



HAL
open science

An ordinal generative model of Bayesian inference for human decision-making in continuous reward environments

Gabriel Sulem

► **To cite this version:**

Gabriel Sulem. An ordinal generative model of Bayesian inference for human decision-making in continuous reward environments. Cognitive Sciences. Université Pierre et Marie Curie - Paris VI, 2017. English. NNT: 2017PA066556 . tel-01897446

HAL Id: tel-01897446

<https://theses.hal.science/tel-01897446>

Submitted on 17 Oct 2018

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.

Université Pierre et Marie Curie

Ecole doctorale Cerveau, Cognition, Comportement

Laboratoire de Neurosciences Cognitives, INSERM U960 / Fonctions du lobe frontal

**An ordinal generative model of Bayesian inference for
Human decision-making in continuous reward
environments**

Par Gabriel Sulem

Thèse de doctorat de Sciences Cognitives

Dirigée par Etienne Koechlin

Présentée et soutenue publiquement le 14 Septembre 2017

Devant un jury composé de :

Pr. Mathias PESSIGLIONE	Président du Jury
Pr. Peter DAYAN	Rapporteur
Pr. Pierre-Yves OUDEYER	Rapporteur
Dr. Mehdi KHAMASSI	Examineur
Pr. Etienne KOEHLIN	Directeur de Thèse

Table of contents

Abstract.....	5
Introduction.....	6
1. Reward	6
1.1 Reward as an elementary drive for behavior.....	6
1.1.1 Representation of rewards in the brain	8
1.1.2 Construction of reward value.....	10
1.1.3 Reward magnitude.....	11
1.1.4 Binary vs. Continuous rewards.....	11
1.1.5 Critiques of neuro-economics	12
1.2 Reward as a learning signal.....	13
1.2.1 Rescorla-Wagner model.....	15
1.2.2 Limits of RW on modeling human data.....	16
1.2.3 Rescorla Wagner and machine learning	17
1.2.4 Theoretical limits of TD LEARNING	19
2. Bayesian Models: Optimal information treatment.....	20
2.1. Definition	20
2.2 Principles of Bayesian inference	21
2.3 Bayesian inference and human behavior	22
2.3.1 Perception	22
2.3.2 Abstract reasoning.....	23
2.3.3 Neuronal implementation of Bayesian inference.....	24
2.4 Learning a generative model.....	25
2.4.1 Comparison of human performance and state of the art algorithms	25
2.4.2 Hierarchical Bayesian models.....	26
2.4.3 Generalization by continuity	27
2.4.4 Limits of Bayesian inference.....	27
3 Bayes and rewards.....	28
3.1 Theoretical introduction to the exploration/exploitation dilemma.....	28
3.2 Bandits problems.....	29
3.2.1 Normative approaches of bandit problems	30
3.2.2 Behavioral account of the exploration-exploitation trade off	31
3.3 Structure Learning.....	33
3.3.1 Learning simple correlations	35

3.3.2	Hierarchical structure	35
3.3.3	Reversal Learning.....	37
	Question and strategy to answer	40
	Algorithmic Models.....	43
	Reinforcement Learning.....	43
	Structured Reinforcement Learning.....	43
	Hierarchical structure.....	44
	Frequentist learning.....	45
	Parametric model.....	47
	Model NEIG: Hierarchical, frequentist and counterfactual	49
	Materials and methods	57
	Experimental paradigm	57
	Experiment 1: Continuous reward distribution	57
	Experiment 2: Discrete reward distribution	61
	Fitting procedure	62
	Model parameters	64
	Model Comparison.....	65
	NEIG and reversal learning	66
	Results.....	69
	Behavior	69
	First model comparisons	69
	Model NEIG results	72
	Other models specific to the task	74
	Discrete task	75
	Generalization to K-bandit reversal task.....	78
	Problem	78
	Extension of the algorithmic models	81
	Extension of the NEIG Algorithm	82
	Experiment	83
	Comparison of the BEST and the EV rule	85
	Experimental Results.....	86
	NEIG fits data on the 3 options task.	88
	Which algorithm is the most efficient in the 3 option task?	91
	Discussion.....	93

Discussion of the model	93
Repetition bias.....	93
Volatility	94
Reference frame	95
Speed.....	95
Sampling.....	96
Out of frame rewards.....	97
Adaptability	97
Biological plausibility.....	99
Noise function.....	99
Reward distributions	99
f-MRI predictions	99
Extension of the PROBE model.....	100
Prediction of the model: no computation of expected values.....	101
Ordinal vs Cardinal.....	101
Limitations and future directions	102
Bibliography.....	104
List of Figures.....	116
Annex	117

Abstract

Our thesis aims at understanding how human behavior adapts to an environment where rewards are continuous. Many works have studied environments with binary rewards (win/lose) and have shown that human behavior could be accounted for by Bayesian inference algorithms. Bayesian inference is very efficient when there are a discrete number of possible environmental states to identify and when the events to be classified (here rewards) are also discrete.

A general Bayesian algorithm works in a continuous environment provided that it is based on a “generative” model of the environment, which is a structural assumption about environmental contingencies which limits the number of possible interpretations of observations and structures the aggregation of data across time. By contrast reinforcement learning algorithms remain efficient with continuous reward scales by efficiently adapting and building value expectations and selecting best options.

The issue we address in this thesis is to characterize which kind of generative model of continuous rewards characterizes human decision-making within a Bayesian inference framework.

One putative hypothesis is to consider that each action attributes rewards as noisy samples of the true action value, typically distributed as a Gaussian distribution. Statistics based on a few samples enable to infer the relevant information (mean and standard deviation) for subsequent choices. We propose instead a general generative model using assumptions about the relationship between the values of the different actions available and the existence of a reliable ordering of action values. This structural assumption enables to simulate mentally counterfactual rewards and to learn simultaneously reward distributions associated with all actions. This limits the need for exploratory choices and changes in environmental contingencies are detected when obtained rewards depart from learned distributions.

To validate our model, we ran three behavioral experiments on healthy subjects in a setting where reward distributions associated with actions were continuous and changed across time. Our proposed model described correctly participants’ behavior in all three tasks, while other competitive models, including especially Gaussian models failed.

Our results extend the implementation of Bayesian algorithms to continuous rewards which are frequent in everyday environments. Our proposed model establishes which rewards are “good” and desirable according to the current context. Additionally, it selects actions according to the probability that it is better than the others rather than following actions’ expected values. Lastly, our model answers evolutionary constraints by adapting quickly, while performing correctly in many different settings including the ones in which the assumptions of the generative model are not met.

Introduction

1. Reward

Most of us are familiar with the concept of reward. In many cultural representations of the Wild West, dangerous criminals are wanted, and a reward is offered for their capture. Alternatively, we are supposed to be rewarded for our “good” actions, “good” here either means something positive for society (helping a blind person to cross the street), or qualifies the action as particularly well realized (sales performances). In all these cases a reward is supposed to increase the motivation to perform an action.

It is also a central concept in psychology as a factor in understanding human behavior. It actually has two different origins, which eventually converged (White, 1989). One originates from the writings of Epicurean philosophers which posit that individual behavior is driven by looking for pleasure and avoiding pain. By extension, in a modern psychological view, anything which attracts individuals or animals is a reward; anything which repulses is a punishment or a negative reward (Young, 1959).

The second origin comes from the field of memory creation (today named, learning). The idea of George Ramsay (Ramsay, 1857), later developed experimentally by Thorndike was that a connection between two events (either two stimuli, or a stimuli and an action) is better remembered if it is associated with or followed by a “satisfying state of affairs” (Thorndike, 1911). Because of its effect on the association of the two events, this “satisfying state of affairs” was later named a reinforcer. Rewards are particularly good reinforcers (Skinner, 1938), and these two terms are now used almost equivalently in the literature.

In this chapter we will successively review reward as a drive for behavior, then as a memory reinforcer and describe theories accounting for both.

1.1 Reward as an elementary drive for behavior

For thousands of years people have tried to understand human behavior. Several disciplines, such as philosophy, economics, cognitive sciences, and more, have developed concepts and theories to explain actions and choices.

In psychology, a central concept is reward. Reward is the cause of approach behavior, and is only defined behaviorally. It has been observed and is understood in evolutionary terms as anything that contributes to maintaining homeostasis (for example food), reproduction, or survival (avoiding pain) is rewarding. In a recent review, Wolfram Schultz defines the function of the brain as recognizing and getting rewards in the environment (Schultz, 2015).

Reward is a subjective notion, as different individuals are attracted towards different things, and can vary across time. Rewards also have a magnitude so that prospective rewards can be compared to drive a choice (Samuelson, 1938).

Several approaches have been proposed to measure experimentally the subjective magnitude of reward. For example, the motivation (Berridge, 2004), or effort (Pessiglione et al., 2007) displayed by an animal or human to get something. This variable can be measured in different ways: conscious

reporting, behavioral choices, or physiological measures, and be used as a proxy for reward. Similarly others have tried to measure the “hedonic” pleasure felt during consumption of a good (Plassmann et al., 2008), (Kahneman et al., 1997). Of course nothing guarantees to get the same value scale when different measures are used.

The field of economics is also very interested in human behavior, particularly through neuro-economics which is a recent subfield aiming at explaining physiological data with economic concepts (Rangel et al., 2008). The central paradigm of economics is the trade-off between cost and gains. Economics postulates the existence of a common scale on which costs and gains can be compared. In particular, all actions have a cost: it can be a physical cost like fatigue, but also an opportunity cost as an action is selected over other potential actions. On the other side, actions have also an expected return in terms of pleasure, survival, money, social prestige, etc. If the return is superior to the cost, the action is performed.

The notion of a personal scale of value on which anything can be compared was first introduced by Bernouilli who spoke of “moral value” (Bernouilli, 1738), and the name “utility” which is still used today in economics was first introduced by Bentham (Bentham, 1781).

Utility is central to explain a choice between several alternatives but there is also no generic way to compute its magnitude. In microeconomics, there are attempts to get a theoretical value. Intuitively, the utility of an umbrella is higher on a rainy day. As a restaurant owner, given some variables (prestige of the restaurant, expected salary, etc.), it could be possible to evaluate the utility of the raw products to buy. But ahead of simple market considerations, these theoretical values are essentially valid to describe the average behavior of a population and there is no real economic model to compute subjective individual “liking” of options. For example, one can generally like red over white wine independent of the relative value of both in the present context (e.g. does it come with meat or cheese?, etc.).

To go further on utility estimation, one can imagine asking people directly to evaluate how much they like or value several items and use their answers as a measure of utility. In behavioral economics, a common measurement procedure consists of attributing an equivalent monetary value to all things. The “willingness to pay” for a good/service is measured through an auction procedure, which converges to an equivalent value in euros. This value is then used as a proxy for utility (Becker et al., 1964). The prediction is, if the measure is valid, that subsequent choices should follow predicted utility.

Are reward and utility similar concepts? We discussed above aspects on which reward and utility are similar. A central difference sits in the summation between positive and negative magnitudes. It is experimentally shown that punishments are not equivalent to the inverse of rewards (D’Ardenne et al., 2008). It has been proposed that rewards and punishments are represented on different scales (Fiorillo, 2013), (Frank et al., 2004), and are encoded by different neurotransmitters.

This difference was first pointed out by economists who spotted non-symmetric behavior between losses and gains (Kahneman and Tversky, 1979). However utilities on their side can be added, so that the sum of an equally positive and negative event is theoretically similar to a neutral event.

The framework of utility is very practical to model costs as they can be discounted from the expected utility. However, there is no equivalent operation for physiological rewards. Therefore, even neurophysiologists are now using the economical concept in their description of biological data: the word “reward” names a positive event which has a “value”. Reward values correspond to subjective utility and can be manipulated as the economic notion. For example Camillo Padoa-Schioppa and John Assad claim that the brain encodes the economic value of rewards (Padoa-Schioppa and Assad, 2006).

One of the difficulties of the economic approach is the need to put a value on everything, including friendship, physical integrity, reading a book. If one accepts to have an arm broken for €10 billion, does it mean that this is the actual utility of one’s arm?

To solve this issue, another school of thought in the field of economics has refused to compute cardinal scales to value everything. Instead, they claim that preferences can only be ordered without a notion of absolute distance between the different possibilities. This order can be inferred through choices (Pareto, 1906). Intuitively, if I prefer A to B, then A will be ranked before B. One of the problems is that human or monkey choices can be non-transitive, A can be preferred to B, B to C, but also C to A. In sane humans, the proportion of non-transitive choice is around 3% (Fellows and Farah, 2007). This invalidates the existence of an absolute order in human minds, but it is a sufficiently low proportion, so that absolute order can be used as an approximation.

Economics and psychology explain choices by a comparison of value. Where and how these values are encoded in the brain?

1.1.1 Representation of rewards in the brain

The rewarding dimension of an object is independent of its sensory properties. We have external receptors like eyes and ears to evaluate different sensory properties of objects, but is there something like a reward receptor in charge of encoding reward value?

A large brain network has been identified as the “reward system”. It is defined as brain structures activated when a reward is obtained. The reward system principally includes fronto-basal ganglia loops and thalamo-cortical loops. It has been initially characterized with the technique of electrophysiology by the different teams of Wolfram Schultz.

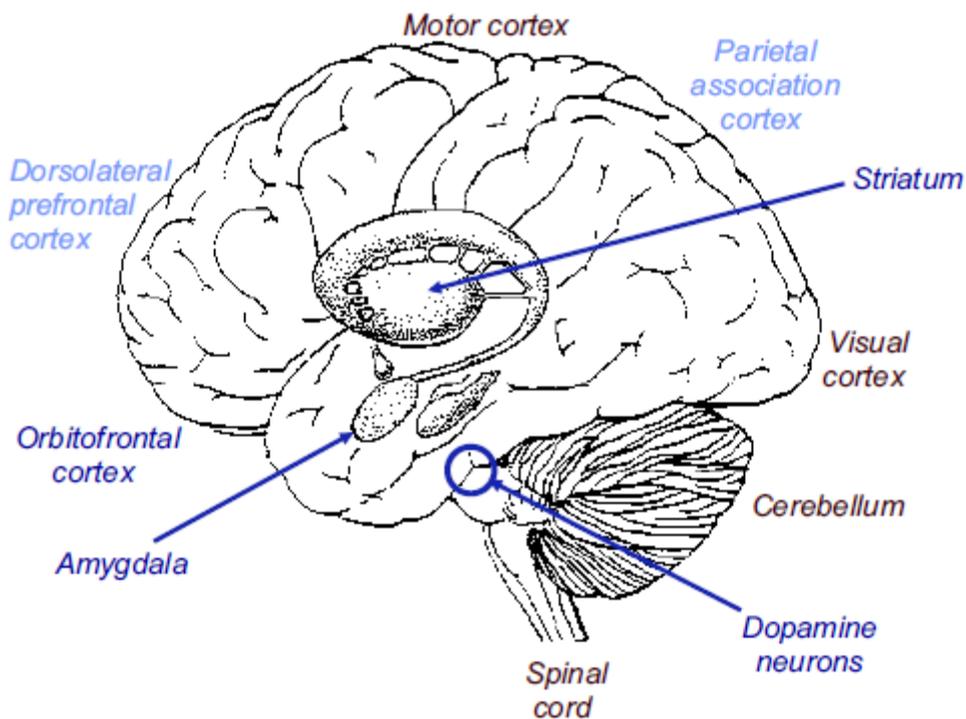


FIGURE 3. Principal brain structures for reward and decision-making. Dark blue: main structures containing various neuronal subpopulations coding reward without sensory stimulus or motor action parameters (“explicit reward signals”). Light blue: structures coding reward in conjunction with sensory stimulus or motor action parameters. Maroon: non-reward structures. Other brain structures with explicit or conjoint reward signals are omitted for clarity.

Figure 1: From Schultz 2015

In particular, dopamine neurons of the Substantia Nigra Compacta (SNc) and of the ventral tegmental area (VTA) were found to scale to the magnitude of the anticipated reward prior to a choice (Morris et al., 2006). These neurons project to the striatum in which the “expected” action-value of all possible options was found to be encoded (Samejima et al., 2005). This same result was also showed in rats (Roesch et al., 2009). Therefore the striatum receives value of all possible options, can pick up the most rewarded and trigger action to obtain it. Importantly, the response of dopamine neurons is independent of the modality of the reward, but depends only on its “expected” value (Ljungberg et al., 1992).

In the cortex, the Orbitofrontal Cortex (OFC) which is particularly well connected to the striatum is known to code the brain equivalent of the utility function (Montague and Berns, 2002). Neurons encoding value have been found in animal OFC by intracranial electrophysiology studies, from the early works of the 1980s (Thorpe et al., 1983). The development of a precise methodology to measure subjective value (from animals’ point of view) by Camillo Padoa-Schioppa and John Assad (Padoa-Schioppa and Assad, 2006) enabled to link their activity to utility in the sense that it is independent of its nature, position, or the action necessary to get it (Padoa-Schioppa, 2011). This result was reproduced in other labs (O’Neill and Schultz, 2010).

In humans, functional-MRI studies have confirmed this result in finding correlations between the activity of OFC and its neighbor, the ventromedial prefrontal cortex, and the “value” of a choice. But the difficulty to define “value” objectively led authors to use close but still different definitions of value. Indeed, Kringelbach & colleagues asked participants explicit pleasantness ratings and used them as “values” that they saw coded in the OFC (Kringelbach et al., 2003). Hilke Plassman et al. have used an economic procedure to measure the “willingness to pay” for a good (Plassmann et al., 2007). Jan Peters et al. worked with real money and fitted a model of temporal discounting to use expected value (Peters and Buchel, 2009). In each study, the activity of OFC was correlated with what was defined as value. The study of Maël Lebreton et al. did not use food or money but different categories of objects, and showed that brain activations (in particular in the OFC) reflected choice procedure by consistent and automatic valuation of items (Lebreton et al., 2009).

Values of reward are encoded in the brain but how are they computed? Are they innate or do they depend upon prior experience? Is it possible for an external observer to anticipate the subjective value of a subject?

1.1.2 Construction of reward value

Approach behavior has been characterized in all species as sensibility to reward. Trees grow their roots where there are the most nutrients, or bacteria follow nutrient gradient by chemotaxis to find food sources (Stocker et al., 2008). According to the behavioral definition of rewards, this corresponds to a sensitivity to rewards.

We mentioned earlier that to favor gene transmission, rewards could have been tuned by evolution towards survival and breeding. This brings the distinction between primary and secondary rewards. Rewards are primary when they directly contribute to help gene reproduction. Following Wolfram Schultz’s recent review, we can put in this category:

- Eating and drinking when needed. This is part of the homeostasis system.
- Behavior associated with reproduction and parental care to transmit genes appropriately.
- Avoidance of “punishments” which could affect survival and future reproduction (Schultz, 2015).

But we have all experienced desire and motivation for things which are not in this list. For humans, primary among them is money. Similarly, it is easy to have animals approaching a-priori neutral objects when these objects have been frequently paired to a primary reward. Frederic Skinner developed a set-up where hungry animals would be put in a chamber (“Skinner Box”) and get food after pressing a lever. A lever is neutral and initially rats are not particularly approaching the lever. But the box is not that big so as they go all around, they eventually press the lever and get food. Skinner observes that these animals will come back to the lever and their pressing frequency will increase as they learn the association between food and the lever. Therefore, animals become motivated to press the lever. This is named operant conditioning or Skinner conditioning (Skinner, 1938). Money or lever pressing are secondary rewards in the sense that it has been learned that they predict the future obtaining of primary rewards.

The same signal can be recorded after similar primary and secondary rewards in the brain (Schultz et al., 1997), even if some areas are specific to the sensory properties of some (primary) rewards (Sescousse et al., 2013).

1.1.3 Reward magnitude

As we explained earlier, approach is the behavioral definition of a reward, and avoidance is the behavioral definition of a punishment (which is different from a negative reward). The rewarding, punishing or neutral status of a stimulus can be very easily determined by observing behavior. However to perform choices, individuals also have a sensitivity to the magnitude of the reward, and are able to choose the greater of two. Importantly, this comparison of magnitude happens when choosing between different quantities of the same good (2 apples are better than 1 apple), but also between goods of different nature (apple or orange?).

It is well known that there is no linear relation between quantities of a good and the magnitude of its reward value. The first introduction of the notion of utility was indeed to introduce a concave relationship between quantity and utility, so that the marginal utility decreases with quantity. This sensitivity is an individual parameter (Lauriola and Levin, 2001).

There is another way to vary reward magnitude of a good: to change the frequency of its delivery. Pascal defined what an optimal choice in this context is, with the notion of expected value. Expected value multiplies the magnitude of a potential gain with the probability to obtain it. In terms of expected value it is equivalent to obtain €10 for sure and €20 with 50% chances (Huygens, 1657). However, the decreasing marginal utility predicts that people are risk averse and prefer the sure option. This means that the value of the reward associated with the sure option has a higher magnitude.

In economics, the framework of expected values has been extensively discussed. Considering the concavity of the utility function, Von Neumann and Morgenstern have made explicit the axioms that should follow an optimal decision maker computing expected utility instead of expected value (Von Neumann and Morgenstern, 1947). Failure to follow these axioms for example the Allais paradox (Allais, 1953) pushed Kahneman and Tversky to develop their prospect theory where they propose that humans have a distorted perception of probabilities (Kahneman and Tversky, 1979). They suggested a new way to compute the magnitude of stochastic rewards: multiplying these distorted probabilities to the utility of the potential gain.

In lab experiments, monetary rewards are often used on humans as we are used to manipulate them and it is easily possible to separately vary magnitudes and probabilities. The field of neuro-economics pushed for similar approaches on animals by varying the magnitude and probability of food or liquid rewards.

1.1.4 Binary vs. Continuous rewards

In the general case, rewards are evaluated on a continuous scale. For example, in the stock market, gains are unpredictable and can be defined with several decimal numbers (+22.47%). The precision is in all cases bounded by human capacities, either reward receptors, or memory limits. In some environments, rewards can only take a finite number of values. For example at a lottery or tombola the prizes are defined in advance and only one of them can be obtained. There are also settings where rewards can only take two values. For example, you can have or miss your train. In this last case, we can say the rewards are binary.

There are interesting simplifications when rewards can be modeled as binary. First, when evaluating a choice the expected value combines the different reward values with the probability of obtaining

them. All this information, two numbers per possible reward, needs to be acquired and memorized. In the case of binary rewards, expected value directly reduced to the probability p of obtaining the reward. This is simpler in terms of memory load, and computational load as no summation across possible rewards is needed.

Also when rewards are discrete or continuous, magnitudes can vary on different orders. And obtaining a high magnitude reward (€1000) may justify the sacrifice of little rewards on the short term (€1). There are therefore trade-offs to operate in order to maximize the total received reward. These trade-offs can be complex as a standard result in economics is that the perception of reward value is not similar for delayed rewards than for immediate rewards (Green and Myerson, 2004). When rewards are binary, only the numbers of rewards need to be maximized without accounting for magnitudes.

Binary rewards also simplify choices in the context of uncertainty as values can only vary on the probability dimension. When magnitudes and probabilities vary independently, it is possible to be presented with a choice where 2 options have the same expected value but not the same uncertainty: do you prefer €10 for sure or €100 with 10% chances? Individual human participants have different preferences for these choices (Harrison and Rutström, 2008), and the concept of risk-liking or risk-aversion is necessary to explain these differences. In the context of binary rewards, choices can be explained without these additional concepts.

Binary rewards are more than 1 or 0. When there is only one possible reward, whatever its magnitude, it can be modeled as a binary reward. Interestingly, in this case, the response of neurons is also rescaled to a unique binary response independently of the true magnitude of the reward (Tobler et al., 2005). The problem with continuous rewards is that it is always possible to receive higher than expected magnitudes. For example your idea of what is the best food in the world might evolve across life while more and more restaurants are visited. The discovery of a new maximum reconfigures expectations and the definition of an average meal.

In the following chapters, we will often make a distinction between binary and continuous rewards as the treatment of continuity asks for more computational tools. In the following sections of this first chapter, we expose simple inference free decision making methods which apply as well to the continuous case as to the binary case. In the third chapter, we will mention complex inference algorithms which are, for the majority of them, designed for discrete rewards. Our PhD work is aimed at understanding how these algorithms can be extended to an environment presenting continuous rewards.

1.1.5 Critiques of neuro-economics

To conclude this part on reward and its motivational dimension, we would like to review critiques addressed to the use of economic paradigm to study animal and human behavior.

The field of neuro-economics developed on the postulate that natural animal behavior is a rational decision problem of utility maximization (Pearson et al., 2014). However, real life environments are usually open. It means that, contrary to laboratory tasks where choices are proposed between a limited numbers of well-defined options, the different real-life options are not always explicit. If we consider that resources (money, strength, etc.) are limited, getting something also means not getting other things. The choice is then between buying an object, being able to use it, and maybe not being

able to buy other objects. Or not buying, not being able to use it, but being able to buy something else. Therefore, the theoretical comparison underlying the act of buying in the utility framework asks to consider an infinite number of possibilities (Köszegi and Rabin, 2007).

Also, experimenters have doubted that utility measures are always relevant to explain behavior. For example, a dissociation was shown in animals between how they “want” something to how they “like” it (Berridge, 1996). In particular, habits (Dickinson, 1985) or rules (Sanfey et al., 2003) can guide action away from (otherwise) preferred outcomes. Therefore, measuring reward values creates a pertinent subjective hierarchy between different goods or services, but which is not enough to explain choices in many common settings.

We will now analyze a completely different use of the reward signal which is not at all present in the concept of utility or in economics. In uncertain environments, rewards act as a reinforcer and drive learning.

1.2 Reward as a learning signal

In this section, we will explain the concept of reinforcement in behavioral psychology. Reinforcement names the effect of rewards on the future repetition of the behavior which enables to obtain them.

The goal of behavior is to maximize reward. When the current environment is well known and predictable, this consists of selecting the “best” action, the one with the highest expected reward. For example, in your favorite bakery you know exactly which bread is the best, and what is your second choice if the first choice is not available. In an unknown environment, what is the best strategy to both learn and maximize reward concomitantly? What is experimentally observed on animals and humans?

Learning experiments on animals have more than a century of history. Thorndike put cats in a puzzle box, and measured how fast they would find the trick to escape, knowing that food is available outside. He noticed that animals would at the beginning find the exit by chance, but once put again in the box, they would reliably be faster to exit until they eventually “learned” to do it immediately. Thorndike suggested that the reward obtained out of the box facilitates the memorization of the correct action (the one which allows getting out) (Thorndike, 1911).

By chance, while he was studying digestion mechanisms in dogs, Pavlov noticed that a neutral stimulus could become so predictive of a reward in animal’s mind, that their reaction to the stimulus would be the same as their reaction to the reward (Pavlov, 1927).

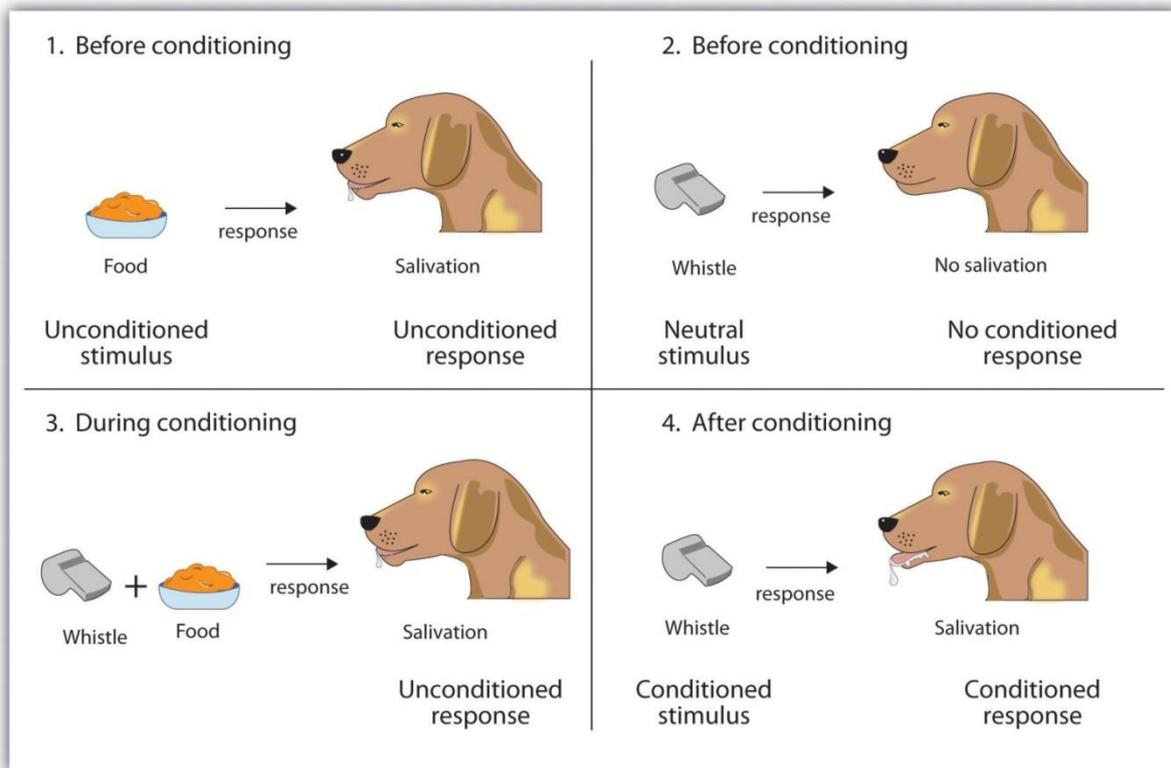


Figure 2: Classical conditioning, whistle sound is associated with food and cause the same response than food. From *Beginning Psychology*, Charles Stangor 2012

In Figure 2 the whistle sound is initially neutral, while food is a rewarding stimulus which causes salivation. After conditioning, the whistle sound by itself causes salivation. The presentation of reward has reinforced the predictive ability of the food by the sound. This is named Pavlovian or classical conditioning.

We already mentioned the work of Frederick Skinner. Interestingly, in operant or Skinner conditioning, it is the action which is predictive of the reward and therefore animals are motivated to perform it. In an unknown environment, conditioning predicts that animals will repeat rewarded actions and avoid punished actions.

Is there a mathematical way to model this reward dependent learning? In particular, is it possible to predict the subjective value of an action, knowing the history of reward and punishment that were obtained after it was performed?

A particular observation was decisive in designing the Rescorla Wagner model which is still a reference today. Kamin describes rat experiments where after conditioning stimulus S1 to reward R (Figure 3 A), a new stimulus S2 is added to S1 and both are paired to R (Figure 3 B). After a new conditioning phase between S1+S2 and R, it is observed on animals that R is not conditioned to S2 alone (Figure 3 C). It means that animals do not look like expecting R when they are presented S2 (Kamin, 1969).

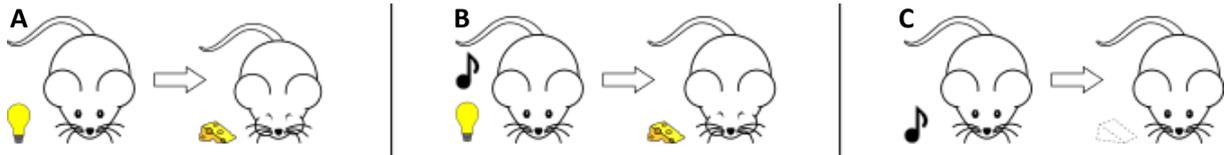


Figure 3: Blocking effect. (A) Light (S1) is associated with food. (B) In a following phase, light (S1) and sound (S2) together are associated with food. (C) Later sound (S2) alone is found not associated with food. From Nicolas Rougier, Blocking, Wikipedia

This result suggests that learning depends critically on the “prediction error”, the difference between what was predicted and what actually happened. Here since R is first conditioned to S1, there is no prediction error when S1+S2 are paired to R, and then S2 do not acquire any predictability of R.

To be comprehensive on this issue, a recent paper, while recognizing the existence and reliability of blocking, contested its universality as they failed to reproduce it in several experiments (Maes et al., 2016).

We will present in the next part the reinforcement model of Rescorla and Wagner.

1.2.1 Rescorla-Wagner model

Rescorla and Wagner published a model accounting for observations on animal learning in particular blocking (Rescorla & Wagner, 1972). It is based on the predictability of reward by an otherwise neutral stimulus, S. When S is repeatedly paired with a reward R, the prediction of R by S will grow until S fully predicts R. Formally it corresponds to a simple equation:

$$V_{T+1}(S) = V_T(S) + \alpha * (R - [V_T(S) + \sum_{Others} V_T(Others)])$$

with $V_T(S)$ the strength of the prediction of R by S at time T. R is the magnitude of the reward (the quantity of reward can vary and can also be zero in case of absence), α is a parameter stable across an experiment measuring the speed of learning, and we include in $V_T(Others)$ the possibility that stimuli happening simultaneously and other than S predicting R, (like S2 in the blocking effect).

Importantly, when the prediction error (here $R - [V_T(S) + V_T(Others)]$) is null $V_{T+1}(S) = V_T(S)$ and no more learning happens. The prediction error can be seen as a measure of the quality of the prediction $V_T(S)$. It is a drive for learning as long as the prediction is not “perfect”.

The strength of the prediction of R by S is named V(S) and is also as stated by the model the rewarding value acquired by S through learning. Therefore, after learning, animals will like and want S as much as they like and want R as long as the association is maintained.

Interestingly, the concept of prediction error has been biologically validated. A series of work by Wolfram Schultz in Fribourg recorded the response of the dopamine neurons in SNc and VTA. Their activity is qualitatively compatible with a prediction error: It is stronger than baseline in case of unexpected reward, on baseline in case of predicted reward, and under baseline in case the reward is expected but not present (Mirenowicz and Schultz, 1994), (Schultz et al., 1997) (Figure 4).

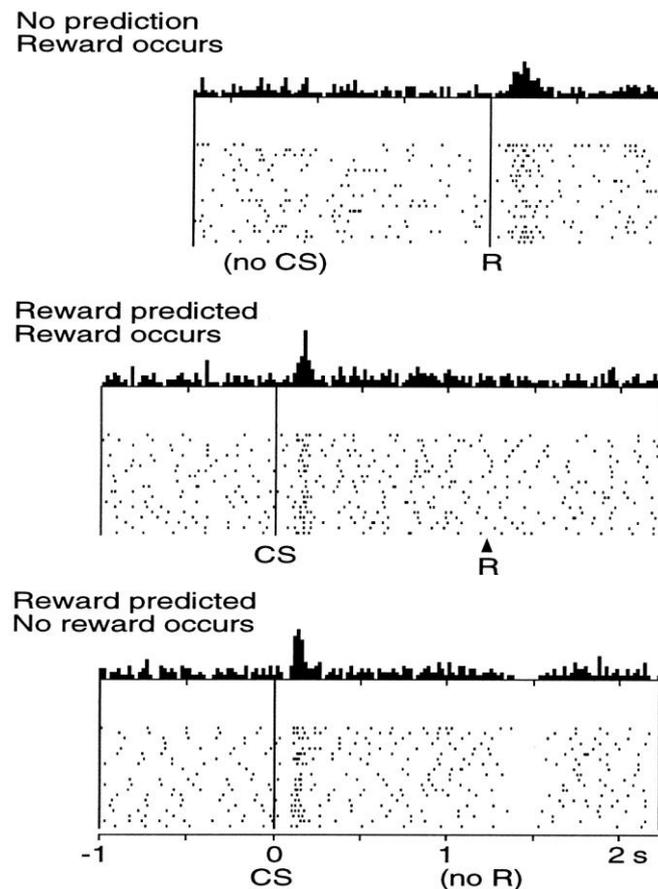


Figure 4: World famous experience, where midbrain dopamine neurons are recorded in an operant conditioning task. Response transfers from reward to its predictive cue. Absence of reward after learning is perceived as a punishment. (Reproduced from (Schultz, Dayan, & Montague, 1997))

It is more difficult to show that this dopamine signal quantitatively match what would be expected by Rescorla Wagner model, as the experiment needs to be perfectly controlled to know precisely what the expected reward is. Experiments validating this idea are more recent but Hannah Bayer & Paul Glimcher showed that at least for positive prediction errors, the dopamine recorded signal corresponds precisely to what Rescorla Wagner would predict (Bayer and Glimcher, 2005).

It is interesting to note that dopamine neurons are both implicated in the signaling of option values and on the quantification of prediction error. Several computational models of neural circuits in particular in the striatum have accounted for this dual role (Collins and Frank, 2014).

1.2.2 Limits of RW on modeling human data

We showed earlier that the Rescorla Wagner model (RW) explains many human and neuronal data in particular blocking. However, some data cannot be accounted for by RW (Miller et al., 1995). In particular, RW predicts that if after presenting always S and R together (conditioning), S is presented sufficiently alone (extinction), the association between S and R is unlearned and we should go back to the initial situation. However animals display evidence of spontaneous recovery (if extinction is recent, the pairing between S and R can re-emerge without relearning) (Pavlov, 1927), or facilitated relearning (if extinction is recent, re-learning happen faster than initial learning) (Frey and Ross, 1968). These two phenomena are not predicted by the Rescorla Wagner model.

Numerous models have tempted to complicate the RW model to vary the learning equation across time and account for all the limits of RW. To cite only few of them, Wagner itself described a mechanism where he varied the update equation of the associative strength $V_{T+1}(A)$ in function of the prior on the strength $V_T(A)$ (Wagner, 1981). Mackintosh (Mackintosh, 1975) proposed a model which tries to identify among all presented stimuli which one is crucial for the reward. Pearce & Hall introduced a new variable named “saliency” applied to the stimulus to modulate learning (Pearce and Hall, 1980). However, due to its simplicity and operational easiness, Rescorla Wagner model is still a reference for the field.

1.2.3 Rescorla Wagner and machine learning

How to learn in a new environment is also major question in the field of machine learning. Psychology and machine learning have inspired each other during the last decades. Indeed machine learning looks for new ideas in animal learning and computer simulations are an idealized learning environment where the properties of psychological models can be tested. While experimental constraints on animal or human testing prevent from going beyond simple problems like learning by trial and errors, in a virtual environment everything can be adjusted, controlled, repeated; and therefore more complex learning problems can be studied. Thus, algorithms evidenced experimentally can be tested on more complex situations for generalization. On the converse it is interesting to test whether optimal and complex theoretical algorithms can explain experimental data.

As we saw above, rewards are desirable for a living organism and 2 different rewards can be compared to each other and ranked by order of preference. Therefore, it is also possible to consider rewards as feedbacks evaluating the quality of the performance. Obtaining 2 apples is “better” than obtaining 1 apple, so it is sensible to bias future choices towards the action which leads to 2 apples. Unfortunately, the behavioral answer to these feedbacks is uncertain as one ignores whether receiving 2 apples is good: there might be an action enabling to win 10 apples!

Therefore, typical problems faced by humans lay between supervised (explicit feedback) and unsupervised (no feedback) learning. They are usually named “Reinforcement learning” problems (Sutton and Barto, 1998). These problems have been transposed in virtual environments, with artificial rewards: the algorithm is rewarded for its actions with points, and it has the instructed goal to maximize his earned points. Models similar to the one of Rescorla Wagner as they use critically the prediction error were developed from the 1950s in machine learning (Minsky and Lee, 1954).

In reinforcement learning problems, the environment is represented as states. In each state, different actions can be performed. Performing an action can bring some reward, but it also changes the environment so that the following decision will be taken in another state. As an example, to study chess playing, all different configurations of the board can be represented as states. Thus, moving a piece changes the state. Other example, when moving physically from A to B, all positions in between can be considered as states where one can decide to go forward backward right or left.

While navigating the state space, the generic problem is to find the strategy maximizing the total reward received. In particular, some states can be poorly rewarded but lead to very highly rewarded states. This problem generalizes simple operant conditioning cases reported before. There, the credit of a reward was entirely credited to the action which immediately leads to it. Here the entire state-path towards a high reward needs to be credited (Minsky, 1961).

A strategy (or policy) is a probability distribution on the different proposed actions in each of the states. One or several strategies are in average better than the others, but usually all strategies cannot be tried one by one. Indeed if the problem is complex, many successive states are visited in a trial and there are, for example, far more possible chess games than atoms in the universe. Therefore, reinforcement learning problems are solved by clever exploration methods able to converge in a reasonable time to the optimal strategy.

State of the art methods for reinforcement learning problems are using the concept of prediction error and are named temporal-difference learning (Sutton and Barto, 1981). These methods are particularly easy to implement while being guaranteed to converge to the optimal solution (Sutton and Barto, 1998, p.138).

The principle of these methods is to compute either the value of all states, $V(s)$, either the value of all actions in all states, $Q(s,a)$. $V(s)$ represents the expected future gain when 's' is visited; $Q(s,a)$ the expected future gain when 's' is visited and 'a' is performed. These values depend of course on the strategy used. Generally, these methods perform two successive steps in a loop. In a first step the strategy is set and for this strategy, V or Q are computed. In a second step V or Q are used to inform how to modify the strategy in order to increase the earned reward. This defines a new strategy and the first step can be performed once again.

For a given strategy V or Q are computed with a formula similar to the Rescorla Wagner rule. Let's suppose we are in state S and following the current strategy, action A is performed. This brings a reward r and leads us to state S' . Then $V(S)$ can be updated as,

$$V_{T+1}(S) = V_T(S) + \alpha * (r + V_T(S') - V_T(S))$$

In the second step, to improve the strategy, one can use directly action values $Q(s,a)$ to increase in each state the probability to choose the action bringing the most reward. Alternatively, state values, $V(s)$ can be used to increase the probability to visit highly rewarded states, but it asks the knowledge of a structural model of the environment: A structural model describes the transition probabilities between states conditioned to actions ($p(S \rightarrow S' | a)$). This model is independent of the strategy used, and can be learned gradually by experience.

The difference between methods using a structural model and others is central to a long standing debate in animal learning: can an animal or human learn and use such a model of the environment? If yes they are said "model-based", if not "model-free"?

There is an interest in knowing the structural model of the environment when different types of rewards are available and that their relative value varies according to current needs or to time. For example one will not search a house identically if he looks for jewels, or if he looks for the toilets. Without the use of a structural model, any modification in the desirability of rewards asks for the change of all the state-action values and computation of a new optimal policy. On the converse, knowing the structural model allows easy re-planning.

To study the difference between model-based and model-free behavior, a famous paradigm extensively studied in animal psychology is reward devaluation. An action, lever pressing is associated with a food reward (Operant conditioning) then through poisoning or satiety, the

subjective value of the reward is decreased. In a third phase, it is tested whether animals are able to inhibit lever pressing (Figure 5).

In a model free algorithm, the initial action has always been associated with a positive outcome, therefore it still have a high value and is not devaluated. A model based algorithm remembers the arrival state and thus the reward at stake. When the reward is devaluated, it can reevaluate action values accordingly and therefore the response will be inhibited.

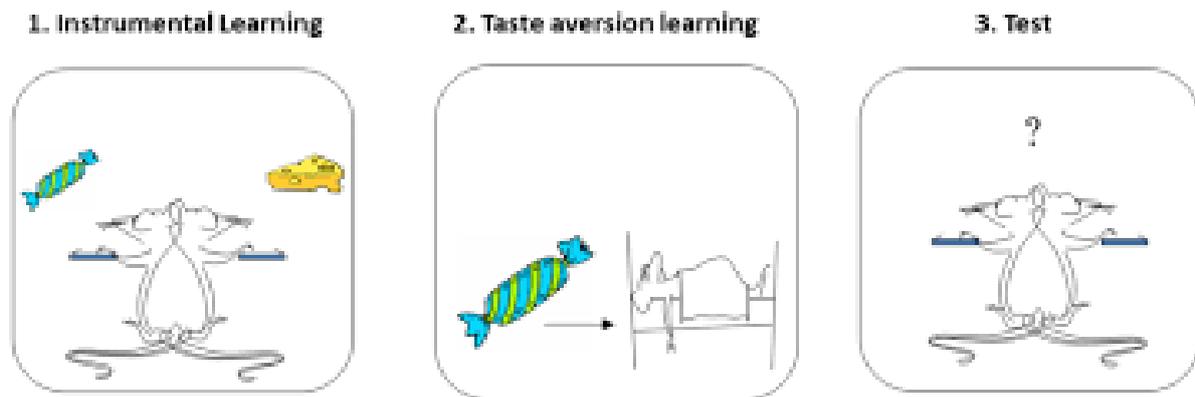


Figure 5: Usual reward devaluation paradigm: a lever is associated with a food item. This food item is then devaluated. Hypothesis 1: There is a model linking between lever pressing to the food item, then if it does not like the food it does not like also the lever. Hypothesis 2: In the learning phase it transfers the value of the food into the value of the associated lever. It likes the lever but does not “record” why. When the food is devaluated, it still like the lever. From B. Balleine

There are debates to understand whether model free or model based methods are more appropriate to describe human behavior and it seems that there is a part of both (Dayan and Berridge, 2014).

Usual problems in machine learning only use one type of reward, therefore model based and model free algorithms are equivalent. In particular, there is an algorithm able to infer the structural model of the environment from $Q(s,a)$ values: the DYNA algorithm (Sutton, 1990).

Among the model-free algorithms solving reinforcement learning, some are even simpler than temporal difference methods mentioned above as they neither construct nor use a strategy to guide behavior in the different states. One of these algorithms is named Q-Learning and was developed by Watkins during his PhD (Watkins, 1989). This algorithm computes states' action values $Q(s,a)$ as TD Learning, but these values do not depend on a particular strategy. Instead, in each state the algorithm chooses with high probability the action with the highest value. It cannot always choose the action with the highest value as there is a mathematical constraint on the algorithm to guarantee its convergence to the optimal behavior: every action must be explored sufficiently (the number of times an action is chosen tends to infinity as the number of total trials tend to infinity) (Watkins and Dayan, 1992). Therefore, frequent sub-optimal choices have to be performed in Q-Learning.

In the results of our experimental work, we will compare Bayesian algorithms to model free algorithms and Q-Learning will be our reference.

1.2.4 Theoretical limits of TD LEARNING

TD Learning algorithms in general, Q Learning in particular do minimal hypothesis on the environment and whatever the environment's structure, it is guaranteed to converge to the true

action values (Watkins and Dayan, 1992). A fundamental limit of these algorithms appears when the environment changes across time. These changes may involve the transition between states (a new street is opened in a neighborhood which modifies itineraries) or the rewarding status of states (one's favorite restaurant can move or change its cook). TD Learning reacts to a change in the environment by forgetting past Q values, and gradually readapting to the new ones. If a new environmental change brings the environment back to the initial situation there is no way to retrieve past Q-values.

In particular, our usual environment is seasonal. Winters and summers are very different as the appropriate actions in the two seasons. However, most winters and most summers are similar so that a strategy developed one winter will work the following year. The prediction of TD Learning algorithms is that Q values will always be relearned from scratch after each change, albeit it would be valuable to retrieve past Q values when the context in which they were valid comes back.

One can imagine that for the same state-action pair, several Q values are memorized in order to account for these different environmental contexts. However, how to decide which set of Q values to use? Which algorithm can manage this selection?

Another limit of TD Learning is its impossibility to generalize its knowledge. If the algorithm learned to play chess, it cannot infer how to play Draught (checkers in American English) and needs to restart its learning from scratch. Indeed a TD Learning algorithm makes no hypothesis on the transition or reward structure of the environment when it starts in a new setting. Yet there are general laws like physical gravity or societal codes which structure most of the situations we face. Therefore, one might expect an efficient learning algorithm to accumulate general information about natural environments which would constitute an a priori strategy when a new setting is encountered. The algorithm would presumably converge faster to the optimal strategy.

Bayesian inference algorithms are an answer to these problems and we will discuss their origin and implementation in the following part.

2. Bayesian Models: Optimal information treatment

2.1. Definition

Bayesian inference is a general statistical method aiming at finding the hidden causes that explain an observation. As an example, when listening at an English speaker (observation), one can guess where he is coming from (what constitutes the "hidden" cause of the accent). To do so, Bayesian inference uses probabilities in order to represent the "belief" or the "confidence" that a particular cause explains the observation. In our former example, the result of the Bayesian analysis can, for example, be that the speaker has 30% chance of being Spanish, 50% of being French and 20% of being Italian.

When they were introduced, Bayesian statistics were a new use of probabilities as probabilities were initially only used to represent known frequencies. The name "Bayesian" refers to Thomas Bayes who proposed a formula to compute the now called beta distribution: a distribution of the probability of success at a lottery given the number of successes so far (Bayes and Price, 1763). This approach was formalized and popularized later by Pierre-Simon Laplace who gave a formulation of what is now named Bayes Theorem (Laplace, 1774).

$$p(\text{cause}_i|\text{obs}) = \frac{p(\text{obs}|\text{cause}_i) * p(\text{cause}_i)}{\sum_j p(\text{obs}|\text{cause}_j) * p(\text{cause}_j)}$$

Therefore, given an observation, the probability that a cause (i) is responsible for the observation can be inferred from knowledge of how likely all possible causes (j) would generate this observation.

What is particularly interesting is the possibility to build up on it when a new independent observation (obs_new) is acquired.

$$p(\text{cause}_i|\text{obs}, \text{obs_new}) = \frac{p(\text{obs_new}|\text{cause}_i) * p(\text{cause}_i|\text{obs})}{\sum_j p(\text{obs_new}|\text{cause}_j) * p(\text{cause}_j|\text{obs})}$$

The independence of the two observations, obs and obs_new enables to use the result of the former computation, $p(\text{cause}_i|\text{obs})$, as a prior for the latter, and this can go on as new information are added. This process makes this method particularly appropriate to data acquired sequentially, where intermediate decisions have to be taken in parallel to data acquisition.

Using probabilities to represent beliefs and combine them through Bayes theorem has been shown by Cox's theorem to correspond to what one can expect from an optimal "logical" reasoning system (Cox, 1946), (Van Horn, 2003). Therefore, Bayesian inference refers as a good paradigm to reason in an uncertain environment, and consequently is a good hypothesis to explain how the brain could perform probabilistic learning (Tenenbaum et al., 2011).

2.2 Principles of Bayesian inference

Bayesian inference can be used to model almost everything. However, three elements need to be defined to use it properly:

- The ensemble of causes: A set of possible causes well identified must be defined. Causes represent "hidden states", in the sense that they are not directly observable but that they still influence observations. It is the goal of the inference to find the hidden state that explains observations the best. In the previous example where one has to guess the origin of an English accent, causes can be nationalities, mother tongues, social origin, etc. However, it can also be the combination of several of them (one exemplar cause would then be: popular Brazilian with Hebrew as a mother tongue). The more causes there are, the more precise the result is (popular Brazilian is more precise than just Brazilian), but the more computations have to be performed to track the likelihood of all causes.

- A likelihood function for each of the causes: The definition of the ensemble of causes is crucial as for all possible causes "j", and for all observations, it is necessary to know $p(\text{obs}|\text{cause}_j)$: how likely is the observation knowing cause "j". This knowledge allows for comparing the likelihood of an observation given different causes. In the example given above (determine the origin of someone according to his accent), the likelihood function allows generating English as spoken by all nationalities considered as possible causes. The likelihood function integrates a huge quantity of information and learning it is usually a major difficulty which can prevent the use of Bayesian inference.

In the example of Thomas Bayes, the likelihood function is trivial. Indeed, for all p, the probability to win at a lottery which has a probability p is p. In machine learning applications, the likelihood

function is complex but will often be given to algorithms as an input. However, for behavioral application, the key question is to understand how brains acquire and represent likelihood functions.

-The prior: The last key element for Bayesian inference is to have a prior probability distribution on the different causes: before seeing any observation, what is a priori the current hidden state? First, observations will be combined with the prior to give a posterior probability distribution on the causes, which will serve as a prior for later acquired observations. The prior can be freely chosen albeit using accurate priors is critical in particular when there is little information in observations. In the English accent example, the prior can be the probability to meet each of the possible nationalities in the present context. An alternative would be to compute priors through another inference on the visual look of the speaker. If no relevant information is available, the prior can also be defined as uniform on the different possible causes. In all cases the prior influences the final result: if one expects someone to be Spanish and not Italian, and that at the same time his accent is ambiguous, he will be classified as Spanish. And the converse if the speaker is expected to be Italian despite the observation is identical.

Thomas Bayes used a uniform prior on the probability of success at a lottery, what was a posteriori justified by Laplace as being the best prior to use when no information is available.

The combination of the likelihood function and the prior constitutes a generative model. It includes all the information about the environment. Thus, the generative model allows a complete simulation of the environment in particular the generation of fictive observations.

In its formula, Bayes theorem gives the same weight to both the prior and the likelihood function. Therefore, the posterior distribution is driven by the more precise of these two functions. Thus, if priors are non-informative (for example a uniform prior), then posteriors will follow the likelihood function, and if observations are non-informative according to the likelihood function (equally likely for every cause), then the posteriors will be aligned to priors. The Bayesian formula accounts for uncertainty and favors sure sources of information.

To illustrate the power of the approach, there are cases where a little observation is revealed through the generative model as crucial information. For example, if Australians have a specific way of pronouncing “y”, only hearing a single word with a “y” is sufficient to be sure that the speaker is Australian. Past information has been stored in the generative model (priors and likelihood function) and Bayesian inference uses this knowledge to interpret efficiently any new observation.

Bayesian inference has been proposed to be the solution used by the brain to deal with multiple uncertain sources of information (Pouget et al., 2013). In the next part, we will go through available evidence validating this hypothesis.

2.3 Bayesian inference and human behavior

2.3.1 Perception

Since the end of the 19th century, scientists have proposed that human perception can be described by Bayesian inference (Mach and Williams, 1897). Indeed Bayesian inference answers to a fundamental problem of any perceptual system. Perception relies on receptors which sense information on dimensions which are not the one on which information will be eventually interpreted. For example, visual neurons represent information by their electrical activity, but the

relevant information is the image that they code. This poses an inverse problem: what is in the space of all possible images the most likely cause of the data coded by the receptor.

In particular, when information is compressed by the receptor, there will be several possible causes to the data and the inverse problem is underdetermined. To solve this issue, priors based on habitual scene structure or previous images will limit the number of possible interpretations of data and constrain interpretation towards the “good” one.

Bayesian algorithms have been developed in the 1980s to create an artificial visual system (Bolle and Cooper, 1984). These approaches were later demonstrated appropriate (Kersten, 1990) and successful (Bülthoff and Mallot, 1988), (Knill and Richards, 1996) at modeling human visual system.

Furthermore Bayesian inference has been evidenced to describe optimal information combination that happens between different human perceptive systems. Van Beers & colleagues showed that when two different sources of perceptive information are integrated, the weighting between the two in the final representation depends upon their relative precision. This constitutes the core of Bayes theorem (van Beers et al., 1999). Quantitative account of this weighting was demonstrated by Ernst & Banks in a visual/haptic combination (Ernst and Banks, 2002).

2.3.2 Abstract reasoning

The use of Bayesian algorithms has not only been reported at the perceptual level. The prefrontal cortex, which operates complex and abstract reasoning in human, has been described as following Bayesian inference rules. Collins & Koechlin have proposed a Bayesian model of reasoning, where an algorithm integrates information across time to find the current hidden state of the world and select accordingly the best available action. Their model treats information optimally except for an account of human memory and computational limitations (Collins and Koechlin, 2012). Neural correlates of the model evidencing a use of probabilistic reasoning have been found with functional Magnetic Resonance Imaging (fMRI). In particular, parts of the prefrontal cortex track the likelihood of the different possible hidden states (Donoso et al., 2014).

To continue on high level reasoning and abstract causes, most social behavior of individuals constitute as well a natural field where Bayesian inference principles can be applied. Indeed social interactions rely on the understanding of intentions of others. Understanding others allows an adequate interpretation of their behavior and therefore an appropriate reaction. Intentions can be considered as hidden states that behavioral and contextual observations help to determine. For example, the sentence “Do we have to do this?” can be interrogative or complaining. To react appropriately, one can either use the tone with which the sentence was pronounced, or the facial emotion displayed when it was pronounced, or both.

In general this ability to decode others intentions is named theory of mind (Premack and Woodruff, 1978). Baker, Saxe & Tenenbaum proposed a Bayesian model of the use of others’ actions to understand their beliefs and desires (Baker et al., 2011). The authors confirmed their model experimentally and showed that human data fit to model predictions and are therefore close to optimal rational behavior. Similar optimality results were found in economic games bearing uncertainty on the intention of other actors (McKelvey and Palfrey, 1992), or on tasks where language ambiguities have to be decoded (Jurafsky, 1996).

These multiple examples of brain mechanisms which can be modeled by Bayesian algorithms suggest it is a general information treatment mechanism in the brain. The resulting questions are 1/whether it is possible for neurons to encode probability distributions and to apply simply the Bayes theorem, and 2/are there neuronal evidences that it is happening?

2.3.3 Neuronal implementation of Bayesian inference

Instead of a fixed value, several authors have proposed that neurons could encode the probability (Anastasio et al., 2000), or the log-probability (Barlow, 1969) of an event. The coding in log probability is particularly interesting as multiplications in Bayes theorem become summations and are simpler to perform neuronally.

Alternatively, it is also possible to encode full probability distributions at the neuronal population level. Indeed, when several noisy neurons encode the same variable, the resulting activity of the population can be interpreted as a probability distribution over this variable (Zemel et al., 1998). Noisy representations of variables have particular properties which make them particularly suited to neuronal combination and Bayesian inference is easy to perform with population codes (Ma et al., 2006). Neural correlates of a Bayesian combination of population codes were found in neuronal recordings of monkeys (Fetsch et al., 2011).

Another different way to easily perform Bayesian inference with neurons is to mimic Monte Carlo methods. Indeed, each neuron of a population can be seen as coding a drawing (sample) of a distribution. Bayes theorem will then apply to all the individual samples to obtain eventually a sampled representation of the posterior distribution (Hoyer and Hyvarinen, 2003).

To justify the presence and the evolutionary maintenance of Bayesian inference in the brain, we presented that beyond perceptual problems, Bayesian inference allows complex abstract reasoning and is crucial in social contexts. It indeed allows individuals to understand each other in a group and is then decisive for social cohesion. A major hypothesis in evolution explains the development of higher cognitive abilities and “intelligence” as an evolutionary pressure due to the social challenges to live in a group (Jolly, 1966), (Humphrey, 1976). As an example, not understanding the negative intention of a pair diminishes the individual survival chances. On the contrary, overestimating negative intentions may lead to anti-social behavior and to an exclusion from the group. Exclusion also diminishes chances of survival. With this hypothesis, sophisticated problem solving capacities (like chess playing) would be a by-product of social abilities needed to understand others (Dunbar, 1998).

After this presentation, major questions are still remaining: Bayesian inference describes how priors and likelihood functions are combined after an observation, but does not tell whether these two basic elements are innate or acquired. Thus, if they are innate why are they evolutionarily relevant and to which implicit hypotheses on the world are they equivalent to? And if they are acquired, what is the acquisition mechanism?

This is a problem of Learning and we will discuss it in the next part.

2.4 Learning a generative model

2.4.1 Comparison of human performance and state of the art algorithms

Amongst the many applications of Bayesian inference, the problem of learning generative models came from studies on how children learn language. The scientific study of language is inherited from Quine and can be presented as an induction problem (Quine, 1960). Indeed children have to use the adequate word, which can be seen as a hidden state, given the situation. The classical observation is that humans are very good at word learning and a few labeled examples are enough to generalize to other instances of the category (Xu and Tenenbaum, 2007).

This is considered as exceptional and impossible to reproduce on artificial learning systems. Indeed, when only a few examples of a word are presented, there are necessarily ambiguities. For example to learn what is a “dog”, only some breeds of dogs will be shown, so the learner has to interpret by itself that “4 legged animals” is too wide and “Dalmatians” is too narrow. Experimental evidence shows that humans perform this task correctly (Carey, 1978).

Human performance outperforms all computer algorithms on word learning or classification tasks. In one example, humans and algorithms had to learn the name of 256 objects. After learning, humans recognized 56% of objects (Eitz et al., 2012), while the best algorithm was at 33% (Borji and Itti, 2014).

It can be argued that humans are favored in this task as they already know thousands of words and can therefore use this prior knowledge to ease learning. François Fleuret and colleagues then designed an inference based scene classification task with abstract images so that prior knowledge was irrelevant and could not be applied. After a few examples, humans and computers had to classify abstract test images in different categories. With 20 exemplar images humans were able to solve most of the problems, while computers were at chance level. Computers needed 10'000 examples to solve some of the problems and their performance was still not close to human performance (Fleuret et al., 2011).

In this task, Bayesian inference computes $p(\text{category} | \text{image})$ thanks to the generative model $p(\text{image} | \text{category})$. Since images and categories are abstract, the generative model is learned only during the presentation of labelled images. How human can be so efficient in this learning? Understanding it is a challenge for Psychology and has potential applications in machine learning and robotics.

In the case of language as in any learning setting, the generic problem is that there are many candidate categories which would match with provided examples and all possibilities cannot be tested exhaustively. The problem is underdetermined. Braun & al describe how the human mind could solve this issue and they illustrate their intuition by a parallel with a simple example in motor learning (Braun et al., 2010).

To produce a target movement, all muscles and all articulations could theoretically be used in all their degrees of freedom. However, only a limited combination of muscles and articulations produce a meaningful action and limiting exploration to this subset accelerates the finding of the correct movement. It is the same when an animal on a picture is named “dog”. A huge number of implicit assumptions about scene organization, pointing, and animal generic definition limit the set of

regularities which will be associated with the “dog” category. Without these assumptions, as for machine learning algorithms, a lot of images would be needed to constrain sufficiently the interpretation to a unique well-specified category.

Generic assumptions about the environment can be learned on a longer time scale and are not specific to a particular learning problem. They eventually capture the implicit and explicit structure of the environment and can be used to constrain any subsequent learning to the smallest solution space where these assumptions are valid.

To illustrate the interest and the potential advantages of structure learning, a machine learning team at Google used their super powerful calculators to find regularities on millions of random images picked from the internet (Le et al., 2013). This can be seen as analog to a young child looking around him for several years. The algorithm is an unsupervised neuronal network and after training each neuron responds to one of the spotted regularities. A few labeled images of say human faces were then presented to the algorithm. This is analog to a child who would be presented with a few exemplar images of a new category. Among all neurons of the net, some were responding particularly well to the human faces presented. These neurons were selected and new images (human faces and other unrelated images) were later presented to the network. The main result is that these selected neurons were detecting reliably human faces. Critically only a few labeled images were presented to the algorithm, what is far less costly to produce than the 10,000 labeled images that would have been needed without the first phase of unsupervised learning. This is a method of self-taught learning (Raina et al., 2007), which proves the validity of the concept of structure learning. In the case of natural images structure represents for example contours, principal colors, usual shapes...

2.4.2 Hierarchical Bayesian models

The algorithm used by the Google team is well understood, but it is interesting to understand how the brain routinely extracts structure all along life. One influential hypothesis proposes the use of Bayesian inference and in particular, hierarchical Bayesian models at this end (Tenenbaum et al., 2011). The general idea is that on very long time scales, knowledge is accumulated to learn how to spot statistical regularities (causal models (Tenenbaum and Griffiths, 2001a)) and how to appropriately transfer knowledge from one problem to the other (generalization (Tenenbaum and Griffiths, 2001b)). Indeed there are always different ways to generalize, or different types of causal structures. Evidence is accumulated to select in a particular environment the appropriate generalization and causal structure. At a lower hierarchical level, these structures are applied to the building of the current generative model which can be tuned faster.

For example, in animal naming, a natural structure is to separate ground animals from flying animals from sea animals. Then among ground animal, separate oviparous from mammals. Then among mammals, bipeds from quadrupeds; among quadrupeds, urban, savannah and forest animals. Among urban animals, dogs, cats, etc.

When several examples are shown to instruct a new category, localizing these examples in the pre-learned structure enables to extract easily what is the general category that needs to be named (Xu and Tenenbaum, 2007). The generalization is also immediate: if a property is shared by several ground mammals, it is reasonable to generalize the property to all of them.

Critically, these structures are a model of the environment and can be erroneous. If so there will be contradictions in subsequent learning. At some point the hierarchical Bayesian model shall spot that it is stuck and shall correct the structure which is located at a higher hierarchical level.

There are at least two interesting questions. Which proportion of the structures routinely used is acquired across life versus innate? Indeed some of the structures of the world are sufficiently robust and could have been programmed evolutionarily. Secondly, as using these structures simplifies learning, exploration.... can we change them into an equivalent list of efficient heuristics about the world? This list could in the future be implemented in a robot to improve its learning abilities.

2.4.3 Generalization by continuity

To conclude this part and partially answer these questions on at least one point, I would like to discuss one form of generalization which takes place in continuous environments. Indeed Bayesian algorithms work very well on discrete environments as each of the discrete value can represent an event or a state and inference performed on it. For example it is possible to compute the influence of multiple factors on the event "being late".

However, when the variable varies on a continuous scale the number of events is multiplied. Theoretically, being 5 minutes late should be different than being 4 minutes 59 seconds late or being 5 minutes 01 seconds late. To group these events together, as we would naturally do, one should exhibit an appropriate generalization principle.

Pavlov already noticed in his experiments on dogs that when a precise stimulus was conditioned to reward, "neighbor" stimuli were also conditioned (Pavlov, 1927 p.113). For example if the conditioned stimulus was a tone with a precise pitch, pitches around it acquired similar properties. Besides he noted that conditioning strength decreased with the distance to the initial pitch.

Shepard proposed in 1987 a universal law of continuous generalization valid for all psychological variables. In Shepard's idea, each dimension of the physical or mental world has an associated psychological space where a metric can be defined. Knowing this metric a stimulus would "generalize" to its surround with an exponential decay (Shepard and others, 1987).

Shepard mainly described the generalization of a single event. Tenenbaum & Griffiths extended the question to understand how multiple examples can be generalized (Tenenbaum and Griffiths, 2001b). They gave the following example: if 3 observations have 60, 30 and 50 as a value, what would be the most surprising between observing 47 and observing 80?

The authors proposed that Bayesian inference could be used to allow generalizations based on other "rules" than a simple linear distance. One of these rules could for example be: only multiples of 10. Among the possible rules, selection would be based on the computation of the likelihood of each of them with a prior which would enhance "common" rules. For example "multiples of 10" would be more common than "multiples of 10 and 6 and 17". However the authors did not propose an algorithm to compute these priors on rules.

2.4.4 Limits of Bayesian inference

Bayesian inference is a general framework which has the ambition to explain all features of cognition (Tenenbaum et al., 2011). However, several critics are regularly formulated concerning this approach.

In particular, there are so many possibilities in the implementation of Bayesian inference that saying that a behavior can be described by Bayesian inference is almost a tautology. Priors, generative models and the precision of inference are not directly observable. Thus, it is always possible to account for a behavior with Bayesian inference by considering a set of (false) priors, a particular (false) generative model, and noisy inference. This does not show that there is a brain mechanism implementing the Bayes theorem (Jones and Love, 2011).

Besides, except a few specific examples, in particular when generative models and priors are Gaussians (more generally when these two distributions are conjugated), the computations associated with Bayesian inference are usually intractable. Modern mathematical methods, like particle filtering or Monte Carlo Markov Chain give an approximate solution to the problem, but they are particularly difficult and heavy to implement.

Additionally, authors including Bayesian defendants have pointed out that this theory gives a great theoretical weight to priors, without proposing a common mechanism to generate them. Collins & Koechlin propose that priors would come from a weighted average of long term memory: the more a hidden state has been encountered the highest is its prior (Collins and Koechlin, 2012).

Lastly, the philosophy of Bayesian inference considers that human are rational and perform optimal computations. The source of variability in behavior would be only due to a misperception of probabilities. However, several studies have explicitly compared the predictions of a Bayesian optimal model to what a simpler association learning algorithm would do. The usual result is that optimality cannot account for behavioral data (see here for classification of the deviations (Sakamoto et al., 2008)).

These deviations to optimality have generated several explanations. Some authors suggested that humans value priors more than what Bayesian inference predicts (Phillips and Edwards, 1966). Others have described a non-optimal balance between exploration and exploitation (Acuna and Schrater, 2008). These explanations keep the principle of Bayesian inference but add biases in the algorithmic implementation.

It has also been suggested that animals and humans have several parallel decision systems to guarantee a minimal level of reward, and that the one obeying to Bayesian inference is only one of them. Another faster and simpler decision system could take over in habitual or perceived-unsolvable tasks (Kahneman and Frederick, 2002).

We will discuss in the last part of this introduction how a generative models of reward can be constructed. As we saw in the first part, reward is a continuous variable which is particularly important as it drives behavior and learning. We will study the interaction between the learning driven by obtained reward and the statistical learning induced by Bayesian inference.

3 Bayes and rewards

3.1 Theoretical introduction to the exploration/exploitation dilemma

At the end of chapter I, we explained that theoretically, when an environment is stable, reinforcement learning algorithms are guaranteed given minor conditions to converge to true action values. Therefore, at the limit, selecting action with the largest action value is the optimal strategy.

However the “limit” condition is theoretical, and is never attained in real cases. There are always a finite number of trials, and sometimes the environment changes. When rewards need to be maximized on a finite number of trials (before the end of the task or before the next environmental change), the optimal solution is different: the learning of option values and the accumulation of reward need to be balanced.

This is known as the trade-off between exploration and exploitation. By exploitation we mean choosing the option which has currently the highest perceived expected value. By exploration, we mean choosing any other option, in the hope to get information which would help gaining more reward on the long term. In particular, while all options have not been tried, there is always a possibility that the best one has not yet been chosen. And when rewards are stochastic, the true value of options is permanently uncertain. Controlling time spent on exploration is a fundamental problem which emerges when rewards need to be maximized within a limited time frame.

The general problem of exploration/exploitation was first introduced in medicine by Dr Thomson. Dr Thomson was wondering which drug to deliver patients knowing the number of successes and failures of each drug on other patients. In particular, he wondered whether there are cases where delivering another drug than the one with the current highest success rate is optimal, in order to gain information which could potentially reveal a new “best” drug. If so, all following patients could benefit from better treatment (Thompson, 1933).

Herbert Robbins formalized the question as a problem of sequential decision making. A decision is taken at each trial about whether to explore or to exploit. This decision depends on the former samples and influences following samples. He described the problem of “bandit” as a simple illustration (Robbins, 1952).

3.2 Bandits problems

In Multi-armed bandit tasks, participants can choose between different slot machines. Choosing a machine draws a reward (binary discrete or continuous) from a machine-specific reward distribution. The goal is to get as much reward as possible during a limited number of trials. Reward distributions are hidden, and according to the experiment can be either:

-stationary: they remain constant all across the game

-changing: in the general case, any machine’s distribution can change at any time without notice (restless). In some design while a machine is not chosen its associated distribution cannot change

-reversing: machine distributions stay stationary but mapping between machines and distributions can change.

“Bandits” refers to slot machines in Casino, but can be seen as a model of several real life problems. For example “foraging” is the concept of looking for food in different places. Each food source has a probability distribution on the quantity of food it provides and this probability can vary across time, due for example to meteorological reasons or to overuse. It is usual to model food sources as bandits. In the same way shops, writers, or holiday destinations can be seen as a variable source of “satisfaction” and modeled as bandits.

In psychology, tasks analog to bandits were tested on animals since 1939 (Brunswik, 1939). The idea was to study how stochasticity affects the learning of stimulus response patterns. Hereafter, mainly economists started using these tasks routinely to measure how human behavior deviates from optimal decision algorithms (Horowitz, 1973).

In the following section we will focus on bandit problems as a paradigm of decision making problems. First, we will review normative solutions to these kinds of problems. Then, we will focus on animal and human data and see how they solve bandit problems.

3.2.1 Normative approaches of bandit problems

Statisticians have developed different normative approaches to solve bandit problems. The general theoretical problem is to find a good measure to quantify information and reward on the same value scale (Howard, 1966). Thus, bandits can be ordered by this concatenated value, and the optimal policy is to choose the best bandit.

In some versions of the bandit problem, there is a known optimal solution. For example in the stationary problem with n trials (Kaelbling et al., 1996), there is a solution based on dynamic programming. It uses the Bellman equation to solve iteratively the problem (Bellman, 1957). However, the complexity of the computation grows exponentially with n and with the number of bandits. Therefore, the solution is generally intractable (Lusena et al., 2001).

Gittins showed that there is also an optimal solution in another version of the bandit problem: when the horizon is infinite, rewards temporally discounted, and bandit's distributions can only change when the bandit is chosen (Gittins, 1979). In this case, bandits can be ordered by a "Gittins index" measuring an integrated value of future reward and information. However, except in trivial cases, computing the index is difficult.

Besides, when real life situations are modeled as bandit problem, it is limiting to impose stationarity of unchosen bandits. For example, when looking for the best restaurant in an area, cooks of visited and unvisited restaurants can change any time. Similarly, the implementation of the dynamic programming solution requires knowing in advance the total number of trials and that is usually not the case.

Whittle extended Gittins indexes to the "non-stationary" (restless) case but it does not constitute an optimal solution (Whittle, 1988). Indeed the "restless" version of bandit problem is a NP-hard problem (Papadimitriou and Tsitsiklis, 1999), and there is no known general solution (Lusena et al., 2001).

Since most of the problems have either no optimal solution or an intractable optimal solution, statisticians started to work on "good enough" strategies, which operate a trade-off between computational complexity (measured for example by the necessary time to converge to a solution) and the "quality" of performance. To measure the quality of performance, two approaches exist in the line of the traditional frequentist versus Bayesian view of probabilities.

In the frequentist approaches the performance of a strategy is measured on a specific bandit problem given the reward distributions of the problem. Lai & Robbins gave a theoretical bound on the best performance achievable (Lai and Robbins, 1985). The resulting challenge is to find the easiest algorithm able to get as close to this performance as possible (for example the upper

confidence bound algorithm (Agrawal, 1995)). In a second step, it is important to check whether the algorithm found shows similar performance when reward distributions are different. If so the algorithm is not “over-fitted” to a specific version of the bandit problem.

In the Bayesian approach, the performance of a strategy is measured globally on the space of possible reward distributions. A prior defines the likelihood of each distribution in the considered environment. This prior weights the importance of the quality of the proposed strategy on this particular distribution in the global performance measure. Therefore, the best algorithm should perform well in all common settings, while its performance in “pathological” cases will count less. The upper confidence bound algorithm has been extended to reach the best performance bound on bandit problems where attributed rewards are drawn from Gaussian (Srinivas et al., 2012) or Binary (Kaufmann et al., 2012) distributions.

The lack of a general solution to the bandit problem illustrates its complexity. Human and animals are however able to interact satisfyingly with a complex and uncertain environment. They can sometimes develop strategies whose performance comes close to the theoretical optimal solution (Krebs et al., 1978). Thus, one can ask: what is the “good-enough” strategy that humans and animals are using?

3.2.2 Behavioral account of the exploration-exploitation trade off

Many families of algorithms can account for a trade-off between exploration and exploitation. We described just above complex algorithms aiming at valuing information to perform optimal choices on the long term. However, a trade-off between exploration and exploitation can also emerge from information-free algorithms based on heuristics. One of the simplest strategy we can imagine is to keep the same bandit when a reward is obtained and switching at random to another bandit when no reward is obtained (Win Stay, Lose Switch) (Robbins, 1952). Here there is neither computation of option value nor combination with exploration value. Instead the algorithm uses a heuristic.

By “heuristic” we mean a simplification of the problem by using additional hypotheses which are true in general, but for which there exist counter-examples. To give a simple example, instead of forecasting weather in the Sahara desert, one can use the heuristic that it is always sunny. It is not always true, but can be a good enough approximation which sometimes performs better than other more complicated strategies.

Several authors have modeled human behavior on bandit tasks to find which algorithm is driving exploration. We will review the different propositions from the most complex, close to the theoretical optimal solution, to the simplest entirely based on heuristics.

3.2.2.1 Informational account

According to Reverdy & colleagues humans use a Bayesian upper confidence bound algorithm. As explained in the former part, this algorithm reaches asymptotically the optimal performance bound. They report that some subjects in their data set have suboptimal performance, but according to them, this is because the Bayesian algorithm uses false priors. They do not compare different models, but show that for each subject, there is a set of priors and free parameters with which the optimal algorithm matches behavior (Reverdy et al., 2014)

Daniel Acuna and Paul Schrater on their side show that human behavior is compatible with the computation of an optimal index to rank bandits like the Gittins index. However they propose that

humans are limited in their memory and forecast abilities, so that the computation of the index is based only on a limited number of past observations, and optimize reward only a few steps ahead (Acuna and Schrater, 2008). Therefore, humans would be optimal on tasks containing a small number of trials, but are intrinsically sub-optimal on longer tasks.

The idea that human undervalues the long term future gain, for example by computing gain only on a few steps ahead, is common in the literature. This “myopia” (cannot see far ahead) undervalues information, and causes overall under-exploration. Horowitz reported it in his PhD thesis (Horowitz, 1973), and the experimental results of Meyer & Shi (Meyer and Shi, 1995) and Anderson (Anderson, 2001) are well explained by myopic under-exploration.

In particular, one myopic choice rule named “Knowledge Gradient” considers the fictive situation where exploration is only possible at the present trial. Subsequent choices will be only exploitative. This is different from the computation of Gittins indexes where it is assumed that it is possible to explore on any future trial. The myopic rule of “Knowledge Gradient” is fictive, so that at the next trial again the same rule is applied. But this simplifies grandly computations as only two alternative course of action need to be compared:

-Explore, Exploit, Exploit, Exploit ...

-Exploit, Exploit, Exploit, Exploit...

This choice rule has been shown to asymptotically converge to the optimal performance bound, and it is of course optimal when only one trial is left. When the time frame is finite “Knowledge Gradient” is a good approximation of the optimal policy (Frazier et al., 2008). Shang & Yu showed that this myopic choice-rule associated with an optimal learning Bayesian algorithm can account for human exploratory behavior in a binary bandit task (Zhang and Yu, 2013).

3.2.2.2 Heuristic account

Beyond these accounts of near-optimality, other research teams have interpreted from their behavioral data that there is no explicit tracking of uncertainty neither a computation of the value of information. They rather defend that humans explore based on heuristics.

In particular, the Q Learning is an information-free algorithm that can be modified to include a choice rule which balances exploration and exploitation. For example, instead of selecting the bandit with the highest expected value (“best bandit”) in every trial, it is possible to perform a proportion of random choices. With probability $1-\varepsilon$ the “best” bandit is selected and with a probability ε , a random bandit is selected (Watkins, 1989). This selection rule is named ε -greedy, and will lead on the long term to sample all possible bandits. In stationary environments, this is a sufficient condition for the convergence of the Q Learning towards the true action values, and consequently the best bandit will eventually be found.

The problem with this heuristic is that exploration is uniform. Indeed, the second-best bandit has the same probability to be selected than the worst bandit. However, in general, the second-best bandit has a higher potential to be eventually revealed as the best bandit. And the goal of exploration is precisely to spot the best bandit as it is the one which will be exploited in the long term. To this end it is more interesting to explore more the second best bandit (furthermore, in short term value, one loses less by choosing the second best bandit compared to the worse bandit). Exploring bandits

proportionally to their current value heuristically mimics what an algorithm valuing exploration would do without performing the complex and heavy computations. This heuristic describes the principle of soft-max selection (or Boltzmann exploration) (Luce, 1959).

To improve again the soft-max selection rule, it is possible to include another heuristic: the more a bandit was chosen, the more certain is its value, therefore the less it needs to be explored. This corresponds to giving a value bonus to ambiguous bandits and can be implemented simply with a count of the occurrences of each choice (Wilson et al., 2011).

Nathaniel Daw and colleagues compared these 3 choice rules, ϵ -greedy, soft-max and “soft-max + ambiguity bonus” on human behavioral data. The soft-max decision rule best explained behavioral data on a 4 bandits task (Daw et al., 2006).

Other heuristics have been described in the literature. Noah Gans & al compared a family of model with increasing complexity on a set of behavioral data. They show that the simplest of all models namely a satisfaction index explained their data the best (Gans et al., 2007). The satisfaction index is an extension of the Win Stay Lose Switch model proposed by Bush and Mosteller (Bush and Mosteller, 1955). This model only tracks a probability to stay and to switch which depends on a linear combination of recent outcomes (in WSLS this probability depends only on the last outcome). This rule had already been found to explain economic observations at the population level (Schmalensee, 1975).

To reconcile the different views, Mark Steyvers and al. proposed in another study that there is individual variability in the algorithm used by humans to solve the exploration-exploitation trade-off. Thus, they describe the behavior of their participants on a continuous scale of complexity: from simple heuristically based behavior to complex computationally optimal behavior (Steyvers et al., 2009).

We can conclude from this set of studies that human behavior is sub-optimal. This is expected as we explained above that optimal algorithms are intractable. To simplify opportunely the computations, different characteristics of the environment can be exploited. This is the role of heuristics which are selected by personal experience or across generations through evolution. The selection of a heuristic balances its gain in computational cost and its loss in performance. In this sense heuristics describe another form of optimality: the best performance achievable given limited computational resources.

3.3 Structure Learning

To accelerate and direct exploration we described in chapter II the interest of learning environment’s regularities. These regularities are statistical dependencies which constitute a structure and can convert observations into information about the task. A structure allows more information to be extracted from each observation and therefore exploration can be shorter; more trials can be devoted to exploitation, and this increases the total accumulated reward. For example, if two bandits are correlated, only one of them needs to be explored. Observations of the first bandit inform the second bandit.

Acquiring the statistical dependencies constituting a structure is long. In the idealized learning problem all the information is available at each trial. In particular the rewards attributed by both chosen and unchosen bandits are displayed. It is well known that humans are able to use information

from non-obtained rewards as showed by the phenomenon of regret (difference between potential and obtained reward) (Lohrenz et al., 2007), (Hayden et al., 2009). Here human would use non-obtained rewards to acquire the task structure. However in the general bandit problem, one only knows the reward he obtained (chosen bandit). Thus, explorative suboptimal choices need to be performed to spot or test regularities. This requires a trade-off between structure learning and exploitation.

There are two main kinds of structures that are relevant for bandit problems.

-First a temporal structure: the reward given by a particular bandit at a given trial can be correlated with the reward given by the same bandit at the next trial. This is for example the case when bandit draws rewards from a stable distribution. Other example: the obtained reward follows a stable drift (performance of a saving with a fix interest rate). It is the temporal structure which allows anticipating the reward of a bandit from its past outcomes. It is for example relevant in a foraging context: if there are fruits on a tree one day and they are not all picked, there are good chances to find them again the next day.

-Secondly, a “spatial” structure: a unique variable at a higher hierarchical level influences two or more bandits (Figure 6). This is for example the case when the rewards attributed by two bandits are correlated. For example the flows of two water sources which come from the same river are correlated. In the same way, in a seasonal environment the quality of many food and water sources depends upon whether it is summer or winter. The river or the season can be model as (hidden) environmental states. Environmental states cause observations, and observations enable to infer the current environmental state from which future rewards of other bandits can be predicted.

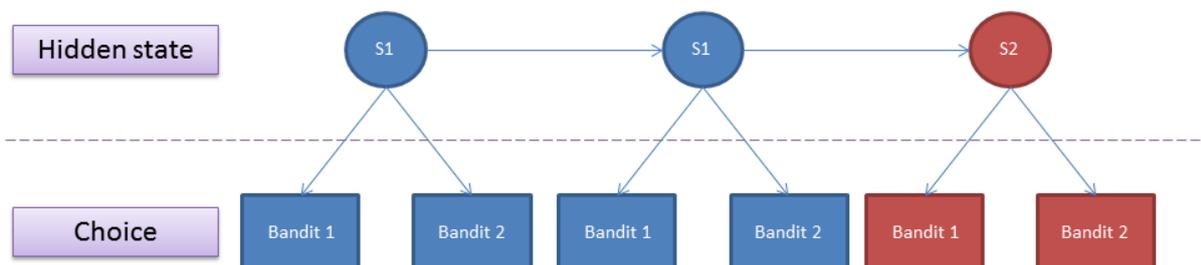


Figure 6: Graphical model of a simple hierarchical structure. Bandits’ reward distribution depend on a same hidden state. This hidden state has also a temporal structure and changes across time.

Human participants are able to use a structure which is explicitly explained to them. For example Elise Payzan-LeNestour and Peter Bossaerts proposed a 6 restless bandit task to participants which were instructed the spatial structure of the task. They show that participants were able to use the structure optimally (Payzan-LeNestour, 2010). However learning the structure of a task is complex, in particular when no prior is available on the general shape of the structure and that learning starts from scratch. In a follow-up experiment, the same authors did not give structural information to their participants and they evidenced that at the end of the experiment, participants had learned another structure than the one of the design (Payzan-LeNestour and Bossaerts, 2011).

This is particularly important in terms of performance. Reverdy et al. measured on their human data that participants which used the right task structure got close to the best theoretical performance on

the task (Lai & Robbins bound: the number of suboptimal choices grows logarithmically with the number of trials). On the other hand, participants using a false structure did not improve during the task and had a non-decreasing probability of doing a suboptimal choice at each trial.

3.3.1 Learning simple correlations

Daniel Acuna and Paul Schrater showed that humans were able to react to a recurrent correlational structure. They proposed to their participants two blocks of 16 bandit games in a row. In each game the reward probabilities of the bandits were different but in one block the probabilities of the two bandits were anti-correlated, in another block, they were independent. They compared human performance to an optimal model able to test and exploit the correlation between the bandits. Participants were able to spot the structure when it is present and to adapt their explorative behavior to account the anti-correlation when it is appropriate (Acuña and Schrater, 2010).

For example in a 2 bandits task, if a strict anti-correlation is found between the two bandits, measuring a reward rate of 70% in one bandit guarantees that it is the best one and the “worse” reward distribution can be re-constructed with a reward rate of 30%.

3.3.2 Hierarchical structure

3.3.2.1 Presentation

Hierarchical structure is the one which has been the most studied in human experiments. It comes from one of the limitations of TD learning mentioned in the first chapter: it does not predict the phenomenon of facilitated reacquisition. As a quick recall, facilitated reacquisition happens when after an association was learned, an extinction procedure is operated. According to TD learning, the association disappears and the situation comes back to as before the initial learning. Experimentally however, animals and humans relearn faster than the first time when the association comes back. We will first show how a hierarchical structure can account for this phenomenon, then discuss behavioral evidences that humans and animals can implement this type of structure. Lastly, we will describe the simplest experimental paradigm probing the use of a hierarchical structure: reversal learning.

3.3.2.2 Hierarchical learning and facilitated reacquisition

Here, the general idea is to explain facilitated reacquisition through the construction of a “learning set” which is constructed during the learning phase, put aside during the extinction, but recuperated when the association comes back (Harlow, 1949).

When different tasks are successively proposed, animals and humans are able to implement task-specific adapted strategies. A simpler alternative would be to monitor a unique strategy which would be constantly adapted to the current task by a mechanism similar to TD Learning. In order to be able to use a specific strategy for each task, a control mechanism is needed to recognize the current task and select the appropriate strategy from a repertoire. Experimental proofs of the existence of a strategy selection system come from an analysis of reaction times when two tasks are intertwined. Participants are slower at the first trial following a task change whatever their level of practice, suggesting the existence of a switch cost to reconfigure the strategy (Rogers and Monsell, 1995).

However, even when there is no explicit cued tasks, humans and animals are able to use a repertoire of strategies. They will find among all environmental cues the one which are behaviorally relevant and use these cues to select an appropriate strategy (Redish et al., 2007), (Collins and Frank, 2013). A

“task set” is defined as the subset of environmental states where one particular strategy is appropriate. This is an extension of the notion of task-specific strategy when tasks are not explicitly cued. Task sets have to be constructed and this discretizes the environment as all possible states will be classified in one of the constructed set. According to the current environmental state, a similar stimulus can trigger different behavioral responses (Koechlin et al., 2003).

Thus, we can explain the phenomenon of facilitated reacquisition by the use of task sets. When a strategy which was working suddenly stops, it suggests a change in the environment. Instead of readapting the now inefficient strategy, it can be saved in memory in the hope to be used again in the future. A new task set is created and a new strategy starts to adapt. If environments are recurrent, this memorization will be useful in the future. If environments are not recurrent, except the cost in memory, it is in average equivalent to adapt from any strategy so there is no opportunity cost to falsely suppose a recurrence.

“Task sets” makes the implicit assumption that in the environment, there is a latent cause which explains why a strategy works (Courville et al., 2004). It departs from classical conditioning where the stimulus is supposed to cause the reward. In the latent cause model, both the stimulus and the reward are caused by a hidden state (here a latent cause), located at a higher hierarchical level (Courville, 2006).

3.3.2.3 *Creating task sets*

To perform structure learning, each state of the world has to be attached to an existing task set or trigger the creation of a new task set. This is a complicated problem of unsupervised learning to which answer Dirichlet process mixtures (Gershman and Niv, 2010). Dirichlet process mixtures use Bayesian inference to compute the likelihood that each task set generated the observations, but include the possibility to create a new task set if none of the existing task sets is likely enough. Unfortunately, the general solution is intractable as after each new observation, all past decisions should be reconsidered. In particular the extent of task sets might need to be reconfigured and some already encountered environmental states moved from one to another task set (Daw and Courville, 2008).

A simple tractable algorithm approximating a Dirichlet process was recently proposed. This algorithm adapts the optimal strategy of each task set through reinforcement learning. It monitors only a limited number of task sets in parallel. And the only decision that can be a posteriori revised is the creation of a new task set (Collins and Koechlin, 2012). This algorithm accounts for human data in a simple behavioral experiment where participants had to find the good mapping between stimulus and responses. The mapping could change from time to time and the new mapping was either already met (recurring task set), or never encountered before (new task set).

A key problem underlying the decision to change task set or to create a new task set is the ability to separate the “natural” variability in the efficiency of a strategy from the variability due to a change of task set (expected vs unexpected variability) (Yu and Dayan, 2005). For example in a task set, a bandit can be rewarded in 70% of the trials. Despite the regular negative prediction errors when no reward is obtained, it is still the same task set. The local perceived success rate can statistically vary due to noise. What is the lowest acceptable reward rate, under which the inefficiency of a strategy is more likely to be due to a change in the environment? Changing too often or too late is maladaptive.

To react appropriately to statistical variability it is important to have a good estimation of the volatility of task sets. Volatility measures how often task sets usually change. If task sets are known to change regularly (volatile environment), even little statistical irregularities shall be interpreted as the sign of a change of task set. If task sets are known to change rarely (stable environment) stronger irregularities and more evidence in favor of a change will be necessary. In a bandit task, Tim Behrens and his colleagues showed that humans are able to infer the volatility of their environment, and to adjust learning accordingly (Behrens et al., 2007). For example, the learning rate of a TD Learning algorithm could be correlated to the inferred volatility. When the environment is volatile recent information has more weight in the decision (high learning rate). When the environment is stable weight can be spread on the full history (low learning rate).

We will conclude this introduction by studying a famous experimental paradigm which uses a very simple hierarchical structure.

3.3.3 Reversal Learning

3.3.3.1 *Presentation*

We will present a particular bandit problem named “Reversal Learning”. There are two (resp. n) bandits and two (resp. n) reward distributions. The “hidden state” is the mapping between bandits and distributions. There are $2!$ (resp. $n!$) possible mappings. One distribution has a higher expected value than the others; therefore the mapping determines also the identity of the current best bandit, and what the current best choice is. Usually the mapping changes without notice during the task and participants have to adapt.

Historically, reversal tasks have been studied for a long time as a test for inhibitive behavior. Indeed, after a reversal, the former best bandit is not the best anymore, and the previously learned response has to be inhibited. Performances of humans and animals have been compared according to their personality traits (Izquierdo and Jentsch, 2012), or their pathology (Swainson et al., 2000). It is now studied as a simple paradigm to understand how humans and animals infer the structure of the task based on the received rewards.

On a reversal task, TD learning algorithm learns and unlearns bandit’s values after each reversal. It will eventually reverse its choice towards the new “best bandit”, but slowly. When a hierarchical representation is applied to the task, 2 (resp. n) task sets will be defined: one per bandit so that in task set i choosing bandit i is the best choice. Observations will be combined to find the most likely task set and thus select the best bandit.

As expected, humans are able to use an instructed hierarchical structure. In a reversal task with two bandits, Hampton & colleagues instructed their participants about the existence of reversals and showed that humans reverse faster than predicted by TD Learning (Hampton et al., 2006). Furthermore, without explicit instruction, this structure can be acquired as experienced monkeys display a similar behavior (Costa et al., 2015). Indeed, opposite to human subjects, monkeys cannot be instructed explicitly the structure and this results show their capacity to acquire the structure.

3.3.3.2 *Normative solution with known reward distributions*

The optimal solution to the reversal task is typically expressed as an inference problem. The task is to find the bandit which has the highest likelihood to draw from the “best” distribution (the one with the highest expected value). To keep it simple, let’s write the equations for the two bandits problem.

We have two bandits A and B, and two distributions, BEST and WORSE. BEST has the highest expected value.

Choice at time T depends on the quantities

$$p(A = BEST | History(1:T - 1)) \text{ and } p(B = BEST | History(1:T - 1))$$

They represent the “beliefs” that A or B draws from the BEST distribution.

History(1:T) stands for all the observations from the beginning of the task to T, here it is the list of actions performed and rewards obtained {reward(1),.....reward(T),action(1),.....,action(T)}.

Without loss of generality let's say that choice A is performed and reward R is obtained.

$$History(1:T) = \{History(1:T - 1), A, R\}$$

To choose at time T+1 one has to compute

$$p(A = BEST | History(1:T)) \text{ and } p(B = BEST | History(1:T))$$

According to Bayes rule

$$\begin{aligned} p(A = BEST | History(1:T)) \\ = \frac{p(A = BEST | History(1:(T - 1))) * p(History(T) | A = BEST, History(1:(T - 1)))}{p(History(T) | History(1:(T - 1)))} \end{aligned}$$

Hierarchical structures are Markovian, what simplifies the equation. The resulting system is the following

$$p(A = BEST | History(1:T)) \propto p(A = BEST | History(1:T - 1)) * p(R | BEST)$$

$$p(B = BEST | History(1:T)) \propto p(B = BEST | History(1:T - 1)) * p(R | WORSE)$$

Knowing that $p(A = BEST | History(1:T)) + p(B = BEST | History(1:T)) = 1$ gives the normalization coefficient.

The update of the beliefs depends crucially on the value of the two reward distributions, BEST and WORSE. Intuitively the probability that the chosen bandit draws from the best distribution depends on how more likely the received reward is in BEST compared to in WORSE (and it does not directly depend on how likely it is in BEST). For example a reward which is as likely in BEST and in WORSE is not informative about the mapping. On the converse a reward which has a probability 0 in one of the distribution is very informative even if it has a very low probability of occurrence in the other distribution.

3.3.3.3 Learning distributions

According to what we explained just above, to act optimally in a reversal learning task, human participants need to know or to have learned all reward distributions: the one with the highest expected value, which is sampled by exploitative choices and all the suboptimal distributions, sampled by explorative choices.

However, we mentioned before that generally human participants are exploring less than in an optimal strategy. This implies that in a reversal learning problem, participants acquire more knowledge about the best distribution than about all other distributions.

The literature has addressed this “sampling” problem. Indeed, observations and judgements are biased as the experienced feedbacks mostly come from one distribution. Negative opinions are less often revised as the responsible bandit is not sampled anymore (Denrell, 2005). The perceived average success rate is higher than its true value due to the experience of more positive events (Fiedler, 2000). To compensate, it has been suggested that humans were able to reconstruct the unobserved feedbacks from any available information (Elwin et al., 2007).

In particular structural information about the correlation between bandits enables the use of the observed reward of the chosen bandit to inform what would have been the other rewards. For example in a 2 bandits reversal task, if a strict anti-correlation exists between the two bandits, measuring a reward rate of 70% in one bandit enable to reconstruct the “worse” reward distribution with a reward rate of 30%. With this information it is then easy to find the highest rewarded bandit. Indeed, for a Bayesian algorithm it is easier to infer whether the reward rate of a bandit is more likely to be 70% or 30%, than to decide whether the reward rate of the same bandit is worth a precise value (here 70%).

In case no spatial structure can be found between bandits, some studies show that humans learn a fictive structure (Yu and Cohen, 2009). Indeed when there is no structure in the environment, nothing can be predicted, and therefore all possible strategies have a similar performance. Therefore, it is always beneficial to try to construct a structure even if it is eventually artificial.

Human participants are however limited in the complexity of the structures they can learn. For example in the study of Payzan-LeNestour & Bossaerts participants were not able to acquire by themselves the structure of their 6 bandits task. In particular, it was difficult for participants to separate residual noise from real environmental changes. Interestingly, in this task human behavior was as well modeled by an information-free reinforcement learning algorithm (Payzan-LeNestour and Bossaerts, 2011).

This can have two interpretations: either there is an explicit decision to stop structure learning when the structure is perceived as too complex, and to apply instead a simpler algorithm like Reinforcement learning. Either human structure learning starts from priors close to the implicit assumptions of Reinforcement learning which are modified only if it improves significantly performance. By “design” human would then always perform better than a Reinforcement learning algorithm.

Question and strategy to answer

In this introduction, we report many studies investigating how Bayesian algorithms can model experimental data and inform about decision making. In more than the three quarters of these studies, the experimental design uses binary rewards (Bernoulli case). To maximize reward, participants have to find the bandit which has the highest reward probability.

In general however, rewards obtained in an environment cannot be considered binary, and are measured on a continuous scale. To maximize reward on a continuous scale, one has to select the bandit having the highest expected value. In contrast to RL algorithms which work indifferently in discrete or continuous environments, several theoretical complications prevent from directly generalizing Bayesian algorithms used in binary case to the continuous case.

1. When rewards are binary, only a few samples allow an estimation of the reward probability of a bandit which is the only relevant information. In the continuous case, each individual reward is received rarely, therefore estimating the probability to obtain each individual reward is very long. Usually a generalization rule will allow clustering rewards together: for example there might be an interval under which two close rewards will be considered identical. But this interval depends itself of the structure of the environment. Winning 60 or 61 million at the lotto is equivalent, but being graded A or B at school is not. Is there a generic generalization rule used? If not how does it depend on reward history?
2. In a binary environment the best option is the one having the highest probability of giving a 1. Therefore, being rewarded after a choice is always desirable and the posterior probability to have chosen the best option will be higher than the prior probability. The converse is true for 0; the posterior probability to have chosen the best option will be lower than the prior probability. In a continuous environment, how desirable is €20? In some settings it is the best which can be expected, in others, it is a low reward. Therefore, a context-specific model is necessary to interpret reward values. A possibility is to learn the reward distributions. Once reward distributions are known, we can extend the notion of desirability to rewards which are more likely drawn in the best distribution than in the others: receiving them increases the posterior probability of being in the best option. We will refer in the following desirability as an “informative value”.
3. In a binary environment, probabilities and values are confounded. A bandit which is rewarded with a probability p has an expected value of p . We explained how Bayesian algorithms are able to easily track the reward probability of a bandit. When rewards are continuous, computing the expected value of bandits requires the computation of an integral, multiplying the probability of obtaining each reward by its magnitude ($\int proba * reward$). Importantly, expected values strongly depend on rare extreme events. If a reward of 1 million is attributed with a probability 0.001, despite its rarity it has a strong influence on the expected value. It is well known that the probabilities of extreme events are not well estimated by humans (Kahneman and Tversky, 1979), (Hertwig et al., 2004). In the first chapter moreover we described some research works questioning the human capacity to

compute expected value. How Bayesian algorithms can decide which action is the most rewarded? Do they compute expected values?

The points mentioned above raise the question whether and how humans are able to perform inference in a continuous environment?

One might solve this problem by postulating that reward distributions have a specific shape parametrized by a few parameters. For example Gaussian distributions are parametrized by their mean and standard deviation. Learning the distributions in this context reduces to finding the parameters which describe obtained rewards the best. This directly extends the case with discrete rewards in which reward distributions were implicitly parametrized by a beta distribution. Once the parameters are estimated, reward distributions are explicit and all 3 points above are easily solved.

The Gaussian shape is a natural candidate to parametrize reward distributions as samples of a noisy process converge to a Gaussian shape (Central Limit Theorem). The Gaussian distribution is moreover part of the exponential family and is particularly suited to Bayesian inference (conjugate prior). However, supposing a Gaussian shape is particularly limiting when the distributions are bi or multi-modal, or when extreme rewards are frequent (heavy tails). This contradicts strongly the bell shape of a Gaussian.

At the opposite, the most elementary solution to the continuity problem is the frequentist approach: to measure frequencies of all possible rewards and to construct step by step the associated reward distribution. However the generalization problem mentioned above holds: how to bin rewards into discrete categories as one might never receive two exactly identical rewards (continuity). There is additionally a memory and computational problem to manipulate resulting distributions as the number of bins is potentially high and could change across time with the variation of environmental contingencies.

Between these two solutions, which one is the best strategy to adopt? In our PhD work, we propose a new solution which is more general than imposing a Gaussian shape, and lighter than tracking a multinomial distribution. Instead of learning the distributions, our solution tracks directly the informative value of rewards. At this end it uses a heuristic counterfactual model able to simulate what would have been the reward obtained with other choices. This heuristic assumes that the best bandit gives a higher reward than all other bandits at every trial. Importantly, our solution does not rank distributions by their expected value and therefore no computation of expected values is needed.

We tested our model on 3 behavioral tasks on human subjects. We wanted in particular to compare our model to 3 alternative models:

1. Human perform no inference and use a Reinforcement Learning algorithm
2. Human parametrize by default distributions as Gaussians
3. Human learn distributions by counting reward frequencies

We will in the next part present in great details the different computational algorithms studied able to perform inference in a continuous environment. Then we will describe our three behavioral tasks

and the fitting methods used to measure how qualitatively and quantitatively each algorithmic model predicts human behavior.

Algorithmic Models

The goal of our work is to find an algorithmic model describing human cognitive processes underlying the learning of option's value in a continuous, uncertain and changing reward environment. These mental representations of option's value guide choice in subsequent decisions. We consider several hypotheses that we will present here with their implementation and their behavioral predictions.

Importantly, for all algorithms presented here, we consider that knowing option's value, decision follows the standard soft-max rule described earlier as modeling human exploration-exploitation trade-off in many behavioral studies (Acerbi et al., 2014), (Daw et al., 2006).

Reinforcement Learning

The default hypothesis is that in continuous environments, humans make no inference about the reward structure of the environment and learn value by a simple reinforcement Learning algorithm. Therefore choice is only based on an estimation of the mean reward of each bandit. We will specifically use the Q-Learning algorithm, for which we give the pseudo-code.

We consider n bandits T trials and $Q_i(t)$ stands for the value of bandit i at trial t .

$t=0$

For all i , $Q_i(0) = 0$, end *%Initialize Q-values*

For $t=1:T$

Choose a bandit J according to *Softmax*($Q_i(t), i = 1:n$)

Obtain Reward R

$Q_J(t + 1) = Q_J(t) + \alpha * (R - Q_J(t))$ *%Chosen bandit*

$Q_{i \neq J}(t + 1) = Q_{i \neq J}(t)$ *%Unchosen bandits*

End

Structured Reinforcement Learning

To improve the performances of Q-Learning, it is possible to include a structure which creates a dependency between the different option values. Thus, observation of the reward of chosen bandit inform about unchosen bandits value.

A simple structure is an anti-correlation structure. One can consider that the sum of the Q-values of the different bandits stays constant over time. Thus, when the value of one bandit increases, the values of other bandits decrease and conversely.

This particular structure is applied in this pseudo-code

$t=0$

For all i , $Q_i(0) = \frac{c}{n}$, end *%Initialize Q-values*

For $t=1:T$

Choose a bandit J according to $Softmax(Q_i(t), i = 1:n)$

Obtain Reward R

$Q_J(t + 1) = Q_J(t) + \alpha * (R - Q_J(t))$ %Chosen bandit

$Q_{i \neq J}(t + 1) = Q_{i \neq J}(t) - \frac{\alpha}{n-1} * (R - Q_J(t))$ %Unchosen bandits

End

Changing the value of unchosen bandits is mathematically equivalent to receiving a fictive counterfactual reward. Counterfactual rewards generally name rewards that would have been obtained if another bandit had been chosen. In the particular case of $n=2$, and applying the anti-correlation described just above, the algorithm is equivalent to the assumption that the reward of the unchosen bandit is $C-R$ where C is the constant value of $Q_1(t) + Q_2(t)$.

This implicitly implies that there is an anti-correlation between the feedbacks of the two bandits. If the obtained reward is high, the counterfactual reward is presumed low and the converse. In the literature at least two behavioral experiments confirm that humans presume a priori this anti-correlation. Participants were pessimistic concerning their own reward when they were showed that the unchosen bandit's feedback was high and optimistic when they were showed that the unchosen bandit's feedback was low (Gu et al., 2011), (Marciano-Romm et al., 2016).

This counterfactual algorithm that will be named in the following Normalized Reinforcement Learning (NRL) has been sometimes described in the literature as a contextual model where the mean reward (here $\frac{C}{2}$) is tracked and action values are represented relatively to this mean (Vlaev et al., 2011), (Palminteri et al., 2015).

To extend the model, it is possible to consider C as a free parameter of the algorithm. Theoretically setting $C = R_{min} + R_{MAX}$, with R_{min} and R_{MAX} being the bounds of the reward available in the present context, enables both bandit values to vary on the complete range $[R_{min}, R_{MAX}]$ but this implicitly supposes that reward distributions are symmetrical by $\frac{C}{2}$. In a non-symmetric environment, it is sometimes more appropriate to use another value of C which can take any value between $[2 * R_{min} \ 2 * R_{max}]$.

Hierarchical structure

In one putative model, humans learn the (latent) hierarchical structure of external contingencies (here observations). In a hierarchical structure the model assumes that the highest rewarded action is dependent on a hidden state which thus needs to be inferred. Only one hidden state is valid at each trial, and the model has a representation of the probability distribution on hidden states (probability that each hidden state is the current one). This probability distribution is updated as observations are sequentially obtained.

Bayesian algorithms are particularly suited to infer hidden states, provided there is a generative model of the environment. Here the generative model is constituted of a prior probability

distribution on hidden states, and a likelihood function:

$p(\text{bandit } J \text{ draws reward } R | \text{hidden state } S)$ for all bandits and all possible hidden states.

Let's suppose that the generative model is available, we give here the pseudo-code of a Bayesian algorithm. We consider k hidden states. In each hidden state bandits have an expected value which can be computed from the likelihood functions. $V_J(S)$ is the expected value of bandit J in hidden state S .

Let $p_S(t)$ be the probability that hidden state S is the current environmental hidden state at time t . $L_J(R|S)$ the likelihood function measuring the probability to receive reward R from bandit J in hidden state S .

The value of each bandit can be computed by accounting for the uncertainty on the hidden state: at time t the value of bandit J is $\sum_S p_S(t) * V_J(S)$.

$t=0$

For all S , $p_S(0) = \frac{1}{k}$, end *%Initialize probabilities*

For $t=1:T$

Choose a bandit J according to Softmax ($\sum_S p_S(t) * V_J(S)$, $j=1:n$)

Obtain Reward R

For $S=1:k$

$p_S(t + 1) \propto p_S(t) * L_J(R|S)$ *%Bayesian update*

end

End

We described in the introductory chapters the efficiency of Bayesian inference algorithms. However when the likelihood functions are not instructed, participants have to learn them online. We will expose 3 different algorithmic models describing the learning of bandit's reward distributions.

Frequentist learning

To learn likelihood functions, the most general model which uses no a-priori hypothesis is to record observed frequencies. This corresponds to a generalization of the Beta distribution to non-binary environments, and is named Dirichlet distribution.

Put simply, when reward R is received after choosing bandit J in hidden state S , the probability to receive R again after choosing J in S in the future is increased.

Observations are ambiguous as there is always uncertainty about the hidden state when they were obtained. Therefore it is not clear to know which likelihood function shall include a new observation.

The normative approach reconsiders after each new observation what was the hidden state at every trial from start and reattribute all observations to likelihood functions based on these recomputed

hidden states. This is named backward induction and is long and terribly costly both in memory and computationally while human cognitive abilities are limited in these domains (Cowan, 2001)(Cowan, 2001). It is therefore not biologically plausible.

An approximation of this process is to perform only forward inductions. Thus, past internal decisions, like the identity of the hidden state, are never reconsidered. This approach has been successfully used in numerous behavioral models (Collins and Koechlin, 2012), (Behrens et al., 2007), (Doya, 2002). In our setting, it corresponds to classifying the observation according to the prior belief about the hidden state. Instead of attributing the observation to the most likely hidden state, one can “spread” the observation in the different hidden states proportionally to the priors $p_S(t)$. For example if at trial t , bandit J is chosen, reward R is received. Consider that there are a priori 70% chances to be in hidden state 1 and 30% in hidden state 2, then $L_J(R|1)$ is incremented with a weight of 0.7, and to $L_J(R|2)$ with a weight of 0.3 [Figure 8 A&B au-dessous p. 53].

Critically this algorithm bootstraps the learning of likelihood functions and the tracking of the hidden state. An observation is classified in one of the likelihood functions based on the prior on the current hidden state. Observations are then interpreted by learned likelihood functions and combined to the prior in order to compute the posterior probability on the current hidden state. This posterior will itself be a prior at the next trial.

The pseudo code of the algorithm is as followed. $Freq(R; J, S)$ is the number of occurrences of R when bandit J is chosen in hidden state S . Thus, $L_J(R|S) = \frac{Freq(R; J, S)}{\sum_r(Freq(r; J, S))}$. The probability distribution on hidden states at time t is $p_S(t)$.

$t=0$

For all S , $p_S(0) = \frac{1}{k}$, end %Initialize probabilities

For all S, R, J , $Freq(R; J, S) = 0$ end %Initialize frequencies

For $t=1:T$

 Choose a bandit J according to $\text{Softmax}(\sum_S p_S(t) * V_j(S), j=1:n)$

 Obtain Reward R

 For all S

$Freq(R; J, S) = Freq(R; J, S) + p_S(t)$ %update frequencies

 end

 For $S=1:k$

$p_S(t + 1) \propto p_S(t) * L_J(R|S)$ %Bayesian update

end

End

One disadvantage of this model is when the number of possible rewards is large it takes a long time to have a representative measure of all frequencies. Additionally, according to the decision rule the most rewarded bandit is most often chosen. Consequently there is a sampling problem, and the likelihood functions of less rewarded bandits are not learned efficiently.

As a side note, instead of approximating backward induction by “spreading” the observations between the different hidden states, it is also possible to attribute entirely the observation to the most likely hidden state. These two classification rules revealed to be equivalent in all our following results.

Parametric model

A simple strategy to compute easily likelihood functions from a few samples is to have a generative parametric reward function associated with each bandit in each hidden state for which we just need to learn parameters. The general problem is to choose the true parametric shape of reward functions of the environment or a sufficiently general shape able to capture common distributions.

Gaussian model

The Gaussian parametric shape corresponds to a model of the environment where bandits give a noisy reward with a stable mean. It is parametrized by its mean m and standard deviation σ :

$$R \rightarrow \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(R-m)^2}{2\sigma^2}}$$

With adequate priors the Gaussian shape is particularly adapted to sequential learning as there are exact formulas to update the most likely m and σ after each new observation. Let $m(t)$ and $\sigma(t)$ be the best estimation of m and σ at trial, t . Reward R is received.

$$m(t+1) = m(t) + \frac{1}{t+1}(R - m(t))$$

$$\frac{1}{\sigma(t+1)} = \frac{a(t+1)}{b(t+1)}$$

With $a(t+1) = a(t) + \frac{1}{2}$ and $b(t+1) = b(t) + \frac{(R-m(t+1))^2}{2}$

However the Gaussian shape is not able to fit bimodal or heavy-tailed distributions.

Other generative parametric models could be used, in particular reward distributions of the exponential family. These distributions (Poisson shape, Gamma shape or exponential shape ...) are conjugated and easy to adapt to Bayesian inference. However heavy-tailed distributions or Gaussian mixtures are not in the exponential family and do not have a conjugate.

In the following we will use the Gaussian shape as a flexible unimodal parametric family of distribution.

Bi-modal Gaussian model

To include the possibility that reward distributions are multi-modal, while keeping computations tractable, it is possible to model these distributions as the superposition of several distinct Gaussian functions. As an example let's consider the case of bimodal distributions that we will use in the following.

A bimodal Gaussian distribution is parametrized by 5 parameters, the mean m_1 and m_2 of the two Gaussians, the standard deviation σ_1 and σ_2 of the two Gaussians and η , the relative weight of each distribution:

$$R \rightarrow \eta \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{(R-m_1)^2}{2\sigma_1^2}} + (1-\eta) \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{(R-m_2)^2}{2\sigma_2^2}}$$

If D_1 is

$$R \rightarrow \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{(R-m_1)^2}{2\sigma_1^2}}$$

And D_2

$$R \rightarrow \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{(R-m_2)^2}{2\sigma_2^2}}$$

The bimodal distribution can be expressed as

$$R \rightarrow \eta D_1 + (1-\eta) D_2$$

Let $m_1(t)$, $m_2(t)$, $\sigma_1(t)$, $\sigma_2(t)$ and $\eta(t)$ be the best estimation of the parameters at trial t . Reward R is received. There is no general exact formula to update the estimation of the parameters. Backward inference needs to be performed and all observations reconsidered from start. As a tractable non-exact alternative, we can extend the formulas of the Gaussian shape to the bimodal case.

Indeed we could update at each trial the parameters of only one of the Gaussian: the one which most likely generated R . To lose as little information as possible, it is alternatively possible to update both distributions and to modulate the strength of the update by the likelihood that R was generated by each. The higher the probability that R was generated by D_i the stronger is the update of the parameters of D_i .

Let $\mu = \frac{p(R|D_1)}{p(R|D_1)+p(R|D_2)}$ the probability that R was generated by the first mode D_1 . We can propose the following parameters update.

$$m_1(t+1) = m_1(t) + \frac{\mu}{t+1} (R - m_1(t))$$

$$m_2(t+1) = m_2(t) + \frac{1-\mu}{t+1} (R - m_2(t))$$

$$\frac{1}{\sigma^2_1(t+1)} = \frac{a_1(t+1)}{b_1(t+1)}$$

$$\frac{1}{\sigma^2_2(t+1)} = \frac{a_2(t+1)}{b_2(t+1)}$$

With $a_1(t+1) = a_1(t) + \frac{\mu}{2}$, $a_2(t+1) = a_2(t) + \frac{1-\mu}{2}$, $b_1(t+1) = b_1(t) + \mu \frac{(R-m_1(t+1))^2}{2}$ and $b_2(t+1) = b_2(t) + (1-\mu) \frac{(R-m_2(t+1))^2}{2}$

Lastly

$$\eta(t+1) = \frac{t * \eta(t) + \mu}{t+1}$$

In the following we will present the results of the algorithmic model described above. Considering separately the two Gaussians of a bi-modal distribution is an approximation for which there are several possible implementations. We could try several of them which appeared all equivalent.

Model NEIG: Hierarchical, frequentist and counterfactual

Using a generative parametric model of reward distributions enables to infer from a few samples the full reward distribution of a bandit in a given hidden state. We propose instead that humans use a generative structural model of hidden states to infer reward distributions of all bandits, chosen and unchosen, from observation of the chosen bandit.

Our hypothesis is based on authors who have argued since early studies on bounded rationality that humans look for satisficing (or good enough) options rather than for options maximizing reward (Simon, 1956). In the context of bandits, given an acceptability threshold, this corresponds to a ranking of bandits according to their probability to deliver a reward above the threshold:

$$\int_{Threshold}^{\infty} L_J(r|S) dr \text{ (Hertz et al., 2017).}$$

We propose that humans assume that the satisficing ranking depends only on the hidden state, and is independent of the threshold. Therefore in a given hidden state, whatever is the level of a satisficing reward, it is always the same bandit which has to be selected.

Mathematically, for all hidden states S , there is a ranking $\{i_1, i_2, \dots, i_n\}$ so that

$$\forall R, \int_R^{\infty} L_{i_n}(r|S) dr < \dots < \int_R^{\infty} L_{i_2}(r|S) dr < \int_R^{\infty} L_{i_1}(r|S) dr$$

This can also be expressed equivalently in term of cumulative distribution:

$$R \rightarrow \int_{-\infty}^R L_J(r|S) dr$$

$$\forall R, \int_{-\infty}^R L_{i_1}(r|S) dr < \int_{-\infty}^R L_{i_2}(r|S) dr < \dots < \int_{-\infty}^R L_{i_n}(r|S) dr$$

When the satisficing ranking is independent of the threshold, bandits' cumulative reward distributions are **ordered** (Figure 7)

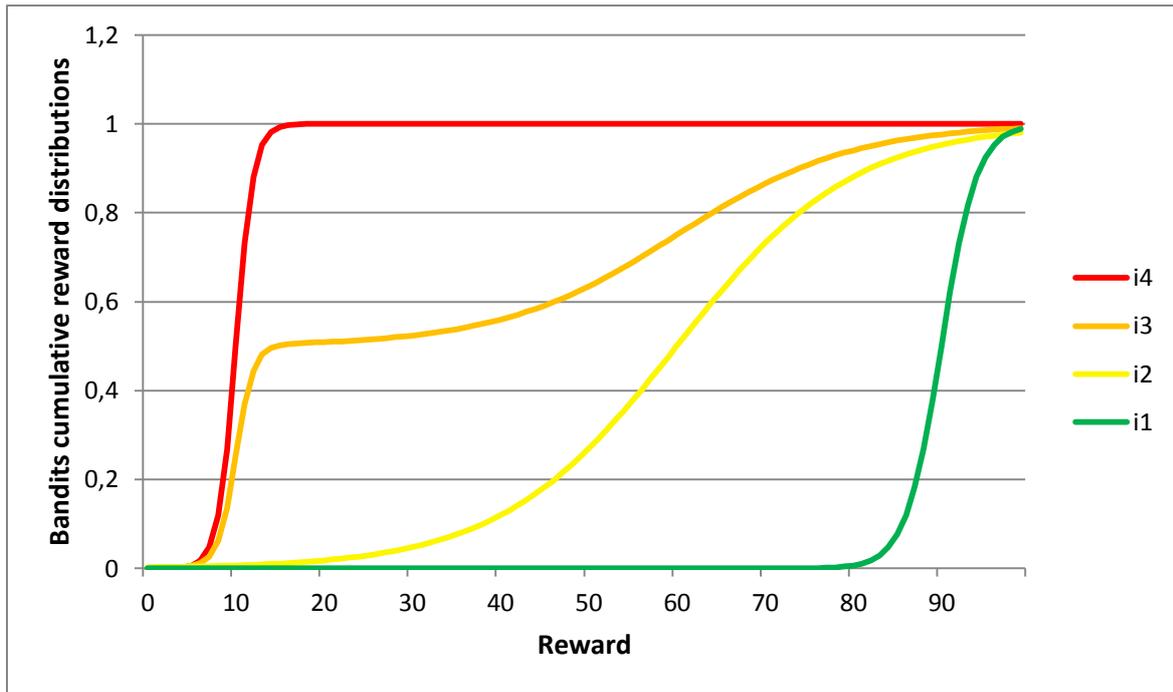


Figure 7 : Example of ordered cumulative reward distributions. The ranking is $\{i_1, i_2, i_3, i_4\}$, the different curves never cross.

It is immediate that when cumulative distributions are ordered as exposed here, distributions' expected values follow the same order:

$$\overline{L_{i_1}(r|S)} > \overline{L_{i_2}(r|S)} > \dots > \overline{L_{i_n}(r|S)}$$

Therefore the target bandit (the one participants want to select) is i_1 .

When bandits' reward distributions are Gaussians with an identical or close to identical standard deviation, it is easy to show that their cumulative functions are ordered with the same ranking as their expected values. But it is not general; there are distributions for which cumulative functions are not ordered (the order changes according to the value of R).

When a reward R is observed it increases the subjective probability that the chosen bandit delivers R again in the future. This modifies the chosen bandit's cumulative reward distribution and can change the satisficing index for some thresholds. We propose that humans modify also unchosen distributions in order to maintain the order of cumulative distributions.

We propose a simple algorithmic solution to adjust unchosen distributions: when bandit ranked i_k , $k \leq n$ is chosen, and reward R observed, the distribution's update system works "as if" the reward of unchosen bandits had also been observed. The fictive counterfactual is lower than R for bandits ranked $[k+1, n]$ (lower rewarded bandits), and higher than R for bandits ranked $[1, k-1]$ (higher rewarded bandits). This transposes the order of cumulative distributions to a single trial and corresponds to the assumption that at each trial

$$R(i_1) > R(i_2) > \dots > R(i_n)$$

This bootstraps on prior beliefs and in a given hidden state, the "target" bandit (the one with the highest rank that participants want to select) does not change. Therefore computing expected values

is not needed. However observations change the probability distribution on hidden states and when the most likely hidden state changes, bandit's distributions change and thus the target bandit.

We give now the pseudo-code of the proposed algorithm in the case of a choice between 2 bandits and 2 hidden states. We will propose a generalization to n bandits later.

We consider two hidden states, S_1 and S_2 and bandits ranking is $\{i_1(S_1), i_2(S_1)\}$ for S_1 and $\{i_1(S_2), i_2(S_2)\}$ for S_2 .

Let $p_1(t)$ (resp. $p_2(t)$) be the probability that the hidden state is S_1 (resp. S_2) at time t.

Let $f_HIGH(R; S_i)$ be the number of occurrences of R for bandit $i_1(S_i)$ in hidden state S_i , and $f_LOW(R; S_i)$ the number of occurrences of R for bandit $i_2(S_i)$ in hidden state S_i . The bandit's reward distribution can be directly obtained after normalization of the number of occurrences.

As a notation, since there are only 2 hidden states, $S_{\bar{i}}$ stands for S_{3-i} (S_1 when $i = 2$, S_2 when $i = 1$)

$t = 0$

Initialize f_HIGH and f_LOW . $\forall (R, i), f_HIGH(R; S_i) = f_LOW(R; S_i) = 0$

Initialize $p, p_1(0) = p_2(0) = 0.5$

For t=1:trial

Choose hidden state S_i according to

$$\text{Softmax}(p_1(t), p_2(t))$$

Choose bandit $i_1(S_i)$

Reward R is received

If $p_i(t) > 0.5$ %The highest rewarded bandit of the most likely hidden state was chosen

%Update f_HIGH with the factual reward

$$f_HIGH(R; S_i) = f_HIGH(R; S_i) + 1$$

%Update $f_LOW(R; S_i)$: the other bandit in the same hidden state with a fictive counterfactual reward inferior to R [Figure 8 A_b]. The total added weight is 1

$$f_LOW(\lfloor R_min R \rfloor; S_i) = f_LOW(\lfloor R_min R \rfloor, S_i) + 1/(R - R_min)$$

end

If $p_i(t) \leq 0.5$ %The highest rewarded bandit in the less likely hidden state which is also the lowest rewarded bandit in the most likely hidden state was chosen

%Update f_LOW with the factual reward

$$f_LOW(R; S_i) = f_LOW(R; S_i) + 1$$

%Update $f_HIGH(R; S_i)$: the other bandit in the same hidden state, with a fictive counterfactual reward superior to R [Figure 8 B_b]. The total added weight is 1

$$f_HIGH([R R_MAX]; S_i) = f_HIGH([R R_MAX], S_i) + 1/(R_MAX - R)$$

end

%Update hidden states likelihood

$$p_i(t+1) \propto p_i(t) * f_HIGH(R; S_i) / (f_HIGH(R; S_i) + f_LOW(R; S_i))$$

%probability to be in the chosen hidden state S_i

$$p_{\bar{i}}(t+1) \propto p_{\bar{i}}(t) * f_LOW(R; S_{\bar{i}}) / (f_HIGH(R; S_i) + f_LOW(R; S_{\bar{i}}))$$

%probability to be in the unchosen hidden state $S_{\bar{i}}$

%Include the possibility that hidden state changes at the following trial (volatility)

$$p_1(t+1) = (1 - \alpha) * p_1(t+1) + \alpha * p_2(t+1)$$

$$p_2(t+1) = (1 - \alpha) * p_2(t+1) + \alpha * p_1(t+1)$$

end

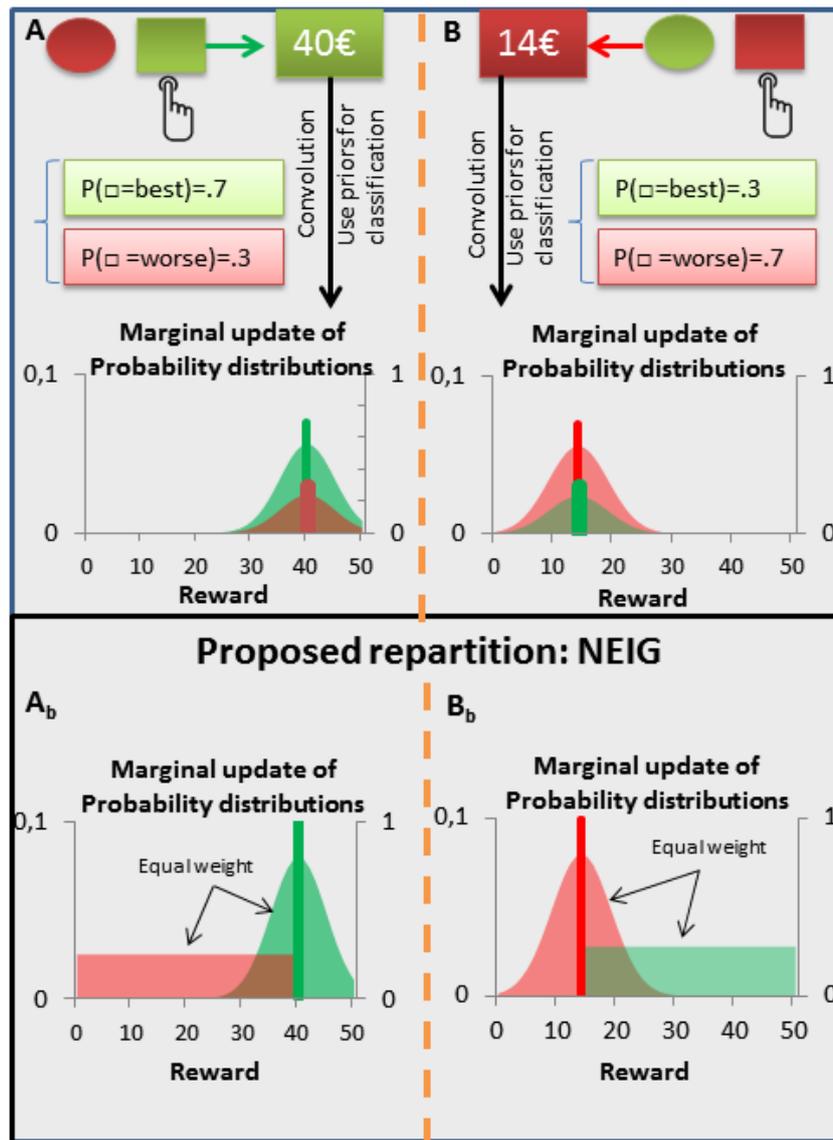


Figure 8 : Bayesian learning of reward distributions. We represent in each case the marginal update of distributions due to a single observation. In (A) and (A_b), a reward of €40 is received. In (B) and (B_b) a reward of €14 is received. This update is modulated by the prior belief that the chosen bandit draws from the reward distribution with the highest expected value (best), or from the reward distribution with the lowest expected value (worse). In (A) and (A_b) the chosen bandit presumably draws from best, in (B) and (B_b) the chosen bandit presumably draws from worse. In (A) and (B) the update follows the optimal solution when only forward inferences are performed. Each distribution is updated proportionally to the likelihood that the obtained reward was drawn from it. (A) A function of total weight 0.7 centered on 40 is added to the best distribution. A function of total weight 0.3 centered on 40 is added to the worse distribution. (B) A function of total weight 0.7 centered on 14 is added to the worse distribution. A function of total weight 0.3 centered on 14 is added to the best distribution. In (A_b) and (B_b) the update follow NEIG algorithm. (A_b) The “perceived” highest rewarded bandit is chosen (exploitation). A function of total weight 1 centered on 40 is added to the best distribution. A uniform function of total weight 1 is added to all possible rewards below 40 in the worse distribution (counterfactual). (B_b) The “perceived” lowest rewarded bandit is chosen (exploration). A function of total weight 1 centered on 14 is added to the worse distribution. A uniform function of total weight 1 is added to all possible rewards above 14 in the best distribution (counterfactual).

Our model is essentially ordinal as rewards are classified according to whether they are above or under observed reward, and not according to their exact magnitude. In the following, we will name this model **NEIG**.

It is interesting to note that NEIG solves the continuity problem as all possible rewards which lay under (or above) one of the observed rewards have been classified in f_{LOW} (or f_{HIGH}). Thus, NEIG has constructed a representation of the likelihood of all rewards in both distributions (and thus a measure of their “informational value”) even when these rewards have never been really observed before.

To account for the intuition that two very close rewards are equally desirable, it is additionally possible to consider that reward representations have a finite precision. Mathematically one can convolve each observed reward with a noise function centered on 0 and decreasing with the distance to 0 before updating f_{LOW} or f_{HIGH} . Therefore the frequency count of rewards close to the observation will also be incremented in the chosen bandit, and become more likely in the future.

In [Figure 9], we show the difference of an exact frequency learning of observations (A) and a learning with observation noise (B). In particular, reward 16 is never observed so has no “information value” in A, while it acquires “information value” from neighbor observations in (B).

This convolution can also be included in the Bayesian model which learns reward frequencies.

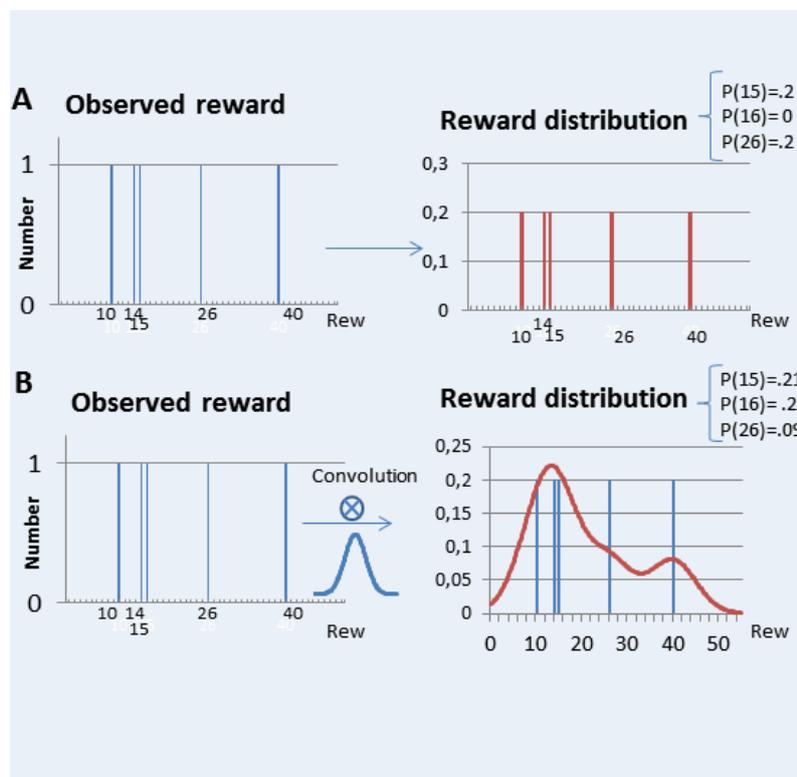


Figure 9 : Observation noise: (A) Construction of reward distributions with an exact frequency count. On the left, received rewards (10, 14, 15, 26, 40). On the right, inferred reward distribution. (B) Construction of reward distribution with an observation noise. On the left, received rewards (idem) on the right convolution with a Gaussian noise function centered of the observed reward with a parametrically adjusted standard deviation.

Behavior of NEIG in simple settings

To give credit to NEIG, we first show that its behavior in simple cases is similar to what is commonly observed on human participants.

We simulated our model on two simple reversal learning tasks using continuous reward scales. One uses (unimodal) Gaussian reward distributions [Figure 10 A]. One bandit draws from the green Gaussian, the other from the red Gaussian, and the mapping flips from time to time.

NEIG algorithm behaves as expected and finds quickly the highest rewarded bandit. We plot here after a flip, how the model readapts and changes its choice to select the bandit drawing from the green Gaussian [Figure 11 A].

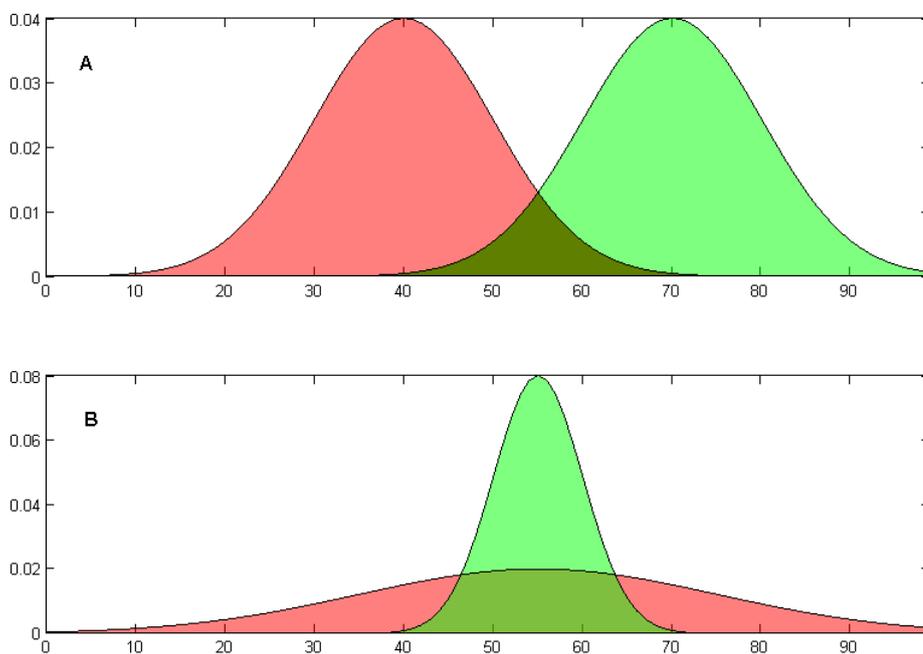


Figure 10 : Hidden reward distributions of 2 reversal learning tasks on which we simulated the NEIG algorithm. A: Unimodal Gaussian distributions. B: Gaussian distributions with different variance.

We proposed a second setting, where both reward distributions have the same mean but a different variance [Figure 10 B]. One distribution is wider than the other, and is therefore riskier. It is known in this setting that participants tend to prefer the narrow “risk averse” distribution (March, 1996), (Niv et al., 2012). This is also how NEIG behaves. In average, it avoids low rewards attributed by the wide distribution and tends then to stay on the narrow one. We see in [Figure 11 B] that the model (in red) reverses its choice after a flip of environmental contingencies to re-select the risk averse distribution.

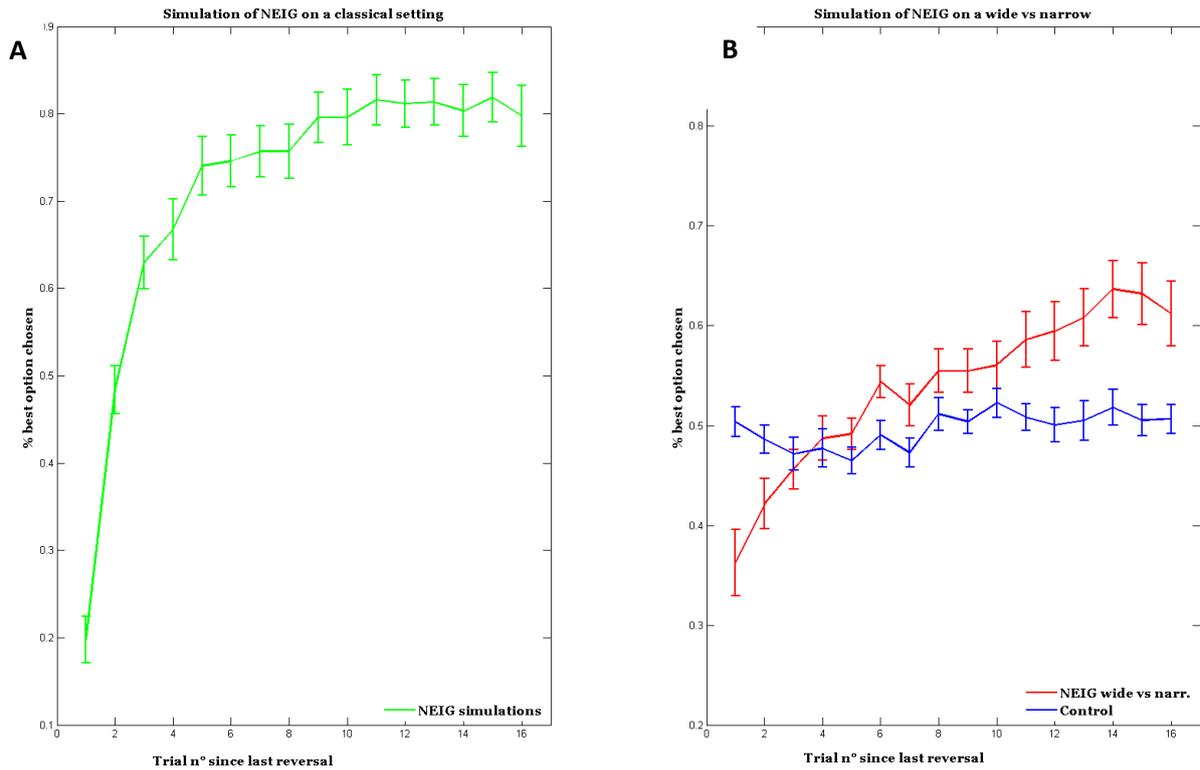


Figure 11 : A: simulation of NEIG’s adaptation curves on a reversal learning task where reward distributions are unimodal Gaussians. NEIG in green switches after reversals to select the highest rewarded distribution around 80% of times. B: simulation of NEIG’s adaptation curves on a reversal learning task where both distributions have the same mean, but a different variance (wide versus narrow). We represent in red the percentage of choice of the smallest variance distribution (risk-averse). We compare with a control (in blue) where the two reward distributions are similar (same mean and variance)

Materials and methods

Experimental paradigm

The goal of our work was to test our proposed model, NEIG. We wanted in particular to compare it with two alternative models, a simple information-free Q-learning, and a “Gaussian” model which parametrize all reward distributions as Gaussians. To decide among the different models, we proposed an experimental paradigm where the different models made different predictions, and observed which one described human data the best.

Reversal learning: The first alternative model, Q-Learning does not learn task-structure. Therefore when there is a structure in an experimental task, we can predict that Q-Learning will not improve its performance across time. On the converse its performance will be stable. Structure-Learning algorithms will on their side catch a part of the structure and improve their performance.

We then proposed to use one of the simplest paradigms using a hierarchical structure: reversal learning. This allowed us to decide whether human participants use Q-Learning or a structure-learning algorithm.

Bimodal distributions: The second alternative model, “Gaussian” model considers that all encountered reward distributions are Gaussians. We therefore avoided drawing rewards from Gaussian distributions in our reversal task. Indeed NEIG, our proposed model, is more general and can adapt to different types of distributions including Gaussians. NEIG and Gaussian models would therefore do similar predictions in a Gaussian environment, but are likely to differ otherwise. We proposed the use bi-modal reward distributions to this end.

Thus, our 3 experimental tasks used the reversal learning paradigm with bimodal reward distributions. The first task used 2 bandits and attributed rewards were valued on a continuous scale: integer reward values from 10 to 99. It will be named “continuous task” in the following. The second task was conducted by a former PhD student, Marion Rouault. It used also 2 bandits, but only 5 different reward values were attributed. It will be named “discrete task” in the following. We will describe these two tasks in this section.

We also generalized our results to a 3 bandits experiment, “3 Option task”, that we will describe later in a dedicated chapter.

Note that considering the experimental paradigm we will also compare two additional alternative models. One is derived from Q-Learning but includes the reversal structure. We described this model earlier as “Normalized RL”. The second is derived from the “Gaussian” model but considers reward distributions to be bimodal and not unimodal.

Experiment 1: Continuous reward distribution

Participants

25 right handed participants (12 males) were recruited through an internet database. They were asked to have a normal or corrected to normal vision and were scanned for no history of psychiatric diseases. To have subjects as naïve as possible, participants were also asked not to have taken part to

any experiment in our lab. In accordance with a French ethics protocol, participants gave informed consent for their participation to the experiment.

Behavioral protocol

Participants received instructions to the task reproduced in annex (Annex 1). In short, participants were told that they could choose between two symbols. Choosing a symbol would draw a reward which depended on the chosen symbol but not on the spatial position of the symbols. They were informed that one symbol was giving better rewards than the other one in average but that the best symbol could vary across time.

All stimuli were white and presented on a black screen using PsychToolBox (Figure 12). At each trial a square and a round would appear centered in height and symmetrical across the middle of the screen.

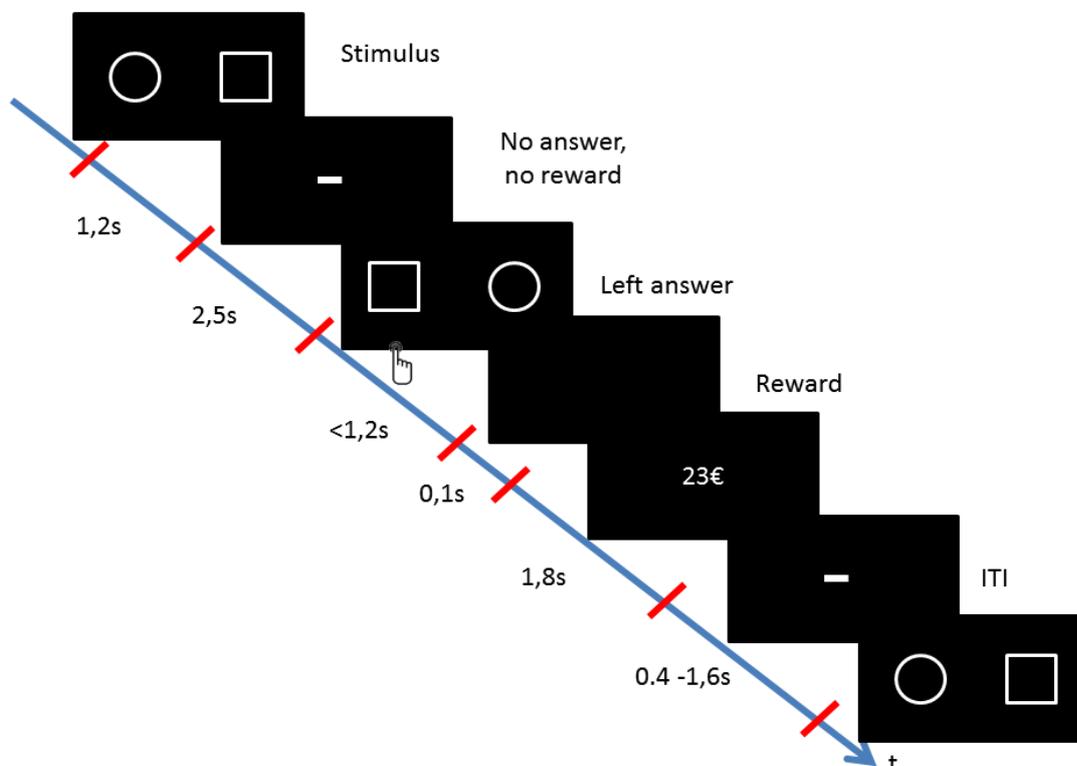


Figure 12: Visualization of the task

The relative position of the two symbols (circle on the right square on the left, or circle on the left square on the right) was pseudo-randomized. Two computer keys were highlighted in blue. Pressing the right key would select the symbol on the right, pressing the left key would select the symbol of the left. Participants were instructed to use their right hand, the index to press left, the major to press right. The experiment contained 936 trials, each lasting 3.5 seconds. Stimuli were displayed for 1s, and participants had to choose an action less than 1.2 seconds after the beginning of the display. If none of the 2 keys was pressed during these 1.2 seconds the trial was considered lost and no

reward was obtained. Participants would see a white underscore on the screen meaning they did not get any reward.

If an action was performed a reward was drawn pseudo-randomly from the distribution associated with the symbol which was on the side of the motor action. 0.1 second later the obtained reward would be displayed in the center of the screen followed by the symbol '€' (euro). Reward would be displayed for 1.8 seconds. The inter-trial lasted between .4 and 1.6 seconds so that independently of the timing of the action, the following trial would start 3.5 seconds after the beginning of the preceding trial.

We said rewards to be “pseudo-randomly” drawn, as they were actually drawn before the experiment so that diverse statistical controls could be performed to equilibrate between participants. These controls will be described in the following of the text.

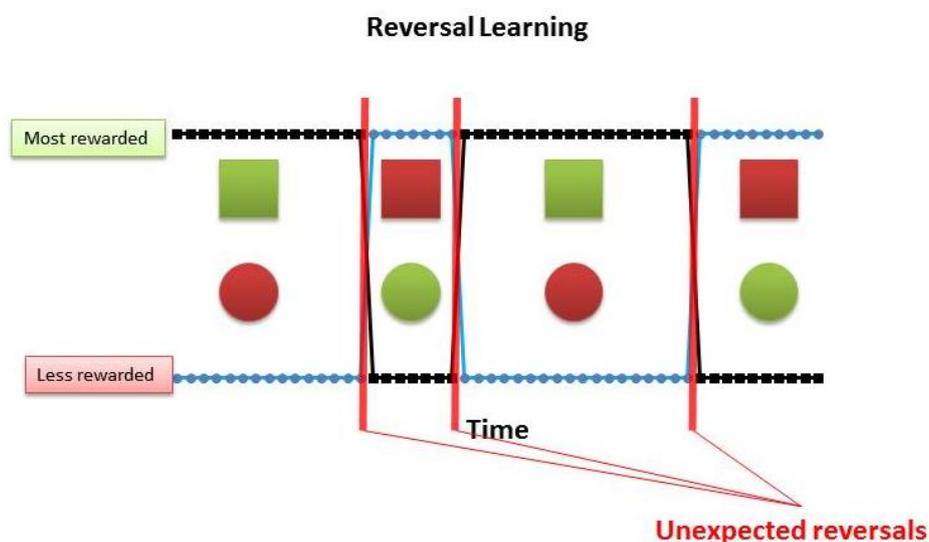


Figure 13: Reversal learning paradigm. At the beginning the square draws from the distribution with the highest expected value. The circle draws from the distribution with the lowest expected value. The mapping between symbols and distributions changes from time to time without notice.

Reward distributions

Our paradigm is a reversal learning task. Thus, from time to time and without notice to participants, the mapping between symbols and hidden distributions switches (Figure 13). A reversal happens at a discrete time T . If at trial $T-1$ the circle was drawing rewards from distribution 1 and the square from distribution 2, then from time T and until the next reversal, the circle will draw from distribution 2 and the square from the distribution 1. Therefore, if distribution 1 is in average more rewarded than distribution 2, participants have to change their choices to keep their level of income.

Reversals occurred on average every 26 trials, never before 18 trials, never after 34 trials. The trials between two reversals are named an “episode”. The experiment was composed of 36 episodes (35 switches) and 936 trials. During an episode the circle and the square would sit the same number of times on the right and on the left of the screen. Rewards proposed by the symbols were pseudo randomized before the test so that the statistics of an episode were as close as possible to the statistics of the real distribution.

In this first experiment, the reward distributions were as follows:

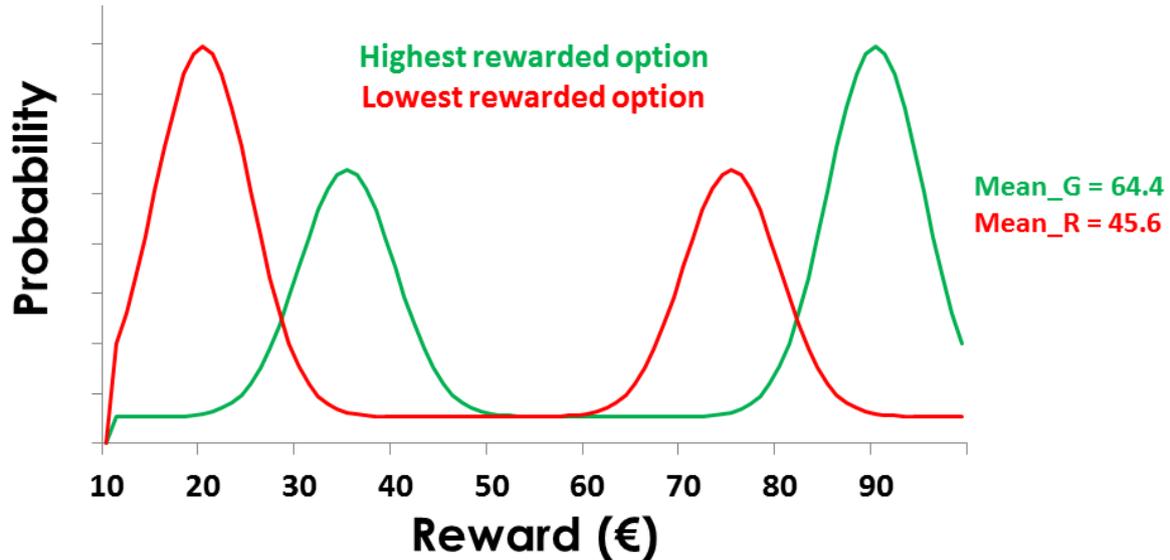


Figure 14 : Distributions of the continuous task. Rewards vary between 11 and 99. Playing at random allows to obtaining €55. The best symbol has a mean of 64.4, the worse symbol 45.6.

In the following we refer to the best distribution as the distribution with the highest mean (named “BEST” in the following). Drawing from BEST is in average more rewarded. Symmetrically, we refer to the worse distribution as the distribution with the lowest mean (named “WORSE” in the following).

Both distributions have two modes: one is “principal” and the other one “secondary”. 54% of the distribution weight sits in the principal mode, 36% in the secondary mode and the remaining 10% is spread on the entire reward space. BEST has a principal mode centered on 90 and a secondary mode centered on 35. WORSE has a principal mode centered on 20 a secondary mode centered on 75. In total, BEST draws an average reward of €64.4, WORSE €45.6, and playing at random gives an average reward of €55 (Figure 14).

Note that in our task, reward values and the informative values of reward are de-correlated. Indeed, receiving a reward around €75 is above the mean of both distributions. Therefore, it can be considered as a high reward. However, given the reward distributions of the design, it is far more likely to receive this reward from the WORSE distribution. Thus, it has a low “informative value” as an omniscient observer would change choice at the next trial. On the converse receiving a reward around €35 is below the mean of both distributions and can be seen as a low reward. However, the omniscient observer would infer that this reward was most likely drawn from the best distribution and keep his choice at the next trial. Thus, 35 has a high “informative value”

Here rewards are not calling the same behavioral response with and without knowledge of the task structure.

Experimental details

The task was broken down in 6 blocks. After each block there was a break and participants chose when to start again. Participants were accurately informed that the presence of a break was independent of the course of the experiment. To limit the influence of the break, before restarting participants were reminded of their last choice before the break and the last obtained reward.

We motivated participants by giving them at each break a bonus related to their performance in the former block. Participants saw their average performance (in the current block and past performance was recalled), and the performance of “another participant” which was always a little bit better than them. The “other participant” was an algorithmic simulation playing exactly the same trials.

Training

Before the task, participants received a training composed of 78 trials which represented 3 episodes (2 reversals). The experimenter stood behind participants during the training and checked that they were adapting to reversals. This evaluation was completed with an oral assessment at the end of the training to check whether participants saw reversals and whether they had additional questions. No critical information was given at that stage; the instructions of the task were only repeated if needed. If participants had not reacted to reversals or did not report to have seen reversals, a second training following the same procedure was conducted.

Experiment 2: Discrete reward distribution

To further separate model predictions (see results), we proposed another task similar to the continuous task. We will for this task only list the experimental differences with the former one.

Crucially, only 5 rewards were available, €1, €2, €5, €8 and €9. The reward distributions were as follows:



Figure 15 : Reward distribution of the discrete task.

Here receiving €2 is a low reward but is more likely drawn from BEST (high “informative value”) and it is on average more rewarded to keep the same choice at the next trial. On the converse receiving €8 is a high reward but the chosen bandit is most likely drawing from WORSE (low “informative value”), and it is therefore in average more rewarded to switch choice at the next trial.

In this experiment each participant performed two sessions of 1040 trials each on two separate days. The two sessions differed on their 5 first minutes which were constructed as priming and in which the reward distributions were different.

In the prime of one session, the probability to obtain each reward in the highest rewarded distribution was linearly proportional to the reward value. Therefore, it was rare to receive 1 or 2; and 8, 9 were the most common rewards. In the lowest rewarded option on the converse it was rare to receive 8 or 9; and 1, 2 were the most common rewards. In this simple setting, a strategy like TD

learning is very efficient, and we wanted to prime participants so that they would use this strategy afterwards during the session.

In the prime of the other session, both reward distributions were identical and uniform over 1, 2, 5, 8 & 9. The prime is therefore structure-free and no strategy works. We expected participants to discard simple strategies and be more attentive to structural information when it eventually appears after priming.

Experimentally, no behavioral difference was found between the conditions, therefore they were pooled in the following analysis.

Lastly, the timing of this task was slightly faster as trials lasted for 3 seconds. There were only 4 breaks and no bonus was given during breaks.

We will now present key tools of our data analysis, in particular our model fitting procedure.

Fitting procedure

A critical point of our analysis consists in comparing different algorithmic models to human behavior. To this end we need measures to evaluate how close a candidate model and human behavior are. Here our behavioral data are a series of choices in response to a multi-armed bandit task.

There are many parameters at play to explain choices and many of them are out of the scope of our study. In particular, some inter-individual differences are due to individual history before the experiment, and it is not what we want to model. For example, participants can be more or less risk takers. This parameter is orthogonal to our question, and we do not want it to influence our comparison between algorithms. Other behavioral observations are not specific to one algorithm in particular as for example when participants select a symbol and press the wrong button. We are similarly not interested in modeling this part of behavior.

We can therefore separate the variables used by an algorithm in two categories. The variables of the first category are computed “internally” by the model during the experiment. For example Q values of a TD Learning belong to this first category. The variables of the second category are used by the studied algorithm but computed extrinsically by another un-modeled process. These variables are “free parameters” and can vary from one participant to another. To simplify we considered these free parameters to be constant across the experiment. We included in this category parameters out of the scope of our study and parameters non-specific to a particular model.

Our first objective was to infer for each participant what the values of these free parameters were when they played the task. For a given participant, we can set these free parameters to an arbitrary value and run (as many times as we want) the algorithm on the same trials than the participant did, then compare the choices of the algorithm to participant’s choices. In particular, we can compute for each trial the probability that the algorithm would have performed the same choice than the participant. On the entire task, we can compute the probability that the algorithm would have played exactly as the participant. This corresponds to the probability of behavioral data (observations) according to the algorithm.

$$p(\text{observations}|\text{algo}, \text{parameters}) = \prod_{i=\text{trials}} p(\text{action}_{\text{algo}}(i) = \text{action}_{\text{participant}}(i))$$

This function is named the likelihood function as explained p. 21 in our introduction to Bayesian inference. The log of the likelihood function is often used to facilitate computations, and we will in the following call this function LLH

$$\begin{aligned} LLH &:= \text{parameters} \rightarrow \log(p(\text{observations}|\text{algo}, \text{parameters})) \\ &= \sum_{i=\text{trials}} \log(p(\text{action}_{\text{participant}}(i)|\text{algo}, \text{parameters})) \end{aligned}$$

Through Bayesian inference, we can compute the probability of a set of parameters knowing participant's data. In particular, if we have no prior on the value of free parameters,

$$p(\text{parameters}|\text{algo}, \text{observations}) \propto p(\text{observations}|\text{parameters}, \text{algo})$$

Therefore, the mathematical problem is to find the values of free parameters which maximize the LLH function. It is a non-trivial optimization problem. Indeed, while it is possible to compute the LLH for all parameter values, there is no exact formula that could be differentiated to find the optima.

An analogy to the problem is to consider a scientist who wants to find the highest point on earth. He can jump to any point on the planet and measure the altitude with an altimeter. One strategy is to always go up. The scientist will arrive on the top of a hill which is a local maximum, but there are no guarantees to be on the highest hill of the planet. Another strategy is to measure altitude everywhere on the planet. He will eventually find the Everest after a lot of measures. However imagine that instead of 2 dimensions, the space to explore has 5 or 6 dimensions. This is the case when there are many free parameters to explore and each parameter is an independent dimension.

Several complex techniques have been developed to use already performed measures to gain information on the shape of the LLH function, and find the global maximum with a minimum number of samples. We used a technique named Slice Sampling. A slice sampler "samples" the parameter space and constructs Markovian "chains" of samples in which the frequency of each set of parameter is proportional to the likelihood function. The technical details of this method can be found in (Neal, 2003).

The slice sampler has a few parameters which we tuned empirically. First we initialized our chain at random in the parameter space, and used 3 different chains. To come back to our analogy, the initialization corresponds to the first point at which our scientist will measure the altitude. The "chain" corresponds to the successive positions he will explore. It can happen that a chain is blocked in a subspace of the parameter space and therefore reveals a local maximum as global maximum: Using multiple chains decreases this risk.

Secondly, the "step-size" corresponds to the longest jump the chain (or the scientist) will do in the parameter space. We set this value at the maximum so that independently of its initial position, the chain can jump to any point of the space. This slows the algorithms as more samples will be rejected - samples are rejected when they are not high enough compared to the current altitude - but enable to explore better the entire space. We constructed chains with 30'000 samples. We observed that this was enough for our tested algorithms which had 3 to 5 free parameters.

After sampling, we took the sample with the highest likelihood in each of our 3 chains, and conducted a gradient ascent on them to reach the local maximum. We considered the highest maximum value as the global maximum.

This slice sampling procedure allowed us for each algorithm and each participant to obtain the best set of free parameters accounting for participant's behavior. It is then possible to compare how well different algorithms describe participant's behavior.

Model parameters

We list in this part for all models presented in the chapter Algorithmic models (p. 43) the number of free parameters and their name. The soft-max choice rule used in every model to select an action corresponds to the following equation:

$$p(a_i) = \frac{\epsilon}{2} + (1 - \epsilon) \frac{e^{\beta \cdot V(a_i)}}{\sum_k e^{\beta \cdot V(a_k)}}$$

with $V(a)$ representing either the Q value in Reinforcement models or the probability to draw from the highest rewarded distribution in Bayesian models.

Reinforcement Learning (RL): 3 free parameters

- Inverse temperature (β): Selection noise, exploration, depends on the difference in Q values.
- Proportion of random choices (ϵ): Selection mistakes, independent of Q values.
- Learning rate (α): Importance of the current trial compared to past trials

Normalized Reinforcement Learning (NRL): 3 free parameters

- Inverse temperature (β)
- Proportion of random choices (ϵ)
- Learning rate (α)

Omniscient Bayesian model (knows reward distributions): 3 free parameters

- Inverse temperature (β)
- Proportion of random choices (ϵ)
- Volatility (α): Inter-trial probability of hidden state change (see p. 36)

Learning Bayesian model (learns reward distributions): 4 free parameters

- Inverse temperature (β)

-Proportion of random choices (ϵ)

-Volatility (α)

-Observation noise (Θ)

Gaussian model (infer distributions' parameters): 3 free parameters

-Inverse temperature (β)

-Proportion of random choices (ϵ)

-Volatility (α)

Bimodal Gaussian model (infer distributions' parameters): 3 free parameters

-Inverse temperature (β)

-Proportion of random choices (ϵ)

-Volatility (α)

NEIG model (proposed model): 4 free parameters

-Inverse temperature (β)

-Proportion of random choices (ϵ)

-Volatility (α)

-Observation noise (Θ)

Model Comparison

Thanks to the fitting procedure described above, we will in the following set the free parameters of each participant for each algorithm to their most likely value. This gives for each model and each participant a measure of the LLH, $p(\text{observations}|\text{model})$ from which we can infer $p(\text{model}|\text{observations})$. This is a way to rank algorithms and find the one that participants are the most likely to have used.

However it is important in this ranking to penalize algorithms which have too many free parameters. Mathematically these free parameters are fitted to participants' behavior before the comparison and therefore adding a free parameter can only improve the fit of a model. Therefore, several methods exist, measuring algorithm likelihood while accounting for the number of parameters. One of the

simplest is the Bayesian information criterion $BIC = -2LLH + k \cdot \ln(n)$ where k is the number of parameters and n the number of trials in the task (Schwarz, 1978).

Additionally to comparing LLH and BIC, it is possible to generate action sequences of the different algorithms playing the same task than participants. These action sequences will have statistics that we can compare to the real behavioral statistics of participants (Palminteri et al., 2017). We particularly analyzed two relevant behavioral measures.

-Adaptation curves: When a reversal occurs, participants gradually change their choice from the formerly highest rewarded symbol to the new highest rewarded symbol. To represent the dynamic of this adaptation, we can plot the proportion of “correct” choices as a function of the number of trials since the last reversal. Algorithms knowing the structure will adapt faster than Reinforcement Learning algorithm.

-Switch curves: To measure participants’ representation of reward, we can measure their propensity to change their choice after receiving each of the reward. For each participant and each reward R , we selected all trials where R was received. Then for all these trials, we looked at the next trial whether participants changed or kept their choice. If participants changed very often, it means that they found R non desirable. If participants kept the same choice often, R was somewhat desirable. Each algorithm uses its own reward representation that we will compare to participants

NEIG and reversal learning

We described the NEIG model earlier (Model NEIG: Hierarchical, frequentist and counterfactual, p.49). NEIG assumes that in each hidden state, cumulative reward distributions are ordered, and that the target distribution (the one that participants want to choose) has the highest probability to deliver rewards higher than a satisficing threshold, independently of the value of the threshold.

In the particular setting of reversal learning, there are only two reward distributions to learn online, as reversals only change the mapping between bandits and reward distributions. Distributions learned in one hidden state can be reused in the other hidden state. With the same notation as before, for all R :

$$f_{HIGH}(R; 1) = f_{HIGH}(R; 2)$$

$$f_{LOW}(R; 1) = f_{LOW}(R; 2)$$

To account for observation noise and spread the weight of the obtained reward on neighbor rewards, we propose the convolution of the obtained reward with a Gaussian function centered on 0 and having a standard deviation Θ which is a free parameter of NEIG. This function is named N in the following algorithm.

We give here the pseudo-code of the algorithm when specifically applied to reversal learning tasks with 2 bandits.

The probability law P_{BEST} that each symbol draws from the distribution with the highest expected value corresponds to the likelihood of the hidden states: $P_{BEST}(\square) + P_{BEST}(\circ) = 1$

$T = 0$

Initialize f_HIGH and f_LOW . $\forall R, f_HIGH(R)=f_LOW(R)=0$

Initialize P_BEST , $P_BEST(\square) = P_BEST(\circ) = 0.5$

For $T=1$:trial

Choose a symbol S according to $Softmax(P_BEST(\square), P_BEST(\circ))$

Reward R is received

$f_OBS = R \otimes N$, the convolution of R with noise function N

If $P_BEST(S) > 0.5$ %EXPLOITATION TRIALS

%Update f_HIGH with the factual reward

$$f_HIGH = f_HIGH + f_OBS$$

%Update f_LOW with a fictive counterfactual reward inferior to R [Figure 8, A_b]

$$f_LOW([R_min R]) = f_LOW([R_min R]) + 1/(R - R_min)$$

%Update P_BEST

$$P_BEST(S) \propto P_BEST(S) * f_HIGH(R)/(f_HIGH(R) + f_LOW(R))$$

$$P_BEST(\bar{S}) \propto P_BEST(\bar{S}) * f_LOW(R)/(f_HIGH(R) + f_LOW(R))$$

If $P_BEST(S) \leq 0.5$ %EXPLORATORY TRIALS

%Update f_LOW with the factual reward

$$f_LOW = f_LOW + f_OBS$$

%Update f_HIGH with a fictive counterfactual reward superior to R [Figure 8, B_b]

$$f_HIGH([R R_MAX]) = f_HIGH([R R_MAX]) + 1/(R_MAX - R)$$

%Update P_BEST (similar to exploitative trials)

$$P_BEST(S) \propto P_BEST(S) * f_HIGH(R)/(f_HIGH(R) + f_LOW(R))$$

$$P_BEST(\bar{S}) \propto P_BEST(\bar{S}) * f_LOW(R)/(f_HIGH(R) + f_LOW(R))$$

%Include the possibility to have a reversal at the following trial (volatility)

$$P_BEST(\circ) = (1 - \alpha) * P_BEST(\circ) + \alpha * P_BEST(\square)$$

$$P_BEST(\square) = (1 - \alpha) * P_BEST(\square) + \alpha * P_BEST(\circ)$$

End

The algorithm is summarized in the following figure [Figure 16]. In (A) we plot the prior distributions. In (C) the posterior distribution. In between a reward was received which was most likely drawn from the highest rewarded distribution (B).

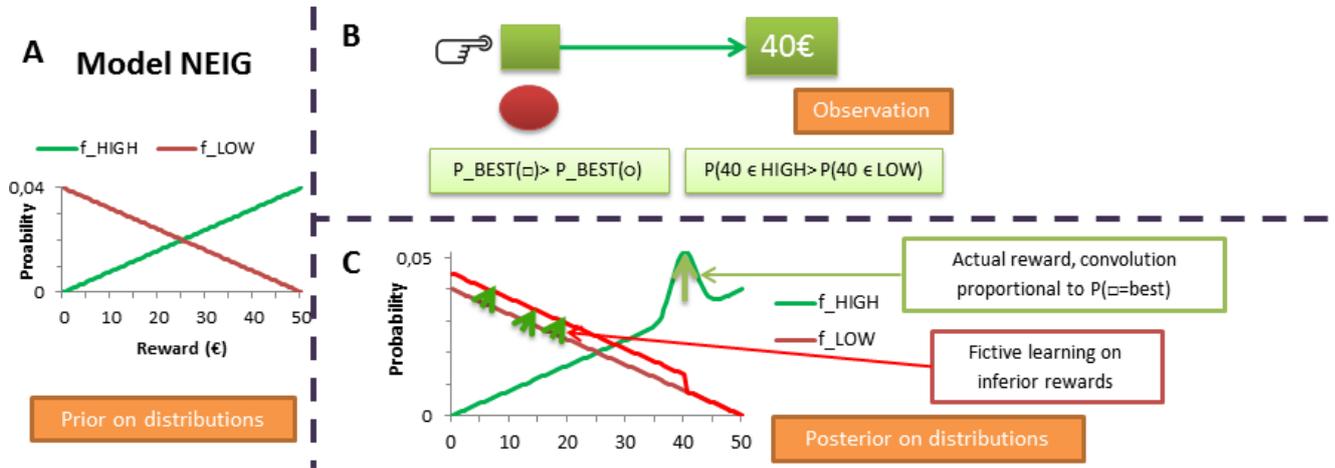


Figure 16: Summary of model NEIG (A) f_{HIGH} and f_{LOW} before observation (priors). Here they are linear. (B) The observation is here more likely to come from the best distribution. (C) f_{HIGH} and f_{LOW} updates following NEIG heuristics. f_{HIGH} is the “factual” distribution, updated through the convolution of the observation with a Gaussian noise function. f_{LOW} , the counterfactual distribution is updated with a step-function on rewards inferior to the observation. Distributions are re-normalized after this operation.

We will now show how the different algorithms described in this section account for human data in our task.

Results

Behavior

We first present behavioral results on the continuous task. Overall participants understood the task. Indeed they selected more often the highest rewarded option and changed their favorite action after a reversal of the task contingencies. We show in [Figure 17 (A)] the proportion with which participants selected the highest rewarded symbol in function of the number of trials following episode onset (last reversal). At the onset of the episode this proportion is 35% and climbs to 66% sixteen trials later (chance level: 50%).

In [Figure 17Figure 18 (B)], we plot in each episode the proportion of trials where the highest rewarded symbol was selected. This shows that the average performance per episode remains similar across session (non-significant influence of the episode number in a regression, $p=0.45$). This either means that participants do not adapt to the task structure, or that this adaptation is very fast or that it is undermined by fatigue causing responses to be noisier.

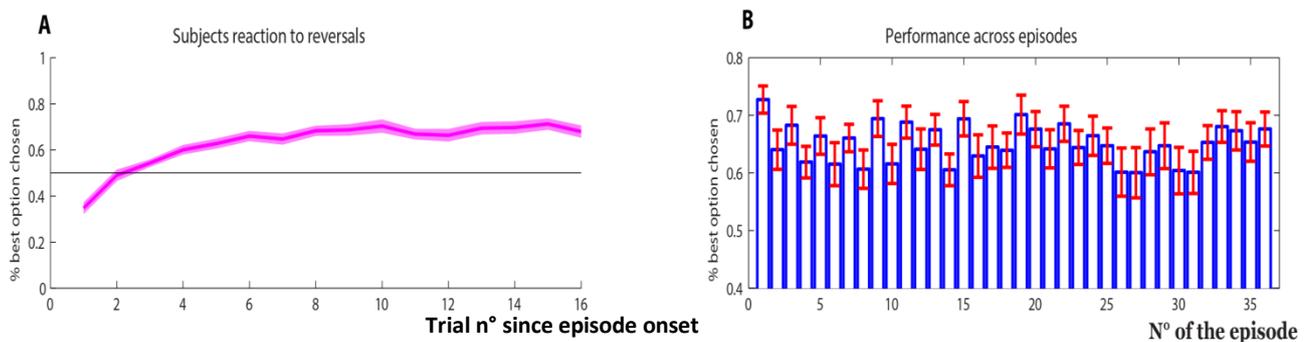


Figure 17 : (A) Average (across subjects and episodes) proportion of highest rewarded choices plotted against the number of trials since episode onset. Light purple represents the standard error. (B) Average (across subjects) proportion of highest rewarded choices plotted against the number of the episode in the task. Red represents the standard error.

We fitted different models to participants' data and present how well models describe human behavior.

First model comparisons

As explained in the material and methods, all proposed algorithms have individual free parameters which were fitted to each participant. We then generated for each algorithm and each participant a "fictive participant" using the fitted parameters and playing the same task as the participant. We then compared the statistics of the 25 simulated participants to real human behavior.

An algorithm can be considered as a good model when at minimum the statistics of the simulated participants are not significantly different than real participants' statistics (Palminteri et al., 2017).

We first compared the performance of three models:

-Q-Learning as representing Reinforcement Learning. It was our information free baseline

-An “omniscient” Bayesian algorithm which knows reward distributions. If participants were able to perform inference and to learn distributions, this model would correctly describe their behavior at least at the end of the task.

-An algorithm which used a Gaussian parametric model of all reward distributions, and inferring the two parameters of each distribution (mean and standard deviation) according to observations (later named Gaussian model).

We fitted the free parameters of these 3 algorithms to participants, and computed how each algorithm readapts after a reversal.

We plot in [Figure 18 (C)] how a Reinforcement learning model (in blue) describes participants’ reversals (in purple). In several points model behavior is significantly different from participants’ statistics ($p=2.10^{-4}$), as Reinforcement Learning does not adapt to a reversal as fast as participants. To test it formally, we conducted a two way repeated measure ANOVA including trial number and model vs data as within subject factors. Here $F=61$, $p<1.10^{-5}$, therefore the predictions of Reinforcement Learning model are significantly different from data.

Counter-intuitively the omniscient Bayesian model also adapted slower than participants in response to reversal [Figure 18 (A), in red] ($F=71$, $p<1.10^{-5}$). This resulted from the parameter’s fit. One of the fitted free parameter, ϵ measures the proportion of random responses and the best fit set this proportion to 15% for the omniscient algorithm. Indeed the omniscient algorithm is so different from what participants are doing that a lot of noise needs to be added in order to degrade model performance and match participants.

As for the Gaussian model [Figure 18 (C), in light blue], it is closer to participants’ behavior, although a systematic significant difference persists ($F=39$, $p<1.10^{-5}$).

To go further we compared algorithms on another behavioral criterion: the switch rate following rewards [Figure 18 (B)]. We represent for each bin of the reward scale the proportion with which participants (in purple) changed their choice at the following trial. As expected for participants, the higher the reward was the less they tended to change.

On this behavioral measure we observed that neither the Reinforcement learning model [Figure 18 (D) t-test, $p=8.10^{-3}$], nor the omniscient Bayesian model [Figure 18 (B) t-test, $p=1.10^{-11}$], nor the Gaussian model [Figure 18 (D) t-test, $p=6.10^{-3}$] presented the expected behavior. In particular, the Gaussian model did not switch as much as participants after low rewards.

Overall, these comparisons show that participants did more than reinforcement learning. It is also clear that participants never got a full knowledge of reward distributions: the “omniscient” model did not describe correctly human behavior even at the end of the task. The Gaussian model worked better, however it did not reproduce the switch behavior of participants for low rewards. It is a major difference as most switches happen following low rewards.

We propose to reject all these models.

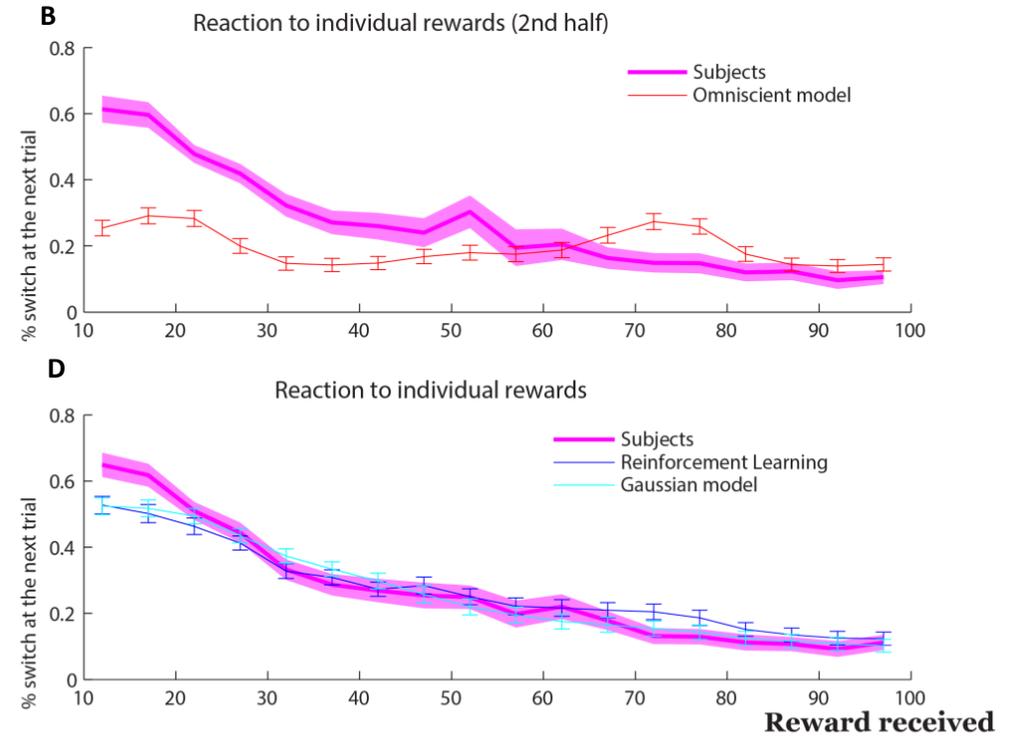
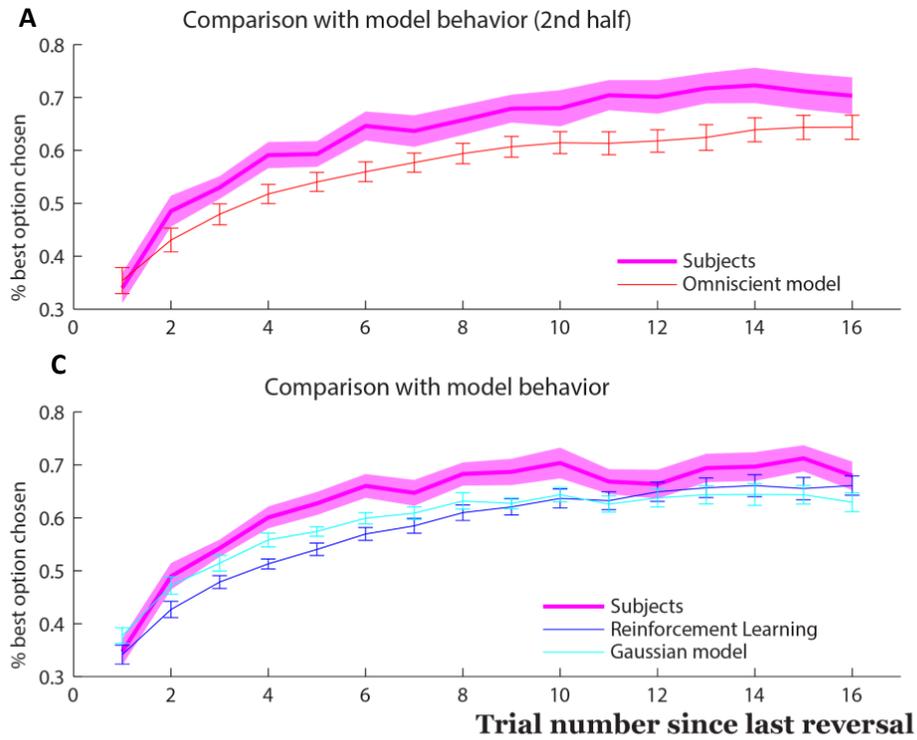


Figure 18 : (A) Adaptation curves of an omniscient Bayesian model (red) on the second half of trials compared to participants (in purple). (B) Proportion of switches at the next trial plotted against received reward at the current trial. Participants are in purple, simulation of the omniscient Bayesian model (fitted to participants) is in red. (C) Adaptation curves of a RL (blue) and Gaussian model (light blue) compared to participants. All trials are considered. (D) Predicted switch rate of a RL and Gaussian model compared to participants

Model NEIG results

We fitted the 4 free parameters of the NEIG algorithm to participants' data, simulated it and compared its statistics to participants' behavior. We show in [Figure 19] the adaptation curve predicted by NEIG (A) and its switch curve (B). In both cases, there was no significant difference between model and participants behavior (2-Way ANOVA: $F=0.02$ in A, t -test: $p=0.24$ in B).

To compare models to each other quantitatively, we computed the likelihood of our dataset for each model. We additionally used the Bayesian Information Criterion (BIC) to account for a different number of free parameters in each model. BIC is an approximate measure of model likelihood and is usually conservative. Therefore, we mention it as an indication and will not use it as a main argument.

In BIC and LLH, we confirm that NEIG explains data better than all algorithms described above (Q Learning, Gaussian model, multi-Gaussian Model, General Bayesian learning algorithm learning frequencies and which makes no hypothesis on the environment) [Figure 19 D].

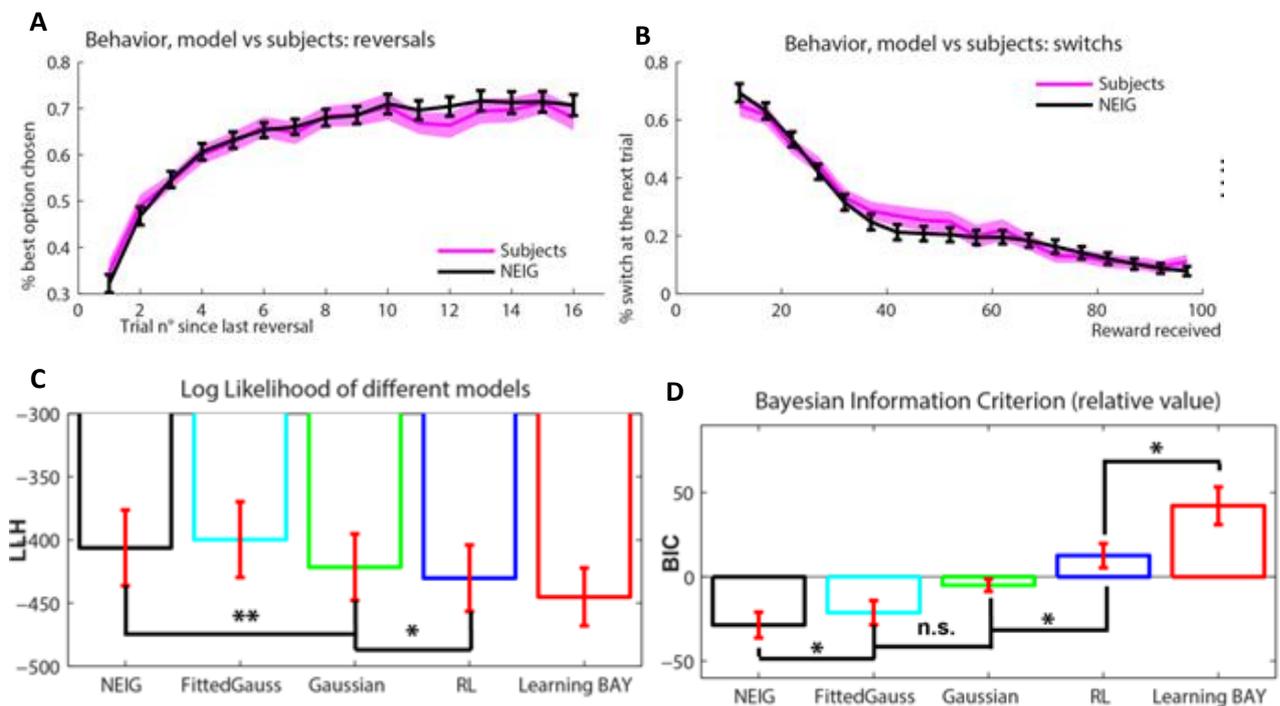


Figure 19 : (A) Adaptation curve of model NEIG (in black) compared to participants (in purple) (B) Switch behavior of model NEIG compared to participants (C) Comparison of the log likelihood (LLH) of different model. The LLH is measured on 936 trials and averaged across 25 subjects. NEIG is better than the Gaussian model ($p=0.009$). The Gaussian model is better than RL, and RL is better than the frequency learner model. (D) To account for parameter number, we computed the BIC. We represent a relative value of BIC where we withdrew inter subject variance. This is just for presentation purpose and does not affect any statistics. In BIC NEIG is significantly better than all models ($p=.04$ compared to Fitted Gaussian, $p=0.03$ compared to Gaussian model).

To characterize better the performance of NEIG, we plot the model internal representation of the different reward distributions at the end of the experiment. In Figure 20 A, we represent the distribution associated with the bandit with the highest expected value (p_{BEST}). No participant gets close to the actual distribution of the design (in bold green) albeit some display "bumps" at the location of the two modes. The representation of the distribution associated with the bandit with the

lowest expected value (p_WORSE , Figure 20 B) is even less similar to the design curve (in bold red). This comes from the fact that participants chose in average 78% of trials the bandit which was the most likely to be associated with the distribution with the highest expected value [range, 57% to 96%], thus this p_WORSE distribution is mostly learned through fictive counterfactual learning rather than real observations.

From these two distributions, we can also represent the “informational value” of each reward: the probability to have chosen the bandit with the highest expected value when this reward is observed.

This probability is measured by the function $R \rightarrow \frac{p_BEST(R)}{p_BEST(R)+p_WORSE(R)}$.

At the beginning of the task, both reward distributions p_BEST and p_WORSE are identical and flat. Therefore, the “informational value” is also flat (in black Figure 20 C). At the end of the experiment the “informational value” is almost a linear function of the reward (in magenta Figure 20 C).

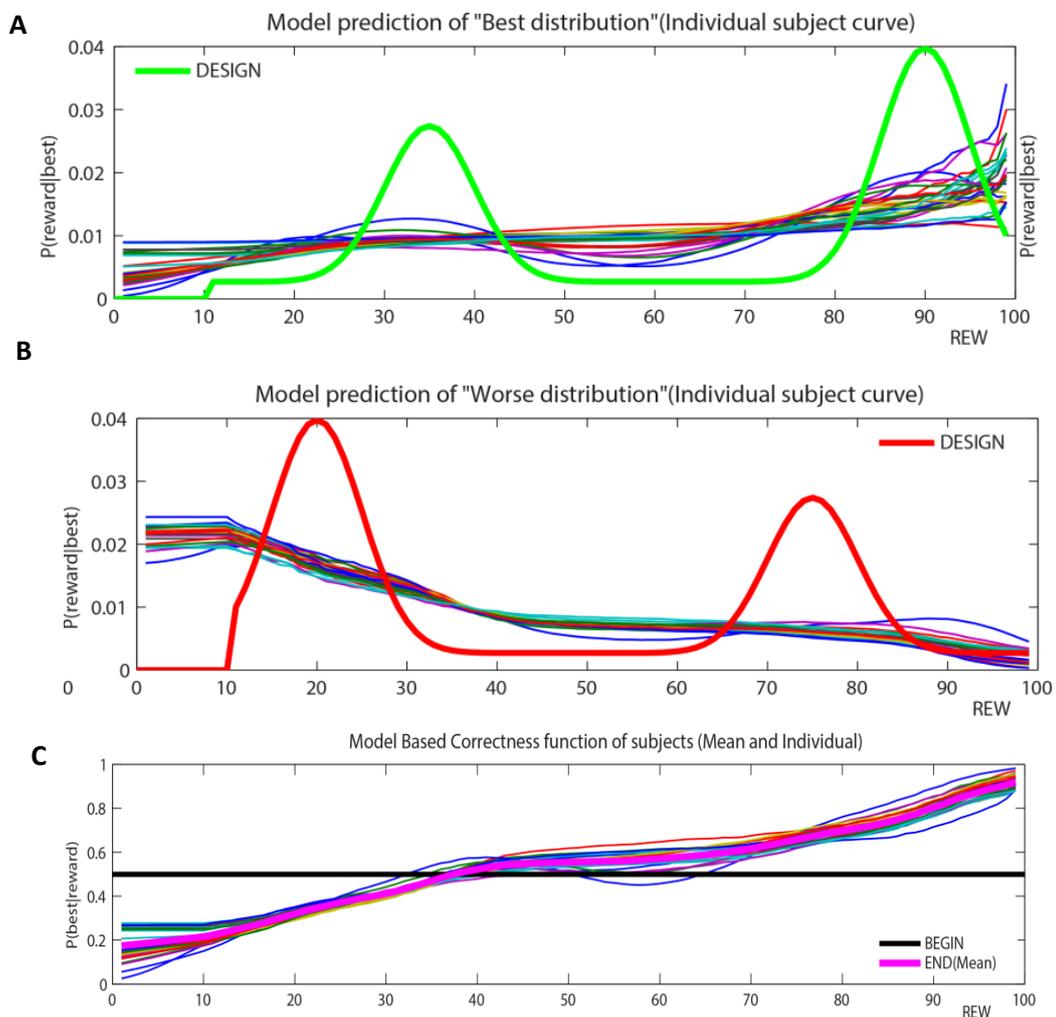


Figure 20 : We show for all participants the model inner representation of reward distributions. (A) Distribution of the bandit with the highest expected value. In bold is represented the bimodal distribution of the design. (B) Distribution of the bandit with the lowest expected value. In bold we represent the reward distribution of the design. (C) From these two distributions, we plot the probability that the observed reward is drawn from the distribution with the highest expected value. When the task begin this probability is flat (in black), at the end of the task it is almost a linear function (in purple)

Other models specific to the task

To probe further our model we compared it with two alternative models specifically adapted to the bimodal reversal learning task that participants perform.

We considered the extension of the Reinforcement Learning that we described p. 43, which integrates the anti-correlational structure between the two bandits in the task. The anti-correlation normalized the sum of the two Q values and we will name this model Normalized Reinforcement Learning (NRL) in the following.

We considered also an extension of the Gaussian model able to parametrically adjust two bimodal reward distributions described (p. 48). As we explained in the presentation of the Gaussian model (p. 47) the latter is a very simple parametric model but intrinsically unable to describe a bimodal environment.

We fitted these two models to participants' data and simulated how fast models change their behavior after a reversal. In Figure 21 A, we show that there is no significant difference between the predictions of Normalized Reinforcement Learning (in blue) and participants' data (2-way ANOVA, $F=0.15$). This similarity is confirmed for the propensity to switch after each possible reward (Figure 21 B, t-test, $p=0.28$).

On the converse the extension of the Gaussian model to bimodal distributions (in cyan) is unable to fit human behavior. We wanted to check that this weak performance was not due to the hypotheses added to reduce the computational complexity of the online adjustment of multi-modal Gaussians. Therefore, we tested a more complex alternative which keeps for both reward distributions the full memory of all obtained rewards. It then re-computes at each trial the best bimodal distribution accounting for the memorized reward plus the immediate reward observed.

This model was long to run and is biologically implausible. Moreover it also fails to account for human behavior (Figure 21 A, in black).

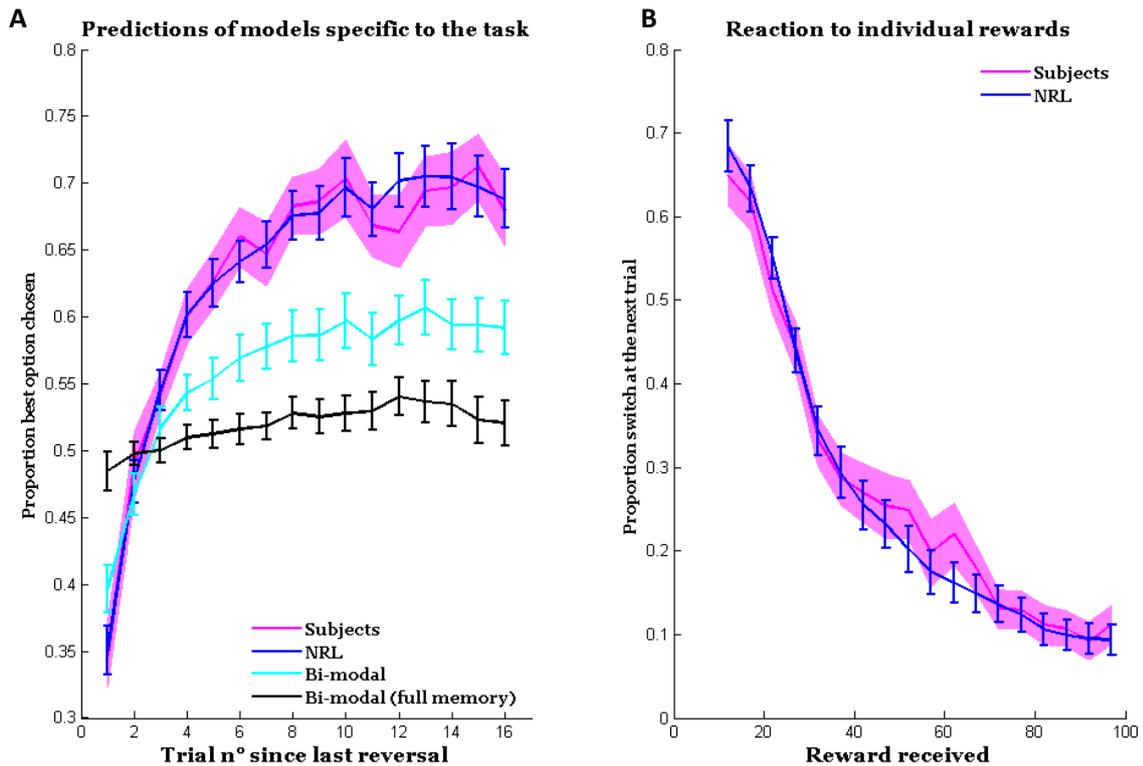


Figure 21 : Human behavior compared to algorithmic models specific to the task. (A) Adaptation curves after reversal. Subjects are in magenta. Normalized RL in blue, The bi-modal parametric model is in cyan, and the bi-modal model which keeps all memory of obtained rewards is in black. (B) Tendency to switch after each individual rewards. Subjects in magenta are compared to Normalized RL in blue.

Normalized Reinforcement Learning is a simple model which describes participants' behavior in this task as well as the NEIG model.

To push the comparison further, we re-analyzed data of another behavioral experiment designed and realized by a graduate student of our lab who already published most of it in her manuscript (Rouault, 2015).

Discrete task

This task is described in the Material & Methods. As a recall it is similar to the former task except that only 5 rewards are possible, 1 2 5 8 & 9. [Figure 22(A)].

We compare here the adaptation curves of real participants to the behavior of model simulations. We show on the same plot NEIG and the Normalized Reinforcement Learning.

We see on the figure that the Normalized Reinforcement Learning is close to what participants are doing but nevertheless differ statistically in several points [Figure 22(B), two way repeated measure ANOVA including trial number and model vs data as within subject factors: $F=6, p=1.10^{-2}$]. The behavior of NEIG is in all points statistically similar to participants ($F=0.07$).

Concerning the switch curves we can also observe that the behavior of normalized Reinforcement Learning is statistically different from participants when receiving reward 8 where participants tend to switch less than the model [Figure 22(C) t-test: $p=5.10^{-4}$].

We can also compare formally the 2 models in terms of LLH. Once more NEIG is statistically better than Normalized Reinforcement learning [Figure 22(D), $p=0.03$]. However NEIG has one extra free parameters compared to NRL: the standard deviation of the noise function with which obtained rewards are convolved. Thus, in BIC, which is a conservative criterion, despite the trend in favor of NEIG, the difference with NRL is not statistically significant.

However we rejected the Normalized Reinforcement Learning as it cannot reproduce the adaptation and the switch curves of participants.

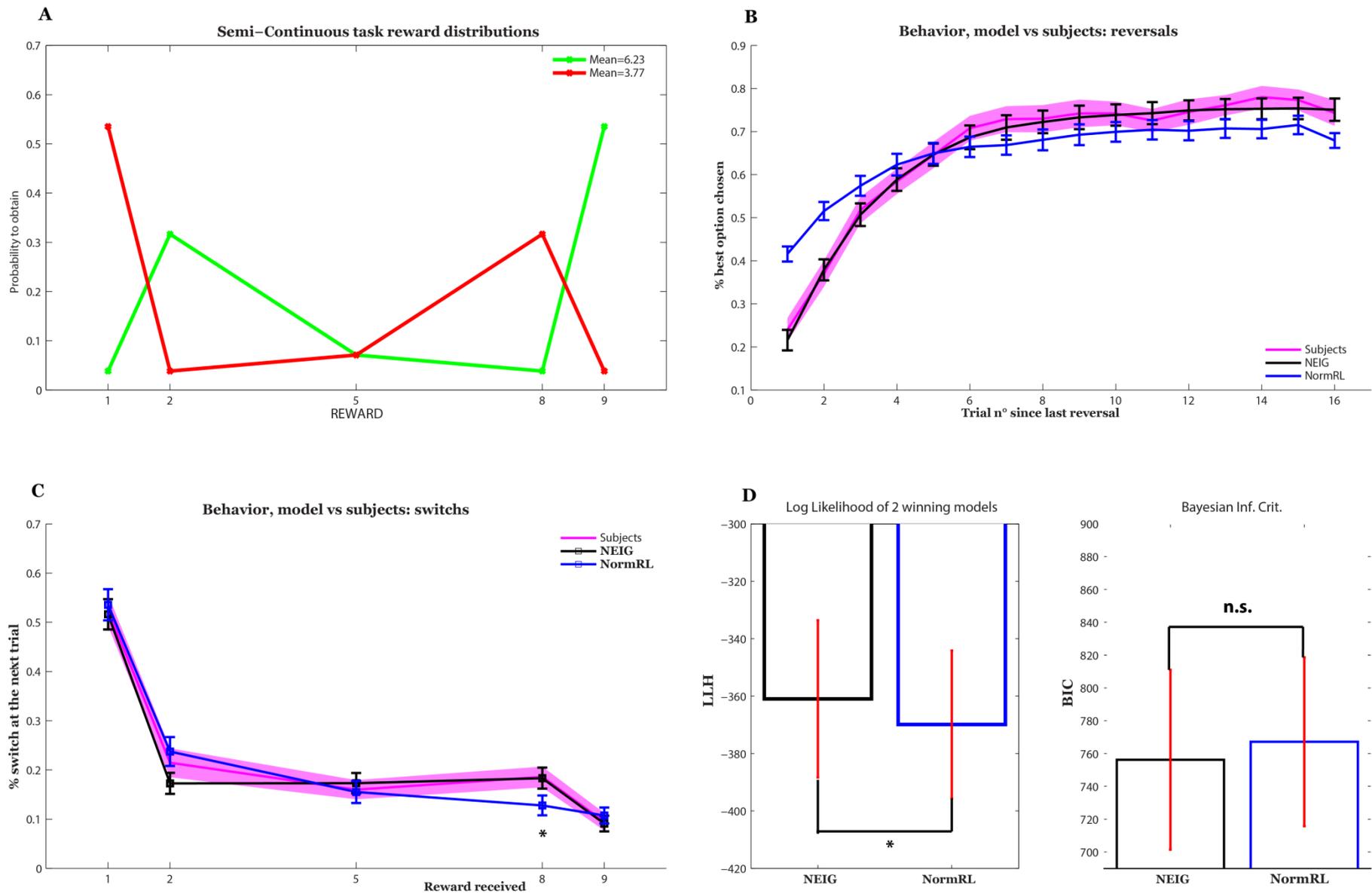


Figure 22 : (A) Semi-Continuous reward distribution used in a similar task. (B) Participants reverse their choice following the environment. NEIG predicts better participants curves than Normalized RL. (C) For switch it might be less clear, but the Normalized RL prediction in 8 is statistically different from participant's responses. (D) In predicting participant's responses, NEIG is better than NormRL ($p=0.03$). In BIC, NEIG has an additional free parameter, so they become statistically equivalent ($p=0.17$)

Generalization to K-bandit reversal task

Problem

In a two-bandit reversal task, reward distributions are stable across hidden states and only the mapping between bandits and distributions changes. This symmetry simplifies the selection rule as selecting a bandit according to its expected value once the uncertainty on the current mapping is accounted for (later named EV rule), is equivalent to selecting the bandit which is the most likely to draw from the reward distribution with the highest expected value (later named “BEST” rule).

To make mathematically explicit the difference between the two choice rule, let p be the probability distribution on the hidden states: $p(S)$ is the probability to be in hidden state S . Each hidden state represents a mapping between symbols and reward distributions. Let $V_i(S)$ be the expected value of bandit i in hidden state S : it is the expected value of the reward distribution mapped to i in S . Let's use the notation $p_i(S)$ as the summed probability of all hidden states where bandit i is mapped to the distribution with the highest expected value.

EV rule: choose i_1 as the maximizer of $\sum_S p(S) * V_i(S)$

BEST rule: choose i_2 as the maximizer of $p_i(S)$

In a 2-bandit reversal task, $p_i(S) = p(S_i)$, $V_1(S_1) = V_2(S_2)$ and $V_1(S_2) = V_2(S_1)$, thus

$$i_1 = i_2$$

When there is in the task $K > 2$ bandits, and that they can all be chosen at every trial, the two choice rules do not always predict the same choice. Indeed there are $K!$ possible hidden states (all possible mappings between bandits and reward distributions) and let's give a toy example with $K = 3$ to demonstrate this difference:

Let Distr_1 (in green, mean reward = 50), Distr_2 (in purple, mean reward = 30), Distr_3 (in red, mean reward = 5), be 3 reward distributions. The 3 bandits are represented by a circle, a square and a triangle. There are 6 possible hidden states and we describe in Figure 23 the current probability distribution over hidden states.

	Hidden state	Probability
a		0.05
b		0.05
c		0.15
d		0.05
e		0.25
f		0.45

Figure 23 : 3 bandits, 3 reward distributions. Example of a probability distribution over hidden states

We can marginalize the distribution over the hidden state to get for each bandit, the probability that it is mapped to the 3 possible reward distributions. In Figure 24, we represent these marginal probabilities and the expected value of each bandit.

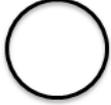
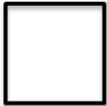
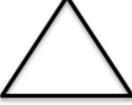
Distribution \ Symbol	Distr 1: Mean=50	Distr 2: Mean=30	Distr 3: Mean=5	Expected Value
	0.3	0.6	0.1	33.5
	0.5	0.1	0.4	30
	0.2	0.3	0.5	21.5

Figure 24: For each bandit, we represent the probability that it is mapped to Distr1, Distr2 and Distr3. We compute the expected value of bands. In this example the square has the highest probability to be mapped to the distribution with the highest expected value, but it has a lower expected value than the circle which is then the best choice to perform.

On this example, the circle bandit has the highest expected value and would be chosen by the EV rule. However, it is the square bandit which has the highest probability to be mapped to Distr_1 and which would be chosen according to the BEST rule.

Albeit it is more optimal to choose with the EV rule, choosing according to the BEST rule simplifies the computations. Indeed using the BEST rule is equivalent to considering the “reduced” hierarchical model where the hidden state is the identity of the bandit drawing from the distribution with the highest expected value. Thus, there are only K hidden states to track (instead of $K!$ with the EV rule). Also the BEST rule enables to learn, memorize and perform inference on only 2 reward distributions. Indeed for this reduced hierarchical model a bandit is either mapped to the distribution with the highest expected value, either mapped to any other reward distribution: whether it is the second best or the worst reward distribution makes no difference.

To explicit the two possible hierarchical models, we consider for simplicity the situation where all reward distributions are known. In Figure 25 left we represent the 3 reward distributions and a mapping between rewards and distributions. This is the hierarchical model associated with the EV rule, the hidden state is the mapping between bandits and distributions and there are as many reward distributions as bandits.

We present on the right the alternative model associated with the BEST rule. The hidden state is the identity of the bandit mapped to the distribution with the highest expected value. The $K-1$ other bandits draw from a unique distribution which is an average of the $K-1$ non optimal distributions.

For $K=2$, $K!=K$ and the two descriptions are equivalent.

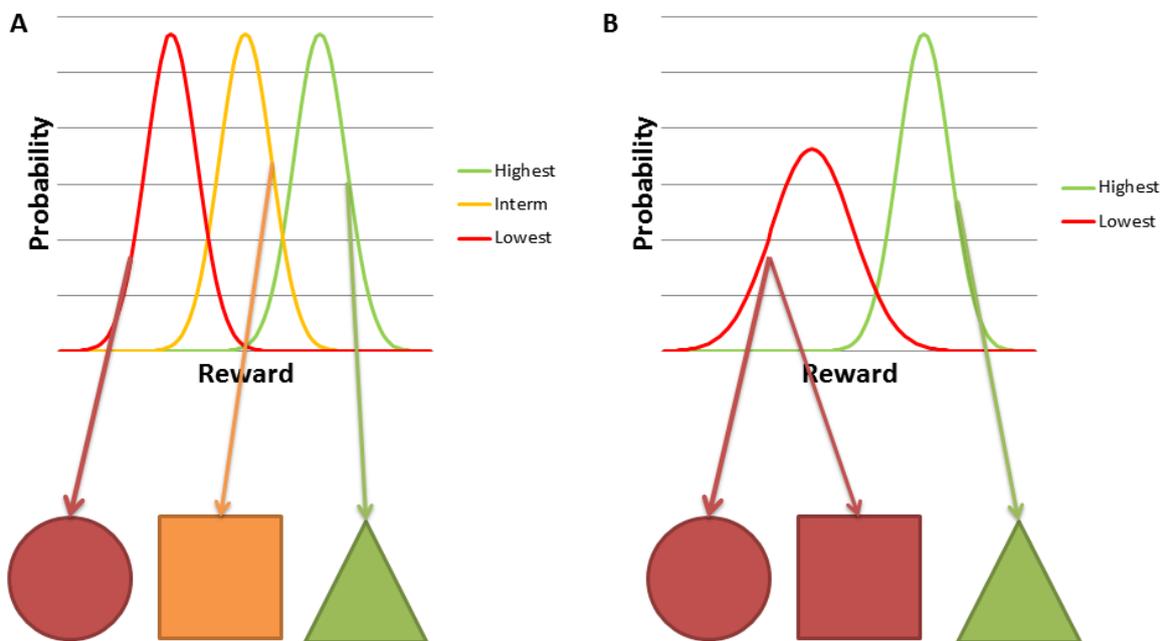


Figure 25 : Illustration of a similar environment, described by two internal model of the world. (A), each bandit draws from a separate reward distribution. The hidden state represents the mapping between the two. (B) One bandit draws from the best distribution, the two other from the “worse” distribution. The hidden state is the identity of the best bandit.

The equivalence between a choice according to the BEST rule and the reduced hierarchical model presented in Figure 25 B, is mathematically due to the independence of hidden states:

$$p(S_i \cup S_j) = p(S_i) + p(S_j)$$

We explicit the mathematical computation in the example described above with 3 bandits and 3 reward distributions. Let's assume that at trial t , \circ is chosen and reward R received. Now the posterior probability that \circ draws from Distr_1 (that we note $p(D1|R)$), Distr_2 ($p(D2|R)$) or Distr_3 ($p(D3|R)$) need to be computed.

One can compute with Bayes rule:

$$p(D1|R) = \frac{p(R|D1) * p(D1)}{p(R)}$$

$$p(D2|R) = \frac{p(R|D2) * p(D2)}{p(R)}$$

$$p(D3|R) = \frac{p(R|D3) * p(D3)}{p(R)}$$

with $p(R) = p(R|D1) * p(D1) + p(R|D2) * p(D2) + p(R|D3) * p(D3)$

When Distr_1 is the target distribution only $p(\circ = D1)$ is relevant to apply the BEST rule. Therefore, instead of computing both $p(D2|R)$ and $p(D3|R)$, one can instead compute

$$\begin{aligned} p(\overline{D1}|R) &= p(D2 \cup D3|R) = \frac{p(R|D2 \cup D3) * p(D2 \cup D3)}{p(R)} \\ &= \frac{[p(R|D2) * p(D2|D2 \cup D3) + p(R|D3) * p(D3|D2 \cup D3)] * p(D2 \cup D3)}{p(R)} \\ &= \frac{[p(R|D2) * p(D2) + p(R|D3) * p(D3)]}{p(R)} = p(D2|R) + p(D3|R) \end{aligned}$$

Therefore, to track $p(\circ = D1)$, it is enough to know the likelihood function $p(R|D2 \cup D3)$ which is the mean of $D2$ and $D3$, the two sub-optimal reward distributions.

We propose that when all bandits are available at all trials, humans make their choice based on the less optimal but simpler BEST rule. According to this assumption, participants need to learn only two reward distributions:

- (i) The reward distribution which has the highest expected value associated with the bandit that subjects intend to choose (target bandit);
- (ii) Another distribution which corresponds to the reward distribution over the $K-1$ bandits that participants intend "to avoid".

Extension of the algorithmic models

All the algorithmic models described on the 2-bandits tasks above can be easily extended to n -bandits tasks. For example the parametric Gaussian model would track the mean and variance of K Gaussian distributions as well as the mapping between these Gaussians and the n bandits [Figure 25 A].

The question of the choice rule used by humans, EV or BEST, is orthogonal to the computational model used to perform inferences on hidden states. For example, the parametric Gaussian model can alternatively be implemented according to the BEST rule. It will keep on tracking the parameters of 2 reward distributions and use inference to find the target bandit: the one which draws from the highest rewarded distribution [Figure 25 B].

The NEIG model is more naturally extended with the BEST rule. Indeed as we explained above in the formal description of NEIG, NEIG computes no expected value as the target bandit associated with a particular hidden state remains constant all across the task. Therefore, the computation of bandit's expected value as proposed by the EV rule is unlikely as it requires the combination of hidden states likelihoods and distributions' expected value, the latter being not computed by the model. Additionally, in the introduction, we underlined the difficulties associated with the computation of expected values in continuous environments to which the design of model NEIG answers.

On the converse, the BEST rule extends the basic principles of NEIG: Keeping two reward distributions enables to separate the reward space in two across the obtained reward, thus classifying all rewards above or below in the counterfactual reward distribution. On the opposite, having K reward distributions would require additional hypotheses to maintain the order of the K cumulative reward distributions.

In the following, we will first make explicit the algorithmic model extending NEIG for K-bandits tasks. We will then propose an experimental paradigm with 3 reversing bandits on which we tested 21 participants. On experimental data we will show that for all algorithmic models studied, the BEST rule describes better participants' behavior than the EV rule. It will then be shown that the NEIG model is again the best algorithmic model of human behavior.

Extension of the NEIG Algorithm

We give the pseudo-code of NEIG in a K-bandit reversal task, selecting bandits according to the BEST rule and tracking only two reward distributions.

Let $f_{HIGH}(R)$ be the number of occurrences of R when the target bandit (the one which most likely draws from the target distribution) is chosen, and $f_{LOW}(R)$ the number of occurrences of R when any other bandit is chosen.

Let N be the noise function which is convolved to the reward. $p_{BEST}(i, t)$ is the probability that bandit i draws from the target distribution at time t. $\sum_i p_{BEST}(i, t) = 1$

T=0

Initialize f_{HIGH} and f_{LOW} . $\forall R, f_{HIGH}(R) = f_{LOW}(R) = 0$

Initialize $\forall i, p_{BEST}(i, 0) = \frac{1}{K}$

For T=1:trial

Choose bandit J according to $Softmax(p_{BEST}(i, t), i = 1:k)$

Reward R is received

Obtain $f_{OBS} = R \otimes N$, the convolution of R with noise function N

If $p_{BEST}(J, t) > 0.5$ %The chosen bandit is the target bandit, EXPLOITATION

%Update f_{HIGH} with the factual reward

$$f_{HIGH} = f_{HIGH} + f_{OBS}$$

%Update f_{LOW} with a fictive counterfactual reward inferior to R. The total added weight is 1

$$f_{LOW}([R_{min} R]) = f_{LOW}([R_{min} R]) + 1/(R - R_{min})$$

end

If $p_{BEST}(J, t) \leq 0.5$ %The chosen bandit is not the target bandit, EXPLORATION

%Update f_{LOW} with the factual reward

$$f_{LOW} = f_{LOW} + f_{OBS}$$

%Update f_{HIGH} with a fictive counterfactual reward superior to R. The total added weight is 1

$$f_{HIGH}([R R_{MAX}]) = f_{HIGH}([R R_{MAX}]) + 1/(R_{MAX} - R)$$

end

%Update hidden states likelihood

$$p_{BEST}(J, t + 1) \propto p_{BEST}(J, t) * f_{HIGH}(R) / (f_{HIGH}(R) + f_{LOW}(R))$$

%probability that chosen bandit is the most rewarded

$$p_{BEST}(i \neq J, t + 1) \propto p_{BEST}(i \neq J, t + 1) * f_{LOW}(R) / (f_{HIGH}(R) + f_{LOW}(R))$$

%probability that chosen bandit is the less rewarded

%Include the possibility that hidden state changes at the following trial (volatility)

$$\forall i, p(i, t + 1) = (1 - \alpha) * p(i, t + 1) + \frac{\alpha}{k-1} \sum_{u \neq i} p(u, t + 1)$$

end

Experiment

We tested 21 participants on a 3 bandits reversal task using a continuous reward scale. 3 symbols were proposed, a square, a circle and a triangle. We kept the idea of bimodal reward distributions that were used in the 2 tasks described in the preceding chapters. In Figure 26 we plot the used reward distributions.

The reward distribution with the highest expected value (67.2) has two modes, one centered on 48, the second centered on 88. An “intermediate” distribution has two modes centered on 35 and 75 and has the same expected value as what playing at random would pay (55). The distribution with the lowest expected value (42) has also two modes centered on 22 and 62 .

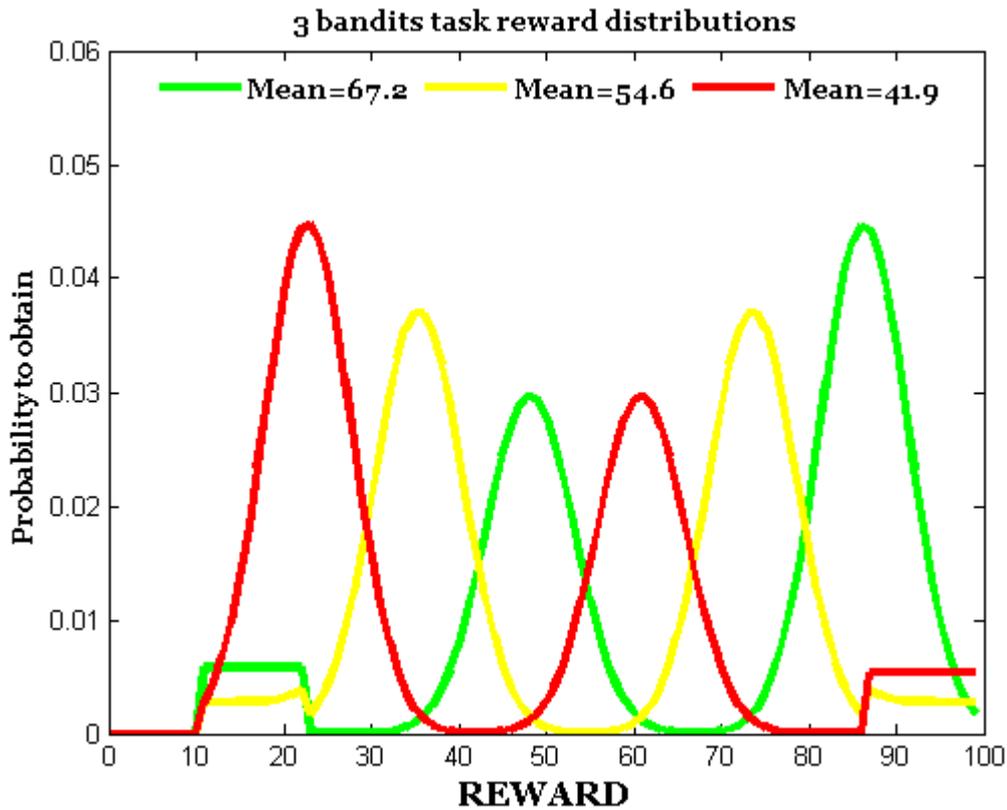


Figure 26: Reward distributions for the 3 bandits' task. Distributions are bimodal and have well separated means.

Our task presented reversals in the mapping between symbols and reward distributions. On Figure 27, we show a typical run: in green the symbol drawing from the highest rewarded distribution, in yellow the symbol drawing from the intermediate reward distribution, in red the symbol drawing from the lowest rewarded distribution.

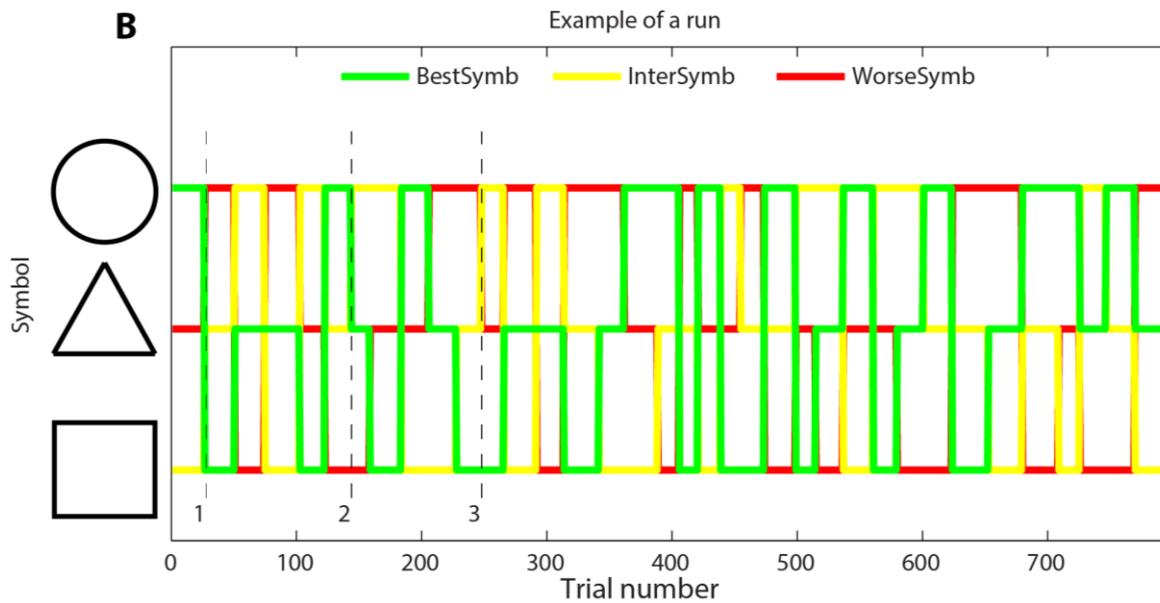


Figure 27 : Typical run of the tasks with 3 bandits. We propose to follow the distribution from which the different symbols draw across time. At the beginning the circle draws from the best distribution, the square from the intermediate distribution and the triangle from the worse distribution. After the first reversal, the square now draws from the best distribution, the triangle from the intermediate distribution, and the circle from the worse distribution. We can distinguish 3 types of reversals. See text for more details.

We can distinguish 3 types of reversals describing how the mapping between the distributions and the symbols change. We highlight in Figure 27 each type of reversal.

In trial 25 (index **1**), the target bandit becomes the bandit associated with the distribution with the lowest expected value (“worse” bandit) after the switch (green to red). These switches are likely easier to detect as the obtained reward presumably drops considerably.

In trial 130 (index **2**), the target bandit preceding the switch becomes the bandit associated with the distribution with the intermediate expected value. These switches are likely less easily detected as the drop of obtained reward is lower.

In trial 250 (index **3**) (\approx trial 250), the target bandit remains the same and only the bandit associated with the intermediate and worse distribution swap. If participants presumably choose the target bandit subjects are unlikely to detect such switches.

The experimental conditions were identical to the continuous task with two bandits described earlier (see p. 58), except that participants performed 1040 trials, and received no intermediate bonuses during each break.

Comparison of the BEST and the EV rule

We presented above the differences between the BEST and the EV rule for choosing a bandit based on beliefs about the current hidden state. We also showed that an algorithmic model using the BEST rule could be described with a reduced hierarchical model tracking only two reward distributions.

Our working hypothesis is that participants are selecting actions using the BEST rule. To test this hypothesis we compared the fit of the 6 following algorithmic models to our data:

-A Bayesian general frequentist learning hierarchical model (described p.45) learning either (a) 2 or (b) 3 distributions. For 3 distributions, we used either the (bB) BEST rule or the (bEV) EV rule. We added to the model a convolution with a Gaussian Noise function to account for continuity (as for model NEIG).

-A fitted Gaussian model where the means and variances of the reward distributions are considered as free parameters and fitted to participants' behavior. We fitted the mean and standard deviation of (u) 2 or (v) 3 Gaussian distributions to represent the environment. Once more, for the 3 Gaussians case, we could use either the (vB) BEST rule or the (vEV) EV rule.

As explained above, (a) and (bB) are theoretically similar, as for (u) and (vB). However, in practice, learning an intermediate distribution can change marginally model behavior.

Experimental Results

Despite the increased difficulty of the 3-bandits task, participants still performed the task and properly adapted to reversals. Figure 28 shows participants' adaptation curve following reversal. At the end of an episode, participants chose the target bandit (associated with the distribution with the highest expected value) on average 60% of trials, while chance level was 33%.

Figure 28 also shows the different adaptation curves following each type of reversals. As expected, participants reacted faster to type 1 switches (when the target bandit becomes associated with the distribution with the worse expected value, in red) than to type 2 switches (when the target bandit becomes associated with the distribution with the intermediate expected value, in yellow): the proportion of choosing the bandit with the highest expected value is significantly larger at trial 2 (t-test, $p=0.01$) following type 1 switches. As expected also subjects continued to select the target bandit with probability 60% when type 3 switches occurred (the worse and intermediate bandit swapped)

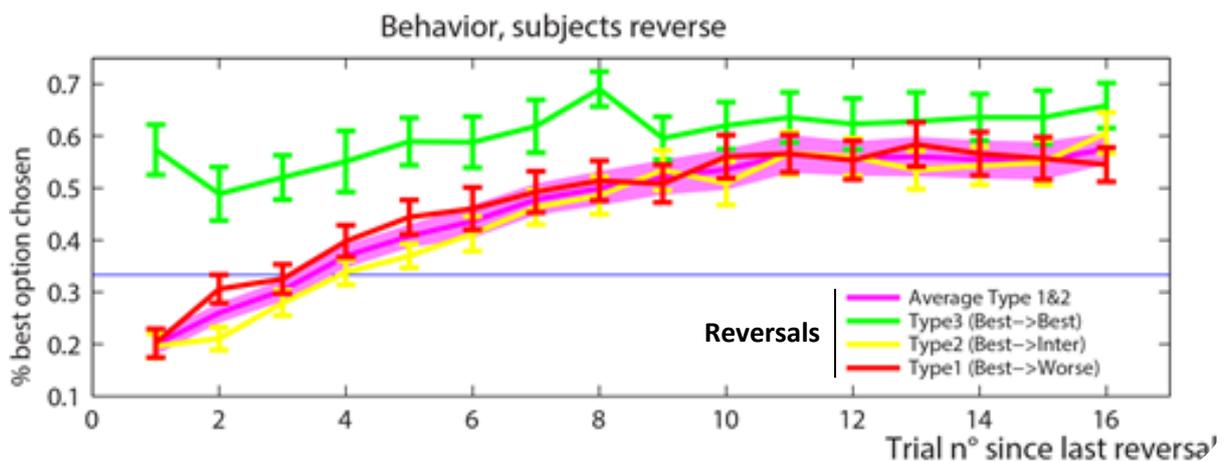


Figure 28 : Subjects reverse after a switch. The chance level is 33%. We show 3 reversal curves associated with the 3 types of switch. In X the number of trials since last reversal, in Y proportion of trials where the bandit associated with distribution with the highest expected value is chosen

Among the 6 alternative models, those associated with the Best rule fit subjects' data better than the others. In Figure 29, we compared the LLH of the 6 alternative models presented above. As expected from our mathematical derivation (p. 81), the frequentist Bayesian learning model (in yellow on the right) using two reward distributions (in black, model a) or 3 reward distributions associated with the

BEST rule (in blue, model bB) had a virtually identical LLH. Moreover these two models fitted better subjects data than the model bEV (in cyan) using 3 reward distributions and the EV selection rule ($p < 1e-3$). This is in accordance with our hypothesis that humans choose according to the BEST rule.

Concerning the fitted Gaussian algorithm (in grey on the left), there were no significant differences between the three possibilities, u, vB & vEV in LLH.

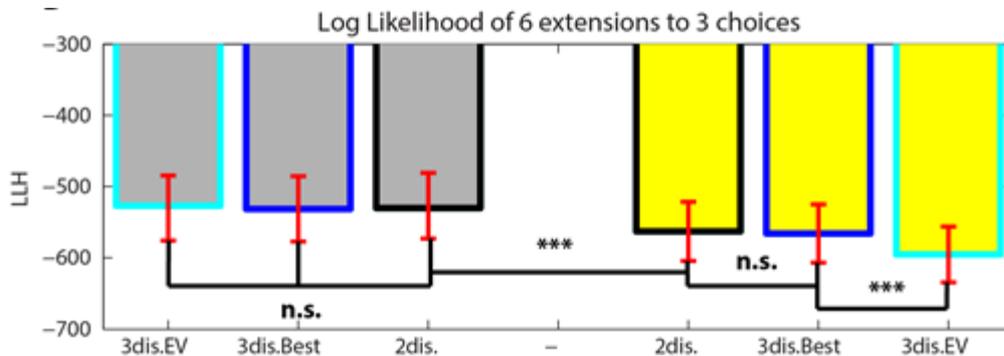


Figure 29 : We compared 6 models in LLH. In grey, non-learning fitted Gaussian. In yellow general Bayesian learning models. These models either tracked 2 distributions (in black), 3 distributions with a BEST decision choice (in blue), or 3 distributions with a Expected Value (EV) decision rule (in cyan). In red, standard error between participants.

To separate the three Gaussian algorithmic models, we simulated the models and compared model simulations to adaptation curves of participants. On this measure, the model using 3 distributions with the BEST rule (vB) cannot reproduce participants' data (two way repeated measure ANOVA including trial number and model vs data as within subject factors: $F=16, p=7.10^{-5}$).

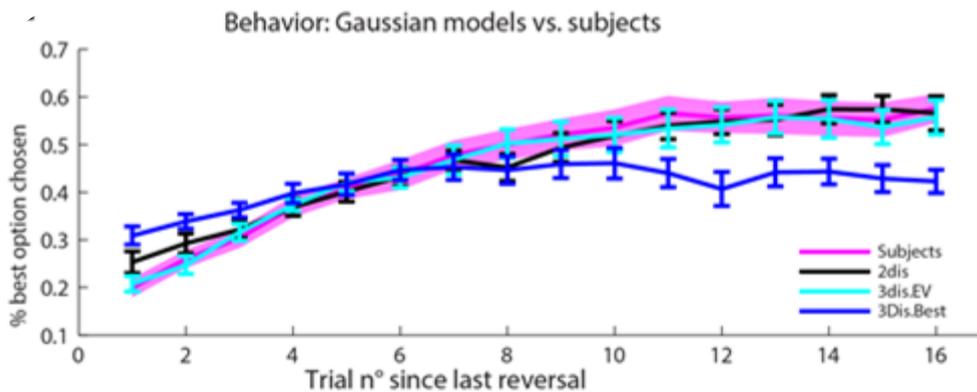


Figure 30: Adaptation curves of participants (in magenta) compared to the prediction of the different Gaussian models

To separate the 2 remaining models, the one using 2 reward distributions, u, and the one using 3 distributions and the EV choice rule (vEV), we used specifically reversals of type 3 (see

Experiment and [Figure 27], the target bandit remains the same and the two others swap).

Theoretically a model working with 2 distributions is indifferent to these reversals and should not react. On the converse a model tracking the mapping associated with the 3 distributions will spot the change.

Specifically on these reversals, the (vEV) model behaved significantly differently than participants [Figure 31] (t-test: $p=0.01$). On the converse behavior of (u) was not significantly different from

participants' behavior (t-test: $p=0.2$). This showed that participants were blind to these type 3 reversals, as predicted by a model working with 2 distributions.

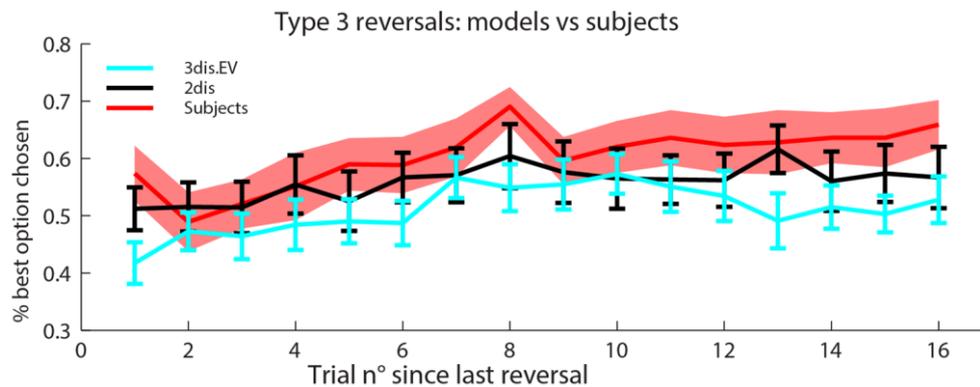


Figure 31 : We look specifically to Type 3 switches (see text), and show the reversal curves of participants (in red), the 2 distributions Gaussian model (in black) and the 3 distributions Gaussian model using an Expected value measure to choose (in cyan)

In both families of algorithms, models working with 2 distributions reproduced human data better. This validates our working hypothesis and justifies our extension of NEIG following the BEST rule. We will in the next part show that this extended NEIG model fits our data on this same 3-bandits task and compare NEIG to the extension of several algorithmic models already tested on the 2-bandits tasks.

NEIG fits data on the 3 options task.

We now compare the behavior of the NEIG model on the 3 option task to the best fitting alternative model analyzed above (fitted Gaussian model working with 2 reward distributions) as well as to the Reinforcement Learning and Normalized Reinforcement Learning models, and to the parametric Gaussian model described p.47.

The overall pattern of results is similar to that observed in the two-bandits tasks. Figure 32 (A) shows that the Bayesian information criterion of the different models was very close and only the Reinforcement Learning fitted significantly lower human data ($p < 1.10^{-3}$). This proximity between models can be attributed to the higher noisiness of human behavior in this task. For example the free parameter ϵ , representing the proportion of random choices doubled its value in this 3 options task compared to the continuous task. This increase was general to all models and as mentioned before, the noisier human behavior is, the harder it is to separate models.

To discriminate models further, we simulated the different models and compared their adaptation to reversals to participants' data. Figure 32 (B) shows that Normalized Reinforcement learning model and the parametric Gaussian model reproduced grossly participant's data but still significantly differed in several points ($F=8$, $p=5.10^{-3}$ for the NRL, $F=4$, $p=4.10^{-2}$ for the parametric Gaussian). In the figure we represent by a star the data points where a significant difference exists. NEIG followed participants' behavior slightly less well than on the 2 option tasks, but nevertheless were not significantly different from participants' data ($F=0.05$).

In Figure 32 (C-D) we plotted the tendency to switch after each reward of the simulations of the different models. On this behavioral measure again, it was difficult to separate models, whose predictions were in no point significantly different from participants' behavior.

We will therefore not use this 3-bandit experiment as an additional argument to separate the Normalized Reinforcement Learning, the parametric Gaussian model, and NEIG. Our results however validate the extension of NEIG which correctly accounts participants' behavior with a mechanism which is as simple as in the 2 option task.

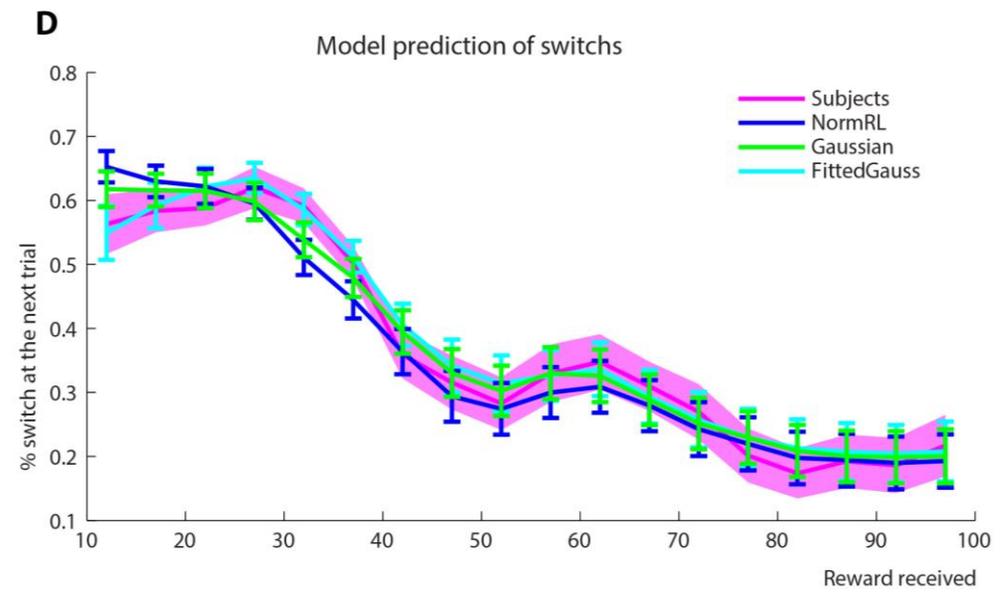
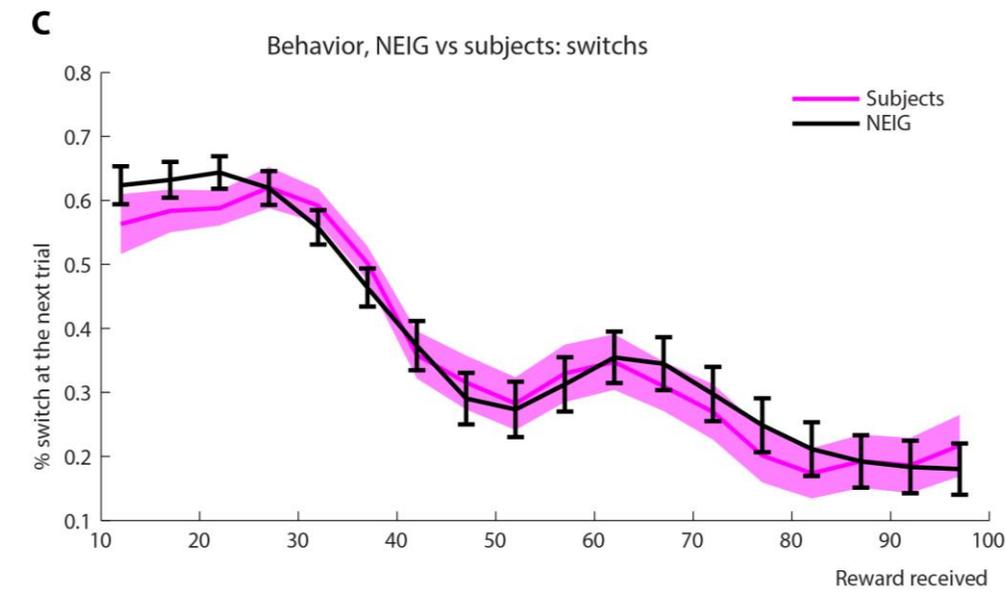
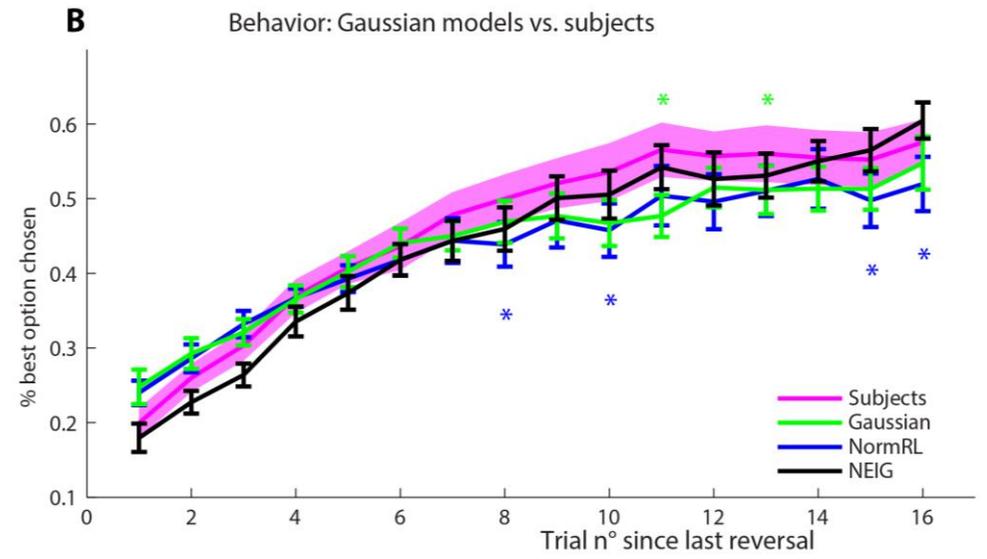
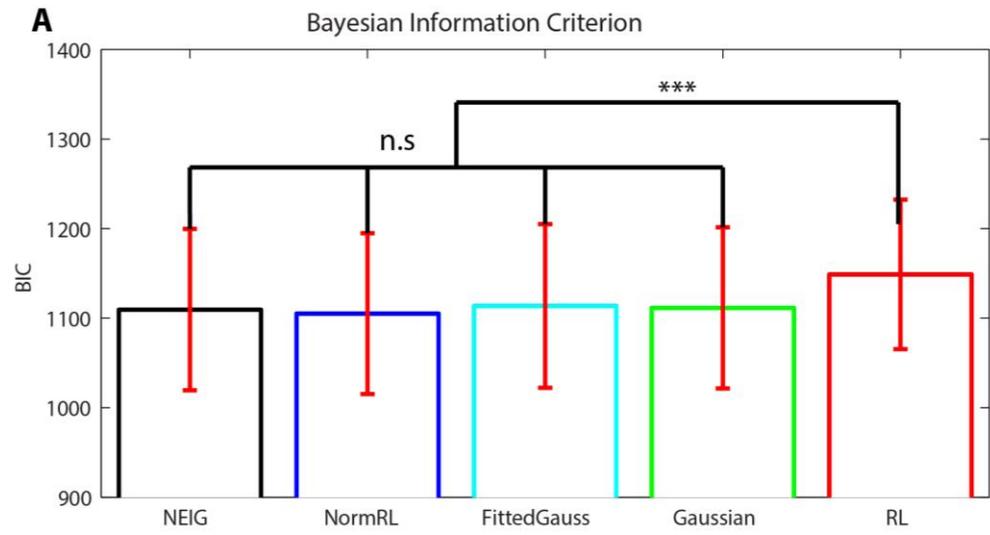


Figure 32 : (A) Comparison of BIC of different models on the 3 option task. NEIG, Normalized Reinforcement Learning, Fitted Gaussian model (tracking 2 distributions, winner of former figure), parametric Gaussian model (self-adjusting) are equivalent on the task, and all statistically better than Reinforcement Learning. (B) We compare the behavior of simulated models to subjects' behavior. Blue stars are trial numbers where the behavior of Normalized RL is significantly different than participants. Green stars are trial numbers where the behavior of the parametric Gaussian model is significantly different than participants. (C) Tendency to switch after each possible reward. Simulation of NEIG compared to subjects. (D) Other models.

Which algorithm is the most efficient in the 3 option task?

To go a little bit further in our analysis of this last experiment, we simulated the different algorithms to test which one would accumulate the most reward. We changed our fitting algorithm: instead of adjusting model's free parameters in order to maximize the correspondence between model predictions and participants' behavior, we tried to maximize for each model the total reward accumulated along the task.

Our goal was to compare these different algorithms on a ratio performance over complexity. It is difficult to propose comparative baselines as there is no known optimal algorithm on this type of task. Therefore, we considered as a lower bound a simple Q-Learning algorithm and as an upper-bound an unrealistic omniscient algorithm which would know from the beginning the task structure and the reward distributions.

To support our comparison between the EV and the BEST choice rule, we simulated the omniscient model using these two choice rules as well as the equivalent model working with the reduced hierarchical model (the hidden state is the bandit associated with the reward distribution with the highest expected value) and only two reward distributions: the one with the highest expected value on one side, the mean of the 2 other distributions on the other side.

Figure 33 shows that interestingly, the omniscient algorithm performs significantly better when it uses the BEST (blue) rather than the EV (cyan) rule. Similarly, as expected by our mathematical derivation (p. 81), the omniscient algorithm using the reduced hierarchical model (in grey) performs as well as the full model using the BEST rule.

The Normalized Reinforcement Learning (in red) and NEIG (in black) cannot compete with this omniscient model. Indeed NRL is based on a false structure of the task and NEIG learns structure from scratch. It is still interesting to note that NEIG is marginally better than NRL, what shows that at least its performance is not capped by the performance of NRL: NEIG can get closer from the real structure of the task than the simple heuristic representation of NRL.

In details, if we compare the performance of NEIG to the performance of Reinforcement Learning and to the best omniscient model, NEIG bridges 57% of the gap.

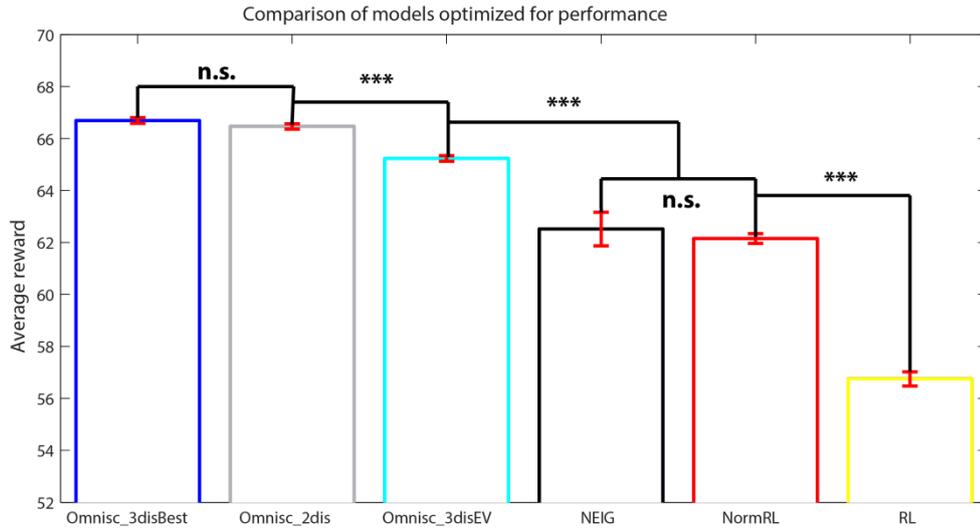


Figure 33: Comparison of the average reward received by several models playing the 3 option task. Mean chance reward is 55, best theoretical is 67.2. Parameters of each model are optimized to maximize performance. The omniscient model (knowing the distribution) performs similarly with 2 or 3 distribution tracked with a choice rule maximizing the probability to choose the best symbol. If the choice rule maximizes the expected value, the model lays significantly below. NEIG (the only learning model of the set) and Normalized RL are equivalent and significantly better than usual Reinforcement learning.

Discussion

Discussion of the model

In Introduction, we explained how building an internal generative model of the world, in particular through the construction of latent hidden states, enables to improve performance in a changing environment. The manipulation of hidden states is more computationally costly than Reinforcement Learning. However we showed evidence that human decision making is based on hidden states representation.

Almost all previously used experimental paradigms and algorithmic models have been based on binary (Win/Lose) rewards. In the present thesis we addressed the important issue of the interpretation of continuous reward values as information (“informational value”) concerning the quality of the action and about the current hidden states. In most real world environments, rewards are indeed valued on a continuous scale.

We showed that the simple intuitive extensions of binary algorithms, namely using a generative model of reward distribution, or building reward distributions through a frequentist count failed to describe sequential human decision making in the presence of continuous reward. Our work demonstrates new mechanisms underlying experience based decision making.

In three behavioral experiments we demonstrated that the best account of human decision making is the NEIG model which combines forward Bayesian inference for evaluating what is the best choice and learning reward distributions, and fictive counterfactual learning to compensate over-exploitation. Relaxing successively these assumptions, namely the use of generative hierarchical structure (Q-Learning), the simulation of counterfactual reward (General Bayesian model), and distribution learning (Gaussian model) failed to account for human decisions. In contrast to these alternative models, NEIG predicts human decision making and its variation across individuals in continuous, semi-continuous environments and in a generalization to 3 way choices.

Critically the NEIG model uses the heuristic that despite the stochasticity of obtained reward, there is a ranking of available actions. This ranking is not only based on action expected values, but on the relative position of their cumulative reward distributions on the full reward space. To keep this ranking constant and uniform on the reward space, when a distribution is modified by an observation, all other unobserved distributions are also modified making it equivalent to the fictive observation of counterfactual rewards (reward that would have been obtained if an alternative action had been performed).

One issue that remains to be addressed is that of mathematically demonstrating that NEIG is an exact probabilistic inference model under the assumption of a constant ranking of cumulative reward distributions.

Repetition bias

In all of the models that we tested on our data, we significantly improved model fits by adding a free parameter encoding a bias towards keeping the same choice as in the previous trial. This bias was implemented to be independent of whether the former choice was exploitative or explorative and

independent of the amount of reward received for this choice. The effect of this bias is orthogonal to our results as it does not modify the advantage of the NEIG model on other models.

Repetition bias has been consistently observed in psychology and interpreted differently. Some authors see it as an over confidence in prior beliefs compared to observations (Nickerson, 1998). This would correspond to a deviation from the optimal update of beliefs assumed by Bayesian inference (Phillips and Edwards, 1966). Other authors interpret this bias as a cost of switching (Hyafil et al., 2009).

Interestingly, the only model which is very sensitive to the presence of this repetition bias is the Normalized Reinforcement Learning model (NRL). Without bias it explains our human data less well than a simple reinforcement learning model. With a repetition a bias, it predicts data much better than a reinforcement learning (with bias) and is close to the best performing models.

Volatility

The parameter of volatility is an essential feature of Bayesian models (Hampton et al., 2006), (Boorman et al., 2009) which enables us to modulate the amount of evidence necessary to infer a change in environmental contingencies. As we explained in the introduction, humans have been shown to adapt to the frequency of environmental contingencies switches (Behrens et al., 2007), (Nassar et al., 2010).

We observed that the volatilities fitted on participants for the NEIG model were higher than expected (around 40%). This was also valid for other models using Bayesian inference. Interestingly, when we ran simulations trying to find the set of parameters maximizing the total amount of reward obtained by the model, we also found that the best performing models had a high perception of volatility (between 0.3 and 0.4).

We concluded that the parameter of volatility in the model does not only capture the subjective perception of the volatility in the environment, but also additional aspects of human behavior. Indeed in Bayesian models, the effect of volatility is to reduce the confidence in the current beliefs and to push the probabilities of the different hidden states towards the state of maximal entropy (uniform probability law on hidden states). This natural property of physical systems corresponds to a loss of information which could be due to an informational leak (Usher and McClelland, 2001) or computational noise (Drugowitsch et al., 2016) that are habitual characteristics of neuronal systems.

In the different Bayesian models we study, we considered that volatility was a constant parameter across the experiment. Finding high volatility values in the fit is likely a bias due to this assumption of a fixed volatility. It implies that to obtain a better fit the model tend to switch too much rather than too little. In the first case, after the model selects a different bandit from participants, it is flexible enough to come back to the initial choice. In the second case, many trials may be necessary to follow the switch of a participant, what is more costly in terms of likelihood.

A possible extension of our work would be to consider that volatility is adaptive and is monitored online to depend on recent outcomes (Behrens et al., 2007). This would allow for the specific tracking of changes in the external contingencies of the environment and facilitate behavioral switches when a change is inferred, while inhibiting switches in the middle of episodes to filter “expected” noise (Payzan-LeNestour and Bossaerts, 2011).

Reference frame

Our model predicts the “informational” value of rewards. Indeed some rewards are more likely in the distribution with the highest expected value than in other distributions. Therefore, obtaining them is positive and participants are more likely to maintain their choice on the next trial. Let’s call these rewards “positive rewards”. Conversely, some rewards are less likely in the distribution with the highest expected value than in other distributions. Therefore, obtaining them is negative and participants are more likely to switch their choice at the next trial. We will call these rewards “negative rewards”.

We could have expected a “satisficing” threshold, under which rewards would be negative and above which rewards would be positive (Simon, 1956). This threshold would be analog to a reference frame (Tversky and Kahneman, 1991) which would have been empirically adjusted across the task. In this frame, rewards inferior to the threshold would be interpreted as punishments and elicit avoidance. Conversely, rewards superior to the threshold elicit approach behavior (Palminteri et al., 2015).

Opposite to this prediction, in the tasks we proposed to our participants, our NEIG model predicts another shape which can be interpreted as presenting two thresholds. Below the first threshold, rewards are negative and elicit avoidance. Above the second threshold, rewards are positive and elicit approach behavior. Between the two thresholds, rewards are “neutral”: neither positive nor negative, and result in participants keeping their current choice, due to a repetition bias discussed above and which has been separately documented (Hyafil et al., 2009)

This division of the reward space in three parts enriches the binary vision defended by a satisficing criterion without bearing the complexity of continuity. Interestingly, NEIG uses binarization of the reward space on each trial to infer counterfactual reward and to build reward distributions. From this repeated binarization emerges in the long run the trinary representation of rewards.

Speed

NEIG model adapts to frequent reversals and, in the long term, also learns reward distributions. We propose here a measure of the speed of convergence of NEIG: how fast does it adjust to reward distributions?

To study the model dynamic, we measured through model simulations the distance between the final representation of distributions, and the representation at trial t . Averaged across participants we plot here [Figure 34] the average (L_1 norm) distance between distributions against the trial number. After 18 trials (less than one episode), the difference is 5%, after 50 trials (approximately 2 episodes), the difference is only 2%.

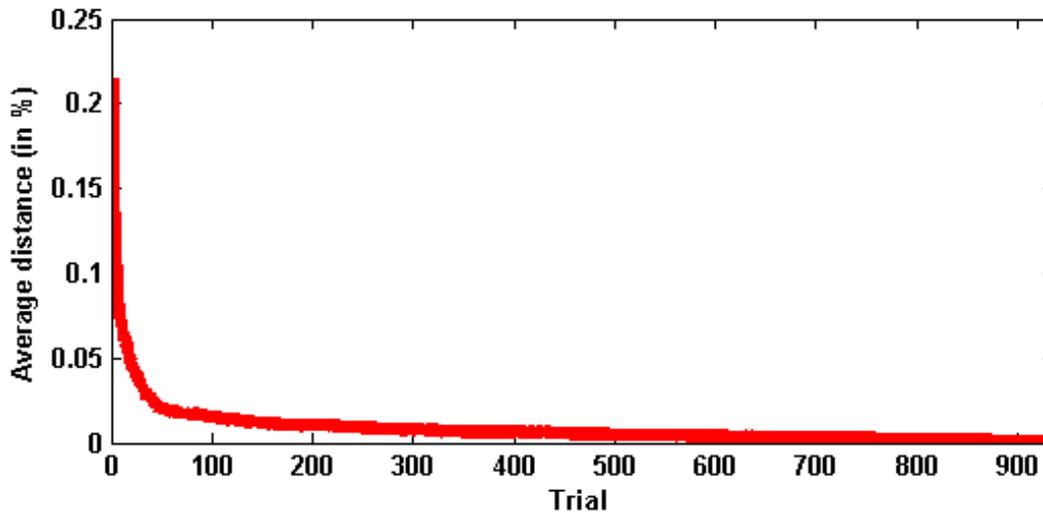


Figure 34 : L_1 norm distance between the final representation of rewards informational value and their current informational value across trials, averaged across participants.

This evolution is very fast as there are 90 possible reward values, and it is rare to have observed the same reward twice across 18 trials. After 50 trials, rewards were observed less than 4 times each. It is however sufficient for NEIG to build a reward “distribution” which enables participants to correctly perform the task. This property explains why we do not observe an improvement in participant’s performance along the task as most of the distribution learning is completed after a few episodes.

This is probably an answer to a constraint of real environments where the number of trials is limited and the lack of reward on several successive trials is detrimental. Therefore, it is likely that humans developed adaptive strategies that can converge over a few trials.

The observation that NEIG adjust quickly to its environment would fit with the idea of the existence of two decision systems (Posner & Snyder, 1975). In the beginning of the experiment, distributions are learned and the effect of additional learning quickly becomes marginal. Therefore, it is possible to imagine the alternative explanation that when the situation becomes habitual, a second system would use and apply learned distributions without further modifying distributions (Adams, 1982), (Daw et al., 2005).

The prediction of NEIG and a dual decision system differ in cases where reward distributions change in the middle of the task. NEIG considers that distributions are updated throughout the task, and that seeing more rewards adds “weight” to distributions and reinforces the confidence in their shape. Therefore, as time goes on, distributions become increasingly difficult to modify if a change occurs. Conversely, the prediction of a dual decision system is that distributions stop being updated at some point, so that the reaction to a change would be similar after 200, 500 or 2000 trials.

Sampling

The main interest of model NEIG is that in the context of forward inference, it solves the sampling problem. Indeed as we explained in the material and methods chapter, choosing mostly the perceived highest rewarded bandit leads to sampling rewards only from the highest rewarded distribution. However the informational value of rewards depends on both the highest rewarded distribution and the lowest rewarded distribution. Many exploratory choices are needed to acquire

knowledge about other distributions, which is costly in terms of short term reward and is not what is usually observed (Daw et al., 2006).

In the continuous task, participants chose on average 78% the bandit that NEIG considered as the highest rewarded [between 57% and 96%]. Despite the low proportion of exploratory choices, fictive counterfactual observations compensated for the learning of the lowest rewarded distribution.

This result is in accordance with studies that compared how rewards and punishments influence behavior. Authors observe that to obtain a similar behavioral effect; approach behavior for rewards, avoidance behavior for punishments; far fewer trials are necessary for avoidance behavior (Guitart-Masip et al., 2012), (Palminteri et al., 2012). This is exactly what our model predicts. Each time a punishment is successively avoided the punishment is still fictively experienced, which reinforces its negative representation and avoidance behavior.

Out of frame rewards

The NEIG model makes specific predictions when rewards from out of the frame of already received rewards are obtained. For example, when rewards scale between 50 and 100, and suddenly 1000 or 10 is received.

For a model which learns frequencies, since this surprising reward has not been observed in any of the reward distributions, it has a null informational value and therefore does not change the belief that the chosen bandit is the highest rewarded bandit. Overall the trial by trial account of volatility decreases this belief at the next trial.

Our model extends observations by heuristically using the natural direction of the reward axis. All rewards inferior to a reward presumably drawn from the highest rewarded distribution are classified in the lowest rewarded distribution and all rewards superior to a reward presumably drawn from the lowest rewarded distribution are classified in the highest rewarded distribution. Therefore, rewards superior to the highest observed reward are necessarily perceived as positive and we can expect participants to keep their choice at the next trial. Conversely, rewards inferior to the current lowest observed reward are necessarily perceived as negative and we can expect participants to change their choice at the next trial.

This effect differs from a frequency learning algorithm, but is similar to the prediction of a Q-learning algorithm.

Adaptability

We explained in our chapter on computational models that to simplify their learning in a continuous environment, some models use strong hypotheses on the structure of the environment as heuristics. The goal is to maximize the number of encountered situations where these hypotheses are accurate, as the model is otherwise unable to adapt.

NEIG is based on the heuristic that reward distributions in the environment are ordered by the position of their cumulative functions. This heuristic is more flexible than imposing a shape on reward distributions. For example it is verified when reward distributions are Gaussian functions with a similar variance and more generally when low rewards are attributed by the distribution to avoid and high rewards by the target distribution.

However this heuristic is false in many environments, and in particular in the different settings of our behavioral experiments. As we showed at the end of our results, NEIG can adapt to these settings and behaves better than a Reinforcement Learning model. We fitted the free parameters of NEIG on simulated data of an optimal observer playing the task, and plotted the reward distributions inferred by NEIG (averaged across 25 fictive runs of the task) at the end of the experiment. This gives us an upper bound on how well NEIG can track design distributions on this task (optimal-NEIG model).

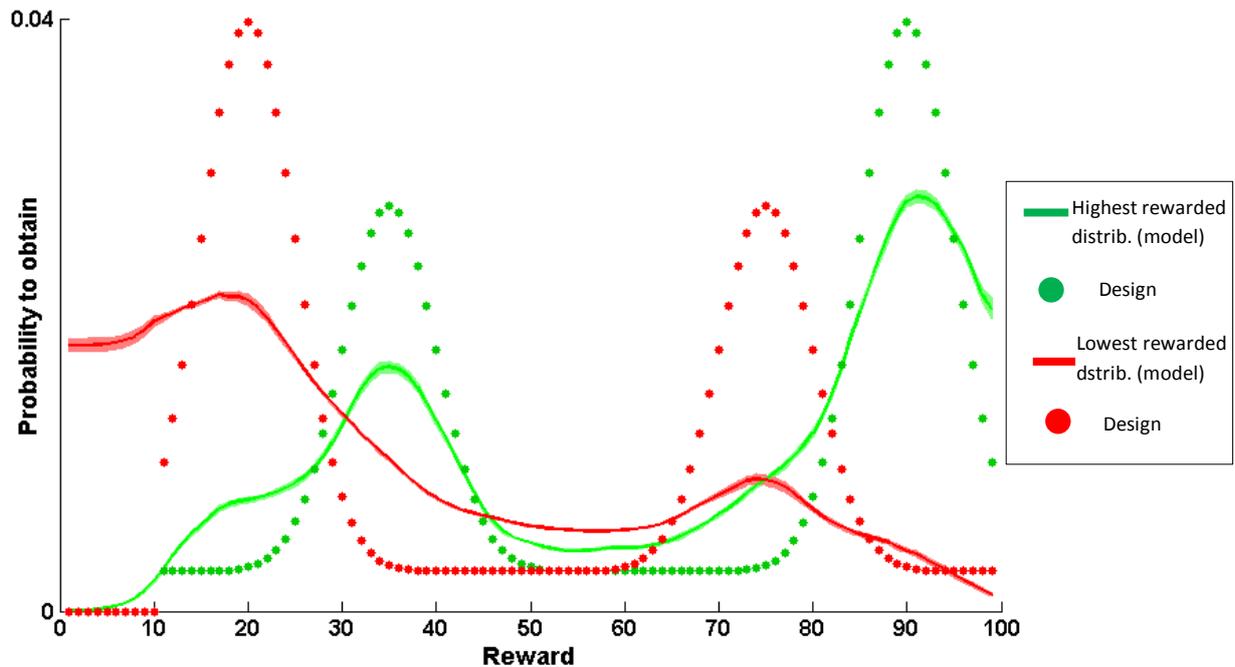


Figure 35 : Reward distributions as inferred by NEIG at the end of the experiment (in bold, error bars across 25 fictive runs of an omniscient observer) vs the design distributions (heavy points).

In Figure 35, we observe that NEIG is able of capturing a bimodal trend and sits the modes at their correct position. However it lacks precision on the relative size of the modes. In particular, in the lowest rewarded distribution which is less often chosen, modes are relatively flat, and the distribution looks like a piecewise constant function. For example 75, by design the most likely reward value of the highest mode of this lowest rewarded distribution, is equally likely in both distributions, and is thus not informative.

This underlines that the heuristic on NEIG is theoretically sufficiently flexible to approximate many different environments, even when these environments do not follow NEIG’s structural hypotheses. However NEIG’s behavior depends on the set of parameters used by the model, and we can explain the differences between participants’ perceived reward distributions and the optimal-NEIG model by the set of parameters they use.

First, as expected, participants have on average an excessive perception of volatility, observation noise and rate of random choice than is needed to match omniscient behavior. More surprisingly, participants explore, on average, less (as reflected by the superior beta rate) than the optimal NEIG model does. Individual participants use model parameters which are distributed across the average, and are thus closer to or further from theoretically optimal parameters.

Biological plausibility

Noise function

Our thesis question relates to the important problem of brains interacting with continuous environments. Biological systems are not infinitely precise: there is a minimal distance between two values they can represent. For example in perception, the notion of receptive field illustrates the finite resolution of human systems: the human eye has an angular resolution of 0.02° .

In the NEIG model, the observation is convolved with a noise function which diffuses the weight of the observation on neighboring rewards. This is compatible with the notion of overlapping receptive fields: the sensitivity of a neuron could be maximal for a given reward, but it could also be sensitive to rewards in the surround. This is something common for example in vision; a V1 cell will respond maximally to a bar oriented with its preferred orientation and minimally to a bar oriented with at the opposite angle. However this response is not binary and decreases in between the two with the angular distance to the preferred orientation (Hubel and Wiesel, 1959).

The notion of receptive fields is also commonly applied to numbers (Dehaene and Changeux, 1993) as neurons tune to numbers have been found in monkeys (Nieder and Miller, 2003). It is also a robust finding that the closer two numbers are, the hardest it is for humans and animals to discriminate them, even after training (Pollock, 1989).

Reward distributions

Concerning reward distributions, our model proposes that they are constructed by the sum of on the one hand, noisy responses to observations, and on the other hand, step-functions modelling counterfactual rewards. It is realistic to imagine neurons encoding step functions: observations would represent a threshold and neurons would fire for rewards under (resp. above) this threshold and not for reward above (resp. below).

We can then imagine a simple neuronal implementation of the model where rewards would be “binned” and mapped to the receptive field of neurons. These neurons would code either the highest rewarded distribution or, the distribution to avoid. Neuronal activation would depend on both prior belief and observation.

The size of reward bins is directly related to the standard deviation of the noise function that we use in our model to convolute the observation. We could imagine that this parameter is itself adjusted to environmental contingencies in order to have tighter bins in areas where rewards are more likely and larger bins in areas where rewards are rare. Adaptive coding of reward value has already been shown in dopamine neurons to depend on the current expected reward and its variance (Tobler et al., 2005). The exact algorithm driving bin size adaptation remains to be understood.

f-MRI predictions

The NEIG model explains how humans are able to perform predictions about which reward to expect following different available choices. Indeed NEIG adapts to environmental contingencies by constructing reward distributions. Through Bayesian inference it can invert reward distribution to have an appropriate behavioral reaction to rewards: switch after negative rewards, stay after positive rewards. We did not have the time to use f-MRI on one of our behavioral tasks however the literature allows us to make several predictions concerning the brain areas implicated in the process.

In particular, we can expect NEIG to be a model of the ventromedial prefrontal cortex (vmPFC). The vmPFC has been found to encode reward expectations associated with a stimulus (Padoa-Schioppa, 2011), (Plassmann et al., 2007). This result holds when the valuation is not only based on direct experience, but also when different information (model-based) is combined (Gottfried et al., 2003). In the NEIG model, reward prediction depends on learned reward distributions, which are based on both direct and counterfactual past experiences. This prediction will be compared to the actual outcome in the reward circuit and it is similarly known that prediction error signals are more complex than the simple application of the Rescorla-Wagner rule and include information about environmental structure (here beliefs and reward distributions) (Nakahara et al., 2004), (Bromberg-Martin et al., 2010).

Other parts of the NEIG model are similar to general Bayesian models. In particular, f-MRI studies have located the computation of beliefs and the representation of other information related to the current hidden state of the world in the (vmPFC) (Donoso et al., 2014), (Wilson et al., 2014).

An interesting question would be to predict where in the brain the counterfactual reward is simulated. According to one hypothesis, this corresponds to a mental simulation which, similarly to self-reflection has been observed in the fronto-polar Cortex (FPC), in particular when participants have to guess the intention of others (Frith and Frith, 2003), or to evaluate an alternative course of action (Boorman et al., 2009), (Donoso et al., 2014). Another hypothesis claims that simulating counterfactual reward is an automatic low level inference which could be implemented as a “model-based” valuation of the alternative action and be also coded in the vmPFC (Jones et al., 2012).

We tend to favor the second option. Indeed the FPC is activated when variables associated with hidden states, other than the current state, are tracked or computed. In the NEIG model, inferring the reward delivered by the alternative symbol is used to construct a representation of the alternative reward distribution in the same hidden state and not for changing variables (like beliefs) in another hidden state. Therefore, all the computations of NEIG are specific to a single hidden state, the current state, and it is parsimonious to expect to find them grouped in the vmPFC.

Extension of the PROBE model

In introduction, we mentioned several algorithmic models describing human decision making in binary environments. These algorithms construct a predictive model including the structure of the environment that depends on latent hidden states of the environment.

In particular, the PROBE model developed by Collins & Koechlin creates a hidden state structure and develops in each hidden state predictions about the result of actions as well as a strategy to exploit the environment (Collins and Koechlin, 2012). In the PROBE model, the current hidden state is estimated through the reliability of its predictions, which also guarantees that the associated strategy would be efficient.

In binary environments predictions are reduced to a simple number: the proportion of positive rewards expected. NEIG can be proposed to extend the PROBE model in a continuous environment. In parallel to the core principles of PROBE, NEIG would construct a simple predictive model of rewards that are more likely to be attributed when the best available action in the current hidden state is performed. The actual action outcome will generate a prediction error inversely proportional to its likelihood: high prediction error when this outcome is presumed rare, low prediction error

when this outcome is presumed likely. This will be interpreted in the hidden state space as validating or invalidating the current hidden state.

To test these predictions one may use the same experimental paradigm as Collins & Koechlin 2012 except that feedbacks will be evaluated along a continuous scale. Performing the “correct” action or the “incorrect” action would provide a reward drawn from two different, albeit partially overlapping, reward distributions.

Prediction of the model: no computation of expected values

In the NEIG model, computing expected values is never needed. Indeed for each hidden state there is a target bandit which is stable across the task. By design, the distribution associated with the bandit with the highest expected value in a given hidden state will maintain a higher expected value than other distributions when a new observation is included with the distribution.

NEIG considers that a change of preferred bandit is always associated with the inference of a change in the latent state of the environment. The alternative hypothesis would be that participants have re-ordered the values of the different bandits in the same hidden state, which would require a permanent computation of bandits’ expected values.

In introduction, we underlined the theoretical difficulties associated with the computation of expected values from probability distributions. We also mentioned how the initial normative idea of basing decisions on the computation of expected values has been modified to match human behavior; accounting for distortions in the representation of values (Von Neumann and Morgenstern, 1947) and of probabilities (Kahneman and Tversky, 1979) was necessary.

Recently papers have shown incidental evidence that challenge the idea that humans multiply a probability and a value to create a representation of the expected gain. Instead, according to Scholl & al, who compared several computational models on several data sets, the model describing human behavior the best applies the heuristic that probability and value are two independent pieces of information which are added to obtain the decision variable (Scholl et al., 2015). This was confirmed in another experiment of Donahue and Lee on monkeys (Donahue and Lee, 2015). The NEIG model is in line with these decision models which do not use expected values.

When we extended NEIG to n options we also departed from the normative solution which would compute the expected value of each bandit to favor a simpler model choosing first the most likely hidden states then the highest rewarded bandit in this hidden state. Again our solution avoids the computation of expected values.

Ordinal vs Cardinal

The NEIG model has an ordinal rather than cardinal representation of rewards. This representation depends on how many times a reward was classified in the highest rewarded distribution versus in the reward distribution “to avoid”. A reward is classified each time it is directly observed, or presumed to be the (fictive) counterfactual reward. The specificity of NEIG is that the counterfactual classification of a given reward only depends on whether it is above or below observation. Therefore, only the relative position of a reward and an observation (order) is relevant and there is no influence of the precise value in this phase. In total, the positive or negative representation of a reward depends on the number of observed samples of the highest rewarded distribution which were

superior or equal to this reward, and on the number of observed samples of the lowest rewarded distribution which were inferior or equal to this reward.

The ordinal representation of rewards is in line with the economic literature of the school of Pareto (Pareto, 1906) Hicks and Allen (Hicks and Allen, 1934), which considers that while choice is predicted by an ordering of options, it is impossible to quantify the degree of preferences of all choices on the same scale.

The only cardinal aspect of the model is the observation noise which diffuses the observation to neighboring rewards. The weight of this diffusion decreases with the numerical distance from the observation consistent with a cardinal representation of the reward space (Piazza et al., 2004).

Limitations and future directions

We have already mentioned in this discussion several limitations of our work. First the NEIG model supposes a stable volatility across the task and does not describe how this volatility is inferred. We concluded that this free parameter was probably capturing more than what volatility is usually referring to. More experimental work with the same design, but by for example varying the frequency of reversals (volatility of the design), or other experimental paradigms aimed at measuring computational noise (Drugowitsch et al., 2016) would be necessary to understand how factors other than obtained reward influence the temporal evolution of beliefs.

We also realized that the adaptation of the NEIG model to reward distributions was very fast: a few tens of trials. Instead of using the same reward distributions across the entire experiment, it would be interesting to cut the task in runs and to vary reward scales and distributions across these runs. This would enable us to study how the same participant adapts to different contingencies and to further test the NEIG model. In particular, we would better understand the role of the observation noise by varying the orders of magnitude of rewards: one run could present rewards between 1 and 100, another between 10 and 20, and a third using integer powers of 10...

One important aspect of the model is the modulation of the counterfactual reward by prior belief. When one believes one has chosen the highest rewarded bandit (exploitation), one presumes the counterfactual reward is inferior and the converse: when one believes one has chosen a bandit other than the highest rewarded (exploration), one presumes the counterfactual reward is superior. However there are only a few trials wherein participants deliberately choose a non-optimal bandit. It is therefore difficult to have a strong confidence in the proposed counterfactual mechanism for exploratory choices. It would then be interesting to force participants' choices on some trials in order to get a representative number of trials in which participants observe rewards from the distribution to avoid.

Following the same idea, we explained in the 3-bandit task that participants were not interested in differentiating the two non-optimal bandits as, on all trials, it was possible to choose the highest rewarded bandit. A 3-bandit task would be closer to real environments if there were trials where only 1 or 2 bandits were available. In particular, trials where the presumed best bandit is unavailable would be most especially interesting. If these "limited-choice" trials are rare, we would expect participants to choose the bandit having the second best probability to draw from the highest rewarded distribution, rather than from the bandit having the best probability to draw from the intermediate distribution. If these limited-choice trials are numerous, there would be a theoretical

interest to start tracking the intermediate distribution, and we could check whether participants are able to do these more complex computations when needed.

We also mentioned above that it would be interesting to use f-MRI methods on a similar task to the one presented here to understand the brain mechanisms behind our proposed algorithms as well as to test the interaction between the NEIG model and the PROBE model on a specific behavioral protocol.

Bibliography

- Acerbi, L., Vijayakumar, S., and Wolpert, D.M. (2014). On the Origins of Suboptimality in Human Probabilistic Inference. *PLOS Computational Biology* *10*, e1003661.
- Acuna, D., and Schrater, P. (2008). Bayesian modeling of human sequential decision-making on the multi-armed bandit problem. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, (Washington, DC: Cognitive Science Society), pp. 200–300.
- Acuña, D.E., and Schrater, P. (2010). Structure Learning in Human Sequential Decision-Making. *PLOS Computational Biology* *6*, e1001003.
- Adams, C.D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B* *34*, 77–98.
- Agrawal, R. (1995). Sample Mean Based Index Policies with $O(\log n)$ Regret for the Multi-Armed Bandit Problem. *Advances in Applied Probability* *27*, 1054.
- Allais, M. (1953). La psychologie de l'homme rationnel devant le risque: la théorie et l'expérience. *Journal de La Société de Statistique de Paris* *94*, 47–73.
- Anastasio, T.J., Patton, P.E., and Belkacem-Boussaid, K. (2000). Using Bayes' rule to model multisensory enhancement in the superior colliculus. *Neural Computation* *12*, 1165–1187.
- Anderson, C.M. (2001). Behavioral models of strategies in multi-armed bandit problems. California Institute of Technology.
- Baker, C.L., Saxe, R., and Tenenbaum, J.B. (2011). Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution. In *CogSci*, pp. 2469–2474.
- Barlow, H.B. (1969). Pattern Recognition and the Responses of Sensory Neurons. *Annals of the New York Academy of Sciences* *156*, 872–881.
- Bayer, H.M., and Glimcher, P.W. (2005). Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. *Neuron* *47*, 129–141.
- Bayes, M., and Price, M. (1763). An Essay towards Solving a Problem in the Doctrine of Chances. *Phil. Trans.* *53*, 370–418.
- Becker, G.M., DeGroot, M.H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behav Sci* *9*, 226–232.
- van Beers, R.J., Sittig, A.C., and Gon, J.J.D. van der (1999). Integration of Proprioceptive and Visual Position-Information: An Experimentally Supported Model. *Journal of Neurophysiology* *81*, 1355–1364.
- 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. *Nat Neurosci* *10*, 1214–1221.
- Berridge, K.C. (1996). Food reward: Brain substrates of wanting and liking. *Neuroscience & Biobehavioral Reviews* *20*, 1–25.

- Berridge, K.C. (2004). Motivation concepts in behavioral neuroscience. *Physiology & Behavior* *81*, 179–209.
- Bolle, R.M., and Cooper, D.B. (1984). Bayesian Recognition of Local 3-D Shape by Approximating Image Intensity Functions with Quadric Polynomials. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6*, 418–429.
- Boorman, E.D., Behrens, T.E.J., Woolrich, M.W., and Rushworth, M.F.S. (2009). How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. *Neuron* *62*, 733–743.
- Borji, A., and Itti, L. (2014). Human vs. computer in scene and object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 113–120.
- Braun, D.A., Mehring, C., and Wolpert, D.M. (2010). Structure learning in action. *Behavioural Brain Research* *206*, 157–165.
- Bromberg-Martin, E.S., Matsumoto, M., Hong, S., and Hikosaka, O. (2010). A Pallidus-Habenula-Dopamine Pathway Signals Inferred Stimulus Values. *Journal of Neurophysiology* *104*, 1068–1076.
- Brunswik, E. (1939). Probability as a determiner of rat behavior. *Journal of Experimental Psychology* *25*, 175.
- Bülthoff, H.H., and Mallot, H.A. (1988). Integration of depth modules: stereo and shading. *J. Opt. Soc. Am. A, JOSAA* *5*, 1749–1758.
- Bush, R.R., and Mosteller, F. (1955). Stochastic models for learning.
- Carey, S. (1978). The child as word learner. In *Linguistic Theory and Psychological Reality*, (Cambridge MA: Halle, M. ; Bresnan, J. ; Miller G.A.), pp. 264–293.
- Collins, A., and Koehlin, E. (2012). Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making. *PLOS Biol* *10*, e1001293.
- Collins, A.G.E., and Frank, M.J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review* *120*, 190–229.
- Collins, A.G.E., and Frank, M.J. (2014). Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological Review* *121*, 337–366.
- Costa, V.D., Tran, V.L., Turchi, J., and Averbeck, B.B. (2015). Reversal Learning and Dopamine: A Bayesian Perspective. *J. Neurosci.* *35*, 2407–2416.
- Courville, A.C. (2006). A latent cause theory of classical conditioning. Carnegie Mellon University Pittsburgh, PA.
- Courville, A.C., Daw, N.D., and Touretzky, D.S. (2004). Similarity and Discrimination in Classical Conditioning: A Latent Variable Account. In *NIPS*, pp. 313–320.
- Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences* *24*, 87–114.

- Cox, R.T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics* *14*, 1–13.
- 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* *319*, 1264–1267.
- Daw, N., and Courville, A. (2008). The pigeon as particle filter. *Advances in Neural Information Processing Systems* *20*, 369–376.
- Daw, N.D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* *8*, 1704–1711.
- Daw, N.D., O’Doherty, J.P., Dayan, P., Seymour, B., and Dolan, R.J. (2006). Cortical substrates for exploratory decisions in humans. *Nature* *441*, 876–879.
- Dayan, P., and Berridge, K.C. (2014). Model-Based and Model-Free Pavlovian Reward Learning: Revaluation, Revision and Revelation. *Cogn Affect Behav Neurosci* *14*, 473–492.
- Dehaene, S., and Changeux, J.-P. (1993). Development of elementary numerical abilities: A neuronal model. *Journal of Cognitive Neuroscience* *5*, 390–407.
- Denrell, J. (2005). Why Most People Disapprove of Me: Experience Sampling in Impression Formation. *Psychological Review* *112*, 951–978.
- Dickinson, A. (1985). Actions and Habits: The Development of Behavioural Autonomy. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* *308*, 67–78.
- Donahue, C.H., and Lee, D. (2015). Dynamic routing of task-relevant signals for decision making in dorsolateral prefrontal cortex. *Nature Neuroscience* *18*, 295–301.
- Donoso, M., Collins, A.G.E., and Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science* *344*, 1481–1486.
- Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks* *15*, 495–506.
- Drugowitsch, J., Wyart, V., Devauchelle, A.-D., and Koechlin, E. (2016). Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron* *92*.
- Dunbar, R.I. (1998). The social brain hypothesis. *Brain* *9*, 178–190.
- Eitz, M., Hays, J., and Alexa, M. (2012). How do humans sketch objects? *ACM Transactions on Graphics* *31*, 1–10.
- Elwin, E., Juslin, P., Olsson, H., and Enkvist, T. (2007). Constructivist coding: Learning from selective feedback. *Psychological Science* *18*, 105–110.
- Ernst, M.O., and Banks, M.S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* *415*, 429–433.
- Fellows, L.K., and Farah, M.J. (2007). The Role of Ventromedial Prefrontal Cortex in Decision Making: Judgment under Uncertainty or Judgment Per Se? *Cereb. Cortex* *17*, 2669–2674.
- Fetsch, C.R., Pouget, A., DeAngelis, G.C., and Angelaki, D.E. (2011). Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience* *15*, 146–154.

- Fiedler, K. (2000). Beware of Samples! *Psychological Review* 107, 659–676.
- Fiorillo, C.D. (2013). Two Dimensions of Value: Dopamine Neurons Represent Reward But Not Aversiveness. *Science* 341, 546–549.
- Fleuret, F., Li, T., Dubout, C., Wampler, E.K., Yantis, S., and Geman, D. (2011). Comparing machines and humans on a visual categorization test. *Proc Natl Acad Sci U S A* 108, 17621–17625.
- Frank, M.J., Seeberger, L.C., and O’reilly, R.C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science* 306, 1940–1943.
- Frazier, P.I., Powell, W.B., and Dayanik, S. (2008). A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization* 47, 2410–2439.
- Frey, P., and Ross, L. (1968). Classical conditioning of the rabbit eyelid response as a function of interstimulus interval. *Journal of Comparative and Physiological Psychology* 65, 246–250.
- Frith, U., and Frith, C.D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 358, 459–473.
- Gans, N., Knox, G., and Croson, R. (2007). Simple Models of Discrete Choice and Their Performance in Bandit Experiments. *M&SOM* 9, 383–408.
- Gershman, S.J., and Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology* 20, 251–256.
- Gittins, J.C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)* 148–177.
- Gottfried, J.A., O’Doherty, J., and Dolan, R.J. (2003). Encoding Predictive Reward Value in Human Amygdala and Orbitofrontal Cortex. *Science* 301, 1104–1107.
- Green, L., and Myerson, J. (2004). A Discounting Framework for Choice With Delayed and Probabilistic Rewards. *Psychol Bull* 130, 769–792.
- Gu, R., Wu, T., Jiang, Y., and Luo, Y.-J. (2011). Woulda, coulda, shoulda: The evaluation and the impact of the alternative outcome: The evaluation and the impact of the alternative outcome. *Psychophysiology* 48, 1354–1360.
- Guitart-Masip, M., Huys, Q.J.M., Fuentemilla, L., Dayan, P., Duzel, E., and Dolan, R.J. (2012). Go and no-go learning in reward and punishment: Interactions between affect and effect. *NeuroImage* 62, 154–166.
- Gurney, K., Prescott, T.J., and Redgrave, P. (2001). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biol Cybern* 84, 401–410.
- Hampton, A.N., Bossaerts, P., and O’Doherty, J.P. (2006). The Role of the Ventromedial Prefrontal Cortex in Abstract State-Based Inference during Decision Making in Humans. *J. Neurosci.* 26, 8360–8367.
- Harlow, H.F. (1949). The formation of learning sets. *Psychological Review* 56, 51.

- Harrison, G.W., and Rutström, E.E. (2008). Risk Aversion in the Laboratory. In *Research in Experimental Economics*, (Bingley: Emerald (MCB UP)), pp. 41–196.
- Hayden, B.Y., Pearson, J.M., and Platt, M.L. (2009). Fictive Reward Signals in the Anterior Cingulate Cortex. *Science* 324, 948–950.
- Hertwig, R., Barron, G., Weber, E.U., and Erev, I. (2004). Decisions from Experience and the Effect of Rare Events in Risky Choice. *Psychological Science* 15, 534–539.
- Hertz, U., Bahrami, B., and Keramati, M. (2017). Stochastic satisficing account of choice and confidence in uncertain value-based decisions. *BioRxiv* 107532.
- Hicks, J.R., and Allen, R.G.D. (1934). A Reconsideration of the Theory of Value. Part I. *Economica* 1, 52.
- Howard, R.A. (1966). Information Value Theory. *IEEE Transactions on Systems Science and Cybernetics* 2, 22–26.
- Hoyer, P.O., and Hyvarinen, A. (2003). Interpreting neural response variability as Monte Carlo sampling of the posterior. *Advances in Neural Information Processing Systems* 293–300.
- Hubel, D.H., and Wiesel, T.N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology* 148, 574–591.
- Humphrey, N.K. (1976). The social function of intellect. In *Growing Points in Ethology*, (Cambridge University Press), pp. 303–317.
- Hyafil, A., Summerfield, C., and Koechlin, E. (2009). Two Mechanisms for Task Switching in the Prefrontal Cortex. *J. Neurosci.* 29, 5135–5142.
- Izquierdo, A., and Jentsch, J.D. (2012). Reversal learning as a measure of impulsive and compulsive behavior in addictions. *Psychopharmacology (Berl)* 219, 607–620.
- Jolly, A. (1966). *Lemur Social Behavior and Primate Intelligence*. Science.
- Jones, M., and Love, B.C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences* 34, 169–188.
- Jones, J.L., Esber, G.R., McDannald, M.A., Gruber, A.J., Hernandez, A., Mirenzi, A., and Schoenbaum, G. (2012). Orbitofrontal Cortex Supports Behavior and Learning Using Inferred but not Cached Values. *Science* 338, 953–956.
- Jurafsky, D. (1996). A Probabilistic Model of Lexical and Syntactic Access and Disambiguation. *Cognitive Science* 20, 137–194.
- Kaelbling, L.P., Littman, M.L., and Moore, A.W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research* 4, 237–285.
- Kahneman, D., and Frederick, S. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In *Heuristics and Biases*, T. Gilovich, D. Griffin, and D. Kahneman, eds. (Cambridge: Cambridge University Press), pp. 49–81.

- Kahneman, D., and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society* 263–291.
- Kahneman, D., Wakker, P.P., and Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *The Quarterly Journal of Economics* 375–405.
- Kaufmann, E., Cappé, O., and Garivier, A. (2012). On Bayesian Upper Confidence Bounds for Bandit Problems. In *AISTATS*, pp. 592–600.
- Kersten, D. (1990). Statistical limits to image understanding. *Vision: Coding and Efficiency* 32–44.
- Knill, D.C., and Richards, W. (1996). *Perception as Bayesian Inference* (Cambridge University Press).
- Koechlin, E., Ody, C., and Kouneiher, F. (2003). The Architecture of Cognitive Control in the Human Prefrontal Cortex. *Science* 302, 1181–1185.
- Köszegi, B., and Rabin, M. (2007). Mistakes in choice-based welfare analysis. *The American Economic Review* 97, 477–481.
- Krebs, J.R., Kacelnik, A., and Taylor, P. (1978). Test of optimal sampling by foraging great tits. *Nature* 275, 27–31.
- Kringelbach, M.L., O’Doherty, J., Rolls, E.T., and Andrews, C. (2003). Activation of the Human Orbitofrontal Cortex to a Liquid Food Stimulus is Correlated with its Subjective Pleasantness. *Cereb. Cortex* 13, 1064–1071.
- Lai, T.L., and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6, 4–22.
- Lauriola, M., and Levin, I.P. (2001). Personality traits and risky decision-making in a controlled experimental task: An exploratory study. *Personality and Individual Differences* 31, 215–226.
- Le, Q.V., Ranzato, M., and Ng, A.Y. (2013). Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference On*, (IEEE), pp. 8595–8598.
- Lebreton, M., Jorge, S., Michel, V., Thirion, B., and Pessiglione, M. (2009). An Automatic Valuation System in the Human Brain: Evidence from Functional Neuroimaging. *Neuron* 64, 431–439.
- Ljungberg, T., P, A., and W, S. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *J Neurophysiol* 67, 145–163.
- Lohrenz, T., McCabe, K., Camerer, C.F., and Montague, P.R. (2007). Neural signature of fictive learning signals in a sequential investment task. *PNAS* 104, 9493–9498.
- Luce, R.D. (1959). On the possible psychophysical laws. *Psychological Review* 66, 81.
- Lusena, C., Goldsmith, J., and Mundhenk, M. (2001). Nonapproximability results for partially observable Markov decision processes. *J. Artif. Intell. Res.(JAIR)* 14, 83–103.
- Ma, W.J., Beck, J.M., Latham, P.E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience* 9, 1432–1438.
- Mach, E., and Williams, C.M. (1897). *Contributions to the analysis of the sensations*.

- Mackintosh, N.J. (1975). A theory of attention: variations in the associability of stimuli with reinforcement. *Psychological Review* 82, 276.
- Maes, E., Boddez, Y., Alfei, J.M., Kryptos, A.-M., D'Hooge, R., De Houwer, J., and Beckers, T. (2016). The elusive nature of the blocking effect: 15 failures to replicate. *Journal of Experimental Psychology: General* 145, e49–e71.
- March, J.G. (1996). Learning to be risk averse. *Psychological Review* 103, 309.
- Marciano-Romm, D., Romm, A., Bourgeois-Gironde, S., and Deouell, L.Y. (2016). The Alternative Omen Effect: Illusory negative correlation between the outcomes of choice options. *Cognition* 146, 324–338.
- McKelvey, R.D., and Palfrey, T. (1992). An Experimental Study of the Centipede Game. *Econometrica* 60, 803–836.
- Meyer, R.J., and Shi, Y. (1995). Sequential choice under ambiguity: Intuitive solutions to the armed-bandit problem. *Management Science* 41, 817–834.
- Miller, R.R., Barnet, R.C., and Grahame, N.J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin* 117, 363–386.
- Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE* 49, 8–30.
- Minsky, and Lee, M. (1954). *Theory of neural-analog reinforcement systems and its application to the brain-model problem*. Princeton University Press.
- Mirenowicz, J., and Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. *Journal of Neurophysiology* 72, 1024–1027.
- Montague, P.R., and Berns, G.S. (2002). Neural Economics and the Biological Substrates of Valuation. *Neuron* 36, 265–284.
- Morris, G., Nevet, A., Arkadir, D., Vaadia, E., and Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience* 9, 1057–1063.
- Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., and Hikosaka, O. (2004). Dopamine Neurons Can Represent Context-Dependent Prediction Error. *Neuron* 41, 269–280.
- Nassar, M.R., Wilson, R.C., Heasly, B., and Gold, J.I. (2010). An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment. *J. Neurosci.* 30, 12366–12378.
- Neal, R.M. (2003). Slice sampling. *Ann. Statist.* 31, 705–767.
- Nickerson, R.S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2, 175–220.
- Nieder, A., and Miller, E.K. (2003). Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. *Neuron* 37, 149–157.
- Niv, Y., Edlund, J.A., Dayan, P., and O'Doherty, J.P. (2012). Neural Prediction Errors Reveal a Risk-Sensitive Reinforcement-Learning Process in the Human Brain. *J. Neurosci.* 32, 551–562.

- O'Neill, M., and Schultz, W. (2010). Coding of Reward Risk by Orbitofrontal Neurons Is Mostly Distinct from Coding of Reward Value. *Neuron* 68, 789–800.
- Padoa-Schioppa, C. (2011). Neurobiology of Economic Choice: A Good-Based Model. *Annual Review of Neuroscience* 34, 333–359.
- Padoa-Schioppa, C., and Assad, J.A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature* 441, 223–226.
- Palminteri, S., Justo, D., Jauffret, C., Pavlicek, B., Dauta, A., Delmaire, C., Czernecki, V., Karachi, C., Capelle, L., Durr, A., et al. (2012). Critical Roles for Anterior Insula and Dorsal Striatum in Punishment-Based Avoidance Learning. *Neuron* 76, 998–1009.
- Palminteri, S., Khamassi, M., Joffily, M., and Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications* 6, 8096.
- Palminteri, S., Wyart, V., and Koehlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences* 21, 425–433.
- Papadimitriou, C.H., and Tsitsiklis, J.N. (1999). The complexity of optimal queuing network control. *Mathematics of Operations Research* 24, 293–305.
- Pavlov, I.P. (1927). Conditioned reflexes (Gleb Vasīlevīch Anrep).
- Payzan-LeNestour, E. (2010). Bayesian learning in unstable settings: Experimental evidence based on the bandit problem. *Swiss Finance Institute Research Paper* 10, 1–41.
- Payzan-LeNestour, E., and Bossaerts, P. (2011). Risk, Unexpected Uncertainty, and Estimation Uncertainty: Bayesian Learning in Unstable Settings. *PLOS Comput Biol* 7, e1001048.
- Pearce, J.M., and Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review* 87, 532.
- Pearson, J.M., Watson, K.K., and Platt, M.L. (2014). Decision Making: The Neuroethological Turn. *Neuron* 82, 950–965.
- Pessiglione, M., Schmidt, L., Draganski, B., Kalisch, R., Lau, H., Dolan, R.J., and Frith, C.D. (2007). How the Brain Translates Money into Force. *Science* 316, 904–906.
- Peters, J., and Buchel, C. (2009). Overlapping and Distinct Neural Systems Code for Subjective Value during Intertemporal and Risky Decision Making. *Journal of Neuroscience* 29, 15727–15734.
- Phillips, L.D., and Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology* 72, 346–354.
- Piazza, M., Izard, V., Pinel, P., Bihan, D.L., and Dehaene, S. (2004). Tuning Curves for Approximate Numerosity in the Human Intraparietal Sulcus. *Neuron* 44, 547–555.
- Plassmann, H., O'Doherty, J., and Rangel, A. (2007). Orbitofrontal Cortex Encodes Willingness to Pay in Everyday Economic Transactions. *J. Neurosci.* 27, 9984–9988.
- Plassmann, H., O'Doherty, J., Shiv, B., and Rangel, A. (2008). Marketing actions can modulate neural representations of experienced pleasantness. *PNAS* 105, 1050–1054.

- Poltrock, S.E. (1989). A random walk model of digit comparison. *Journal of Mathematical Psychology* 33, 131–162.
- Pouget, A., Beck, J.M., Ma, W.J., and Latham, P.E. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience* 16, 1170–1178.
- Premack, D., and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1, 515–526.
- Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A.Y. (2007). Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th International Conference on Machine Learning*, (ACM), pp. 759–766.
- Ramsay, S.G. (1857). *Principles of Psychology* (London).
- Rangel, A., Camerer, C., and Montague, P.R. (2008). Neuroeconomics: The neurobiology of value-based decision-making. *Nat Rev Neurosci* 9, 545–556.
- Redish, A.D., Jensen, S., Johnson, A., and Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling. *Psychological Review* 114, 784–805.
- Reverdy, P.B., Srivastava, V., and Leonard, N.E. (2014). Modeling human decision making in generalized Gaussian multiarmed bandits. *Proceedings of the IEEE* 102, 544–571.
- Robbins, H. (1952). Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*, (Springer), pp. 169–177.
- Roesch, M.R., Singh, T., Brown, P.L., Mullins, S.E., and Schoenbaum, G. (2009). Ventral Striatal Neurons Encode the Value of the Chosen Action in Rats Deciding between Differently Delayed or Sized Rewards. *J. Neurosci.* 29, 13365–13376.
- Rogers, R.D., and Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General* 124, 207.
- Sakamoto, Y., Jones, M., and Love, B.C. (2008). Putting the psychology back into psychological models: Mechanistic versus rational approaches. *Memory & Cognition* 36, 1057–1065.
- Samejima, K., Ueda, Y., Doya, K., and Kimura, M. (2005). Representation of Action-Specific Reward Values in the Striatum. *Science* 310, 1337–1340.
- Samuelson, P.A. (1938). The numerical representation of ordered classifications and the concept of utility. *The Review of Economic Studies* 6, 65–70.
- Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., and Cohen, J.D. (2003). The Neural Basis of Economic Decision-Making in the Ultimatum Game. *Science* 300, 1755–1758.
- Schmalensee, R. (1975). Alternative models of bandit selection. *Journal of Economic Theory* 10, 333–342.
- Scholl, J., Kolling, N., Nelissen, N., Wittmann, M.K., Harmer, C.J., and Rushworth, M.F.S. (2015). The Good, the Bad, and the Irrelevant: Neural Mechanisms of Learning Real and Hypothetical Rewards and Effort. *J. Neurosci.* 35, 11233–11251.

- Schultz, W. (2015). Neuronal Reward and Decision Signals: From Theories to Data. *Physiological Reviews* 95, 853–951.
- Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599.
- Schwarz, G. (1978). Estimating the Dimension of a Model. *Ann. Statist.* 6, 461–464.
- 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 & Biobehavioral Reviews* 37, 681–696.
- Shepard, R.N., and others (1987). Toward a universal law of generalization for psychological science. *Science* 237, 1317–1323.
- Simon, H.A. (1956). Rational choice and the structure of the environment. *Psychological Review* 63, 129.
- Skinner, B.F. (1938). *The behavior of organisms: An experimental analysis* (BF Skinner Foundation).
- Srinivas, N., Krause, A., Kakade, S.M., and Seeger, M.W. (2012). Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting. *IEEE Transactions on Information Theory* 58, 3250–3265.
- Steyvers, M., Lee, M.D., and Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology* 53, 168–179.
- Stocker, R., Seymour, J.R., Samadani, A., Hunt, D.E., and Polz, M.F. (2008). Rapid Chemotactic Response Enables Marine Bacteria to Exploit Ephemeral Microscale Nutrient Patches. *Proceedings of the National Academy of Sciences of the United States of America* 105, 4209–4214.
- Sutton, R.S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning*, pp. 216–224.
- Sutton, R.S., and Barto, A.G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review* 88, 135–170.
- Sutton, R.S., and Barto, A.G. (1998). *Reinforcement learning: An introduction* (MIT press).
- Swainson, R., Rogers, R.D., Sahakian, B.J., Summers, B.A., Polkey, C.E., and Robbins, T.W. (2000). Probabilistic learning and reversal deficits in patients with Parkinson’s disease or frontal or temporal lobe lesions: possible adverse effects of dopaminergic medication. *Neuropsychologia* 38, 596–612.
- Tenenbaum, J.B., and Griffiths, T.L. (2001a). Structure learning in human causal induction. *Advances in Neural Information Processing Systems* 59–65.
- Tenenbaum, J.B., and Griffiths, T.L. (2001b). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences* 24, 629–640.
- Tenenbaum, J.B., Kemp, C., Griffiths, T.L., and Goodman, N.D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science* 331, 1279–1285.

- Thompson, W.R. (1933). On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika* 25, 285.
- Thorndike, E.L. (1911). *Animal intelligence* (New York,: The Macmillan company,).
- Thorpe, S.J., Rolls, E.T., and Maddison, S. (1983). The orbitofrontal cortex: neuronal activity in the behaving monkey. *Exp Brain Res* 49, 93–115.
- Tobler, P.N., Fiorillo, C.D., and Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science* 307, 1642–1645.
- Tversky, A., and Kahneman, D. (1991). Loss Aversion in Riskless Choice: A Reference-Dependent Model. *The Quarterly Journal of Economics* 106, 1039–1061.
- Usher, M., and McClelland, J.L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychol Rev* 108, 550–592.
- Van Horn, K.S. (2003). Constructing a logic of plausible inference: a guide to Cox's theorem. *International Journal of Approximate Reasoning* 34, 3–24.
- Vlaev, I., Chater, N., Stewart, N., and Brown, G.D. (2011). Does the brain calculate value? *Trends in Cognitive Sciences* 15, 546–554.
- Von Neumann, J., and Morgenstern, O. (1947). *Theory of games and economic behavior*, 2nd rev. ed (Princeton, NJ, US: Princeton University Press).
- Watkins, C.J., and Dayan, P. (1992). Q-learning. *Machine Learning* 8, 279–292.
- White, N.M. (1989). Reward or reinforcement: what's the difference? *Neurosci Biobehav Rev* 13, 181–186.
- Whittle, P. (1988). Restless Bandits: Activity Allocation in a Changing World. *Journal of Applied Probability* 25, 287.
- Wilson, R.C., Geana, A., White, J.M., Ludvig, E.A., and Cohen, J.D. (2011). Why the grass is greener on the other side: Behavioral evidence for an ambiguity bonus in human exploratory decision-making. *Neuroscience*.
- Wilson, R.C., Takahashi, Y.K., Schoenbaum, G., and Niv, Y. (2014). Orbitofrontal Cortex as a Cognitive Map of Task Space. *Neuron* 81, 267–279.
- Xu, F., and Tenenbaum, J.B. (2007). Word learning as Bayesian inference. *Psychological Review* 114, 245–272.
- Young, P.T. (1959). The role of affective processes in learning and motivation. *Psychological Review* 66, 104.
- Yu, A.J., and Cohen, J.D. (2009). Sequential effects: superstition or rational behavior? In *Advances in Neural Information Processing Systems*, pp. 1873–1880.
- Yu, A.J., and Dayan, P. (2005). Uncertainty, Neuromodulation, and Attention. *Neuron* 46, 681–692.
- Zemel, R.S., Dayan, P., and Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural Computation* 10, 403–430.

Zhang, S., and Yu, A.J. (2013). Cheap but clever: human active learning in a bandit setting. *Ratio* 12, 14.

List of Figures

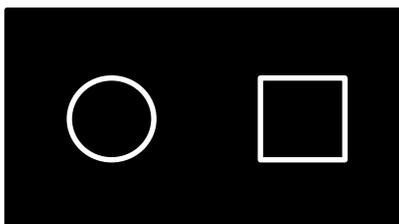
Figure 1: From Schultz 2015	9
Figure 2: Classical conditioning, From Beginning Psychology, Charles Stangor 2012	14
Figure 3: Blocking effect. From Nicolas Rougier, Blocking, Wikipedia.....	15
Figure 4: World famous dopamine neurons experiment, (Reproduced from (Schultz, Dayan, & Montague, 1997))	16
Figure 5: Usual reward devaluation paradigm: From B. Balleine	19
Figure 6: Graphical model of a simple hierarchical structure.	34
Figure 7 : Example of ordered cumulative reward distributions.....	50
Figure 8 : Bayesian learning of reward distributions.	53
Figure 9 : Observation noise and distribution learning.....	54
Figure 10 : 2 reversal learning tasks on which we simulated the NEIG algorithm..	55
Figure 11 : NEIG's performance on the tasks	56
Figure 12: Visualization of the timing of the task	58
Figure 13: Reversal learning paradigm.....	59
Figure 14 : Reward Distributions used in the continuous task.....	60
Figure 15 : Reward distribution of the discrete task.....	61
Figure 16: Summary of model NEIG.....	68
Figure 17 : Behavioral results (continuous task).	69
Figure 18 : Model comparison (continuous task)	71
Figure 19 : Performance of model NEIG (continuous task).....	72
Figure 20 : NEIG learns reward distributions	73
Figure 21 : Additional model comparison (continuous task)	75
Figure 22 : Behavioral and model results (discrete task)	77
Figure 23 : Hierarchical structure with 3 bandits	79
Figure 24: Illustration of the difference between BEST and EV rules.	79
Figure 25 : Comparison of 2 hierarchical models	80
Figure 26: Reward distributions used in the 3 bandits task.....	84
Figure 27 : Exemplar run of the 3 bandits task.....	85
Figure 28 : Behavioral results (3 bandits task)	86
Figure 29 : Model comparison (BEST rule vs EV rule).	87
Figure 30: Distinction of the different Gaussian models	87
Figure 31 : Type 3 switches and Gaussian models.....	88
Figure 35 :Model Comparison (3 options task).	90
Figure 33: Optimal performance of models on the 3 options task	92
Figure 34 : Convergence rate to reward distributions of model NEIG.....	96
Figure 35 : Optimal NEIG learns reward distributions.	98

Annex

Instructions given to participants for the continuous task (in French):

Cette expérience est un jeu où il vous sera possible de gagner entre 10 et 40 euros. Votre gain dépendra de vos performances. L'expérience dure environ 55 minutes.

Un cercle et un carré s'affichent à l'écran à intervalles régulier.



Vous devez choisir l'une de ces deux formes en appuyant sur *une* des touches ‘,’ ou ‘;’ (marquées en bleu sur le clavier) avec les doigts de votre **main droite**. Pour choisir la forme présentée à gauche, appuyez sur la touche gauche (‘,’). Pour choisir la forme présentée à droite, appuyez sur la touche droite (‘;’).



Selon **la forme** que vous avez choisie, il apparaîtra à l'écran le montant que vous remportez pour cet essai. Ce montant dépend uniquement de la forme choisie et non de sa position à droite ou à gauche.

Si vous ne répondez pas dans le temps imparti, l'essai est considéré comme perdu, vous ne remportez aucune récompense pour cet essai. N'appuyez sur une touche qu'une seule fois par essai : seule la première réponse est prise en compte.

C'est à vous de trouver quelle forme rapporte le plus d'argent afin d'essayer d'en gagner le plus possible. Attention : la meilleure forme peut changer de temps en temps.

L'expérience s'interrompra six fois en cours de déroulement pour vous laisser le temps de vous reposer. Utilisez-les ! Il n'y a aucun lien entre la survenue d'une pause et le déroulement de l'expérience. N'oubliez pas ce que vous faisiez avant la pause, l'expérience reprendra son cours là où vous l'avez laissée quand vous appuierez sur la touche « Entrée ».

Durant la pause vous observerez la moyenne de votre récompense depuis la dernière pause. Cela vous permettra de suivre vos performances. En fonction de votre score vous vous verrez attribuer un bonus à chaque pause, qui représente une partie de votre rémunération. Vous verrez aussi en parallèle les performances d'un autre sujet. Il est bien évidemment possible de faire mieux que cet autre sujet.

A la fin de l'expérience, votre rémunération dépendra des bonus accumulés lors des 6 périodes. Vous recevrez également un super-bonus calculé en fonction de votre récompense à 20 essais tirés au hasard au cours de la tâche.

Vous allez d'abord réaliser un entraînement. A l'issue vous pourrez poser des questions à l'expérimentateur qui vérifiera que vous avez bien compris la tâche.

NB : Merci de ne pas parler de l'expérience à des personnes qui voudraient la passer après vous : nous devons étudier vos réponses sans à priori.