **112**, 15530–15532 (2015).
- Kanai, R. & Rees, G. *Nat. Rev. Neurosci.* **12**, 231–242 (2011).
- Wang, X.-J. & Krystal, J. H. *Neuron* **84**, 638–654 (2014).
- O’Doherty, J. P., Hampton, A. & Kim, H. *Ann. NY Acad. Sci.* **1104**, 35–53 (2007).
- Gläscher, J. P. & O’Doherty, J. P. *Wiley Interdiscip. Rev. Cogn. Sci.* **1**, 501–510 (2010).
- Cohen, J. D. et al. *Nat. Neurosci.* **20**, 304–313 (2017).
- Patzelt, E. H., Hartley, C. A. & Gershman, S. J. *Personal. Neurosci.* **1**, e18 (2018).
- Daw, N.D. in *Decision Making, Affect, and Learning: Attention and Performance XXIII* (eds. Delgado, M.R., Phelps, E.A. & Robbins, T.W.) Chapter 1 (2011).
- Corrado, G. & Doya, K. *J. Neurosci.* **27**, 8178–8180 (2007).
- Chen, X., Holland, P. & Galea, J. M. *Curr. Opin. Behav. Sci.* **20**, 83–88 (2018).
- Joiner, J., Piva, M., Turrin, C. & Chang, S. W. C. *NPJ Sci. Learn.* **2**, 8 (2017).
- Dayan, P. & Daw, N. D. *Cogn. Affect. Behav. Neurosci.* **8**, 429–453 (2008).
- Maia, T. V. & Frank, M. J. *Nat. Neurosci.* **14**, 154–162 (2011).
- Palmiteri, S. & Pessiglione, M. in *Decision Neuroscience* (eds. Dreher, J.-C. & Tremblay, L.) 291–303 (Academic Press, 2017).
- Rescorla, R.A. & Wagner, A.R. in *Classical Conditioning II: Current Research and Theory* (eds. Black, A.H. & Prokasy, W.F.) 64–99 (Appleton-Century-Crofts, 1972).
- Sutton, R.S. & Barto, A.G. *Reinforcement Learning: An Introduction*. (Cambridge University Press, 1998).
- Garrison, J., Erdeniz, B. & Done, J. *Neurosci. Biobehav. Rev.* **37**, 1297–1310 (2013).
- O’Doherty, J. et al. *Science* **304**, 452–454 (2004).
- Daw, N. D., O’Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. *Nature* **441**, 876–879 (2006).
- Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J. & Frith, C. D. *Nature* **442**, 1042–1045 (2006).
- Montague, P. R., Dolan, R. J., Friston, K. J. & Dayan, P. *Trends Cogn. Sci.* **16**, 72–80 (2012).
- Robinson, O. J. & Chase, H. W. *Comput. Psychiatr.* **1**, 208–233 (2017).
- Huys, Q. J. M., Maia, T. V. & Frank, M. J. *Nat. Neurosci.* **19**, 404–413 (2016).
- Stephan, K. E. et al. *Neuroimage* **145 Pt B**, 180–199 (2017).
- Harvey, A., Watkins, E., Mansell, W. & Shafraan, R. *Cognitive Behavioural Processes Across Psychological Disorders: A Transdiagnostic Approach to Research and Treatment*. (Oxford University Press, 2004).
- Poldrack, R. A. *Dev. Cogn. Neurosci.* **11**, 12–17 (2015).
- Norman, D. A. & Bobrow, D. G. *Cogn. Psychol.* **7**, 44–64 (1975).
- Navon, D. & Gopher, D. *Psychol. Rev.* **86**, 214–255 (1979).
- Humphreys, M. S. & Revelle, W. *Psychol. Rev.* **91**, 153–184 (1984).
- Navon, D. *Psychol. Rev.* **91**, 216–234 (1984).
- Matthews, G., Warm, J.S., Reinerman, L.E., Langheim, L.K. & Saxby, D.J. in *Handbook of Individual Differences in Cognition: Attention, Memory, and Executive Control* (eds. Gruszka, A., Matthews, G. & Szymura, B.) 205–230 (Springer New York, 2010).
- Yarkoni, T. & Braver, T.S. in *Handbook of Individual Differences in Cognition* (eds. Gruszka, A., Matthews, G. & Szymura, B.) 87–107 (Springer New York, 2010).
- Cox, K. M. & Kable, J. W. *J. Neurosci.* **34**, 16533–16543 (2014).
- Louie, K. & Glimcher, P. W. *Ann. NY Acad. Sci.* **1251**, 13–32 (2012).
- Padoa-Schioppa, C. J. *J. Neurosci.* **29**, 14004–14014 (2009).
- Rangel, A. & Clithero, J. A. *Curr. Opin. Neurobiol.* **22**, 970–981 (2012).
- Kerr, N. L. *Soc. Psychol.* **2**, 196–217 (1998).
- Nuzzo, R. *Nature* **526**, 182–185 (2015).
- Munafò, M. R. et al. *Nat. Hum. Behav.* <https://doi.org/10.1038/s41562-016-0021-2017>.
- Poldrack, R. A. et al. *Nat. Rev. Neurosci.* **18**, 115–126 (2017).
- Hajcak, G., Meyer, A. & Kotov, R. J. *Abnorm. Psychol.* **126**, 823–834 (2017).
- Hedge, C., Powell, G. & Sumner, P. *Behav. Res. Methods* <https://doi.org/10.3758/s13428-017-0935-1> (2018).
- Pedroni, A. et al. *Nat. Hum. Behav.* **1**, 803–809 (2017).
- Button, K. S. et al. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
- Szucs, D. & Ioannidis, J. P. A. *PLoS Biol.* **15**, e2000797 (2017).
- Abi-Dargham, A. & Horga, G. *Nat. Med.* **22**, nm.4190 (2016).
- Caceres, A., Hall, D. L., Zelaya, F. O., Williams, S. C. R. & Mehta, M. A. *Neuroimage* **45**, 758–768 (2009).
- Plichta, M. M. et al. *Neuroimage* **60**, 1746–1758 (2012).
- Nord, C. L., Gray, A., Charpentier, C. J., Robinson, O. J. & Roiser, J. P. *Neuroimage* **156**, 119–127 (2017).
- Renvall, V., Nangini, C. & Hari, R. *Sci. Rep.* **4**, 3920 (2014).
- Mueller, S. et al. *Neuron* **77**, 586–595 (2013).
- Laumann, T. O. et al. *Neuron* **87**, 657–670 (2015).
- Honey, G. & Bullmore, E. *Trends Pharmacol. Sci.* **25**, 366–374 (2004).
- Bestmann, S. & Feredoes, E. *Ann. NY Acad. Sci.* **1296**, 11–30 (2013).
- Polania, R., Nitsche, M. A. & Ruff, C. C. *Nat. Neurosci.* **21**, 174–187 (2018).
- Poldrack, R. A. & Gorgolewski, K. J. *Nat. Neurosci.* **17**, 1510–1517 (2014).
- Barch, D. M. et al. *Neuroimage* **80**, 169–189 (2013).
- Miller, K. L. et al. *Nat. Neurosci.* **19**, 1523–1536 (2016).
- Van Essen, D. C. et al. *Neuroimage* **80**, 62–79 (2013).
- Wilson, R. C. & Niv, Y. *PLOS Comput. Biol.* **11**, e1004237 (2015).
- Cooper, S. R., Jackson, J. J., Barch, D. M. & Braver, T. S. *Neurosci. Biobehav. Rev.* **98**, 29–46 (2019).
- Friston, K. *Annu. Rev. Neurosci.* **25**, 221–250 (2002).
- Hunt, L. T. & Hayden, B. Y. *Nat. Rev. Neurosci.* **18**, 172–182 (2017).
- Silver, R. A. *Nat. Rev. Neurosci.* **11**, 474–489 (2010).
- Heeger, D. J. & Ress, D. *Nat. Rev. Neurosci.* **3**, 142–151 (2002).
- Logothetis, N. K. *Nature* **453**, 869–878 (2008).
- Gordon, E. M. et al. *Neuron* **95**, 791–807.e7 (2017).
- Turk-Browne, N. B. *Science* **342**, 580–584 (2013).
- Bearden, C. E. & Thompson, P. M. *Neuron* **94**, 232–236 (2017).
- Smith, S. M. & Nichols, T. E. *Neuron* **97**, 263–268 (2018).
- Insel, T. R. & Cuthbert, B. N. *Science* **348**, 499–500 (2015).
- Woo, C.-W., Chang, L. J., Lindquist, M. A. & Wager, T. D. *Nat. Neurosci.* **20**, 365–377 (2017).
- Hariri, A. R. *Annu. Rev. Neurosci.* **32**, 225–247 (2009).
- Congdon, E., Poldrack, R. A. & Freimer, N. B. *Neuron* **68**, 218–230 (2010).
- Gershman, S. J., Pesaran, B. & Daw, N. D. *J. Neurosci.* **29**, 13524–13531 (2009).

97. Gläscher, J., Hampton, A. N. & O'Doherty, J. P. *Cereb. Cortex* **19**, 483–495 (2009).
98. Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J. P. *Neuron* **66**, 585–595 (2010).
99. Palminteri, S., Boraud, T., Lafargue, G., Dubois, B. & Pessiglione, M. *J. Neurosci.* **29**, 13465–13472 (2009).
100. Palminteri, S., Khamassi, M., Joffily, M. & Coricelli, G. *Nat. Commun.* **6**, 8096 (2015).
101. Pessiglione, M. et al. *Neuron* **59**, 561–567 (2008).

Acknowledgements

During the preparation of this work, M.L. was supported by a NWO Veni (Grant 451-15-015) and a Swiss National Found Ambizione grant (PZ00P3_174127). M.L. also acknowledges the support of the Bettencourt-Schueller Foundation. S.P. is supported by an ATIP-Avenir grant (R16069JS), the Programme Emergence(s) de la Ville de Paris, the Fyssen foundation, and the Fondation Schlumberger pour l'Education et la Recherche. The Institut d'Etude de la Cognition is supported

financially by the LabEx IEC (ANR-10-LABX-0087 IEC) and the IDEX PSL* (ANR-10-IDEX-0001-02 PSL*).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41562-019-0681-8>.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence should be addressed to M.L.

Peer review information: Primary Handling Editor: Mary Elizabeth Sutherland

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

© Springer Nature Limited 2019

Appendix D

[Re] Adaptive properties of differential learning rates for positive and negative outcomes

[Re] Adaptive properties of differential learning rates for positive and negative outcomes

Sophie Bavard¹ and Héloïse Théro¹

¹ Laboratoire de Neurosciences Cognitives Computationnelles (ENS - INSERM), Département d'Études Cognitives, École Normale Supérieure, PSL Research University, 29 rue d'Ulm, 75005 Paris, France

sophie.bavard@gmail.com, thero.heloise@gmail.com

Editor

Olivia Guest

Reviewers

Xavier Hinaut
Benoît Girard

Received Feb, 20, 2018

Accepted Jun, 14, 2018

Published Jun, 14, 2018

Licence [CC-BY](#)

Competing Interests:

The authors have declared that no competing interests exist.

 [Article repository](#)

 [Code repository](#)

A reference implementation of

→ Cazé, R. D., & van der Meer, M. A. (2013). Adaptive properties of differential learning rates for positive and negative outcomes. *Biological cybernetics*, 107(6), 711-719. <https://doi.org/10.1007/s00422-013-0571-5>

Introduction

Reinforcement learning represents a fundamental cognitive process: learning by trial and error to maximize rewards and minimize punishments. Current and most influential theoretical models of reinforcement learning assume a unique learning rate parameter, independently of the outcome valence (Sutton and Barto [14], O'Doherty et al. [10], Behrens et al. [1]). However human participants were shown to integrate differently positive and negative outcomes (Frank, Seeberger, and O'Reilly [3], Frank et al. [4], Sharot, Korn, and Dolan [13]). This motivated the reference article to implement a modified version of the reinforcement learning model, with two distinct learning rates for positive and negative outcomes (Cazé and Meer [2]).

They have shown that although differential learning rates shifted reward predictions and could thus be seen as a maladaptive bias, this model can outperform the classical reinforcement learning model on tasks with specific outcome probabilities. Following Cazé and Meer [2]'s predictions, a subsequent empirical article have modeled human behavior on these specific tasks (Gershman [7]). The question is still an active research area, as various articles have further investigated the difference learning rates bias (Garrett and Sharot [5], Moutsiana et al. [9], Shah et al. [12], Garrett and Sharot [6], Lefebvre et al. [8], Palminteri et al. [11]).

A link to the pdf version of the reference article was posted on the last author's laboratory website (<http://www.vandermeerlab.org/publications.html>), but the corresponding code was not available (https://github.com/vandermeerlab/papers/tree/master/Caze_vanderMeer_2013). We believe that an openly available code repository replicating the results of Cazé and Meer [2]'s paper can be helpful to the scientific community. We therefore implemented the model and analysis scripts using Python, with numpy, random and matplotlib libraries.

Methods

We first implemented our scripts on Matlab, as we were more familiar with this language, and then adapted them on Python.

We used the modeling description of the reference article to implement our replication. They used standard Q-learners with a softmax action selection rule (Sutton and Barto [14]), and their precise description enabled us to implement them with low difficulty. But we found four ambiguities in the simulation procedure.

First, the authors described their analytical results to be valid for “ $Q_0 \neq \{-1, 1\}$ ” in section 2, but did not specify what value of Q_0 they used in all the following simulations. We chose to use $Q_0 = 0$, as this initial value is the middle point between the two possible outcomes (i.e., -1 and 1). As we replicated all the original figures, even the dynamics in the beginning of the learning curves (see Figures 2 A, 3 and 4 B), we believe the reference article must have used similar initial Q-values.

Second, regarding the parameter setting for Figure 1’s simulations, the ratio of α^+ over α^- was said to be either 0.25, 1 or 4, but they did not specify what were the exact values of α^+ and α^- used. We thus set them according to the following description of the pessimistic, rational and optimistic agents in section 3, i.e.,:

- $\alpha^+ = 0.1$ and $\alpha^- = 0.4$ for the ratio of 0.25
- $\alpha^+ = 0.1$ and $\alpha^- = 0.1$ for the ratio of 1
- $\alpha^+ = 0.4$ and $\alpha^- = 0.1$ for the ratio of 4

Third, the number of iterations made to generate Figures 3 and 4 were not indicated, and we assumed the authors used the same number as in Figures 1 and 2 (i.e., 5,000 runs).

Finally, in the reinforcement learning framework, the probabilities to choose each action are computed, then used to select an action through a pseudo-random generator. In the reference article, it was sometimes unclear whether the analyses were performed on the probabilities of choice, or rather the proportions of implemented choices. For example Figure 2’s legend indicated: “Mean probability of choosing the best arm”, suggesting that the probabilities themselves were used. However, when commenting the figure in section 3, the authors appeared to say that the actual choices were rather used: “the optimistic agent learns to take the best action significantly more than the rational agent”. For our analyses, we started by using the probabilities of choice, as this would lead to more clear, less noise-corrupted results. However we then obtained very smooth learning curves, and were unable to reproduce the spikiness of the original Figures 2, 3 and 4. We thus computed the proportions of implemented choices for all our figures.

Results

We numbered our figures in the same way as the reference article.

All our figures reproduced the patterns of the original results. We were even able to replicate the fine-grained details of the learning curves, like the early bumps in performance in the high-reward task (Figures 2 A, 3 and 4 B, right panels, around 50-100 trials). In Figure 1, the mean and the variance of the Q-values were also very similar as the ones in the original figure.

The only discrepancy we found was in Figure 4 A. Although the general pattern was replicated, our learning curves appeared smoother than in the reference article. As the number of simulations were not explicitly specified for this figure, we cannot know if this is due to us running a higher number of simulations than the reference article, or from another difference in model implementation.

Conclusion

All the figures in Cazé and Meer [2] have been successfully reproduced with high fidelity, and we confirm the validity of their simulations. Overall the whole replication

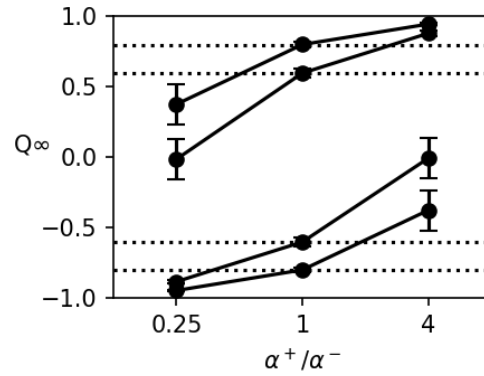


Figure 1: Average estimated Q-values after 800 trials averaged for different ratios of α^+ and α^- . The dotted lines represent the underlying average reward: 0.8, 0.6, -0.6, -0.8. The error bars represent the variance of the estimated Q-values.

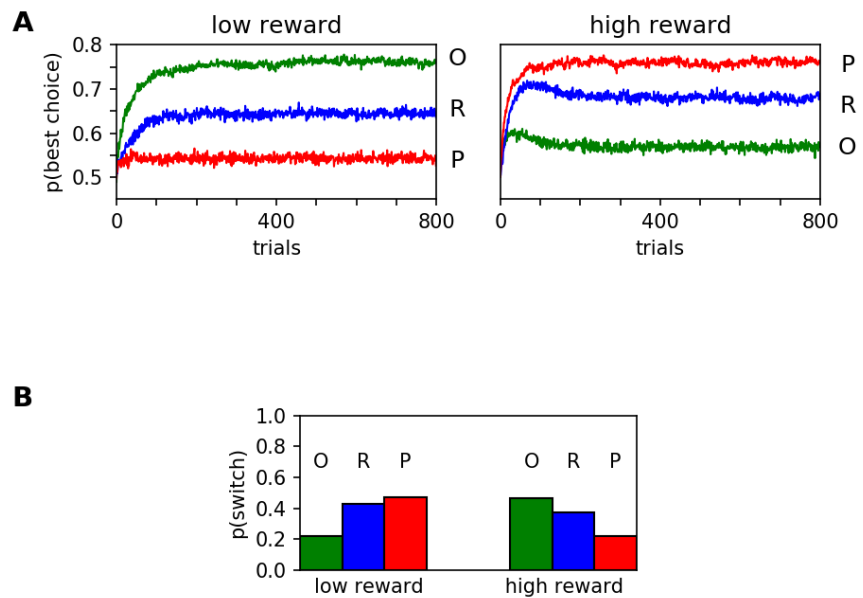


Figure 2: A. Performance, i.e. proportion of choices for the best action, for the three agents: Rational (R, $\alpha^+ = \alpha^-$, blue line), Optimistic (O, $\alpha^+ > \alpha^-$, green line) and Pessimistic (P, $\alpha^+ < \alpha^-$, red line). In this figure and the following ones, the left (resp. right) panel corresponds to the low-reward (resp. high-reward) task. **B.** Proportion of action switch after 800 trials for each agent, in the two different tasks.

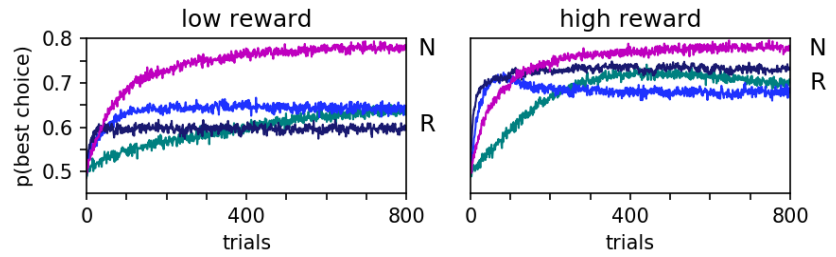


Figure 3: The performances of the Meta-learner (N) are shown in *purple* and those of the Rational agents (R) in different colors of blue (in *teal* for $\alpha = 0.01$, in *royal blue* for $\alpha = 0.1$ and in *navy blue* for $\alpha = 0.4$).

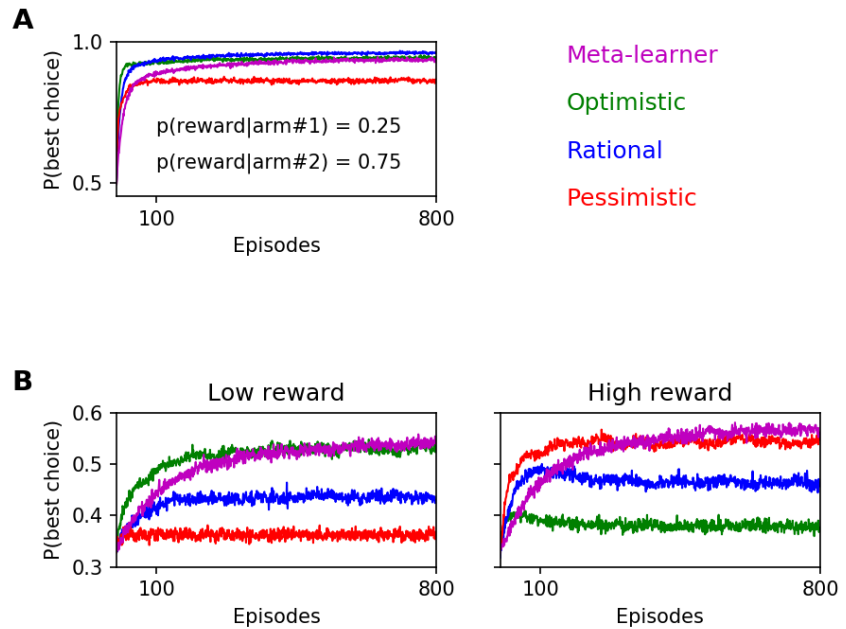


Figure 4: The performances of the Meta-learner, Optimistic, Rational and Pessimistic agents **A.** in a task where the probabilities of reward are 0.75 and 0.25 for the two choices. **B.** in a “three-armed bandit” task.

procedure was smooth: the models were implemented with low difficulty, and the simulations were quite straightforward apart from a few obscure details. We hope this replication can foster future research in the domain.

References

- [1] Timothy EJ Behrens et al. "Learning the value of information in an uncertain world". In: *Nature neuroscience* 10.9 (2007), p. 1214.
- [2] Romain D Cazé and Matthijs AA van der Meer. "Adaptive properties of differential learning rates for positive and negative outcomes". In: *Biological cybernetics* 107.6 (2013), pp. 711–719.
- [3] Michael J Frank, Lauren C Seeberger, and Randall C O'Reilly. "By carrot or by stick: cognitive reinforcement learning in parkinsonism". In: *Science* 306.5703 (2004), pp. 1940–1943.
- [4] Michael J Frank et al. "Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning". In: *Proceedings of the National Academy of Sciences* 104.41 (2007), pp. 16311–16316.
- [5] Neil Garrett and Tali Sharot. "How robust is the optimistic update bias for estimating self-risk and population base rates?" In: *PLoS One* 9.6 (2014), e98848.
- [6] Neil Garrett and Tali Sharot. "Optimistic update bias holds firm: Three tests of robustness following Shah et al." In: *Consciousness and cognition* 50 (2017), pp. 12–22.
- [7] Samuel J Gershman. "Do learning rates adapt to the distribution of rewards?" In: *Psychonomic bulletin & review* 22.5 (2015), pp. 1320–1327.
- [8] Germain Lefebvre et al. "Behavioural and neural characterization of optimistic reinforcement learning". In: *Nature Human Behaviour* 1.4 (2017), p. 0067.
- [9] Christina Moutsiana et al. "Human frontal–subcortical circuit and asymmetric belief updating". In: *Journal of Neuroscience* 35.42 (2015), pp. 14077–14085.
- [10] John O'Doherty et al. "Dissociable roles of ventral and dorsal striatum in instrumental conditioning". In: *science* 304.5669 (2004), pp. 452–454.
- [11] Stefano Palminteri et al. "Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing". In: *PLoS computational biology* 13.8 (2017), e1005684.
- [12] Punit Shah et al. "A pessimistic view of optimistic belief updating". In: *Cognitive Psychology* 90 (2016), pp. 71–127.
- [13] Tali Sharot, Christoph W Korn, and Raymond J Dolan. "How unrealistic optimism is maintained in the face of reality". In: *Nature neuroscience* 14.11 (2011), p. 1475.
- [14] Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*. Vol. 135. MIT Press Cambridge, 1998.

RÉSUMÉ

L'apprentissage par renforcement est un processus cognitif fondamental, qui se manifeste au quotidien depuis notre naissance. Grâce à l'expérience, nous apprenons par essais et erreurs à maximiser le nombre d'événements plaisants (récompenses) et à minimiser le nombre d'événements désagréables (punitions ou "récompenses négatives"). Dans le cadre de l'apprentissage par renforcement, l'une des questions les plus fondamentales est de savoir si les valeurs sont apprises et représentées sur une échelle absolue ou relative (i.e., dépendante du contexte). La réponse à cette question est non seulement cruciale d'un point de vue théorique, mais est aussi nécessaire pour comprendre pourquoi la prise de décision chez l'humain diverge des modèles normatifs et donne lieu à des comportements sous-optimaux, tels que ceux observés dans de nombreux troubles psychiatriques tels que l'addiction.

Afin de répondre à cette question, nous développons des modèles computationnels afin de prendre en compte la dépendance au contexte dans l'apprentissage par renforcement chez l'humain. Dans cette thèse, à travers deux expériences impliquant des tâches probabilistes, nous avons montré que des volontaires sains apprennent les valeurs de façon relative. Cette dépendance au contexte implique par ailleurs des choix sous-optimaux lorsque les options sont comparées en dehors de leur contexte d'apprentissage, ce qui suggère que les valeurs économiques sont normalisées en fonction de l'intervalle généré par les valeurs présentées. De plus, nos résultats ont confirmé que cette adaptation implique des erreurs systématiques et est d'autant plus grande que la tâche est facile. Les analyses comportementales ainsi que les simulations de modèle convergent vers la validation d'un modèle générant une adaptation au contexte progressive. En conclusion, nos résultats montrent que les valeurs ne sont pas représentées sur une échelle absolue, ayant des conséquences positives et négatives. Afin de faire le lien entre – une altération de – ce processus et des troubles psychiatriques impliquant la récompense, nous avons réalisé une méta-analyse sur le biais de valence qu'on observe dans plusieurs maladies. Nos résultats préliminaires suggèrent que les volontaires sains apprennent aussi bien des récompenses que des punitions, ce qui n'est pas le cas des patients souffrant de certaines pathologies comme la maladie de Parkinson ou l'addiction. Dans une expérience à grande échelle avec une approche transnosographique utilisée en psychiatrie computationnelle, nous n'avons pas trouvé de lien direct entre les paramètres de notre modèle et les différentes dimensions des symptômes, dont les troubles obsessionnels compulsifs, l'anxiété sociale, et l'addiction. Des travaux complémentaires permettront d'améliorer nos techniques computationnelles pour mieux prendre en compte la variance comportementale. A long terme, ces analyses pourront potentiellement aider à développer des outils pour mieux caractériser les phénotypes pathologiques et les troubles comportementaux, afin d'améliorer le traitement des patients au niveau individuel.

MOTS CLÉS

Apprentissage, prise de décision, modélisation, contextualisation, psychiatrie computationnelle

ABSTRACT

Reinforcement learning is a fundamental cognitive process operating pervasively, from our birth to our death. The core idea is that past experience gives us the ability of learning to improve our future choices in order to maximize the occurrence of pleasant events (rewards) and to minimize the occurrence of unpleasant events (punishments). Within the reinforcement learning framework, one of the most fundamental and timely questions is whether or not the values are learned and represented on an absolute or relative (i.e., context-dependent) scale. The answer to this question is not only central at the fundamental and theoretical levels, but also necessary to understand and predict why and how human decision-making often deviates from normative models, leading to sub-optimal behaviors as observed in several psychiatric diseases, such as addiction.

In an attempt to fill this gap, throughout the work carried out during this PhD, we developed existing models and paradigms to probe context-dependence in human reinforcement learning. Across two experiments, using probabilistic selection tasks, we showed that the choices of healthy volunteers displayed clear evidence for relative valuation, at the cost of making sub-optimal decisions when the options are extrapolated from their learning context, suggesting that economic values are rescaled as a function of the range of the available options. Moreover, results confirmed that this range-adaptation induces systematic extrapolation errors and is stronger when decreasing task difficulty. Behavioral analyses, model fitting and model simulations convergently led to the validation of a dynamically range-adapting model and showed that it is able to parsimoniously capture all the behavioral results. Our results clearly indicate that values are not encoded on an absolute scale in human reinforcement learning, and that this computational process has both positive and negative behavioral effects. In an attempt to explore the link to -an impairment of- this process in reward-related psychiatric diseases, we performed a meta-analysis based on the valence bias observable in several pathologies. Preliminary results suggest that healthy volunteers learn similarly from rewards and punishments, whereas it is not the case for pathologies such as Parkinson's disease or substance-related disorders. In a large-scale experiment, coupled with a transnosographic approach used in computational psychiatry, we found that the parameters of our model could not be directly linked with different dimensions of psychiatric symptoms, including obsessive compulsive disorders, social anxiety, and addiction. Further work will improve our modeling tools to better account for behavioral variance. In the long term, these analyses will potentially help to develop new tools to characterize phenotypes of several pathologies and behavioral disorders, as well as improve patients' treatment at the individual level.

KEYWORDS

Reinforcement learning, decision making, computational modeling, context-dependence, computational psychiatry