



**HAL**  
open science

# The cognitive roots of strategic learning in repeated social interactions : from cognitive neuroscience back to behavioral economics

Thibaud Griessinger

► **To cite this version:**

Thibaud Griessinger. The cognitive roots of strategic learning in repeated social interactions : from cognitive neuroscience back to behavioral economics. Cognitive Sciences. Université Paris sciences et lettres, 2017. English. NNT : 2017PSLEE093 . tel-03383031

**HAL Id: tel-03383031**

**<https://theses.hal.science/tel-03383031v1>**

Submitted on 18 Oct 2021

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

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

# THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres  
PSL Research University

Préparée à l'Ecole Normale Supérieure

**The cognitive roots of strategic learning in repeated social interactions - from cognitive neuroscience back to behavioral economics**

**Ecole doctorale n°158**

Ecole Doctorale 3C Cerveau, Cognition, Comportement

**Spécialité Sciences Cognitives**

**Soutenue par Thibaud GRIESSINGER  
le 18 décembre 2017**

Dirigée par **Giorgio CORICELLI**  
Codirigée par **Mehdi KHAMASSI**

## COMPOSITION DU JURY :

M. ATTANASI Giuseppe  
Université de Lille 1, Rapporteur

M. RUFF Christian  
University of Zurich, Rapporteur

Mme. ZALLA Tiziana  
Ecole Normale Supérieure, Examinatrice

M. BROVELLI Andrea  
Université Aix-Marseille, Examineur

M. CORICELLI Giorgio  
Ecole Normale Supérieure, Directeur de thèse

M. KHAMASSI Mehdi  
Université Pierre et Marie Curie,  
Codirecteur de thèse



## ABSTRACT

Social interactions rely on our ability to learn and adjust on the spot to the other's behavior. Strategic games provide a useful framework to study the cognitive processes involved in the representation of the other's intentions and their translation into the most adapted actions. In the last decades, the growing field of behavioral economics provided evidence of a systematic departure of human's behavior from the optimal prescription formulated by game theory. Based on recent advances in cognitive sciences, we hypothesized that characterizing the source of heterogeneity in behavior might provide key insights to understand the boundaries over human social learning, and therefore deviation from mutually beneficial interactions.

We first address the question of the interplay between the game environment and the heterogeneity in formation of high-order beliefs over the opponent's behavior through strategic learning. We show that in a competitive repeated interaction, the payoff structure of the underlying game can influence the engagement in strategically sophisticated learning and explain deviation from game optimality (equilibrium). Our data suggest that participants in a disadvantaged role are constrained in their learning sophistication, and thus in the overcoming of their position, by their own cognitive capacities. Their opponents, albeit advantaged, still need to engage in strategically sophisticated learning but to follow and adjust their behavior in order to maximize their earnings. This study provides the first evidence of the key implication of strategic learning heterogeneity in equilibrium departure and provide insight to explain the emergence of a leader-follower dynamics of choice. In addition our results suggest that a cost-benefit analysis might drive the engagement of strategic players in a more sophisticated learning process. In a second step, we investigated the hypothesis that the depth of strategic learning is not the only factor in play to grasp the other's mind during competitive interaction, but that the capacity to detect and exploit patterns in her behavior is also important. We found that not only subjects were able to detect patterns in the opponent's behavior, but that the capacity to do so was not correlated to a lower engagement in sophisticated strategic learning, therefore suggesting that humans can combine information from both types of learning to improve belief accuracy during social decision making.

## TABLE OF CONTENTS

<b>Chapter I - Introduction</b> .....	4
I- From decision-making to learning: a neuroeconomic perspective .....	6
A) The value-based framework in decision making .....	6
B) Learning the value, learning from value .....	9
C) Structure inference and (probabilistic) beliefs .....	16
II- Deciding in a social world .....	20
A) Social decision making .....	20
B) <i>ToM</i> and the Social brain hypothesis .....	23
C) Social learning .....	27
III- Social learning in strategic interactions .....	39
A) Strategic interaction .....	39
B) The Neuroeconomics of strategic learning .....	50
IV- Synthesis and working hypothesis of the present thesis .....	57
<b>Chapter II - The interplay of learning sophistication and strategic asymmetry in social competitive interactions (Exp.1-3)</b> .....	60
I - The interplay of learning sophistication and strategic asymmetry in social competitive interactions. [Griessinger T., Khamassi K. and Coricelli G. (in prep.)] .....	60
A) Introduction .....	61
B) Exp. 1: Model simulation and prediction .....	64
C) Exp. 2: Human against human .....	66
D) Exp. 3: Human against computerized opponents .....	76
E) Discussion .....	80
F) References .....	85
II - Strategic learning in repeated game interactions: Methodological considerations .....	89
A) Additional discussion .....	89
B) Additional results Exp.2 .....	92
<b>Chapter III -Transfer effect in strategic learning (Exp.4)</b> .....	94
A) Introduction .....	94
B) Methods .....	96
C) Results .....	96
D) Discussion .....	104
<b>Chapter IV - From strategic learning to Pattern detection in competitive Interactions (Exp.5,6)</b> .....	108
I - Pattern detection in strategic dyadic interactions (Exp.5) .....	108
A) Introduction .....	108
B) Methods .....	112
C) Results .....	115
D) Discussion .....	125
II - Pattern detection and Strategic sophistication in Rock-Paper-Scissor (Exp.6) .....	127



A) Introduction .....	127
B) Methods .....	129
C) Results .....	132
D) Discussion .....	135
III - Conclusion of Experiments 5 and 6.....	138
<b>Chapter V - General Discussion</b> .....	140
A) Conclusion of the experimental studies (Exp1-6) .....	141
B) Future research directions .....	149
C) General conclusion .....	151
Appendix II - Griessinger, T., & Coricelli, G. (2015) .....	153
Appendix II - Supplementary Information [Griessinger T., Khamassi K. and Coricelli G. (in prep.)] .....	154
Appendix III - Supplementary Information 2 .....	179
Appendix IV - Supplementary Information 3 .....	186
References .....	190

# - Chapter I -

## Introduction - Scientific Background

### Foreword

Everyday social interactions are considered a key foundation of our development and cognitive abilities (Frith & Frith, 2010). With the progress of social neuroscience in the last decade, many investigations have focused on the brain mechanisms underlying our capacity to grasp the mental states of others (Frith & Frith, 2012). However, different theories have been proposed to explain how humans can engage in such “Theory of Mind”. Moreover, even if some brain regions have been identified as specifically recruited in this cognitive process, no clear consensus has been reached about the underlying computations and specific cognitive mechanisms involved (Mahy et al, 2014; Stanley & Adolphs, 2013). Faced with this inconsistency it has recently been proposed to rethink the way social interactions are studied in laboratory. Some authors have suggested that the use of static tasks exposing participants to so-called social stimuli (such as faces, or stories about fictive characters) might limit our understanding of the cognitive processes involved during real social interactions (Di Paolo & De Jaegher, 2012; Hari et al, 2015; Przyrembel et al, 2012). To shed light on the *dark-matter of social interaction* they hypothesized that knowing others might not be limited to perceiving them, but also to engage with them. In this line of inquiry, dynamic interaction between humans appears to be a cornerstone for understanding how humans grasp other minds. As Schilbach (2014) points out, “social interactions are characterized by intricate reciprocal relations with the perception of socially relevant information prompting (re-) actions, which are themselves processed and reacted to”. However, analyzing data emanating from such ecological inter-individual interactions is a methodological challenge (Hari et al, 2015; Lee & Harris, 2013; Schilbach et al, 2013).

The special case of strategic interaction might embody this complex systemic problem in the specific decision problem it constitutes. Indeed, in such a social setting, the outcome of one’s action depends directly on what the other individual in the interaction decides. In this particular type of social exchange it thus appears crucial to anticipate the other’s actions in order to adjust our own behavior and to maximize the outcome of the interaction. Game theory models strategic interactions as games representing decisions between agents where one’s payoffs depend on the other’s actions. This normative framework provides precise theoretical solutions for optimal behavior embodied in the premises of rationality (such as the notion of Nash equilibrium). It also provides a benchmark for the analysis of its behavioral

departures. However, it has been empirically shown that in practice humans do not always/systematically follow this prescription of optimality (Camerer, 2003). One hypothesis, which encounters growing support, lies on the idea that following the (optimal) action profile (by which no player can increase her payoff by changing her action given the other players' actions) requires that both players should hold correct beliefs over their opponent's behavior and best-respond to it (Camerer et al, 2014). The ability to do so might therefore be somehow cognitively and/or contextually constrained, leading to sub-optimal behavior (Crawford et al, 2013).

As game interactions extend the model of individual decision-making to the understanding of the interactions in multi-agents situations, recent research in neuroeconomics proposed to combine the computational approach from cognitive neuroscience with the experimental framework provided by behavioral game-theory to unravel the brain mechanisms underlying human decision-making during social interactions (Lee, 2008).

In this thesis we propose to take a step back to understand how the heterogeneous deviations observed in human behavior from the normative prescription of game-theory could be informed by the various ways humans might learn in repeated (strategic) interactions. We believe that a better understanding of how the two fields of cognitive neuroscience and behavioral economics can be combined, in order to explain how human make decisions in repeated interactions, might provide crucial insights for a better understanding of the cognitive processes implicated in Theory of mind.

In the following we will review recent advances in the field of neuroeconomics and outline perspectives for a neuroeconomic approach of strategic interactions. First, we will briefly introduce the neuroeconomics of non-social decision-making and the cognitive mechanisms underlying human learning (I). Then we will describe how this framework can, and has been extended to social interactions (II). Finally, we will present the framework in which this thesis is rooted (III). We will demonstrate how the theoretical framework of game theory can be used to better understand the cognitive mechanisms underlying decision-making in strategic interactions. Finally we will review how behavioral economics and cognitive neuroscience have tackled the question of social learning during strategic interactions to introduce the questions that we investigate in this thesis (IV).

## I- From decision-making to learning: a neuroeconomic perspective

### **A) The value-based framework in decision making**

At the heart of choice is believed to be the concept of subjective value. Within this framework, value is considered to be the main drive of action selection. Basically a reward, by nature attractive for an individual, has a positive value, and conversely a punishment, repulsive, a negative one. Historically the concept of value in philosophy, and later in psychology, is rooted in the notion of pleasure, so that the value of an option corresponds to the amount of pleasure that is eventually obtained once chosen, and by extension corresponds to its attractiveness or desirability (Mill, 1901).

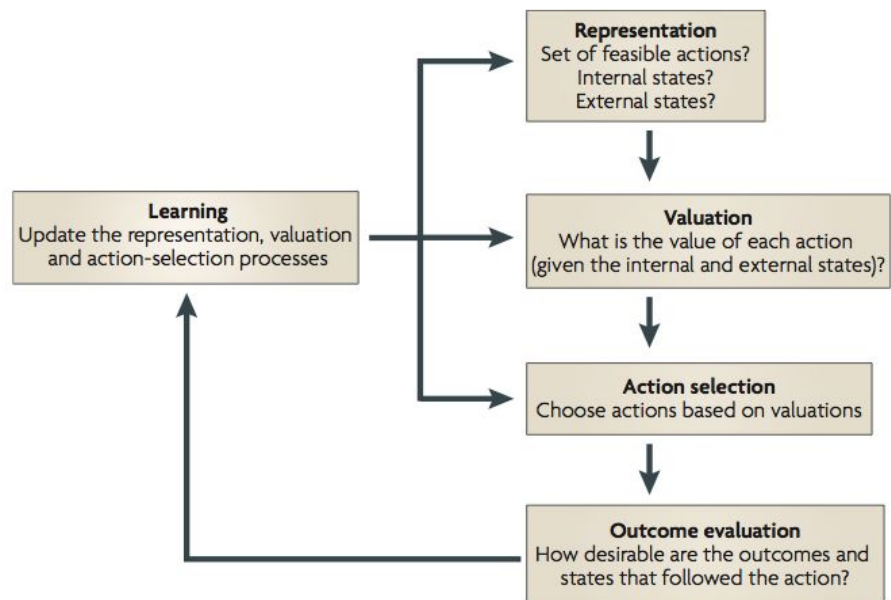
Inspired by this concept, but rebutted by the blurriness of introspection, economists ought to build a decision theory based on observations. The rationale was the following: if the value is the hidden variable driving action then from choices such quantity could be inferred and value objectively measured (Padoa-Schioppa, 2008). Samuelson, (1938) first proposed that individual preferences can be revealed through choices made in situations of risk and uncertainty. Based on this idea Von Neumann & Morgenstern (1947) proposed the expected utility theory. EUT is a set of axioms which aims to ensure that if the preferences expressed by an agent in a probabilistic environment are consistent, then a “utility” function, a cardinal measure of the expected value driving her choices, can be computed from her ordinal preferences. To fill the gap between the generalization of the notion of value proposed by the EUT and the predictability of an agent’s choices, Savage (1954) finally reversed this relation by stating that a rational agent’s choices should be the outcome of a utility maximization process. Note that within the neoclassical framework in economics a slightly different concept of value is typically confounded, namely the motivation to engage in a costly or effortful action in order to obtain a good, such as exchanging money against a good, or giving up on an option to obtain another outcome (O’Doherty, 2014).

In psychology, two types of rewards (i.e. outcome values) are usually distinguished: food and drink are considered as primary, in opposition to secondary rewards which are non-crucial for the survival of the individual (Schultz, 2015). In economics, utility represents a common currency for reward, and choices are often paired to monetary outcomes as money allows an independency from the nature of the expected reward. Finally, it is worth noting that in the framework of EUT, subjective value commonly encompasses both the amplitude of the expected reward and its availability, the probability to receive it. The EUT thus provides a useful mathematical object that unites the properties of the subjective (expected) value driving the choice behavior. Psychology later influenced economics, by refining the typical (average) shape of the utility function through the incorporation of some insights from empirical

observations, such as risk aversion, probability distortion or the effect of framing (Kahneman & Tversky, 1979).

Following the formalism of choice and expected value provided by the decision theory in economics, two complementary lines of inquiry have emerged in the last decades. With the rise of modern psychology, some economists aimed to experimentally test the predictions made by the EUT in order to “improve the realism of the psychological assumptions underlying economic theory” (Fumagalli, 2016), and behavioral economics became a growing field. In a second time, the raise of the cognitive neuroscience field lead economists to move a step further and collaborate with cognitivists to use neuroscientific tools to “advance economic modeling by building more predictive and explanatory models of choice” (Camerer et al, 2004). These two domains thus aimed at gathering more precise knowledge about choice behavior in order to explain the (often) observed deviation from theoretical rationality (Camerer, 2003). Still, to date they remain two separate fields of research with distinct methodologies and priors.

In 2008, a group of prominent neuro-economists proposed a unified model of value-based decision-making breaking the choice process into computational steps hypothesized to be subserved by different brain structures (Rangel et al, 2008) (**Fig1**). The key feature of their model is that it posits the existence of two different value computations. The first valuation process takes place at the time of the outcome and corresponds to the (affective/hedonic) value of the reward, once experienced. The second is the (state-dependent) value of each available action that drives, through direct comparison, the choice process. This value corresponds to the reward expected to be received. In this framework, such a value signal is viewed as the computational equivalent of the theoretical expected utility, and is thus considered to be sensitive to the same psychological effects included in the EUT (risk, uncertainty, framing).



**Figure 1.** The unified model of value-based decision making proposed by Rangel et al (2008). They describe five main computational steps: first, the construction of a representation of the decision problem, which entails identifying internal and external states as well as potential courses of action; second, the valuation of the different actions under consideration; third, the selection of one of the actions on the basis of their valuations; fourth, after implementing the decision the brain needs to measure the desirability of the outcomes that follow; and finally, the outcome evaluation is used to update the other processes to improve the quality of future decisions. (reproduced from Rangel et al, 2008)

---

However, two types of valuation systems operating at the time of choice can be distinguished: the habitual system storing cached values, which corresponds to the automatic association learned between a stimulus and an action, and a goal-directed system that computes at the time of choice a value for each action depending on the reward expected to be received at the time of the outcome (Balleine & O’doherly, 2010; Dickinson, 1985). This latter system is thought to be a more flexible system as the expected value signal generated is the product of the computation of the action-outcome and stimulus-reward contingencies. Such computation requires the knowledge of the structure of the environment as well as a costly prospection process to infer potential distal outcomes. As we will see in the next section, a distinction between these two systems can be done at the brain level (Dolan & Dayan 2013).

Now, how these two systems relate to the notion of expected utility in EUT remains unclear. In their framework Rangel et al suggest that goal directed values correspond to the subjective value as defined

by the economic theory (Rangel et al, 2008). Still, the notion of expected utility is not action dependent, but relates to the value of the outcome directly. In fact, a series of neuroscience studies lead by Padoa-Schioppa et al (Padoa-Schioppa, 2011). showed for the first time that some neurons in the orbitofrontal cortex (OFC) (in monkeys), specifically encode the value of a good per se, independently of the associated action (Padoa-Schioppa, 2013). Similar evidence of common value signal were found in the ventromedial prefrontal cortex (vmPFC) in humans (Levy & Glimcher, 2012). As if values in the economic sense were abstract constructs (i.e. independent of any choice modalities of physical properties) encoded at the time of choice by the brain in order to drive action selection. Such a finding brings the economic concept of value closer to the goal-directed value presented in the value-based framework, except that for the former the value is attached to a possible outcome, while the latter is attached to an action-outcome association (an action representing here any measurable behavior leading to an outcome). If this difference may seem negligible, it actually does matter from a theoretical perspective (Padoa-Schioppa & Schoenbaum, 2015).

Indeed, there is one element which is not modeled by the EUT: the variation of values over time. As preferences are considered as stable, EUT does not consider that values are learned or that they can be readjusted over time. As presented in **Fig.1**, in the value-based decision-making framework, the (expected) values are computed from past exposure to the actual (experienced) value, and thus emerge through learning (Rangel et al, 2008). In this framework, subjective values driving our behavior are not just *there* when it comes to make a decision, but are they instead computed through experience, and can thus be manipulated.

## **B) Learning the value, learning from value**

As previously introduced decision-making can be seen as the cognitive process of evaluation and comparison of the different options available, reducing our behavior to essentially a reward (utility) maximization process. In this value-based framework, a temporality emerges in decision-making as the utility maximization concept in economics is split into two steps: the anticipation, or prediction of the reward to come, and the experience of it. The loop formed by the shaping of the value through experience and the prediction of experience through evaluation is the core of a related cognitive function: learning. At the beginning of the last century, two concomitant studies lead psychologists to first formalize what they called instrumental learning, or operant conditioning. Pavlov reported that a behavioral response, such as salivation in dogs, can be eventually induced by any initially neutral (conditioned) stimulus (e.g. sound), if a primarily rewarding (unconditioned) outcome (e.g. food) was repeatedly paired to it (i.e. presented shortly after). This observation first suggested that a behavioral response (salivation) can be

associated to a stimulus; however, no voluntary action is involved in this process. A bit later, Thorndike reported that animals, who managed by chance to experience a reward following an action, would be more likely to reproduce that same action in order to get the reward. Such observation, labeled as “law of effect”, extended the first evidence of (Pavlovian) conditioning, to actions, and not just stimuli. It further led to extensive work on operant conditioning, most notably by Skinner, showing that the amplitude of the reward associated to an action and the timing of its consecutive presentation can modulate the strength of the behavioral reinforcement and eventually influence learning

Nevertheless, the behaviorist movement that led the empirical studies on operant conditioning focused on the observable while neglecting the hidden variables that could have explained the association between reward and action, and thus any notion of subjective value. Later on, psychologists observed that in such an instrumental setting, two types of behaviors could be elicited following a procedure of devaluation of the stimuli (extinction) once the association had been established (learned) (Dickinson, 1985). When the action had been paired to a rewarding stimulus (in the sense that the stimulus has become sufficient to automatically -- psychologists say “habitually” -- trigger the action without evoking any anticipatory representation of the outcome), diminishing the value of the reward predicted by the stimulus did not seem to drive away the selection of this stimulus. In other words, the animal has become insensitive to outcome devaluation and just keeps on habitually performing the same behavioral response in the same context, similarly to humans’ tendency to persist with their previous credit card pin code after a recent notification of change by the bank. In contrast to this situation, when the animal had learned that the stimulus was paired to a reward and then the action also paired to the reward (in the sense that the animal is able to mentally anticipate and represent the outcome it will get as a consequence of the action), then a dissociation seemed to have been operated as the devaluation led to a decrease of the rewarding stimulus selection (Dickinson & Balleine, 1994). The interpretation here is that the anticipation of the outcome permitted by the action-reward association can be mentally combined with the fact that now the outcome has been experimentally devalued (e.g. the animal has been given so much food that it is now satiated) so that the animal will stop to perform the action because it no longer desires the outcome. Later on, lesion studies in animal models showed that such a dichotomy between habitual and goal-directed control in instrumental learning has a brain counterpart since lesion to different brain regions differentially impair habitual and goal-directed processes (Killcross & Coutureau, 2003; Yin et al, 2005).

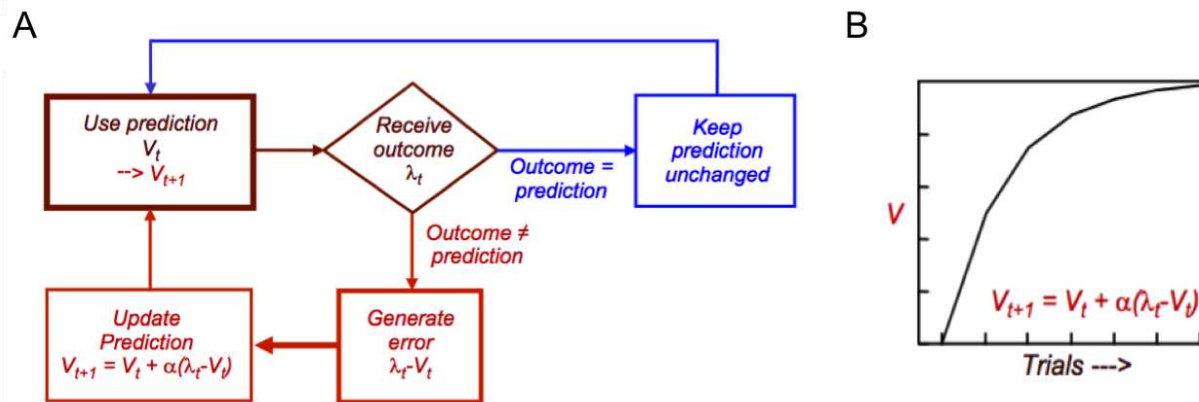
Researches on animal learning were later extended by the rise of computational tools coming from the field of dynamic programming, allowing the modelling of the potential computational processes and hidden variables in play. At the origin, the concept of prediction error formulated by Rescorla-Wagner (Rescorla, 1972), who proposed that the discrepancy between the prediction of the choice outcome -- the



expected reward at the time of choice (subjective value  $V(t)$ ) -- and its realization (experienced reward once the action chosen,  $r(t)$ ), was the main component driving learning:

$$V(t+1) = V(t) + \alpha(r(t) - V(t)) \quad (1.1)$$

In this model, the (signed) prediction error, is weighted by a learning rate  $\alpha$ , a parameter modulating its update influence on the subjective value of the selected action that lead to the reward ( $r(t)$ ) (**Fig.2.A**). From choice to choice, the value  $V$  driving the decision is modulated, in a trial-and-error fashion (**Fig.2.B**).



**Figure 2. Reward Prediction Error, as a main drive for learning** (adapted from Schultz, 2015)

Sutton & Barto (1998) then famously combined this concept of (reward) prediction error to dynamic programming to solve the decision problem formalized mathematically as a Markov Decision Process (MDP). A MDP is composed by four elements, a series of states ( $s$ ), an action set ( $a$ ) available in each state, a (probabilistic) transition function linking states together depending on the action chosen in each state, and a reward function ( $r$ ) which associates to each ( $s,a$ ) pair a reward value. Such a framework presents two advantages: first, any uncertain but controlled environment an animal or a robot faces can be simply modeled as an MDP; second, in this framework the action taken in a state fully determines the next state. Moreover, an MDP can be solved through dynamic programming to algorithmically determine the best policy (which action to select in a each state of the task) an agent should follow to maximize its total cumulative sum of reward over the long-term. Reinforcement learning algorithms were proposed to model an agent's optimal policy when the generative model of the environment (reward or transition functions of the MDP) is unknown. For instance, Q-learning (Watkins & Dayan, 1992) estimates the optimal decision at each state through the comparison of the expected reward associated to each state-action pair ( $Q(s,a)$ ). The key feature here is the computation of so-called Q-values, which are

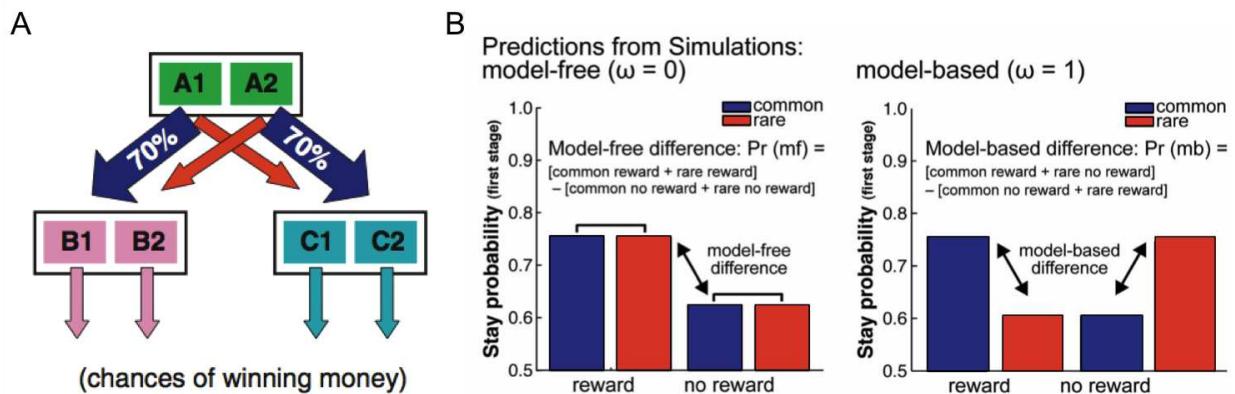
updated, at a certain rate similar to the alpha parameter in Equation 1.1, proportionally to a reward prediction error: how much the reward the agent expected to receive by selecting an action in a state is different from the actual reward received once the action has been performed. Such a teaching signal is the (reward) prediction error (RPE).

The suitability of this framework for psychology and neuroscience, is obvious, given the way decision experiments are usually designed. Indeed, in the laboratory, the experimenter imposes a structure to the probabilistic environment in which a human, or any other animal, evolves: the world is divided into sequential choices, or trials, in which a set of possible options represented by a recognizable stimulus are presented -- through (state,action) pairs --, each leading to an outcome with a different probability (reward). RL algorithms then allow to reverse-engineer the choice series to estimate the subjective values that drove the agent's behavior (Forstmann et al, 2011).

An impressive body of evidence shows that animal learning can be captured by reinforcement learning models (Schultz, 2015). These models are even more compelling that this learning process is biologically rooted: the (phasic) firing of dopamine neurons projecting to the ventral striatum (part of the basal ganglia) encodes a signal which strikingly resembles the reward prediction error driving value-based learning in reinforcement learning models (Schultz, 2015, 2016).

In the Q-learning model, the action values computed are attached to each available stimuli (i.e. through (state,action) pairs). This is because this model assumes that the agent has no structural knowledge of its environment, i.e. while people usually assume that the model already knows from the beginning of the simulation the total number of states and actions which are relevant for the task (and is moreover perfectly able to recognize them without ambiguity), the model nevertheless is not provided with nor tries to learn the probabilistic transitions between these states in the MDP that defines the task. In other words, the model is not able to do prospective inference by estimating the next state and outcome it is likely to reach after performing a given action in a given state. Instead, in the Q-learning model action-outcome and outcome-reward associations are conflated into unique state-action-reward values which are simply locally compared in a given state to decide which action to perform without estimating the consequences of the action (reactive behavior). Therefore if the reward or transition function changes throughout the task, like in reversal learning experiments or in outcome devaluation paradigms, the model will take a long time to update the Q-values and re-adjust the behavior to the new contingencies, pretty much in an habitual fashion (Wilson, 2014). These cached values synthesize the whole world for the simulated animal. But as the early literature of cognitive map suggests, animals can also develop through experience a predictive mental representation of their environment (Redish, 2016). In that case, state transitions can be conceptualized as a map or tree when the number of states remains tractable as often

in experiments. From this internal model of the task, outcomes can be distinguished from actions which allows the updates of the expected reward values of all the state-action pairs encountered to be propagated along the tree, resulting in a more flexible and efficient behavior when a change in the contingencies of the environment happens. The process is goal-oriented because it searches within the tree the best compromise between path length and magnitude of long-term reward values, rather than simply reacting to immediately perceived stimuli. Nevertheless, such a tree-search-based inference process takes time before making a decision (Viejo et al., 2015), which could partly explain why the two learning processes seem to co-exist in the brain: because each stable and familiar situation where the behavior can be automatized makes the brain save the time and energy associated to the tree-search process (Khamassi et al, 2016). Such RL models were labelled as model-based (because a map or a tree constitute a partial approximate model of the task), in opposition to Q-learning and other classical RL models that are considered as model-free (Daw et al, 2005; Doya et al, 2012; Daw et al, 2005; Sutton & Barto, 1998) (Fig.3).



**Figure 3. Task environment in which a dissociation between MF/MB can be observed**

A) Classic two stages (sequential choices) task used to differentiate between behavior generated through model-free and model-based reinforcement learning (RL). The first step state (green) leads, with a transition probability associated to each of the available action in this state to either one or the other second step state (pink, turquoise). (reproduced from Doll et al, 2012) B) Average stay probability expected for each of the two RL models. Left panel: Simulations show that model-free decision-making is reflected in a main effect of reward. That is, stay behavior on the first choice depends on whether behavior on the previous trial was rewarded or not. Model-free behavior is independent of the transition probability structure. Right panel: Model-based behavior is reflected in an interaction between transition on the previous trial and reward on the previous trial. That is, model-based behavior takes the model-free information as well as knowledge of the transition structure into account. Typically a hybrid version of the

model in which the omega parameters represents the arbitration weight between the two systems is fitted to the choice series of the participant. (adapted from Eppinger et al, 2013)

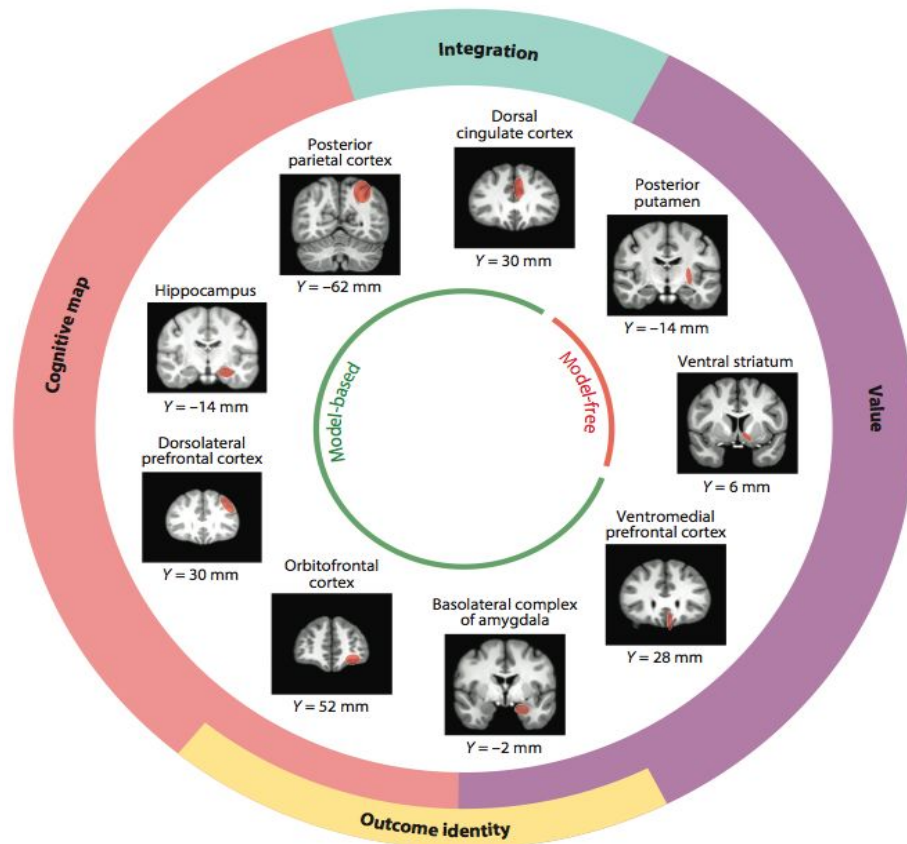
---

In the last two decades, the information theory from cognitive sciences has deeply influenced the field of neurosciences. The emerging field of computational neuroscience uses computational models such as RL algorithms to capture human behavior, to estimate the hidden variables behind their actions and eventually to investigate how the brain performs such computations through model-based fMRI analyses<sup>1</sup> (Frank, 2015; O'doherty et al, 2007). For instance, available evidence suggests that outcome values are encoded<sup>2</sup> in the vmPFC (Lebreton et al, 2009; McNamee et al, 2013; Reber et al, 2017), while the prediction error signals are detected in the striatum as input from dopamine neurons (Chase et al, 2015; Gläscher et al, 2010). Computations specifically related to goal-directed learning have been found to be correlated to the BOLD signal in the prefrontal cortex (PFC). For instance, the so-called state prediction error (SPE), which signals the accuracy of expected state transitions, has been found to be encoded in the dorsolateral (dl)PFC (Stalnaker, 2015). Moreover, it has been recently proposed that the orbitofrontal cortex (OFC, including the ventromedial (vm)PFC) might encode a predictive model of the task, mapping actions to outcomes in a state-structure representation of the model-based learning system (Doll et al, 2015; Schuck et al, 2016; Stalnaker, 2015). Such a model construction process would be made possible by the reception of retrieving signal from the hippocampus, which is believed to encode cognitive maps across domains (Wikenheiser & Schoenbaum, 2016). **Fig.4** summarizes a possible mapping of the different reinforcement learning computations in the human brain.

---

<sup>1</sup> Model-based fMRI consists in identifying the brain areas in which the Blood-oxygen-level dependent (BOLD) signal recorded by the scanner (a proxy of the underlying neural activity) covaries with the trial-by-trial value of the (individually) estimated variable of a computational model such as prediction error or Q-value in this case. (for in-depth explanation, see O'doherty, 2007)

<sup>2</sup> Using "encoded by" or even "correlates to the activity of" [a certain brain structure] to summarize that a significant correlation between a variable and the BOLD signal recorded in a specific voxel using MRI have been established is a simplification, given that such findings supports this hypothesis. It is however a common stretch of language.



**Figure 4. An implementation of RL in the human brain.** Schematic mapping specific neuroanatomical loci to the implementation of different functions underlying model-based and model-free control (reproduced from O'Doherty et al, 2017).

As briefly mentioned above, the habitual model-free learning process is considered as less flexible than the model-based one, but also less costly (computationally) as it requires only the storage of state-action cached values. On the other hand, the model-based system appears to be more flexible but also more engaging (since it requires to maintain and update a cognitive map of the task). The question of the modality of interaction between these two systems is thus central, in order to promote the influence of the right system on decision-making at the right moment, and thus benefit from the respective advantages of each system (Keramati et al, 2011). Initially authors have proposed that the two systems operate in parallel and that the brain selects the outcome of one of the two to drive decision-making based on the reliability of each system's predictions when faced with the uncertainty of the task (Daw et al, 2005;

Gläscher et al, 2010; Lee et al, 2014). However in practice, humans often display a pattern of average choice, as if they were employing a combination of the two (see Daw et al, 2011 and Fig.3.B). Nevertheless, such an average combination of the two might also be partly due to the fact that subjects' behavioral tendencies are often measured on average during a block of trials [cite Akam Costa Dayan, PLOS CB], while several more recent studies have clearly identified sudden shifts in the balance between the two systems to explain the trial-by-trial dynamic evolution of choices and reaction times (e.g. Viejo et al, 2017). Other recent studies have refined the initial hypothesis of parallel systems, and favored a more hierarchical system, model-free by default, that transiently engages in model-based computations, using action-outcome contingencies of the mental state map to refine the value computation (Deserno et al, 2015; Gershman et al, 2014; Zsuga et al, 2016). In fact recent evidence seems to point the nature of this system interaction in the direction of cognitive control mechanisms (Gershman et al, 2014; Khamassi et al, 2011; Otto et al, 2014), suggesting that a monitoring of the relative benefit, in terms of choice accuracy, to engage into costly model-based computations is performed (Kool et al, 2017; Pezzulo et al, 2013). It is worth noting however that the existence of such a meta-controller in the brain has not yet been established, and that other findings suggest that a cooperation between the two systems might also take place (Dollé, 2010; Kool et al, in press).

### C) Structure inference and (probabilistic) beliefs

Now that we have seen how the brain can learn in an uncertain (probabilistic) environment by using a representation of the well-structured task, one question arises: how does an agent deal with an unknown (not cued) MDP structure? As we previously mentioned, inferring the (state) structure of the environment through trial-and-error is challenging, and might require a model-based learning system dedicated to the encoding of the action-outcome contingencies that might operate independently of the model-free value-based learning systems (Gershman et Niv, 2010; Tenenbaum et al, 2011).

Learning the probabilistic structure of the task requires to form beliefs about the underlying hidden variables, such as the action-outcome contingencies or the state space, that generate the observations (rewards), and constantly update these internal representations (i.e. estimated probabilities) (**Fig.5.A**). In the last decade the Bayesian framework has encountered a growing interest to model learning, as it provides an upper bound on optimal levels of performance that can be reached by computational models of how an individual learns about and infers the underlying latent causes generating observable phenomena within its environment (Griffiths et al, 2012; O'Reilly et al, 2012; Zednik, 2016). At its heart, the Bayes theorem:

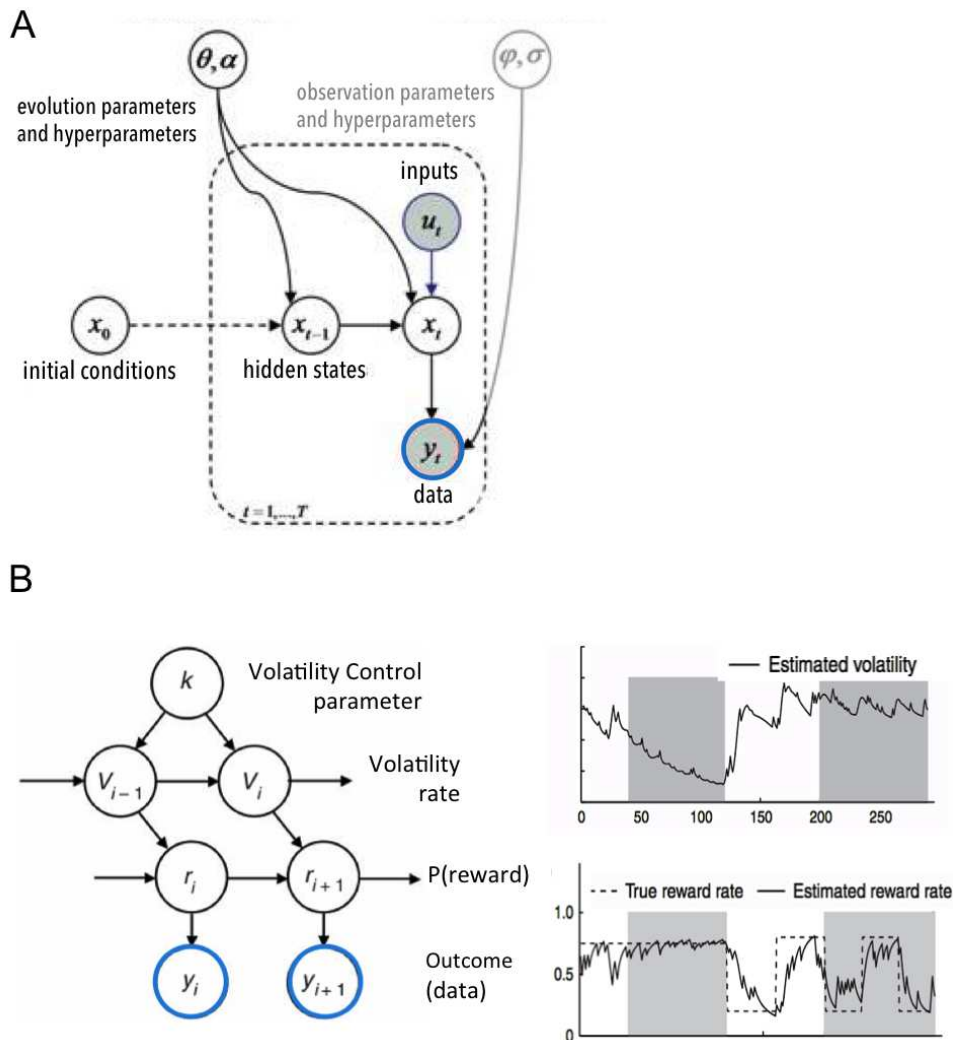
$$p(m|d) \propto p(d|m)p(m) \quad (1.2)$$

This (simplified) equation can be used to compute the probability that a particular model ( $m$ ) is correct given a dataset ( $d$ ). Such posterior probability can thus be seen as the belief that a hypothesized latent cause generates the observations accumulated so far. According to the theorem, the belief over this cause can be approximated by the estimation of both the probability of the observations given this cause (likelihood,  $p(d|m)$ ) and the probability of the cause itself (prior,  $p(m)$ ).

Such a learning framework has been proven to be useful to approximate human's perceptual performances (Pouget et al, 2013) and has been adapted to (value-based) decision problems in neuroeconomics. Hampton et al (2006) first suggested that humans faced with two choices in a learning task in which the probability of reward associated to each option flipped at regular interval (reversals) might use Bayesian inference to track these predictable changes in the state transition. Behrens et al (2007) went a step further and exposed participants to a similar two-armed bandit task<sup>3</sup> in which the reward probability associated to each option was stochastically and independently drifting from trial-to-trial. They modeled human choice using a generative model which represents the hidden variables generating the observations. By inverting the model through Bayes optimization, a Bayesian learner, starting with given priors over the variables probability distribution, can infer at each time point the sufficient parameters of the variable probability densities generating the observations. **Fig.5.B** reproduces the seminal (generative) model used by Behrens et al to show that humans can track the volatility of a stochastic task (e.g. a two-armed bandit), which corresponds to the trial-to-trial amplitude of change in the underlying reward probabilities. Such Bayesian agents can thus estimate quite efficiently the hidden variables structuring an uncertain decision environment.

---

<sup>3</sup> In two-armed bandit tasks, participants choose between two slot machines. Choosing a machine leads to a reward (binary discrete or continuous) drawn from a hidden probability distribution.



**Figure 5 - Bayesian Structure Learning**

A) An illustration of a generative model of a choice environment. (Adapted from <http://mbb-team.github.io/VBA-toolbox/wiki/>) B) (adapted from Behrens et al, 2007)

Nevertheless, this framework poses two constraints to structure learning. First, a Bayesian learning process requires the encoding of the probabilities, or at least of the sufficient statistics (mean and variance) of the hidden variable (Gaussian) probability distributions of the task structure, which become quickly intractable when transposed to more complex decision problems (Jones & Love, 2011). Second, clear evidence is still lacking that the brain can perform such a type of computation, still recent studies begin to suggest otherwise (Diaconescu et al, 2017; Ting et al, 2015). Another appealing use of such a modelling strategy is that, in the case where the generative model embodies in itself the structure of the



MDP (state transition function, action-outcome probabilities), the Bayesian agent, having correct beliefs over the latent causes of the task, becomes an optimal learner, which can thus be used to provide a benchmark for model comparison (by capturing departure from optimality) (Tauber et al, 2017).

However, despite their computational complexity, Bayesian models do not always outperform RL models when it comes to capturing human choice behavior (Geana & Niv, 2014; Niv et al, 2015). Some authors have recently proposed that humans could learn action-outcome contingencies through Bayesian inference and use the computed beliefs over the task structure to guide model-based learning (Collins & Franck, 2013; Gershman, 2017; O'Doherty et al, 2015; Starkweather et al, 2017). For instance, Collins & Koechlin (2012) proposed a tractable way to combine the two learning models in rich choice environments, based on the notion of stored chunks of action-outcome association representation (task-set). In their model, the RL process is duplicated for each different recognized task-set, and a Bayesian module tracks the reliability (accuracy) of the selected task-set given the environmental contingencies. In their model, when the current task-set does not lead to a satisfying level of performance, exploitation is stopped in favor of exploration which then either leads to the selection of a more appropriate task-set or to the creation of a new one. Imaging data (Donoso et al, 2014) showed that the striatum correlates to the exploitation of the current task-set, the dlPFC to its rejection, while the vmPFC was tracking the adequacy of the currently exploited one and the dorsal part of the dorsomedial (dm)PFC<sup>4</sup> was triggering the switch between exploitation and exploration. Interestingly, such an interaction between the accuracy of the action-outcome representations driving learning and the level of cognitive control (as assessed by the degree of remapping of the contingencies representation) has been shown by Bahlmann et al who observed an interaction between the dlPFC, encoding the level of engaged cognitive control, and the dmPFC, encoding the amplitude of the expected value (Bahlmann et al, 2015).

Taken together, these results suggest that the task structure representation in value-based learning might be subserved by the medial (m)PFC and that more lateral parts of the PFC might be involved in the shaping of such model to drive efficient learning.

Another way to reduce the computational weight of structure inference and to relax the Bayesian brain hypothesis is to consider that some types of heuristics can be used. For instance, Palminteri et al (2015) placed participants in a context in which they were implicitly primed for task (state) structure through trials in which feedback to both the chosen and unchosen options were provided to the participants – the feedback on the unchosen option thus constituting a fictive or counterfactual outcome. The authors showed that in the presence of feedback on the unchosen option, humans could efficiently use outcome

---

<sup>4</sup> activation peak was found specifically in the dorsal part of the anterior cingulate cortex (ACC) which correlated with the volatility parameter in Behrens et al 2007. Note that the higher the volatility of a choice environment, the higher the learning rate in a reinforcement model, which leads to a stronger trial-by-trial update, and thus to new expected values reflecting more the recent choice history of the agent than the past (Khamassi et al, 2013)

information to indirectly infer in which state they were by learning the value of the context (positive / negative) and re-frame the reward received according to the context: a positive reward could be experienced as subjectively negative if it was lower than the average reward of the current context, and conversely. In another study (Lefebvre et al, 2017), the authors highlighted the signs of use of an optimistic heuristic in humans learning in an uncertain environment. In their study, some participants (half) faced with multiple two-armed bandits displayed a stronger update (higher learning rate) for the expected value of an option that lead to a better than expected feedback (positive reward prediction error) than a negative one (see also Kuzmanovic & Rigoux, 2017). In addition, the same team identified an opposite trend (lower learning rate) when participants were also learning from the (provided) unchosen outcome, thus suggesting the existence of a sort of confirmation bias in learning (Palminteri et al, 2017). The way the representation of the hypothetical outcome that would have been obtained if one had made a different choice alters decision-making has been well studied in both economics and cognitive neuroscience (Coricelli & Rustichini, 2010). However, it requires for the subjects to know the task structure in order to infer the counterfactual outcome. In the case where capturing this task structure is difficult, some studies suggested that heuristics can be employed to compensate. In fact, a recent study (Gershman et al, 2017) showed that humans could even use imagination, even if suboptimal, to explore the possible structures of the task and drive model-based learning.

## II- Deciding in a social world

### **A) Social decision making**

The first section of this introduction has mostly focused on non-social tasks, where a subject learns in interaction with its environment through the sole evaluation of the outcome of her own actions. Nevertheless, social interactions constitute an important part of our daily life, which drove an important body of the decision-making literature to investigate choices made in a social environment (Rilling & Sanfey, 2011). Social decisions may be of different nature than nonsocial ones. Neuro-economists have questioned whether the value-based decision-making framework can be useful to shed light on the different components of social cognition, with one particular question at the front: what actually makes a social decision different from a non social decision from a cognitive point of view?

At first, one could simply ask if a decision with consequences for only oneself differs when made in a social or in a non-social context. In 2011, Zaki et al (2011) used functional MRI to test the famous influence (or conformity) effect first noticed by Asch in 1956 (Bond & Smith, 1996), according to which

peer choices might modulate (non-social) value-based decisions. They found that value-related brain regions (OFC, striatum) would be more activated when evaluating a face attractiveness after being informed that average participants ratings were congruent, than when incongruent. A year later, Lebreton et al (2012) directly tested whether the evaluation of a good can be affected by the implication of another person. They hypothesized that the value of a good would be enhanced when participants are (visually) informed that someone else would also be interested in choosing that good. They found that this was indeed the case, extended for goods of different natures, and presented evidence that brain areas previously labeled as part of the mirror neuron system (MNS) -- for their similar activation when an action is either performed or observed -- might modulate the activity of regions typically involved in non-social valuation (vmPFC, striatum). But this social contagion effect has been found to go beyond good evaluation, and to also shape individual preferences. Chung et al (2015) recently showed that observing someone else making a decision between risky options (gambles) influenced participants in their choices. Participants chose the risky gamble more often after observing two other persons picking the risky option than when choosing alone. This influence effect can be computationally operated by increasing the subjective value of the gamble chosen by others in observation trials. The authors linked the importance of this preference-related influence effect during valuation to the vmPFC activity, while the size of the discrepancy between one's preference and the other's choice correlated to the BOLD signal in the ACC and the Insula. Unlike the other two previous studies, however, this experiment did not reveal any involvement of brain regions usually implicated in social cognition tasks in preference modulation.

Bault et al (2010) showed that in probabilistic settings (choosing between two lotteries), observing someone making more risky decisions also pushed participants to become more risk-seeking. Their study however shows that the task itself, which displayed the outcome obtained by the other after making a choice in the same environment (state), leads to a social comparison effect. The computation of the difference between one's experienced reward in comparison to her opponent's (relative gain or loss) was correlated to BOLD activity in the striatum, while obtaining a higher reward than the other lead to higher activity in the temporo-parietal junction (TPJ, a region involved in social cognition tasks requiring self/other perspective switching) and a more dorsal part of the PFC (mPFC).

Another study by Strombach et al (2015) investigated preferences when the decisions were directed towards a counterpart. Participants were asked to choose between keeping a given amount of money and sharing a fixed amount with another person, this person varying from close relative to total stranger. The authors showed that the subjective value signal encoded in vmPFC was modulated by the activity of the TPJ, which varied in the task with the social distance of the receiver. Taken together, these results suggest a common subjective value computation process during non-social decisions made in a social

context, modulated by socially relevant information even when not necessary for goal achievement (expected value maximization).

Through an extensive review of studies investigating the neural correlates of value-based decision making in social contexts, Ruff & Fehr (2014) suggested that the available evidence point towards a common cognitive machinery subserving both social and non-social value-based decisions, modulated by the integration of signals provided by regions usually found to compute socially-specific features. In order to test this hypothesis of a common valuation system to social and non-social decisions, the two authors distinguished between different types of social frames in which the value-based decision-making model proposed by Rangel et al (2008) could be implemented. Indeed, while social decisions might be, as previously seen, non-social decisions made in a social context, they nevertheless can also be directed towards (or at least involved in) the consideration of another person's behavior, and might thus require the evaluation of the counterpart, their decisions, or the nature of the social interaction *per se*.

Note that a reward can be of a social nature, such as a smile or a positive verbal feedback. Then interrogating if such rewards are computed similarly to money, good or food can be considered as part of the value-based framework. Smith et al (2014) addressed this question in a study in which male participants were asked to pay money to see (subjectively) attractive pictures of women's faces. Their fMRI analysis revealed that the values of the social stimuli correlated with the BOLD signal recorded in the vmPFC. Functional connectivity revealed that the correlation in activity between vmPFC and brain areas related to social cognition such as TPJ, covaried with the willingness to pay to see socially attractive stimuli, suggesting signal exchange between these areas during the social evaluation process. However, their study did not tackle the issue of how to compute the value of rewards of social nature *per se*. Nevertheless, recent results suggest that the same striatum structure also encodes both the experience of non-social (money) and social reward stimuli such as social approval (Davey et al, 2010), but also social reputation (Wake & Izuma, 2017), and even the ability to engage in facial mimicry in response to a counterpart (Hsu et al, 2017).

As Ruff and Fehr pointed out, social decisions can also involve the evaluation of another person's reliability or attractiveness, for instance when asked to make a charitable donation. In a sense, the study by Strombach et al posits that the closer a recipient is from the participant, the more valuable s/he appears, and therefore the more willing they would be to donate some of their money. Here again, common brain mechanisms have been identified when participants were asked to pay for a good, or donate to charities. For instance, Hare et al (2010) showed that the BOLD signal in the vmPFC correlates with the magnitude of the donations, while functional connectivity analyses suggested that areas such as TPJ might mediate these social value signals. A recent study (Tusche et al, 2016) managed to disentangle the specific trial-by-trial involvement of TPJ activity during charitable decision-making, and concluded that TPJ

encodes a perspective-taking computation modulating social valuation towards more charitable choices, while the Insula correlates, independently at the individual level, with the empathetic motivation to donate. Thus, similar cognitive mechanisms seem to encode experienced or expected reward values in social and non-social contexts. However, the adequacy of the value-based framework for social decisions can be extended beyond self-oriented choices. An important body of evidence indeed suggests that similar brain structures encode experienced or expected rewards when a counterpart performs a task. The striatum has been found to be also activated when observing someone else experiencing a (non-social) reward following a choice (Mobbs et al, 2009). Besides, Nicolle et al (2012) showed that making a choice for a counterpart which required to take into account her preferences activated the vmPFC similarly than when making a choice for oneself. Interestingly, Apps et al found a distinction between two parts of the mPFC (ACC), one encoding similarly the effort one or someone else needs to engage in order to obtain a non-social reward, while the activity of a more anterior part of this area correlated with the net value (benefit - cost) of the counterpart only (Apps & Ramnani, 2014). Altogether, these studies suggest that a common value-based decision-making process implicated in non-social decisions is also recruited during social decisions, from non-social choices in a social context to the representation of the other's decisions. Evidence however suggests that a set of socially dedicated areas such as TPJ, Insula and mPFC could encode socially-relevant information then used to modulate one's choice process. Something about social decisions might actually be special.

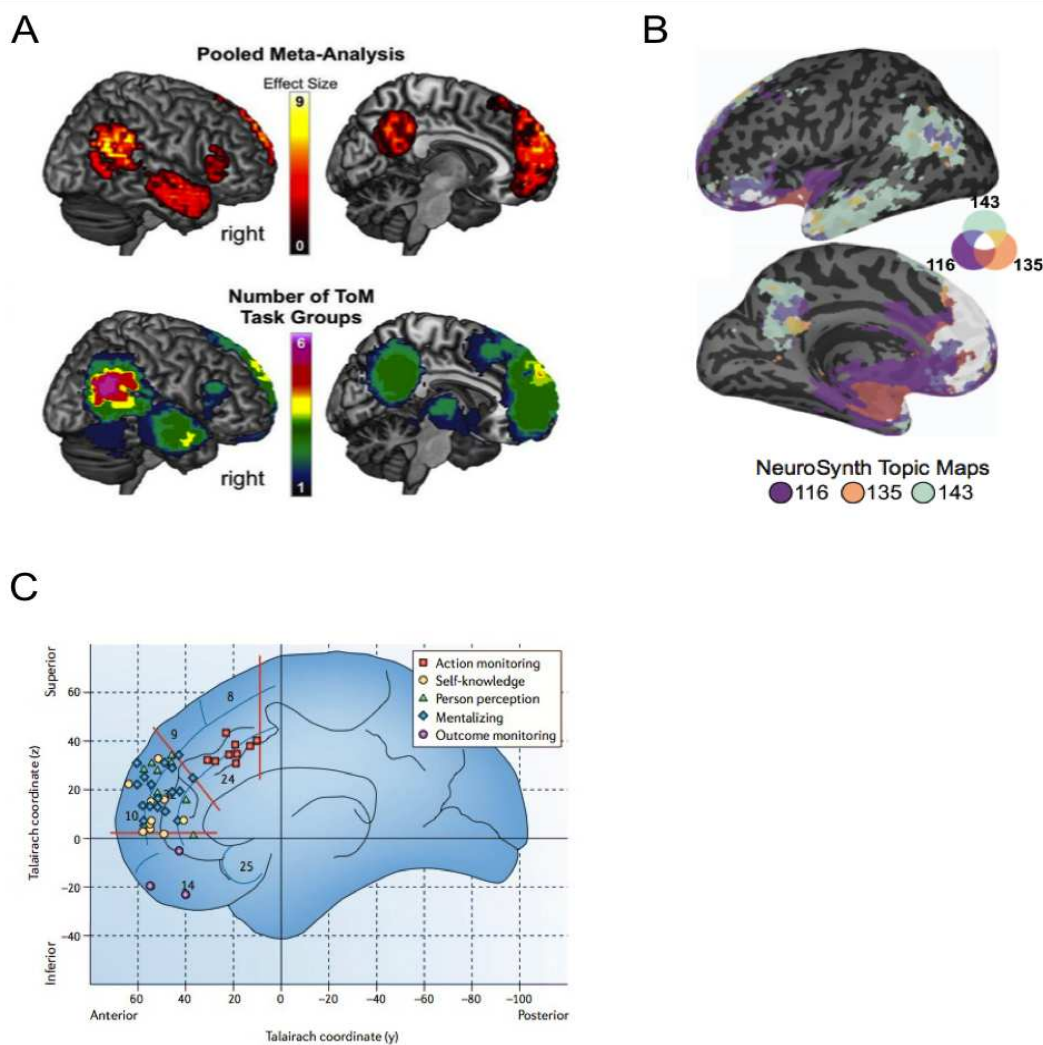
## **B) ToM and the Social brain hypothesis**

Theory of mind, refers to the ability to infer the mental states of other individuals, such as beliefs or intentions. This cognitive function has been the center of heavy debates since almost 40 years, from the question of its human-specificity to its neural underpinnings (Mahy et al, 2014). In the last decades, the ways to capture this cognitive function have moved from simple static experiments, such as the false-belief task<sup>5</sup> (Baron-Cohen et al, 1985), to the development of sophisticated computational models capable of capturing some of its key aspects (Baker et al, 2017). Many hypotheses have been made concerning the cognitive processes at its core. Some authors have suggested that displaying ToM requires a complete simulation of the other person using innate brain apparatus or dedicated modules (simulation theory) (Gallese & Goldman, 1998). In contrast, others argue that this capacity relies on the

---

<sup>5</sup> The standard version of the false belief task has been mainly used with children to assess their ability to engage in Theory of mind. The task consists in a series of pictures in which they first see a character, Sally, who leaves a valuable item (such as a chocolate bar) in a box before leaving her room. Then, while Sally is away, Anne comes and removes the item from the box and places it in another content (such as a basket). Finally, the children see Anne leaving and Sally returning to the scene. Participants are asked to tell where Sally will look for the item (or where she thinks it is).

development of a representation of the other in the same way we learn the contingencies of an external phenomenon (theory theory) (Gopnik & Wellman, 1994). However, the accumulated empirical evidence does not seem to allow strong rejection of one hypothesis in favor of the other (for a recent review see reference Mahy et al, 2014). Indeed, ToM is puzzling for cognitivists. On the one hand, neuroscience studies investigating ToM lead to quite heterogeneous results depending on the task, leading meta-analyses to provide little information about its underlying cognitive processes (Molenberghs et al, 2016; Schurz et al, 2014). On the other hand, a set of brain areas including TPJ and mPFC have been consistently found to correlate in activity with the ToM involvement in a variety of tasks (this paradox is presented in **Fig.6**) (Rushworth et al, 2013; Spunt & Adolphs, 2014); as if there was actually a *social brain* hidden within the brain (Dunbar, 2002). Some authors have recently tried to use a different type of analysis using MRI, starting from how people organize their representations about other's mental states. They clustered the variety of mental contents reported to be linked to thinking about a counterpart by an important pool of Internet participants into relevant dimensions or social mental states, and scanned subjects who engaged in the representation of situations located along these parameterized dimensions. The authors found some regularities in the brain activity, along with the mPFC or the TPJ, but no clearer distinctions emerged (Tamir et al, 2016).



**Figure 6 - A social brain?**

A) Meta-analytic results of 73 ToM studies from Schurz et al, 2014. (adapted from Schaafsma et al, 2015) B) Topic maps (Yarkoni et al., 2011; <http://neurosynth.org>): emotion (116), social games and interactions (135), mentalizing (143). (adapted from Stanley & Adolphs, 2013) C) Mapping of medial frontal cortex activations observed during action monitoring, social cognition and outcome monitoring. The meta-analysis suggests that social cognition tasks, which involve self-knowledge person perception and mentalizing activate areas in the anterior rostral MFC (arMFC). (adapted from Amodio & Frith, 2006).

When it comes to investigating the specific cognitive processes underlying ToM, one main difficulty lies in its nature: social interactions are hard to control, and hence hard to measure in a laboratory setting (Schilbach et al, 2013). Additionally, ToM encapsulates a variety of cognitive processes, not all

domain-specific, which is a conceptual problem (Frith & Frith, 2012). Recently a wake-up call was given by authors who suggested that, while faced with an *unsolvable* debate, the black box should be opened. ToM should be decomposed into its (hypothesized) sub-functions in order to understand how each one is encoded by the brain, how specific their computations are and how these ToM subprocesses interact (Schaafsma et al, 2015). Identifying the different processes involved in ToM, and how the related signals are encoded at the brain level might help disentangling these different subfunctions and shed light on their specificity (Spunt & Adolphs, 2017). For instance the mPFC and the TPJ have been shown to encode different computational signals given the task performed. The former has been linked to various cognitive functions such as representing other's preferences (Amodio & Frith, 2006), or distinguishing between the thoughts and feelings of the self and the others (Jenkins & Mitchell, 2011) (see Fig.6.C), while the latter was found to encode perspective-taking (Santesteban et al, 2011), as well as moral judgment (Koster-Hale et al, 2013), or the representation of someone else's emotions (Saxe & Houlihan, 2017).

A recent experiment by Kanske et al (2016) proposed that two of the subfunctions commonly associated to ToM (Schurz et al, 2014), empathy and mentalizing, might be subserved by two distinct processes. In a previously validated paradigm in which participants had to answer either empathy-oriented or mentalizing-oriented questions, they showed that the performance in each task did not correlate within subject, and that two distinct networks were recruited, including anterior insula in the empathy network and TPJ in mentalizing network, that interacted with each other in case of conflict. Besides, Koster-Hale & Saxe (2013) proposed that the brain computation involved in mentalizing can be investigated through the predictive coding framework. The authors suggested that the other's mental states can be considered as "unobservable, internal causal structure" driving the observable actions. Indeed, it has been shown that such mental states might be inferred from action observation, like confidence in a choice (Patel et al, 2012). In this framework, grasping the other's intentions or beliefs (but also preferences and even personality traits) can be reduced to a prediction problem, requiring inference over the other's behavior. Social learning, seen as a paradigm where subjects aim at using information provided throughout the interaction to improve the prediction over other participants' next action, thus appears as a key cognitive process underlying mentalizing.

The value-based computational approach presented previously could thus provide a useful framework to understand how the human brain dynamically updates beliefs about others in a continuously changing social environment. Moreover, the similarity of the processes involved in the computation of value-based choices in non-social and social settings suggests that social learning might operate upon the same cognitive mechanisms as non-social learning. Geşiarz & Crockett (2015) recently highlighted this parallel and proposed to extend the reinforcement learning framework to social decisions such as prosocial



behavior (donation, cooperation), arguing that “brain circuits specialized for prosocial behaviors, if such circuits exist, could either be embedded within the general-purpose [value-based learning] systems or constitute an input and output for them”.

### **C) Social learning**

Social learning can take at least two forms depending on the nature of the interaction between individuals. Learning from other persons requires implementing the information relative to their actions in one’s own learning process. Learning about the other, also called vicarious learning, requires learning someone’s behavior through observation. A learning type in-between could be seen as learning about a specific attribute of the other’s behavior in order to improve self-oriented decision-making.

#### **1) Learning from others**

Previous studies have shown that humans can learn in a RL fashion the action-outcome contingencies of a probabilistic learning task through the observation of a counterpart’s learning in a similar environment (Nicolle et al, 2011). Prediction error signals driving expected value update have been found to correlate with BOLD signal in similar areas during learning a task in isolation and during learning by observation (Cooper et al, 2012).

When faced with a decision problem, individuals can observe others’ decisions and try to incorporate this information into their own learning process. Burke et al (2010) developed a task where participants were faced with a two-armed bandit. During each trial the choice of an unknown confederate in the same task was displayed. By modulating the amount of information available about the other’s decision-making process, either complete (both actions and rewarding outcome of the other are observable), incomplete (other’s actions but not outcome displayed) or in isolation, they show that the more information about the other player’s decision process they had, the better they performed. They fitted the participants choices using a reinforcement learning (RL) model that computes, in complete feedback, a reward prediction error from the other’s decisions meant to refine, through simulation, the expected value from their own choice. This other related PE was found to correlate with BOLD signal in the OFC and striatum. In the incomplete information condition, the observational learning only influences imitation based on the relative advantage it represents for the participant’s cumulated rewards. The teaching signal in this condition, the action prediction error signal, was found to correlate with activity of the lateral prefrontal cortex (IPFC). Crucially,

Burke et al's results show that when it is available, individuals use the information provided by the observation of a confederate's decision to improve their own learning process. Conversely, when only actions are available they mainly use information provided by imitation to drive their choice.

Simply imitating the other's behavior is computationally efficient, and as a matter of fact this type of learning occupies an important place in animals' social learning strategies (Heyes & Galef, 1996). Also, a specific network has been identified in the human brain dedicated to action mirroring, an automatic process (Van Overwalle & Baetens, 2009). A hypothesis which has been put forward is that humans imitate the other's behavior only when this enables to maximize the probability to obtain an expected reward. Recently, Vostroknutov et al (2017) showed that in a similar task with incomplete information over a confederate behavior evolving in the same decision environment (a two-armed bandit in which reward probabilities change through time), humans were capable of using different types of observational learning strategies. Some subjects would simply imitate the action of the confederate but only when it had lead to increased reward in the past trials, as could be accounted for by the RL framework, similarly to what Burke et al showed. However, other participants imitated only when the opponent's past actions were constant, and when this mainly selected action corresponded to the best response according to their own learning process. The authors also showed that the latter, more sophisticated strategy, was used more among participants with a higher IQ (as measured by a suitable additional reasoning task). However, these (high IQ) participants were found to abandon this sophisticated observational learning process for a simpler imitation strategy when the IQ score (high only) of the other confederate was given to them. This result suggests that an information relative to the other can be considered as a good enough proxy to evaluate the usefulness of its behavior for one's own learning strategy.

However, in a social situation, some sort of intentional communication is usually possible between individuals, such as verbal communication or signalling (Pezzulo, 2013). In that case, an individual faced with an uncertain environment could for instance receive information directly from a counterpart, in the form of an advice. In 2008, Behrens et al showed that humans were able to integrate such information in their learning process to improve their performance in a changing two-armed bandit environment. The authors used a learning model combining a model-based RL and a Bayesian learning model learning the generative model of the task (as instructed to the participants) (see Fig.6.B). The Bayesian model was keeping track of the probability of change in the reward probability driving the choice environment (volatility) but also of correctness of the advice received by the counterpart, which was controlled by the experimenter in order to keep the task solvable. The (model-based) fMRI analysis revealed that the prediction error computed by the model over the information provided by the confederate (misleading advice) correlated to the amplitude of BOLD signal in the dmPFC and rTPJ, while the prediction error on the (own) value-based choice correlated with the activity of the striatum and the vmPFC. As previously

shown in the nonsocial domain, the dmPFC (ACC) was found to encode the choice environment volatility (change in action-outcome associations) over time, expect that here this correlation was stronger for subjects weighing more the advice in their learning process. Similarly the encoding of expected reward value derived from the other advice encoded in the vmPFC was found to be modulated by the volatility signal in the dmPFC. Biele et al (2011) investigated the effect of trustworthy advices on learned expected values in a probabilistic four-option-task, they showed that participants' choices were best modelled by a RL model including a (subjective) bonus to the experienced reward after following advice. At the brain level following an advice lead to a greater activity of the striatum at the time of the outcome, but also a reduced BOLD value-related signal in the OFC when following the advice did not lead to the reward.

Learning from another individual might thus require evaluating the trustworthiness of this person, either by mere observation -should I imitate or not when uncertain -, or when receiving an advice regarding of what choice to make or what learning strategy to adopt.

## 2) Learning about others

How humans learn about someone else encompasses a large number of different topics of investigation. We will distinguish here two main lines of interests: learning about someone's attributes and learning about someone's behavior.

As previously mentioned, two theories have been proposed in developmental and then cognitive psychology, to explain mentalizing, or inference about someone's intentions, beliefs or traits. At one side, the simulation theory posits that this inference process is rooted in our own cognition, so that we start from our knowledge about how we decide or about our own behavior to make sense of the behavior of the other. On the other side, the so called theory theory proposed that we learn from scratch about the other by making predictions and forming representations about someone's behavior as much as we learn about the world surrounding us (Apperly, 2008). Investigating the cognitive mechanisms implicated in learning about others, can shed light on this debate (Joiner et al, 2017; Mahy et al, 2014).

When it comes to judge someone, including across cultures, two main dimensions or traits, can be distinguished: warmth, and competence (Judd et al, 2005). Warmth is associated to the perceived intent including trustworthiness, helpfulness and sincerity, and appears crucial to maximize chances of survival. Competence on the other hand corresponds to traits such as the abilities to be creative, intelligent or skilled. From an evolutionary perspective, competence can be seen as important as warmth when it

comes to ask for help or to cooperate (Fiske et al, 2007). Albeit complex, both dimensions require accurate estimation, so that forming impressions should involve specific learning abilities. Similarity and dissimilarity have been proposed as an effective proxy for impression formation, and have been linked to the mPFC, the former to the ventral part of the mPFC, the latter to the dorsal mPFC (Mitchell et al, 2006). Recently, Ma et al (2013, 2016) used a specific fMRI design protocol (repetition-suppression) to identify the brain areas encoding both warmth and competence traits when representing someone. They found that both traits correlated with the BOLD signal of the vmPFC. However, in the protocol used by the authors, participants were asked to read sentences eliciting the mental representations of those traits, but the learning processes subserving the formation of impression regarding the others remained unexplored. In the study conducted by Mende-Siedlecki et al (2012), participants were presented with different faces, either alone or associated to a brief characteristic such as the description of a behavior, with a certain valence that switched after some consecutive trials. They found that among the network displaying higher activity when the description were attached to faces -- this network including TPJ and lateral PFC --, only the dmPFC showed an increased BOLD response after a switch suggesting an update in impression formation about the presented face.

Moreover, in a similar task Hughes et al (2017) showed that when participants were informed that the presented face corresponded to someone from the in group (same university), but not the out group, they failed to update the impression regarding this face once a negative description was attached, which translated into a lower activation (compared to actual update in ingroup) in these brain areas. Still, the type of stimuli used to initiate representation of someone else's traits were written descriptions, which limits the social aspect of the task. Boorman et al (2013) used a decision-making task in which participants had to predict the evolution of an asset with a value statistically fluctuating over time, either by betting directly on its next state (higher/lower than previous trial), or by betting over someone else (among 3 different but recognizable confederates) they could watch perform the same task. The learning model best fitting the participants' behavior was a Bayesian learner based upon a similar predictive model as introduced by Behrens et al (2007). The model could track both the performance level of the observed confederate and the probability of the asset to go up and down. The authors showed that the beliefs, representing in the Bayesian framework the trial-by-trial updated probability over someone's expertise (accuracy) in the task, correlated with the BOLD signal in the rmPFC (the expected value of the asset and the reward prediction error over its evolution correlated respectively to the activity of the vmPFC and striatum). They then distinguished two types of action prediction error: one at the time of their prediction outcome, when the confederate's action is revealed, encoded in the rTPJ and dmPFC; and one at the time of the other's choice outcome (update of the belief over its expertise), encoded in the dorsolateral PFC.

Other studies have recently proposed that humans can also infer other's preferences using a similar Bayesian inference scheme (Jern et al, 2017). Devaine & Daunizeau (2017) extended this framework to capture human's ability to learn someone else's biases in costly and uncertain value-based choice environment. They notably found that individuals exhibited a strong egocentric bias at start, before updating their impression over the other.

To test specifically how the computation underlying trait learning interact with the cognitive processes implicated in value-based learning, Hackel et al (2015) designed an experiment that aimed at distinguishing the proportion of trait-related information from reward-related information in a probabilistic two-choice task (a two-armed bandit) combining social and non-social stimuli. They showed that when informed about both the generosity (propensity to share money from a pool in social context, or pay out from a maximum payoff scale in the non-social context) and the reward (experienced outcome), participants rely more on the generosity information. Moreover, they showed that the prediction error related to this trait correlated with BOLD signal in the network previously implicated in social impression updating (IPFC, parietal cortex, and TPJ). In contrast, the reward prediction error correlated with striatal activity.

While forming one's impressions by learning someone else's traits is crucial in social interaction, inferring her intentions and beliefs through the observation of her trial-by-trial behavior appears as a key function of mentalizing (Joiner et al, 2017). Gershman et al (2016) tested whether humans are capable of learning a strategy employed by the other from her observed behavior. They developed a series of experiments in which participants were asked to predict habitual actions in the other's behavior, by variables such as action repetition or decision time, to more sophisticated patterns in choice history suggesting the use of a model-free system. They showed that humans are able to infer habitual control in the other's behavior. In addition, they presented evidence that individuals are more compliant (i.e. they blame less the other), when the negative outcome of their behavior is believed to have been generated through a habitual rather than a deliberative process. Together these results suggest that humans are able to build through learning a representation of an observed model-free agent.

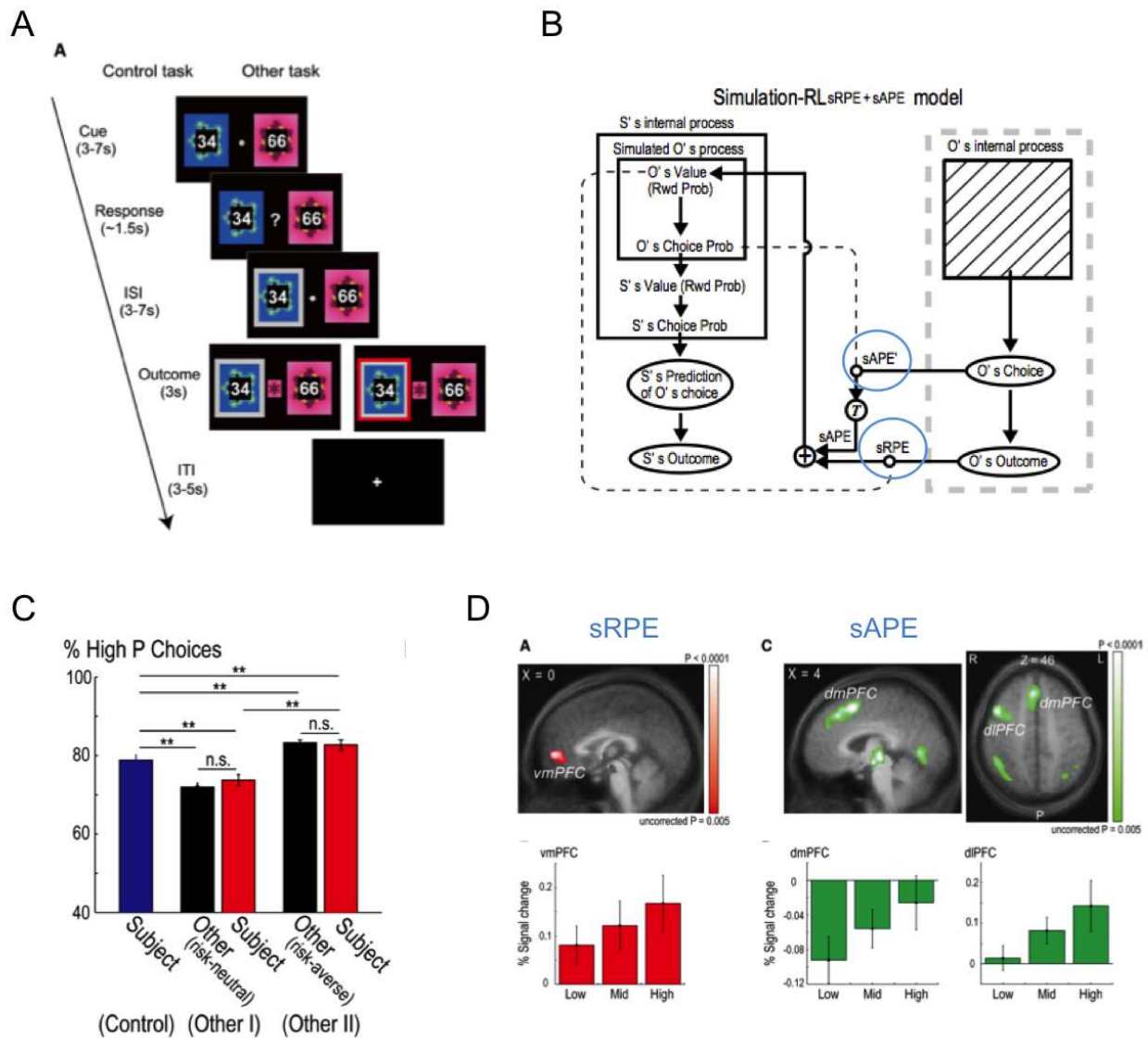
Seid-Fatemi & Tobler (2014) completed this scheme by showing that the blocking effect<sup>6</sup>, considered as characteristic of an efficient (model-free) reinforcement learning strategy (Tobler et al, 2006), could be learned in others by humans. Participants were asked to predict the outcome of a confederate's choice in a probabilistic two-choice task, in contrast to learning themselves the reward contingencies in the same task. The authors showed that participants displayed similar effect when predicting the other's choice, and

---

<sup>6</sup> Blocking corresponds to the absence of association between a stimulus and a reward when another present stimulus had already been paired to it, and so when the reward is already fully predictable by this other stimulus.

this effect was not due to learning in isolation on behalf of the other. Moreover, they found that the dorsomedial (dm)PFC displayed higher activity during blocking in the social condition of the experiment only, therefore suggesting that humans are able to efficiently model someone else's learning behavior.

Suzuki et al (2012) hypothesized that when observing a model-free learner, humans employ model-based reinforcement mechanisms to learn the action-outcome contingencies of that other person in order to predict her behavior in a probabilistic environment. The authors developed a paradigm in which participants had to learn the reward contingencies of a two-armed bandit in two conditions. In one condition they had to solve the decision problem for themselves; in the other condition they were informed trial after trial of the choice that the other made in the same task, and had to place a bet on her next choice (**Fig.7.A**). Manipulating the learning model generating the choices of the computerized confederate, and testing different models that could explain participants' choice behavior, they were able to specifically show that humans could predict the other's choice through the learning of a model of the other's behavior based on the computation of a reward and an action prediction error. These two signals (uncorrelated and both equally predictive of the other's choice), were used to generate at each trial an estimation of the simulated other's choice probability, which would then be integrated within their own (model-based) reinforcement learning process to drive their next prediction of the other's choice (**Fig.7.B**). Using (model-based) fMRI analyses, Suzuki et al were able to present evidence for a specific dissociation in the PFC between the subject's choice process and the representation and update of the other's choice behavior. In their task the BOLD signal in the vmPFC correlated (simulated) reward prediction error (rPE) of the other (participant's rPE correlating with striatum activity), while the activity of the dlPFC and dmPFC respectively correlated positively and negatively with the action prediction error (**Fig.7.D**) (which also correlated to activation in the TPJ). Again, these results provide more evidence that different parts of the human's brain are able to decompose the observed behavior of someone else into different meaningful variables that can be then combined into a prediction of the other's forthcoming actions.



**Figure 7 - Learning the other's action-outcome contingencies through observation**

A) Illustration of the experimental tasks. In both tasks, subjects chose between two fractal stimuli, and the stimulus chosen by the subject was indicated by a gray frame. In the Control task, the “correct” (rewarded) stimulus of the subject was revealed in the center. In the Other task, the rewarded stimulus of the other was indicated in the center, and the other’s choice was indicated by a red frame. B) Best fitting model: *Simulation-RL* (*sRPE+sAPE*). The large box on the left indicates the subject’s internal process; the smaller box inside indicates the other’s (O’s) internal decision making process being simulated by the subject. At the time of decision, subjects use the learned simulated-other’s value to first generate the simulated-other’s choice probability (O’s Choice Prob), based on which they generate their own value (S’s Value) and the subject’s choice probability for predicting the other’s choice (S’s Choice Prob). Accordingly, subjects then predict the other’s choice. Once the outcome is shown, subjects update the

simulated-other's value using the simulated-other's reward and action prediction errors (sRPE and sAPE), respectively; sRPE is the discrepancy between the simulated-other's value and the other's actual outcome, and sAPE is the discrepancy between the simulated-other's choice probability and the other's actual choice, in the value level. C) Similar data averaged across all trials in a separate experiment. The two Other task conditions, Other I and Other II, correspond to the other's choices modeled by the RL model using risk-neutral and risk-averse parameters, respectively. D) Neural activity in the vmPFC correlated significantly with the magnitude of the sRPE at the time of outcome. Neural activity in the dmPFC and dlPFC correlated significantly with the magnitude of the sAPE at the time of outcome. (adapted from Suzuki et al, 2012)

---

In another study employing a similar probabilistic task, Sul et al (2015) showed that when participants were asked to learn to maximize either the earning for themselves or the money for someone else (or both), the value-related signal driving their choice was segregated along the mPFC (**Fig.8.A**). Indeed, self-related values correlated with BOLD signal in the vmPFC, while other-regarding values implicated a more dorsal part of the mPFC, which was more active for prosocial participants than for more selfish ones. These results mimic the findings obtained by Christopoulos & King-Casas (2015) who investigated how expected reward values are encoded while learning in a two-armed bandit task in which choice outcomes were displayed for both the participants and someone else, since two distinct action-outcome contingencies had to be learned through reinforcement, one for the participants' own payoffs, and one for the other. The authors showed that BOLD signal in the mPFC correlated to the other's reward prediction error that was then used to update other-regarding values (**Fig.8.B**). They moreover showed that the strength of this teaching signal was modulated by the social value orientation score of participants, indicating their level of prosociality<sup>7</sup>.

These findings thus suggest that humans are able to build a representation of the other's choice behavior, and learn in a reinforcement fashion through the use of this model. However, it remains unclear how these computations related to the other interact with the (egocentric) cognitive process in social reinforcement learning.

The study by Christopoulos & King-Casas indeed suggests a key feature of decision in social context: the influence of one's behavior over the other. Two recent studies provide interesting insights on that matter. Suzuki et al (2016) investigated the preference contagion effect we briefly described in the last section, in a learning setting in which humans had to make decisions between an uncertain probabilistic option and a safe one that varied (independently) in risk and expected reward, and then to switch to prediction of a

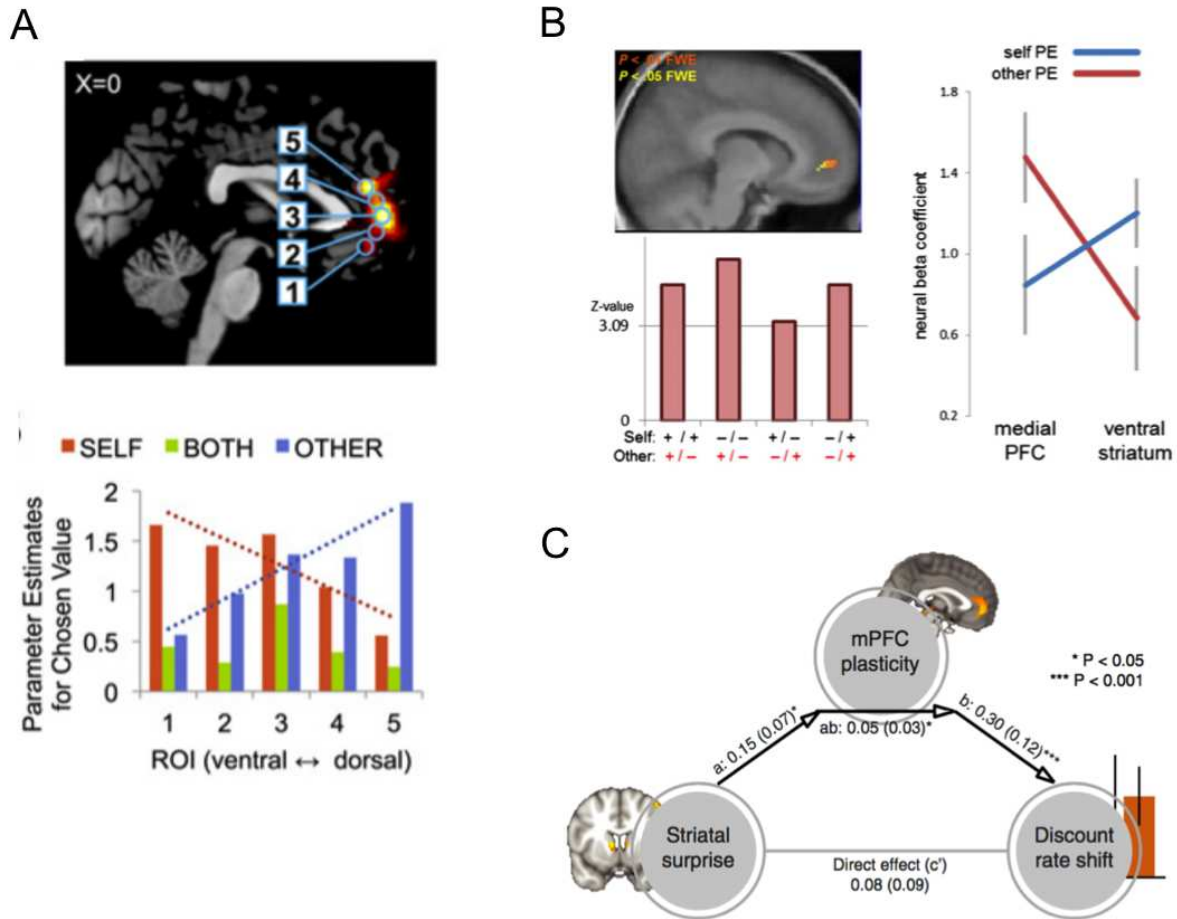
---

<sup>7</sup> "In this separate [SVO] task, participants chose between allocations of an endowment between oneself and an anonymous social partner, thus revealing the Cooperative, Individualistic, or Competitive orientation of each participant."



confederate's choice in the next series of trial. Using a Bayesian learning model, the authors showed that participants were able to track the preference of the other towards risk, and to use this signal (found to be encoded in the striatum) to predict their choice. Moreover, they found that participants' own relation to risk was modulated by such learning, hence leading to a contagion effect, and that such modulation, independent of their own learning process, was correlated with the activity of the dlPFC. In another study using a sophisticated fMRI paradigm (repetition-suppression), Garvert et al (2015) were able to show that the change in configuration of the mPFC activity correlated to the contagion effect, captured by the shift in discount rate in one's own intertemporal decisions towards the learned time-discounting preference of a confederate. Their results also suggest that the amplitude of such mPFC plasticity modulates the correlation observed in the activity of the striatum and the shift in discount rate (**Fig.8.C**).

Taken together these results suggest that humans are able to learn specific traits of another person through minimal information regarding their (choice) behavior. They are also able to learn someone else's action-outcome contingencies on-line, from the observation of her trial-by-trial decisions. A learned internal model of the other's learning process is then used to implement a (model-based) reinforcement learning strategy, using the prediction made over the other's choices to guide their own learning process. Brain imagery data suggests the existence of a specific network dedicated to the learning of the action-outcome contingencies underlying the confederate behavior, which includes (dorsal) mPFC and lateral PFC. Moreover, it seems likely that the prediction error signals computed in these areas modulate the (non-specific) value-related signals driving the learning process and subserved by brain regions like vmPFC and striatum.



**Figure 8 - Interaction between social and nonsocial learning processes in the mPFC**

A) Spatial gradient for self- and other- regarding value computation within the MPFC. Dotted lines indicate linear fits of the spatial gradient for self (red) and other (blue) conditions. (adapted from Sul et al, 2015) B) Prediction error signal for other-value encoded in the mPFC. Left panel: Preference-dependent prediction errors associated with updating of other-value were estimated across four experimental conditions and subsequently regressed to hemodynamic activity. The four conditions correspond to the conditions in which other-value differed between the two available options. Right panel: Beta values representing fitted responses to 'self' [blue] and 'other' [red] PE in medial prefrontal cortex and ventral striatum. (adapted from Christopoulos & King-Casas, 2015) C) The striatal correlate of the surprise about the novel other's choices predicted plasticity in the mPFC (path a), and the mediator (mPFC plasticity) predicted the shift of subjects' own discount rate toward the discount rate of the novel other (path b, controlled for the striatal surprise signal). There was a significant mediation effect (path ab), indicating that mPFC plasticity formally mediates the relationship between striatal surprise and the shift in discount rate. (adapted from Garvert et al, 2015)

Nevertheless, it is still unclear if these cognitive processes are specifically involved during social decision-making or if they subservise sophisticated forms of learning. Moreover, as recent studies point out, the distinction between the two forms of learning remains blurry: a contagion is observed between social and non-social learning processes when the task requires to alternate between the two to make appropriate (rewarded) decisions. In fact, during social interactions, the outcome of one's behavior often interacts with the outcome of the other, and therefore influences the contingencies of the choice environment itself.

### 3) Learning through interactions with others

Most of our social interactions require both learning about and from another person, and often at the same time. Indeed, when the outcome of one's decision impacts the environment of the other, or directly the outcome of her own action, humans must engage in active learning to infer the other's intentions and beliefs, to adapt their own behavior to these predictions and to update from the feedback they receive in return.

Several recent studies have tackled this complex problem using a variety of ecological tasks. For instance, Suzuki et al (2015) scanned participants while they engaged in a consensus task with a small group of (unknown) humans. The goal was that after a series of trials all the participants agreed on choosing one among two displayed food items, thus requiring to take into account the initial subjective values (preferences) over the two goods and then, given the feedback received from the other participants, adapt their own behavior in order to obtain the desired good. The authors showed, using a Bayesian learning model, that subjects' preferences correlated with BOLD signal in the vmPFC, while the tracking of the probability that the group chooses an item correlated with activity in the rTPJ. Crucially, they found that the trial-by-trial probability of choice computed by the model through the integration of these teaching signals was correlated with the recorded activity of the ACC. Hertz et al (2017) recently aimed to investigate teaching or influence behavior over the other's learning process. They set up a task in which participants could advise a learner on what choice to make in order to maximize his final earnings. Importantly, the scanned subjects were advising in competition against another adviser, which triggered learning over the optimal choice strategy that would make the learner listen to them. The authors showed that the participant's advising choice was mainly influenced by the relative accuracy of their advice (in comparison with the one of their competitor), and by the actual learner's decisions. They showed that the former hidden variable correlated to the BOLD signal in the mPFC, while the latter matched the activity of the rTPJ. This experiment, in a sense, represents the other side of the Behrens et

al (2008) study on learning a non-social task by weighted advices received by a counterpart, who found that the prediction error over the advisor's choice correlated with the BOLD signal in dmPFC and rTPJ.

Still, in social interactions, interacting with another individual often implies interdependence between the two individuals' behaviors. From a decisional standpoint, predicting the other's behavior during an interaction in which one's outcome depends on the other's decision and vice versa, may require more cognitive resources/computations than simply learning action-outcome contingencies hidden in someone's choice behavior when interacting in isolation (i.e. not in reciprocal interactions). If in both settings (social and non-social) an individual would have to learn the *generative model* in order to apply model-based reinforcement learning and maximize her earnings, in an interactive environment, the hidden structure that the other constitutes changes depending on her own actions. In machine learning terms, such an influence of another individual's behavior on the outcome of one's own actions in the world makes the Markov Decision Process *non-stationary*, which requires to constantly update one's internal model of task.

Therefore, from a value-based decision-making point of view, social interactions might be either seen as a changing outcome rewards environment or more crucially as an adapting one, reacting to one's own choices and in which the term "interaction" takes on its full meaning (Hari et al, 2015) (**Fig.9**).

---

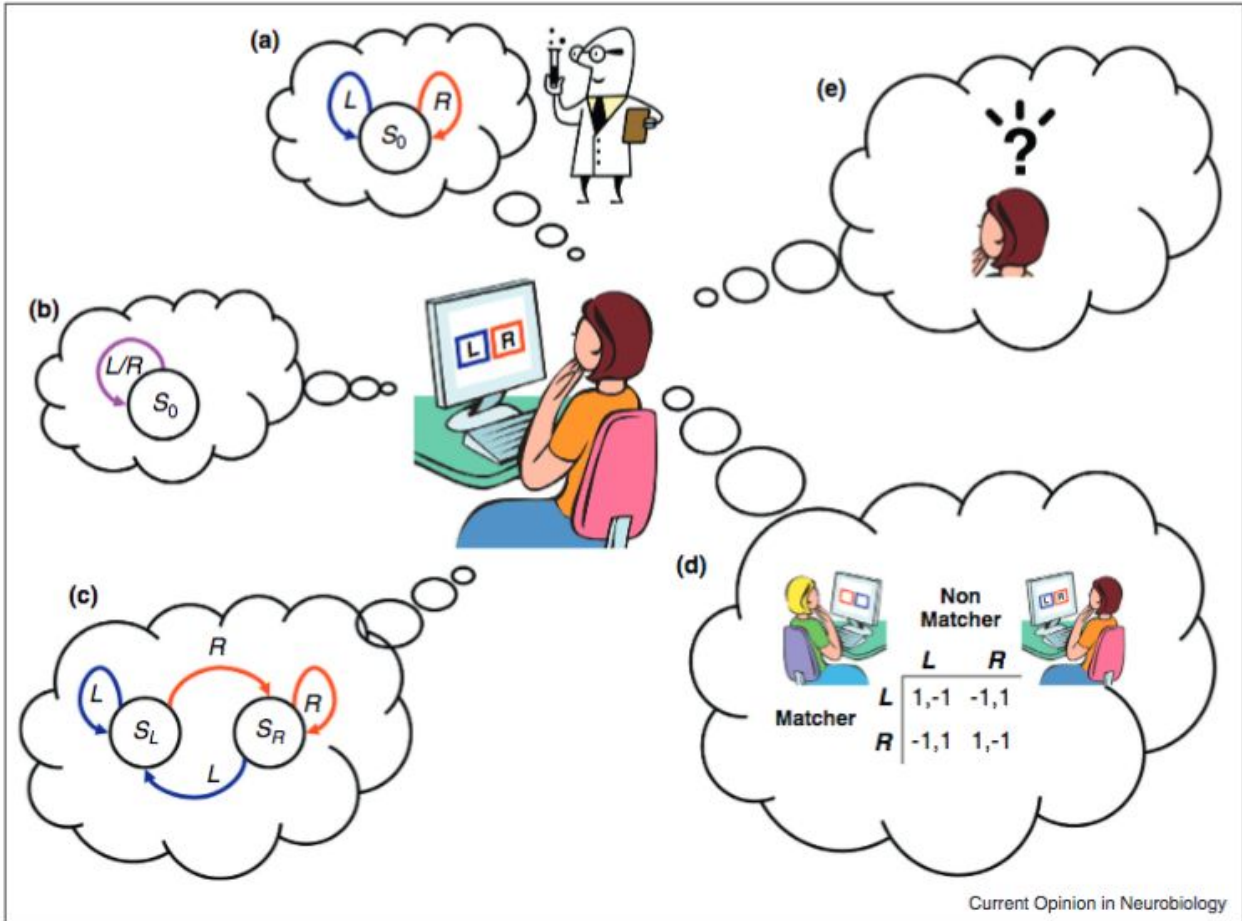


Figure 9 - Towards strategic learning (adapted from Shteingart & Loewenstein, 2014)

### III- Social learning in strategic interactions

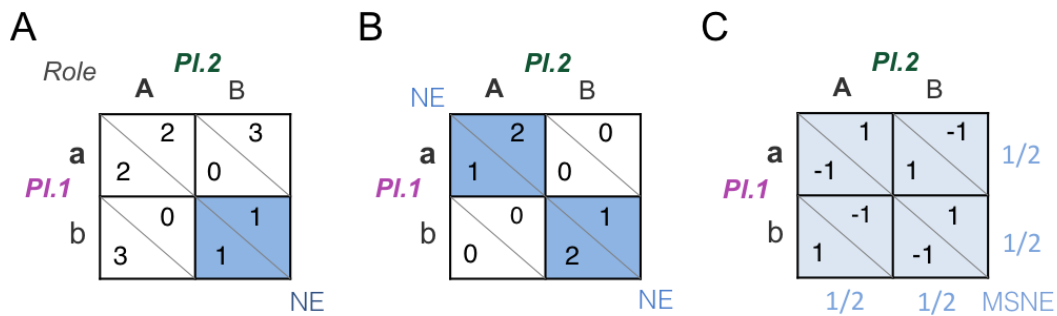
#### A) Strategic interaction

##### 1) Behavioral Game Theory

In economics, people call “strategic interaction” any social situation where one individual’s outcomes depend both on their own actions and on the ones of the other person(s). In other words, in strategic interactions outcomes are jointly determined. Game-theory<sup>8</sup> models strategic interactions as games, with consists in a set of players (roles), a set of strategies (actions) for each player, a set of

<sup>8</sup> By *game theory*, we actually imply noncooperative game theory, in which no direct communication is permitted between players, preventing coalition.

possible outcomes (often monetary gains/losses), and a function that links the outcome to each strategy profile (probability distribution over the actions of each player). Games are played simultaneously, with no communication possible. Therefore, no other information than the one provided by the game structure is available to the players. Games can take multiple forms; we will focus here only on normal-form games with complete information. Since this thesis focuses on dyadic interactions, two-player games can be represented as a payoff matrix. **Fig.10** illustrates three games widely used in the literature.



**Figure 10. Strategic interactions modeled as 2x2 normal form, complete information, games.**

A) Prisoner's Dilemma B) Battle of the Sexes C) Matching Pennies. The numbers in the matrices indicate monetary outcomes available to each player depending on the combination of actions made by the two players: a or b by Player 1 ("PI.1"); A or B by Player 2 ("PI.2"). NE: Nash Equilibrium. MSNE: Mixed Strategy Nash Equilibrium. Numbers written outside the matrices represent the probability distribution over each action prescribed by the MSNE corresponding to a *mutual best-response* profile.

This formalism represents two main advantages. First, it models quite efficiently strategic interactions in a form comparable to a Markov Decision Process, second it provides a well-defined framework for solution concepts. Game theory indeed makes prescriptions about *optimal* play in any normal-form games, under the concept of equilibrium. The well-known Nash Equilibrium (NE) for instance is a strategy profile that insures each player involved to have no incentive to deviate from the strategy prescribed by the equilibrium. The probability distribution over each action prescribed by this equilibrium corresponds to a *best-response correspondence* profile, or *mutual best-response* profile, which states that at this point (strategy profile) each player best responds to the strategy of the other.

The classic example of prisoner's dilemma, illustrated in **Fig.10.A**, consists in a symmetric game in which the two players are confronted to the same games, in which they have two actions. Typically a way to represent this strategic interaction would be to see the players in the role of prisoners, kept in separate cells with no possible communication, and to whom a proposition is made to choose between either to

cooperate (action a/A) or to defect (action b/B). In this game there is one equilibrium point which corresponds to the dominant strategy profile: for each player the payoffs associated to the action b/B, no matter what the other player chooses, is always superior to the one linked to the action a/A. Therefore, choosing b/B in this game will always lead the player to a better outcome (given each of the two actions the other player chooses). Therefore, the strategy profile (b,B) corresponds to the best-response for which none of the two players will increase their outcome by deviating unilaterally from it.

The game illustrated in **Fig.10.B**, and usually called “battle of the sexes”, could be represented as a player being the husband and the other the wife who have two distinct preferences for spending the night doing a common activity (like going to the movie theater vs. going to a concert). In this game, there is no dominant strategy but two equilibrium points which correspond to the strategy profile for which the two characters do something together instead of following their absolute preference and going out by themselves. However, it could be the case that a game has no pure (Nash) equilibrium: none of the two available actions leads to a situation which corresponds to a mutual best-response. This is the case for the matching pennies game illustrated in **Fig.10.C**. In this game the Nash equilibrium is mixed, meaning that the prescribed strategy profile is a probability mixture over all the actions, so that if each player chooses each action with a probability of 1/2, the strategy profile ([1/2: a,1/2; b],[1/2: A,1/2; B]) corresponds to a mutual best response. A mixed strategy equilibrium (MSNE), thus requires that each action is played randomly with a fixed probability. A pure NE, can thus be seen as a special case of MSNE, where the corresponding strategy profile states that one action should be played with a probability of 1.

The notion of Nash equilibrium thus makes two assumptions: (1) that the two players involved are self-interested and maximize their utility over the payoffs of the game by best responding to the beliefs<sup>9</sup> they hold about the game structure and the strategy profile of the other; and (2), that each player holds correct (certain and accurate) beliefs over the strategy of the other so that her best response corresponds to the mutual best-response profile. It is important to note here that this solution concept can be interpreted in two ways: it could be seen as prescriptive, i.e. what (rational) people should do, or descriptive, what people would do in such a situation. This latter assumption has been extensively tested with humans playing games in laboratory (Camerer, 2003). However, the behavioral results consistently showed that humans do not follow the solution prescription made by the game theory, and often deviate from theoretical distributions. Moreover, when the aggregated choices in laboratory seem to fit the

---

<sup>9</sup> Here we talk about “beliefs”. However, the exact premise is that rational players must hold “mutual recognition [or knowledge] of rationality”. The notion of knowledge can be seen as analogous to “belief”, which embodies in itself the notion of probability (see chapter I) allowing for noise in the information received (in case the environment is stochastic for instance), as in the computational process underlying the representation of knowledge. For neuroscientific consideration of this notion see Wyart & Koehlin (2016).

theoretical predictions, a slight change in the game (payoff) structure can lead to a strong deviation from NE. A classic example by Goeree & Holt (2001) is illustrated in **Fig.11**.

---

		<i>Left (48%)</i>	<i>Right (52%)</i>
Symmetric Matching Pennies	<i>Top (48%)</i>	80, 40	40, 80
	<i>Bottom (52%)</i>	40, 80	80, 40
		<i>Left (16%)</i>	<i>Right (84%)</i>
Asymmetric Matching Pennies	<i>Top (96%)</i>	320, 40	40, 80
	<i>Bottom (4%)</i>	40, 80	80, 40
		<i>Left (80%)</i>	<i>Right (20%)</i>
Reversed Asymmetry	<i>Top (8%)</i>	44, 40	40, 80
	<i>Bottom (92%)</i>	40, 80	80, 40

**Figure 11. Empirical deviation from Nash prediction illustrated in a one-shot Matching Pennies game.** When the game is symmetric (top matrix) the average choice (%) fit the mixed-strategy Nash equilibrium (1/2, 1/2). In the center matrix, MSNE predicts that row's decision probabilities should not change (the row player should ignore the unusually high payoff of 320 (cents) and still choose Top or Bottom with probabilities of 1/2), and since column's payoffs are either 40 or 80 for playing Left and either 80 or 40 for playing Right, row's decision probabilities must equal 1/2 to keep column indifferent between Left and Right, and hence willing to randomize. However a strong empirical deviation is observed in this case leading Goeree & Holt to conclude that in practice, the MSNE prediction "only works by coincidence, when the payoffs are symmetric". N=50. (reproduced from Goeree & Holt, 2001)

---

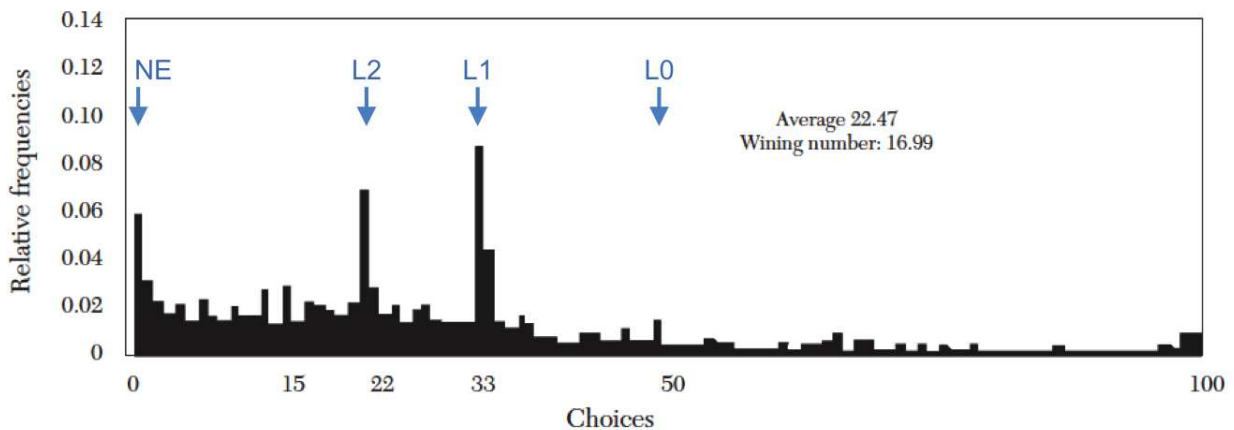
This puzzling result led economists to two different paths concerning game theory (and in general neoclassical economics). The first consisted in assuming that the theoretical solution has no predictive power, that it is *only* a mathematical solution to games -- games could thus be seen as abstract objects not meant to be implemented in reality (Fumagalli, 2016). The second consisted in considering that the theory makes accurate prescription for rational play but that something prevents humans to reach optimality. This second idea, initiated in the 60's, was strongly inspired by the departure from behaviorism and the emerging theory in psychology according to which human can be seen as information-processing entities (Camerer et al, 2011). Following this idea of a bounded rationality, behavioral game theory aimed



at running controlled experiments as well as using field data, in order to test what restrained humans from displaying rational behavior, such as equilibrium play.

Following this line of inquiry, behavior game theorists formulated the hypothesis that one (or both) of the premises of rational game play should be relaxed. In other words, either humans for some reason do not maximize their utility over their payoffs and do not best respond to their correct beliefs over the strategy of the other player, or they do attempt to maximize but best respond to incorrect beliefs. Over the last decade, different models have been proposed to test these assumptions, broadly labeled as non-equilibrium models of strategic thinking. The first type of departure has been essentially modeled as noisy and stochastic choice. For instance, Quantal Response Equilibrium (QRE) (McKelvey & Palfrey, 1995) relaxes the assumption of best response and considers errors in choices (keeping the assumption of (statistically) accurate beliefs and equilibrium responses). In interactive settings, a small amount of noise can have a large effect, and QRE models that incorporate stochastic elements in the analysis of interactive decisions can explain 'anomalous' behaviors (i.e. deviations from rationality) in several experimental games. According to QRE models, individuals are more likely to select better than worse actions, but they are often unable to select the very best one. This type of models thus posits that humans can compute correct beliefs but fail to implement them properly in their choice. This assumption however is purely theoretical, and not data-driven.

Based on the observation of frequency peaks in a famous game called "p-beauty contest", in which an important number of players choose a number from 0 to 100 in order to get the closest to the average of the chosen numbers (or a multiple of an announced value  $p$ ) (**Fig 12**), another class of bounded rationality models have been proposed. The Level- $k$  models (Nagel, 1995; Stahl & Wilson, 1995) and the Cognitive Hierarchy (CH) models (Camerer et al, 2004; Ho et al, 1998) maintain the rational assumption of best response to beliefs, but relax the assumption of 'correct' beliefs (and rational expectation about beliefs). This class of models considers the presence of heterogeneous players in terms of a hierarchy or level of strategic sophistication: level-0 players are strategically naive (e.g. they play randomly, or do not fully consider the incentives of the game), while higher-level players iteratively best respond (i.e. respond optimally) to a distribution (Poisson for CH, and as  $k-1$  for Level- $k$  models) of lower-level players (e.g. L1 players best respond to L0 ones; L2 best respond to a distribution of L1 and L0; and so on). According to this model, high-level reasoners (L2 or higher) expect the others to behave strategically, whereas low-level reasoners (L1) choose based on the expectation that others will choose randomly. Empirically this type of models has been proven to quite efficiently capture departures from Nash equilibrium (Camerer et al, 2015) (see Crawford et al, 2013 for an extensive review of empirical evidence).



**Figure 12. Empirical deviation from Nash prediction illustrated in a game of matching pennies**

Participants choose a number between 0 and 100. The winner is the person whose number is closest to  $2/3$  times the average of all chosen numbers. The level- $k$  model (iterated best response) predicts that a naïve player (level 0) chooses randomly. A level 1 (low-level) player thinks of others as level 0 reasoning and chooses 33 ( $=2/3 \times 50$ , where 50 is the average of randomly chosen numbers from 0 to 100). A more sophisticated player (level 2, high level) supposes that everybody thinks like a level 1 player and therefore he or she chooses 22 ( $=(2/3)^2 \times 50$ ). Zero is the equilibrium solution of the game. (adapted from Bosch-Domènech et al, 2002)

Modelling behavioral departure from an (informational) optimum should not only take into account the behavioral data but should also be (biologically) plausible (Chater et al, 2017). In that sense, this last class of models relaxing the assumption of rationality -- which assumes that players best respond given their incorrect beliefs -- is appealing from a cognitive point of view. Indeed, the computational cost of forming high-order beliefs over someone else's behavior can be quite high when no other information than the game structure (i.e. the payoff matrix) is provided. Congruently, behavioral studies in game theory have found a correlation between the level of strategic sophistication observed experimentally and reasoning or memory capacities measured in additional tasks (Carpenter et al, 2013; Gill & Prowse, 2012).

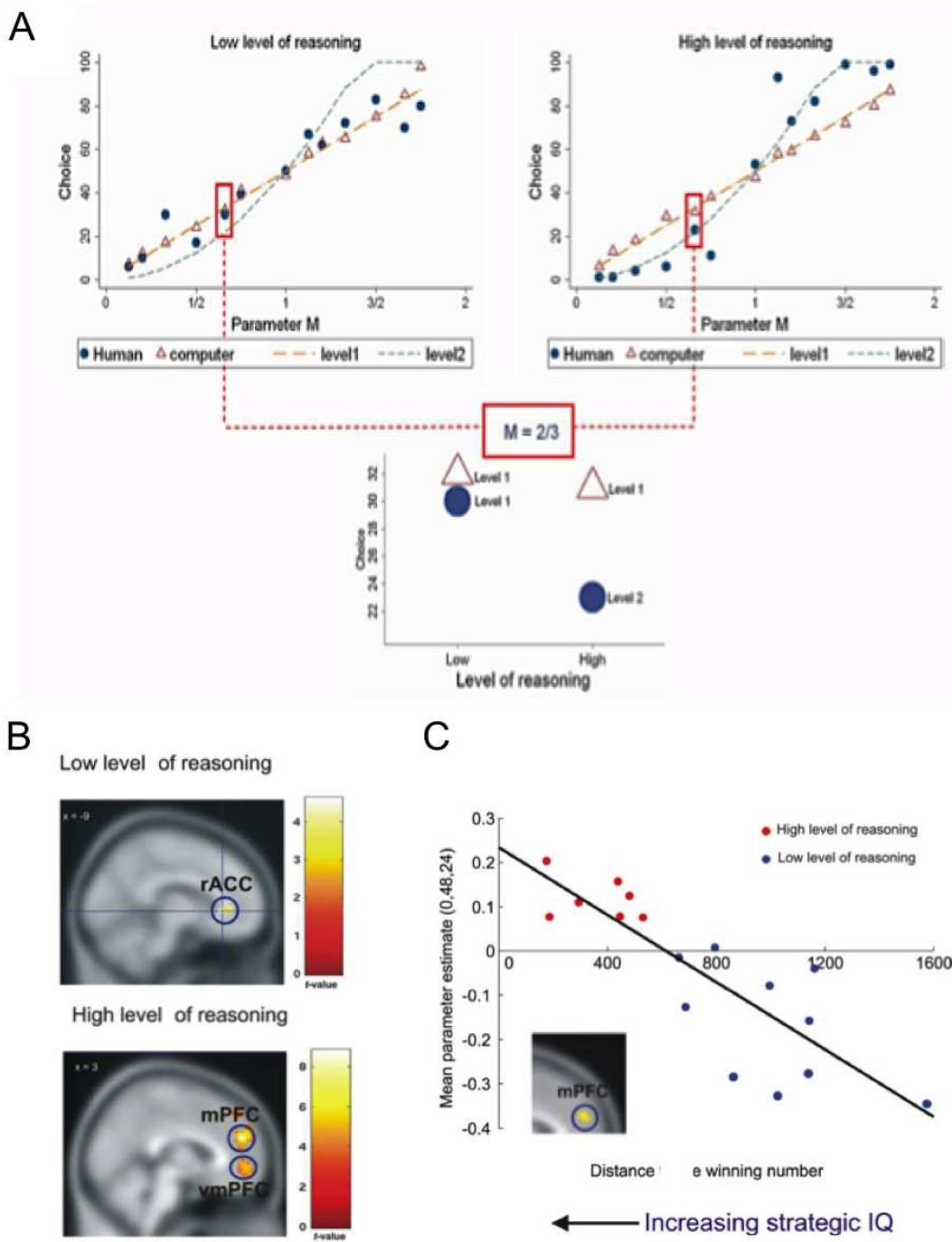
The goal of a new field called “neuroeconomics” was thus to expand the knowledge and theories developed in behavioral economics by measuring and manipulating variables that behavioral economists were unable to observe (Camerer et al, 2004; Camerer, 2008). The hope was to combine theoretical concepts and empirical data into a unified framework for modelling choice that might be able to reconcile the prescriptive and descriptive sides of economics (Glimcher et al, 2005). In this line, neuroeconomics studies have employed neuroscience techniques to test the premises of bounded rationality against the

cognitive processes involved in game behavior. This is the approach that we will present in the next section and which we employed for the experimental work performed during this PhD work.

## 2) The mind in the game

Cognitive neuroscience offers multiple tools to uncover the cognitive processes involved in decision-making during one-shot game play. Early on, neuroeconomists used fMRI to investigate what brain regions subserve equilibrium play. Batth & Camerer (2005) scanned participants while playing a series of (one-shot) dominance solvable games with another individual outside the scanner. At each trial the participants had to state their choice, their belief over the choice of the other player (first-order belief), and their belief about the belief the other player may hold about themselves (second-order belief). By simply contrasting the average BOLD signal change between each condition they showed that stating belief vs. choosing activated more prefrontal areas such as the ACC and dIPFC (along with the posterior cingulate cortex), while stating second- vs. first-order belief lead to more activity in the insula and the inferior frontal gyrus (IFG). By contrasting trials in which players played the NE, compared to out-of-equilibrium play, they suggested that the striatum was a key area for equilibrium play.

To investigate more precisely the brain areas involved in equilibrium play, Coricelli and Nagel (2009) took advantage of the continuous choice data that the p-beauty contest game offers to identify BOLD signal variations among different types of players (i.e. different levels of strategic sophistication). Participants played the (one-shot) game multiple times with different p values, against either another human or a computer. In line with the CH theory, their choices revealed different levels of players (0, random, L1, L2 or higher) in the human condition only. Using the heterogeneity of strategic sophistication levels observed in their population, they were able to identify brain regions in which BOLD signal was higher in the human vs. computer condition, and this for different types of players (**Fig.13.A**). They found that the mPFC, (rostral)ACC, posterior cingulate cortex and TPJ/STS, a network of brain areas known to be recruited in mentalizing (see section II.B), was more active when playing a game against a human than a computer. Moreover, high-level players presented a specifically higher activity in the mPFC and dIPFC compared to low-level players, and a higher mPFC and vmPFC activity when opposed to human vs. computer (**Fig.13.B**). Taking advantage of the parametric nature of their choice data, the authors showed at the subject-level that the activity of the mPFC linearly correlated with the propensity to play on average closer to the NE. (**Fig.13.C**).



**Figure 13. Level of strategic sophistication in the p-beauty contest game correlates with the activity of the mPFC.**

A) 26 choices of 2 (representative) participants for each parameter value  $M$  in the human (blue dots) and computer (triangles) conditions, separately. (Left) the choices of one participant representing a so-called low-level type. In both

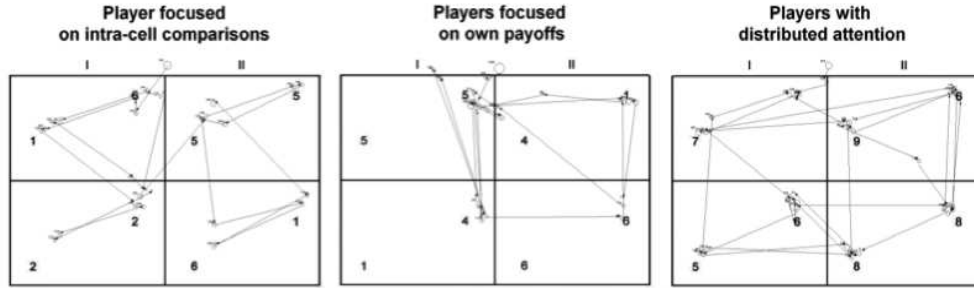
the computer condition (triangles) and the human condition (blue dots) she chose near the theoretical (CH Model) level 1 line (brown line with choices equal to  $50 \cdot M$ ). (Right) the choices of one high-level type participant. In the computer condition she chose near the theoretical level 1 line. In the human condition she chose near the theoretical level 2 line (blue line with choices equal to  $50 \cdot M_2$ ). Below is plotted the choice of the 2 participants for the computer and human conditions for M2/3. B) fMRI results. (adapted from Coricelli & Nagel, 2009)

---

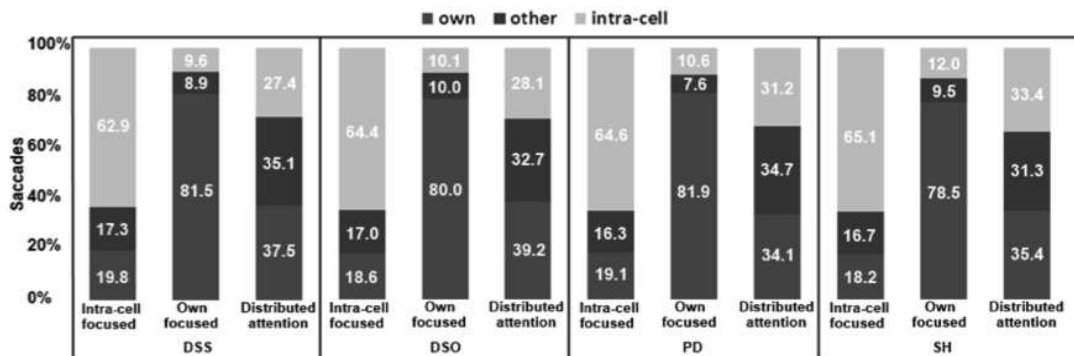
In another study, Batth et al (2010) used another way to capture the brain correlates of strategic sophistication. They asked participants to play a series of one-shot sender-receiver games (without feedback), in which they had to make selling proposals (price) to an anonymous counterpart who would in return propose a buying price. If the former price did not exceed the latter, the deal was accepted and the buyer received the difference. However, this information was not revealed to the participants. They classified their participants depending on the strategy they used in this game, they considered that a strategic player would make proposals negatively correlated with the actual value to trick (through deception) the seller, and would thus maximize their earnings. The authors found that the right (r)dIPFC and (r)TPJ were more active in the brain of the strategic players compared to the other types of players.

Another technique has been used to uncover the cognitive processes underlying equilibrium-play in one-shot games: eye-tracking. Polonio et al (2015) used different classes of 2x2 games from competitive to cooperative games, among which two types of dominant solvable (DS, one action dominates the other) games: a DS self (DSS), and DS other (DSO). To reach the NE in the DSO type of game, participants had to eliminate the dominated strategy (like in the prisoner's dilemma). In the DSS games however, doing so did not lead to the Nash. Instead, participants had to switch perspective and consider that the other player had a dominant solvable strategy that would lead her to a choice, to which the participants should best respond. The authors managed to identify three types of eye-movement patterns made by the participants when choosing in these (randomly ordered) games (without feedback), that revealed their information processing strategy (**Fig.14.A**). They showed that participants were not only consistent across games in the way they analyse the payoff matrices, but also that their eye-movement patterns correlated to their ability to reach the NE (**Fig.14.B**). For instance, participants who constantly paid attention to their own payoff only failed more in finding the NE in the DSO game compared to the ones who paid first attention to their payoffs, then to the payoffs of the other player and then compared them to their own ones (**Fig.14.C**). In a second study (Polonio & Coricelli, 2015), the authors showed that the participants classified as strategically sophisticated presented a high congruence between their patterns of eye movements (visual attention) and choices, and their stated beliefs about the other player's strategy.

A



B



C

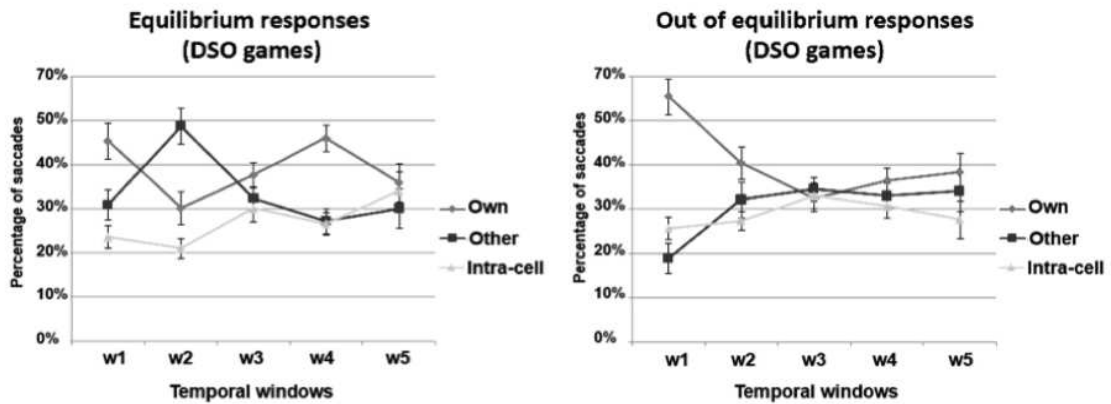


Figure 14. Equilibrium and non-equilibrium play in normal form games: an eye-tracking study.

A) Examples of analysis performed by three column players classified as players focused on intra-cell saccades, players focused on own payoffs and players with distributed attention. Lines indicate the saccades; circles, the fixation location. B) Proportion of each type of saccade for the three clusters (players who analyzed the games in similar ways) in four classes of games: Dominant Solvable Self (DSS), Dominant Solvable Other (DSO), Prisoner's Dilemma (PD), Stag Hunt (SH). C) Temporal pattern of visual analysis (mean and standard error) in DSO games for players having distributed attention (level 2), grouped by equilibrium responses (Panel A) and out of equilibrium

responses (Panel B). Each window ( $w$ ) is based on a sequence of four consecutive saccades. (adapted from Polonio et al, 2015 + early draft)

---

These results suggest that a specific cognitive process subserves the higher level of strategic reasoning in one-shot games; consistent with the hypothesis that L2 or higher levels imply recursivity (reasoning about reasoning others) and the fact that a strategic player considers the impact of his or her own behavior on the behavior of the others. It also shows that the observed departures from game optimality are consistent with the CH/level- $k$  models: participants with a higher level of strategic sophistication formed more accurate beliefs over the other player, leading them to choose closer to the NE prescription. Along the same line, recent results in psychology have suggested that humans at early age use common sense modules to root inferences, or beliefs over a conspecific's behavior. Jara-Ettinge et al (2016) proposed that humans quite automatically assume that the other would act to maximize her utility, by taking into account action costs and benefits. This "naive utility calculus" might be at the heart of the iterative process of strategic reasoning in one-shot games.

So far we focused on one-shot games. However, a game can be repeated, and feedback about the outcome of the choice can be provided once played. This is generally how our general social interactions are structured: in a dynamic fashion with knowledge of the outcome of our actions based on previously experienced similar interactions.

Behavioral game theory also studied how humans played in laboratory in repeated game interactions in which information about the past history of choices and outcome was revealed. In this situation, data usually display a convergence of aggregated choices towards MSNE distribution (Fudenberg & Levine, 2009). However, MSNE posits that not only should (rational) players' aggregated choices follow the prescribed distribution, but that they should also randomize over their action set. This randomization process is crucial since it allows one to not be easily predictable. Thus, in repeated game interactions, playing the MSNE strategy ensures that not only the expected payoff would be optimal in case the other also follows the theoretical prescription, but also, in accordance to the mutual best-response premise, that one's choices should not be exploited by the other since this would lead the other player to best respond differently and not to follow the equilibrium play anymore. And in fact, a second important result, systematically found in empirical studies of repeated games, is that players do not choose randomly from an independent and identically distributed distribution, their choice series often displaying an over-alternation bias (Camerer, 2003). These results pose the question on the predictive nature of the MSNE play.

A modern interpretation proposed by Camerer (Camerer, 2003) is that "players need not actually

randomize, as long as other players cannot guess what they will do”, concluding that a MNSE can be seen as an “equilibrium in beliefs”.

These results thus provide a hypothesis to the question, not tackled by the theory, of how an equilibrium might arise during repeated (strategic) interactions: learning.

Indeed, when the game is repeated and when choice feedback is provided, like most of our social interactions, learning becomes possible as space is given to update and adjust beliefs through predictions over the opponent’s behavior.

## **B) The Neuroeconomics of strategic learning**

In a published opinion paper untitled “The neuroeconomics of strategic interaction” (**Appendix I**), we drew a parallel between the bounded rationality presented in the previous section and the social learning models developed in cognitive neuroscience. In the following section we will extend this point.

To take into account the empirical dynamics of play observed in repeated games with feedback, economists first turned to psychology and proposed to implement the model of reinforcement learning. Erev and Roth famously reported an early work (Erev & Roth, 1998; Roth & Erev, 1995) where, in a variety of games, even a simple reinforcement model with only one parameter (controlling the determinism of the action selection<sup>10</sup> of the agent, and no learning rate parameter, i.e.  $\alpha=1$  in eq(1.1)), could approximate the directions of the subjects’ aggregate choices. This result led behavioral game theorists to the conclusion that humans could actually use the past experience in a repeated game to inform their subsequent decision. However, two criticisms were formulated to such a reinforcement learning framework applied to economics. First, the convergence of play displayed by the model was much slower than the one displayed by participants in the laboratory (Erev & Roth, 2014). Second, learning through reinforcement implies that only the outcomes obtained in the previous games are considered. Thus an important part of the information relative to the other player’s choices is omitted. In other words, by nature, a reinforcement learning model only adapts to the past own plays. It does not embody any inferential process and is thus not “strategic”.

---

<sup>10</sup> Usually a logistic function is used to model a stochastic action selection process. The function, also called softmax, transforms the subjective value of an action, relatively to the value associated to the rest of the action set (Luce, 1977), through an exploratory parameter ( $\beta$ , the inverse temperature) which regulates the sigmoid slope, and the amount of exploratory choices. A large  $\beta$  corresponds to almost deterministic choices (greedy strategy, the action with the highest value is selected), whereas a smaller  $\beta$  leads to noisier action selection and ultimately (when  $\beta=0$ ) to random choice.



A second class of learning models has thus been considered: the belief-based models. Such models posit that players form beliefs over the strategy profile of the other player through the observation of her past play, and best respond to it. This classic belief-based model is the “fictitious play” which computes at time  $t$  the frequency of each action played by the other player since  $t=0$ , and then best responds to it (Brown, 1951). This model has been relaxed in two ways; First, by introducing noise in the strict best response to beliefs (cautious fictitious play) making an agent that has converged towards NE to mimic a QRE model -- Cheung & Friedman (1997) for instance replaced the deterministic action selection rule by a logistic function; Second, and more significantly, by relaxing the assumption of perfect memory. Indeed, the weighted variation of the fictitious play (Cheung & Friedman, 1997; Fudenberg & Levine, 1998) incorporates a decay parameter which weights more the recent play in the computation of the beliefs over the other strategy profile. The probability  $P$  that the other player plays action  $A$  is computed at each game (trial  $t$ ) through:

$$P_A(t+1) = (C_A(t) + \sum_{x=1}^{t-1} (\eta^x \times C_A(t-x))) / (1 + \sum_{x=1}^{t-1} \eta^x) \quad (1.3)$$

This probability thus corresponds to the weighted frequency of the action  $A$  selected by the other player in the past (at each trial,  $C$  is 1 when it was the chosen action, 0 otherwise). The parameter  $\eta$  controls for the slope of the decay, or in other words the size of the memory, so that the model is a classic fictitious play with infinite memory (all past actions are considered the same way) when  $\eta = 1$  and a cournot adjustment model (only considering the last play) when  $\eta=0$ . In Belief-based models, beliefs thus represent a probability distribution over the action-set of the other player.

This model has been shown to converge towards NE. However, depending on the type of model used and the type of games played, it did not appear clear to economists that a belief-based model (essentially with full memory) was more suited to capture human choice behavior than a reinforcement model (Battalio et al, 2001). The study by Nyarko & Schotter (2002), showing that belief-based models could not capture all the variance observed in the beliefs directly stated by the participants, convinced behavioral game-theorists that a hybrid model might be more appropriate.

Camerer and Ho (1999) developed the experience weight attraction model (EWA) with the goal to merge the two approaches into one single learning model that would weight the relative influence of the beliefs and the reinforcements in the human’s choice behavior. Essentially for a 2x2 game (as in Fig.10), the EWA can be simply represented as following: the attraction value  $A_{p1}^c$  is the expected payoff of the action  $c$  that is chosen at trial  $t$  by player 1:

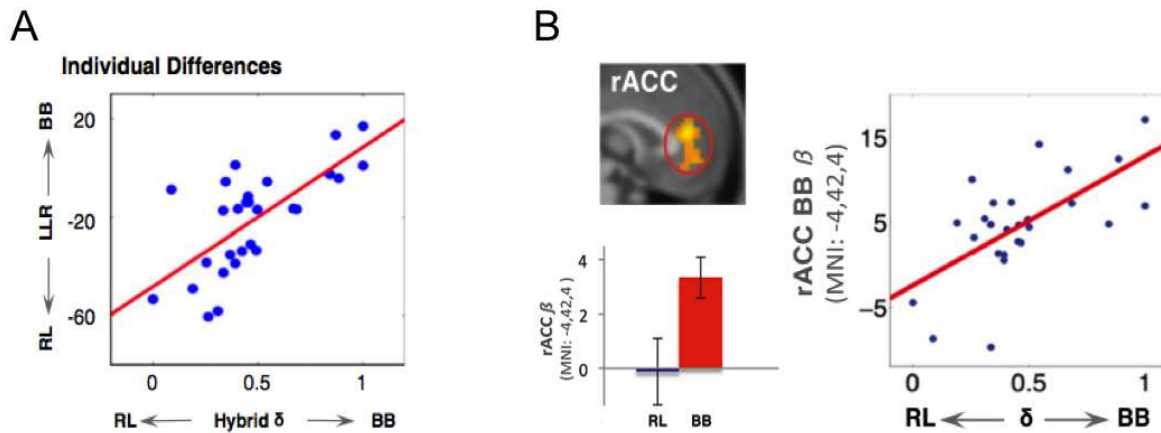
$$A_{p1}^{ac}(t+1) = ( ((\varphi \times N(t) \times A_{p1}^a(t-1)) + R_a(t)) / N(t-1) \quad (1.4)$$

While the action  $u$ , left unchosen is also updated counterfactually:

$$A_{p1}^u(t+1) = ((\varphi \times N(t) \times A_{p1}^u(t-1)) + (\delta \times R_u(t))) / N(t-1) \quad (1.5)$$

In this model,  $N(t)$  represents the weight placed on previous experience, and is updated as  $N(t) = (\rho \times N(t-1)) + 1$  ( $\rho$  acts as a decay parameter). The parameter  $\varphi$  represents the belief about the speed of adaptation of the opponent: a small  $\varphi$  means that the agent believes that her opponent depreciates past values faster. Thus  $\varphi$  acts as a learning rate. And  $\delta$ , the main parameter of the model, is the weight between foregone payoffs and actual payoffs when updating attraction values. The model thus reduces to an RL model when  $\delta = 0$ , and expands to a belief learning model (weighted fictitious play) when  $\delta = 1$ . The key insight of this hybrid model is to consider belief learning as equivalent to a mode whereby actions are reinforced by foregone payoffs in addition to received payoffs as in (model-free) reinforcement learning. For the authors, this parameter can be considered as an inclination towards beliefs or, in their own words, as a “simulation” of outcomes under alternative competitive scenarios (i.e. “counterfactual thinking” in the terminology of psychology”).

Empirically, this hybrid model has been proven successful in capturing human choice behavior in games (Camerer et al, 2002). The question posed again by the neuroeconomists is how such a model matches the cognitive processes involved during learning in a repeated game. Zhu et al (2012) investigated this question directly using fMRI. The authors made participants interact with another human in a competitive 4x5 game (the patent race). They fitted the learning model to each participant’s individual choices to find the combination of parameters that ensures the EWA to capture the best their choice series. The authors found that, at the (sampled) population level, the hybrid model fitted better the participants’ choices than alternative versions of the model reduced to either RL or Belief-based (BB) (**Fig.15.A**). Moreover, they showed that the reward prediction errors values (from their reformulation of the EWA equations) correlated for both RL and BB to the BOLD signal in the striatum. But more crucially, their fMRI analysis revealed that the  $\delta$  parameter positively correlated to the activity of the dmPFC (rACC, similarly to Coricelli & Nagel, 2009), therefore suggesting that this brain region is implicated in the computation of beliefs during repeated strategic interactions (**Fig.15.B**).



**Figure 15. Belief-based learning in the brain**

A) Individual variation in the relative weights placed on RL and belief learning can be captured by using parameter  $\delta$  of the hybrid model (EWA). As  $\delta$  increases, behavioral fit of belief learning improves relative to that of the RL. B) Left panel: Neural activity in the rACC is correlated only with belief and not with RL prediction error, error bars indicate SEM. Right panel: Between-subject neural response to the belief prediction error in rACC is correlated with individual differences in behavioral engagement of belief learning. (adapted from Zhu et al, 2012)

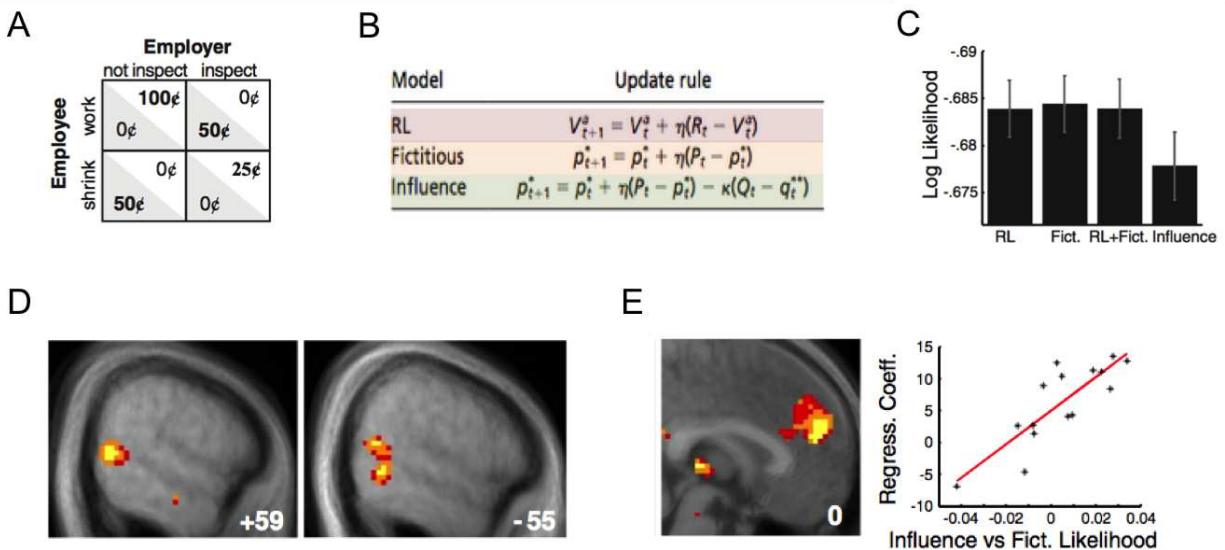
Several studies have replicated this result in non-human primates playing competitive games against a computerized algorithm which varied in their level of strategic sophistication. In a series of in-depth work, Lee et al managed to show that the activity of the neurons recorded directly in the monkey's mPFC correlated with the computation of beliefs during repeated interactions. They notably showed (Abe & Lee, 2011) that when confronted to a belief-based algorithm similar to a fictitious play in a rock-scissor-paper game, the animal's choice behavior deviated from model-free reinforcement learning to take into account the forgone outcomes (reward that would have been obtained if chose otherwise) in their learning process. Their study showed that more neurons in the dIPFC (compared to OFC) computed the hypothetical expected values. In a more recent study (Seo et al, 2014), the same team provided strong evidence that when faced to such a belief-based algorithm, exploiting statistical biases in their choices, monkeys managed to deviate from RL and to engage in sophisticated learning to override their computerized opponent, playing on average the frequency prescribed by the MSNE (and maximizing their earning even more than predicted by NE play). Moreover, the authors showed that more neurons in dIPFC (compared to other areas like dIPFC, ACC or striatum) encoded specific switching patterns revealing the implementation of strategic learning. Such an ability to refine beliefs by exploiting statistical regularities in the opponent's behavior during competitive repeated interactions has been recently

proposed in behavioral game theory. Spiliopoulos (2012, 2013) developed an extension of the weighted fictitious play that takes into account deterministic patterns in the opponent's choice series that might emerge from randomization failure. Instead of estimating the probability of choice from the computation of the (decayed) frequency of each past action, the pattern-based fictitious model computes joint probabilities over each action based on the combination the two and even three observed past actions. Using empirical data on repeated competitive games, Spiliopoulos showed that the pattern version of the fictitious model was more successful at capturing participants choices than the original non-pattern one.

Together these results strongly suggest that humans (and non-human primates) can form beliefs over the opponent's behavior in strategic repeated games, and that they can use these beliefs to improve their learning and best respond in a competitive interaction. Moreover, the computation of such probabilistic beliefs seems to take place in the mPFC, a region previously implicated in action-outcome representations of the other behavior (Suzuki et al, 2012) (section II.B,C). However, as suggested by Seo et al (2014), forming accurate beliefs about the behavior of another human requires to not only engage in adaptive learning, but also to infer strategic beliefs from the other play in an iterative fashion. Indeed, in competitive interaction learning the action-outcome contingencies in another player requires to take into account the strategic nature of the interaction. A strategic learner should form beliefs that take into consideration that her own choices can influence the behavior of the opponent, which can as well form beliefs over her own behavior.

Hampton et al (2008) tested this hypothesis of higher-order belief learning, by extending a fictitious model so that it considers that the other player is also using a belief-based learning strategy. Their model would first compute the probability of their own actions based on their past (decayed) frequency of choice, and then incorporate this simulated belief in its computation of the probability of choice of such (fictitious) agent (**Fig.16.B**). They showed that on average this *Influence* model fitted better the behavior of participants engaged in a 2x2 (competitive) repeated game against another human, compared to a weighted fictitious, a RL or a hybrid model (EWA) (**Fig.16.C**). Using fMRI, the authors showed that the BOLD signal in the mPFC correlated more to the expected reward value computed by the influence model compared to the two others. The key parameter of their Influence model is the parameter that represents how much the simulated belief of the opponent is incorporated in the participants' learning process. This parameter  $\lambda$  thus corresponds to how much the influence of their past choices on the other player's behavior is considered in their own learning process. In other words, it captures their belief that the opponent is actually following a belief-based strategy. The TPJ/STS was the brain area in which the change in BOLD signal was correlated to this parameter (**Fig.16.D**). They found that the dmPFC was more active on average for subjects (relatively) better fitted by the Influence model (**Fig.16.E**). They thus hypothesized and then experimentally confirmed that the (influence) teaching signal encoded in the

TPJ/STS would correlate more with the dmPFC (along with the striatum found to be correlated to influence RPE), at the time of the outcome (learning update). These results shed light on the role of the so-called mentalizing network during competitive interactions through its implication in the strategic learning process.



**Figure 16. Strategic learning in the brain**

A) Payoff matrix of the 2x2 (Inspection) game. B) The RL model updates the value of the chosen action ( $V_{t+1}^a$ ) with a reward prediction error as the difference between received rewards and expected rewards ( $V_t^a$ ), where  $\eta$  is the learning rate. The fictitious play model instead updates the probability of the opponent's action with an (action) prediction error between the opponent's action ( $P_t^*$ ) and expected strategy ( $p_t^*$ ). The influence model extends this approach by also including the influence that a player's own action ( $Q_t$ ) has on the opponent's strategy. C) Model comparison shows that the influence model, which incorporates the effects of players' actions influencing their opponents, has a better fit to subjects' behavior than either the RL or fictitious models or these two models combined (EWA). D) At the time of outcome, the influence update of the inferred opponent's strategy shows significant correlations with BOLD signal change in the TPJ/STS. E) The activity of the mPFC correlates with the (individual) difference in fit between the Influence and the Fictitious model, capturing the degree to which a subject believes her actions are influencing her opponent's choice behavior. (adapted from Hampton et al, 2008)

In a recent study Hill et al (2017) went a step further to causally test the role of the rTPJ in the computation of the influence teaching signal in strategic learning. Using a (non-invasive) brain stimulation technique (TMS) they were able to first inhibit the activity of the right TPJ in participants who then played the Hampton's 2x2 game (Fig.16.A) while being scanned. They first replicated the original study showing

a better fit of the Influence model compared to BB or RL at the (sampled) population scale. They also replicated the fMRI results showing a correlation between the BOLD signal in the dmPFC and the (relative) fit of the Influence model. But more crucially, they managed to show evidence for a causal implication of the rTPJ in the computation of the Influence teaching signal.

These results thus suggest that the TPJ, a core brain area of the mentalizing network, along with the mPFC, subserves higher-order belief formation by encoding the consideration in the learning process of the influence of a strategic individual's behavior on her opponent. The mPFC has been extensively implicated in belief formation in social learning (Apps et al, 2016), and its activity has been shown to correlate with the level of strategic sophistication required to reach (initial) equilibrium play in one-shot games (Coricelli & Nagel, 2009). This suggests that common computations might be required to form higher-order beliefs during strategic decision-making (Griessinger & Coricelli, 2015).

Taken together with the results presented in the two previous sections of this introduction, evidence points towards similar cognitive mechanisms involved in the representation of action-outcome contingencies in a non-social probabilistic environment and in the behavior of another person during dynamic social interactions. The formation of beliefs over someone else's behavior appears to drive choices in a model-based reinforcement fashion. However, a crucial difference arises when the strategic nature of the (repeated) game interaction is taken into account. The model of the behavior of the opponent needs to be fed with computations related to the influence of one's own behavior, leading to the formation of higher-order beliefs and recruiting mentalizing areas specifically implicated in social setting, such as the TPJ.

Nevertheless, it remains unclear if the formation of such higher-order beliefs leads to more accurate choices, and how this cognitive process relates to equilibrium play in real human-human interactions. Indeed none of the studies mentioned above investigating strategic (or belief-based) learning during human competitive interaction reported how this learning process subserved equilibrium play (see also Devaine et al, 2014). Hill et al mentioned that participants with a higher influence (best) parameter increase their earnings. However, no evidence was reported of an effect of the modulation of the TPJ activity on performance. As highlighted by Zaki & Ochsner (2011), the question of accuracy when it comes to social interaction and mentalizing is usually overlooked. However, the question of belief accuracy is at the heart of the (behavior) game-theoretical preoccupations.

Another related but also overlooked characteristic of the research on strategic interaction is the inter-individual variability. The CH/level-k models posit that players might differ in the accuracy of their beliefs, and correlational studies have suggested that engaging in such a strategically-sophisticated behavior during one-shot games is computationally costly.

Previous researches in computational neuroscience do indeed suggest that humans vary in their ability to engage in model-based learning, and the research in neuroeconomics presented here show important variance in belief-based (**Fig.15**) as well as higher-order strategic learning (**Fig.16**). The question thus remains open of what drives such a high heterogeneity in strategic learning.

#### IV- Synthesis and working hypothesis of the present thesis

We have seen that the value-based approach initiated in economics and further developed by the growing field of neuroeconomics has been proven very useful to understand the cognitive mechanisms at play during decision-making. Along this line, the framework of reinforcement learning, formalized in machine learning, has provided neuroscientists with an important toolbox to identify the computations performed by the brain during repeated interactions with an uncertain environment. It has also fed cognitive scientists with a major insight about cognition: the concept of prediction error. This signal, which has been found to be encoded by dedicated structures of the brain, reflects the discrepancy between the expected outcome of an action and the outcome actually experienced once realized, and drives behavioral adaptation (i.e. learning) towards (subjective) value maximization. Recent studies in the field suggest that different learning systems interact on top in order to adjust action selection in the unknown. From probabilistic learning to the use of heuristics, humans seem capable of forming beliefs over the action-outcome contingencies constituting their choice environment and implement these representations in a common value-based frame to guide their decision-making process through reinforcement. Moreover, this computational apparatus has been shown to be transposable to the social realm.

The last decade of research in neuroscience has shown that social information computed in dedicated areas (such as the TPJ) are integrated, voluntarily or not, in the same decision-making process, recruiting similar brain circuits as in non-social environments. But even more importantly, a growing amount of studies now suggest that humans (and non-human primates) can even use prediction error signals to learn action-outcome contingencies in the behavior of a conspecific, and use such a representation of the other's intentions to adjust their own choices through reinforcement. Patterns of activation in the medial prefrontal cortex (mPFC) suggest that this area plays an important role in the computation of the prediction error signals which subserves the update of beliefs over a non-social probabilistic environment as well as the behavior of another person.

Our ability to infer intentions and form beliefs over the mental state of other individuals, broadly labeled “mentalizing” in psychology, might thus relate on prediction and involve shared learning mechanisms across domains. Nevertheless, interacting with another person implicates that our own actions might also affect her behavior, like in strategic interactions, and that the action-outcome contingencies might thus depend on the outcome of our own decision-making process. Therefore the beliefs one forms over the contingencies in the choice behavior of another individual might as well incorporate information about the influence of our own choices. Game theory provides solution concepts for optimal (subjective value maximizing) choices in strategic interactions, modeled as games. (Mixed Strategy) Nash Equilibrium (MSNE) for instance prescribes a specific probability distribution over the actions available in a (dyadic) strategic interaction with another person. However, MSNE relies on the assumption that both individuals (players) best respond to accurate beliefs over the other player (mutual best response). When the game is not repeated, game-theorists have proposed that forming accurate beliefs over the other player might require to engage in iterative thinking, or sophisticated reasoning. However, when the game is repeated and when this information from previous play can be used to inform the subsequent choice, several learning models have been proposed in the behavioral game theory literature, from reinforcement to probabilistic learning. The latter type of model, labeled belief-based learning as it computes the beliefs over the action-outcome contingencies in the behavior of the other, has been found to also recruit the mPFC area. Still, these models take into account neither the strategic nature of the interaction nor the interplay between the two players’ choices. Recently, higher-order beliefs models have been developed, in which the influence of one’s choice is taken into consideration in the learning of the action-outcome contingencies of the other’s behavior. Using a simple Influence model which computes a second-order action prediction error, recent neuroeconomics experiments have shown that the TPJ, a brain area found to encode specific socially-related signals like perspective-taking, encodes such a teaching signal to update (high-order) beliefs over the opponent’s behavior in a competitive game. This type of higher-order inference model can be seen as a dynamic (learning) equivalent to the iterative thinking models proposed in one-shot games, as it might allow a more accurate prediction of the other’s behavior.

On the one hand, heterogeneity in the departure from MSNE play has been systematically observed in behavioral game theory in both one-shot and repeated games. On the other hand, the higher order learning models developed recently capture average behavior quite effectively but still, important variance is reported in individual learning. However, to our knowledge, no study has yet reported how the dynamic computation of higher order beliefs driving learning in repeated game interactions relate to game-theoretical prescriptions.

In this thesis, we will first present a behavioral investigation of the interplay between higher-order belief learning (strategic learning) and optimal play by taking advantage of the heterogeneity observed in human



behavior during repeated (competitive) strategic interactions. This study, composed of three experiments, is presented in chapter II. We then investigated whether an individual's level of strategic learning engagement during a strategic game is affected by the level of the previously encountered player in the same game. This study is presented in chapter III. Finally, we investigated the possibility that statistical redundancies in the other's choice behavior (i.e. choice patterns) can be exploited by humans in order to improve the accuracy of their beliefs. This study, divided in two experiments, is presented in chapter IV. In the last chapter (V), we will discuss the implications of these studies for the field of neuroeconomics but also game theory, and draw a general conclusion of the work conducted in this thesis.

- Chapter II -

**The interplay of learning sophistication and  
strategic asymmetry in social competitive interactions**

**(Exp. 1, 2, 3)**

I - The interplay of learning sophistication and strategic asymmetry in social competitive interactions. [Griessinger T., Khamassi K. and Coricelli G. (in prep.)]

## The interplay of learning sophistication and strategic asymmetry in social competitive interactions

This section is adapted from:

Griessinger Thibaud, Khamassi Mehdi\*, Coricelli Giorgio\*. The interplay of learning sophistication and strategic asymmetry. About to be submitted. [\*the two authors equally contributed to this work]

### A) Introduction

Inferring someone's' intention is key to adjust our behavior and maximize the outcome of a social interaction (Schaafsma et al, 2015 [1]). It enables to establish shared action plans and efficient motor coordination in cases of cooperation (Pacherie & Khamassi, in press [2]). It also enables anticipation of an opponent's actions in cases of competition such as in strategic games.

Recently, emphasis has been placed on one particular feature of this mind reading ability: using the past experience to predict the near-future behavior of a conspecific (Koster-Hale & Saxe, 2013 [3]). Strategic interactions during competitive games have been proved a useful experimental paradigm to capture the behavioral dynamics revealing such theory of mind in human and non-human primates (Lee, 2008 [4]), as they consist in social situations where one's choice outcome critically depends upon the action of the other.

Game-theory provides formal solutions to strategic interactions, modelled as games, through the concept of Nash Equilibrium (NE) and its refinements (Nash, 1950 [5]). The so-called Mixed Strategy Nash Equilibrium (MSNE), for instance, prescribes a probability distribution over possible actions that ensures to each involved agent that they would have no incentive to deviate if they all follow it. In practice, however, humans typically deviate from this solution concept (Camerer, 2003 [6]), and when the aggregated choices seem to fit the theoretical prescription, slightly changing the payoff structure of the game might lead to a strong departure from MSNE (Goeree & Holt, 2001 [7]). Nevertheless, patterns of aggregated choices somehow appear to converge towards MSNE when a game is repeated (Fudenberg & Levine, 2009 [8]). Following MSNE requires for each subject to show some level of randomness in their behavior. This enables subjects in practice to approximately stick to the action probability distribution prescribed by the MSNE without displaying a trivial repetitive behavior which would have entailed the risk

to be detected/predictable by the opponent. This empirical finding is surprising as humans have been systematically proved to be bad randomizers (Gauvrit et al, 2017 [9]). Indeed serial dependency in between actions is usually observed (Shachat, 2002 [10]), leading authors to suggest that learning might lead to MSNE (Erev & Roth, 1998 [11]; Fudenberg & Levine, 1998 [12]). However, the way human subjects progressively learn to reach MSNE is still little understood.

One hypothesis, which encounters growing support, lies on the idea that convergence towards MSNE distribution requires that both players, who aim to maximize their earnings, should hold correct beliefs over their opponent's behavior and best-respond to it (Camerer et al, 2004 [13]). According to this hypothesis the ability to do so must be somehow constrained, leading to sub-optimal behavior (Barros, 2010 [14]).

In case of non-repeated interactions, like in one-shot games, the ability to form accurate beliefs lies on the capacity to engage sufficiently in iterative thinking. However, important variability in the population has been observed (Crawford et al, 2013 [15]), which has been recently linked to specific differences in the computation of the information relative to the other's behavior (Polonio et al, 2015 [16]). When the game is repeated and choice feedback is provided, like most of our social interactions (Schilbach, 2013 [17]), learning becomes possible as space is given to update and adjust beliefs through predictions over the opponent's behavior. Such facilitative effects of game repetition on MSNE convergence has been reported empirically (Fudenberg & Levine, 2009 [7]). Humans are known to be able to track intentions in others (Dennett, 1987 [18]), however one might wonder if actually we are able to engage in sufficiently sophisticated learning to form accurate beliefs about an opponent's behavior.

Research in cognitive neuroscience suggests that, in probabilistic tasks, humans learn to adjust their decision based on expected values computed from previously experienced outcomes (model-free reinforcement learning, RL) but also through the incorporation of (probabilistic) beliefs over the action-outcome contingencies underlying the structure of such environment (model-based RL) (Doll et al, 2012 [19], 2015 [20]). Indeed in such tasks, reward convey at least two types of information, often correlated: the affective (hedonic) value embodied in the monetary reward and the information (predictive) value about the architecture of the world (O'Doherty, 2014 [21]). The ability to use the latter to maintain and update, using prediction accuracy, a mental representation of the choice environment has been found to encompass social interactions as well (Joiner, et al, 2017 [22]; Ruff & Fehr 2014 [23]). In fact, by pairing the normative framework of game theory to the computational approach in neurosciences, recent research has shown that during repeated games, humans can engage in model-based learning using an iterative computation of the strategic information provided by the interaction (Devaine et al 2014 [24]; Hampton et al, 2008 [25]). Such strategically sophisticated computations drive higher order beliefs that incorporate the level of influence of one's past actions on her opponent's choice behavior, thus allowing for more accurate predictions (Griessinger & Coricelli, 2015 [26]). Such hierarchy of belief computation

has been observed at the brain level, with common brain structures implicated in model-based (non-social) and belief-based (social) learning computations (medial prefrontal cortex) (Lee & Seo, 2016 [27]), and higher-order belief (strategic) learning incorporating signals from areas involved in theory of mind (temporo parietal junction) (Hill et al, 2017 [28]).

Crucially, all these studies reported important heterogeneity in the level of engagement in belief-based learning, linked to variation in overall performance. However, none of the previous studies directly investigated the relationship between the human's ability to engage in strategic learning and the observed deviation from best response distribution, and ultimately MSNE. Taken together these results yet suggest that the human's propensity to follow optimality prescription from game-theory requires to disengage from reward-oriented, model-free, learning and fully engage in belief-based learning.

We hypothesized that, depending on how the reward structure interact with the MSNE prescription in a repeated strategic game, human performance in the game may be differently affected so that it does not necessarily reflect an individual's general level or ability of strategic learning. Previous studies suggest that the amplitude of the payoffs interferes with the propensity to follow the MSNE (Goeree & Holt, 2001 [7]), and that the symmetric nature of a game might facilitate the belief formation over the opponent's behavior through perspective taking (Beckenkamp et al 2007 [29]; Feldman et al 2010 [30]).

We developed a novel 2x2 competitive game setting, asymmetric in payoff structure but symmetric in payoff amplitude and expected payoff, so that the two players would earn the same if they both follow the MSNE distribution. The payoff matrix was however designed to lead to *strategic asymmetry* where one player's highest rewarded action would happen to be, at the informational level, the one the MSNE prescribes to choose the most (advantageous role), while for the other player the attractive action (focal point) would be the one she should choose the less (disadvantageous position). If following the optimal distribution of choice is conditioned on the ability to consider the strategic structure beyond the payoffs value to engage in belief learning, then our game should lead to strategic asymmetry.

We made the secondary hypothesis that in the repeated version of our stage game, humans with different individual strategic learning level (SL) would differ in their capacity to overcome this asymmetry and lead to observable differences in the final earnings between the advantageous and disadvantageous roles in the game.

We ran 2 distinct experiments with the same game setting: In the first one human subjects play against each other, while in the second we specifically manipulate the level of subjects' computerized opponent. Beforehand, we simulated agents interacting repeatedly through our competitive game, all modeled as simple learning algorithms varying from reward-based to sophisticated belief-based computations (Hampton et al, 2008 [25]) and developed to capture different levels of strategic learning sophistication

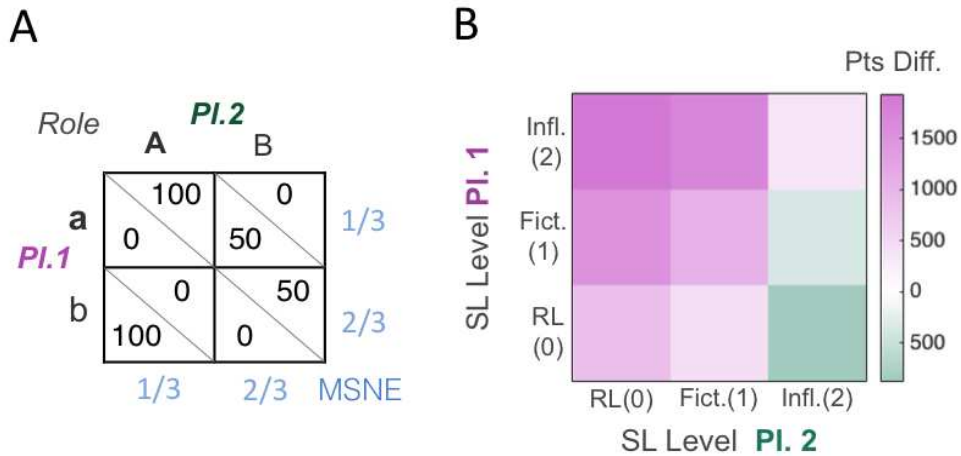
(SL). As anticipated we show that, at the population level, the game payoff matrix lead to a strong strategic asymmetry, such that the agents playing in the disadvantaged position see their loss reduced only when they engage in higher SL level than their opponent.

The first, unconstrained experiment allowed us to observe behaviorally this strategic asymmetry. Since subjects in each role did not differ in their SL level *per se*, we were able to show that the strategic asymmetry of repeated game was indeed causing the observed difference in total outcome, and that, as hypothesized, the observed deviation from game optimality (MSNE) is mainly driven by the individual propensity to depart from reward-based learning and engage in sophisticated belief-learning. These results were refined and replicated in the second experiment, showing that individuals in the disadvantageous position were driven by loss reduction and constrained by their own SL learning capacity, while subjects in the advantageous position were mainly adjusting their best response to the estimated behavior of their opponent. Importantly for our computational hypothesis, behavioral results from both experiments matched the predictions made in the simulation. Strikingly, only subjects endorsing the disadvantageous role (hence pressed towards their own limit) showed a SL level which was stable across opponent, as if the reduction of the strategic asymmetry was cognitively bounded (Friedenberg et al, 2016 [31]). These findings thus provide a possible explanation for the discrepancy between previous studies in which no correlation between SL level in strategic games and cognitive abilities was observed (Devaine et al, 2014 [24]).

## B) Exp.1: Model simulation and prediction

The game is a two by two (two players, two actions) (payoffs) asymmetric game, with a unique Mixed strategy equilibrium (**Fig. 1.A**). The expected payoffs at the mixed strategy Nash equilibrium are the same for both players.

To make predictions about the effect of the strategic asymmetry of our payoff matrix on subjects' behavior we simulated different computerized agents playing a repeated version of our game in each of the 2 roles, with different levels of strategic sophistication. Such simulation analysis allows us to not only test the robustness of our design but also to make precise predictions regarding the effect the individual level of strategic learning would have on the dynamics of play of humans interacting through this experimental setting (see Supplementary Information for details on the computational modelling and the simulation analysis).



**Figure 1.** Strategic characteristic of the game: simulation of play between 2 agents varying in their Strategic Learning level (SL) shows strong asymmetry in total earnings between the 2 roles.

(A) Payoff matrix of the repeated game, in points. In light blue the action probabilities prescribed by the mixed strategy equilibrium (MSNE). (B) Each agent modeled by either one of the 3 models of increasing strategic complexity, or SL level (i.e. Level 0: Q-Learning, 1: Fictitious play and 2: the Influence model) played the game in one of the 2 role. Every Player1-Player2 model combination was simulated 100 times playing against each other the 100 repetitions of the game. Agents endorsing the role of Player 1 won more points on average than Players 2. In fact Player 2 agents won more that their opponent only in the situation where they were playing SL Level 2 and Player 2 agent a level below or more.

To mimic inter-individual variation of strategic learning we used 3 computational models varying in their level of strategic sophistication (SL): a simple reinforcement algorithm learning only from the outcomes obtained through its past choices; a fictitious play best responding to the probability of each opponent's choice computed from its history of actions; and an Influence model, i.e. a 2nd order fictitious taking into account the influence of its own past choices in the computation of the opponent's probability of play (Hampton et al, 2008 [25]). Each simulation consisted in 2 computerized agents, endorsing one of the 2 roles and modeled by one of the 3 models, playing against each other during 100 repetitions of the stage game. Our simulation results reveal an important advantage of Players 1 over Players 2 in our game. Not only agents playing as Player 1 performed better than Players 2 in the game, but Players 2 had to be consistently higher SL level than their opponent in order to win more points (**Fig. 1.B**). To insure that this game propriety does not depend on the tuning of our simulations we replicated the simulation analysis with different proxies of the SL level such as the parameter  $\lambda$  of the Influence model. This parameter captures the weight of the second order fictitious update in the computation of the opponent's action probability. Our additional simulation analyses systematically show that the only way for Players 2 to

outperform their opponent is to engage in a higher level of Strategic Learning (**Supplementary Figure Fig. S1**).

Altogether this preliminary analysis confirms the strategic asymmetry of our payoff matrix, revealing the strong advantage of player 1 over player 2 in the sub-optimal domain. This setting allows us to clearly test how the individual level of strategic sophistication is affected by the strategic asymmetry of the repeated competitive interaction.

## C) Exp.2: Human against human

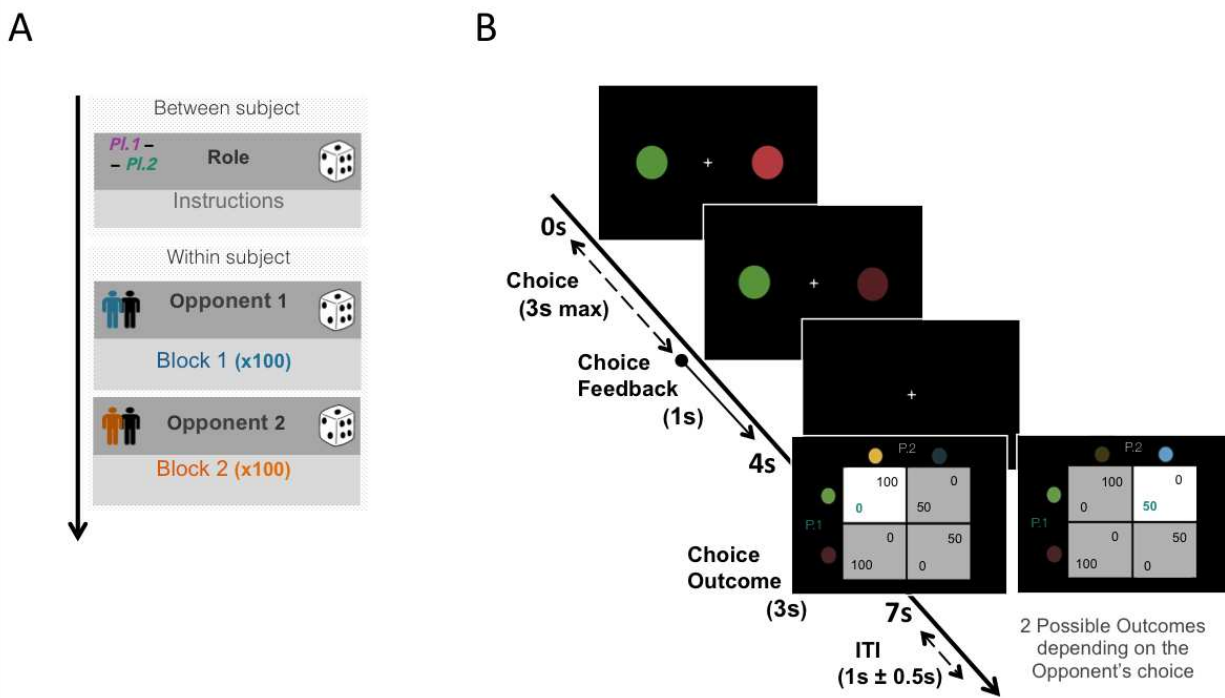
### 1) Material and Methods

#### **a) Population**

64 participants (29 male, 35 female; ages  $27.1 \pm 9.4$ ) took part in the experiment. They were students at Lyon University, who had previously joined the recruitment system on a voluntary basis. These volunteers gave written informed consent for the project which was approved by the French National Ethical Committee. All participants were right-handed, medication-free, with normal eyesight, no history of neurological disorders.

#### **b) Experimental Design**





**Figure 2.** Experimental design of experiment 1. (A) We manipulated 2 variables: the role (within subject level), and the opponent (between subject level). At start subjects were randomly assigned to one of the two roles in the game (player 1 or 2). After being instructed, they were randomly paired to an anonymous counterpart during a block of 100 repetitions of the game, and to a different counterpart during the second trial block. (B) Trial Structure: At each trial the two game actions, represented by the randomly assigned colors, were presented for 3s to each player. The choice was made by pressing the corresponding button (left or right). 4s after the trial onset, both players were simultaneously provided with the outcome feedback of their choice for 3s (the cell matrix matching to the 2 players choices was highlighted and the points won displayed in turquoise).

The first experiment consisted in a repeated interaction against another anonymized participant. One of the 2 roles was randomly attributed to each participant at the beginning of the experiment. Each subject interacted with two different human opponents, one after the other in two trial blocks of 100 repetitions of the stage game with complete choice feedback (**Fig. 2.A**). These two opponents were randomly selected among the participants assigned the opposite role to the subject. Points earned at each trial were accumulated through each block and summed up to determine their final payoff which would ultimately be converted to euros according to a predetermined rule. Each subject was initially instructed of the 2 stimuli representing her two available actions in the task, the payoff structure of the game, and trained to learn the stimulus-outcome contingencies of the payoff matrix. Each action was made of a different colored

circle randomly picked from 4 possible colors (all controlled for luminance). The 4 colors were randomly assigned to each pair of subjects in the first block, and kept unchanged in the second interaction block (thus constraining the re-matching random procedure in the second block). At each trial, both subjects had 3 seconds to select one of the 2 colors displayed at the left and right of the screen (randomized order across trials), the chosen one was highlighted for 1 second as choice feedback. 4s after the trial onset, both players were simultaneously provided with the outcome feedback of their choice and the one of their opponent for 3s. The outcome feedback screen consisted in the payoff matrix (note that depending of the role endorsed in the game the matrix was flipped so that subjects were always presented as row player), with the cell corresponding to the matching of the 2 players choices highlighted and the points won by the subject displayed in turquoise (**Fig. 2.B**). This display ensured minimal framing effect, while controlling for participants' awareness of the underlying payoff structure of the game.

We also provided to the subjects an additional task which consisted of a series of four different types of 2x2 static (one-shot) games (Polonio et al, 2015 [16]). The goal was to test the endogeneous hypothesis of strategic learning sophistication developed in Griessinger & Coricelli, 2015 [26]. We hypothesized that participants with a SL level in the repeated game (captured by our computational approach from the game behavior in the main task) would also display a higher strategic reasoning (SR, expressed as their capacity to conform to equilibrium play when a game is not repeated and no feedback is provided). All the subjects came back a second time to the lab a week later to complete a series of cognitive tasks. Both the additional experiment and cognitive tasks are detailed in Supplementary Information.

### **c) Computational modeling**

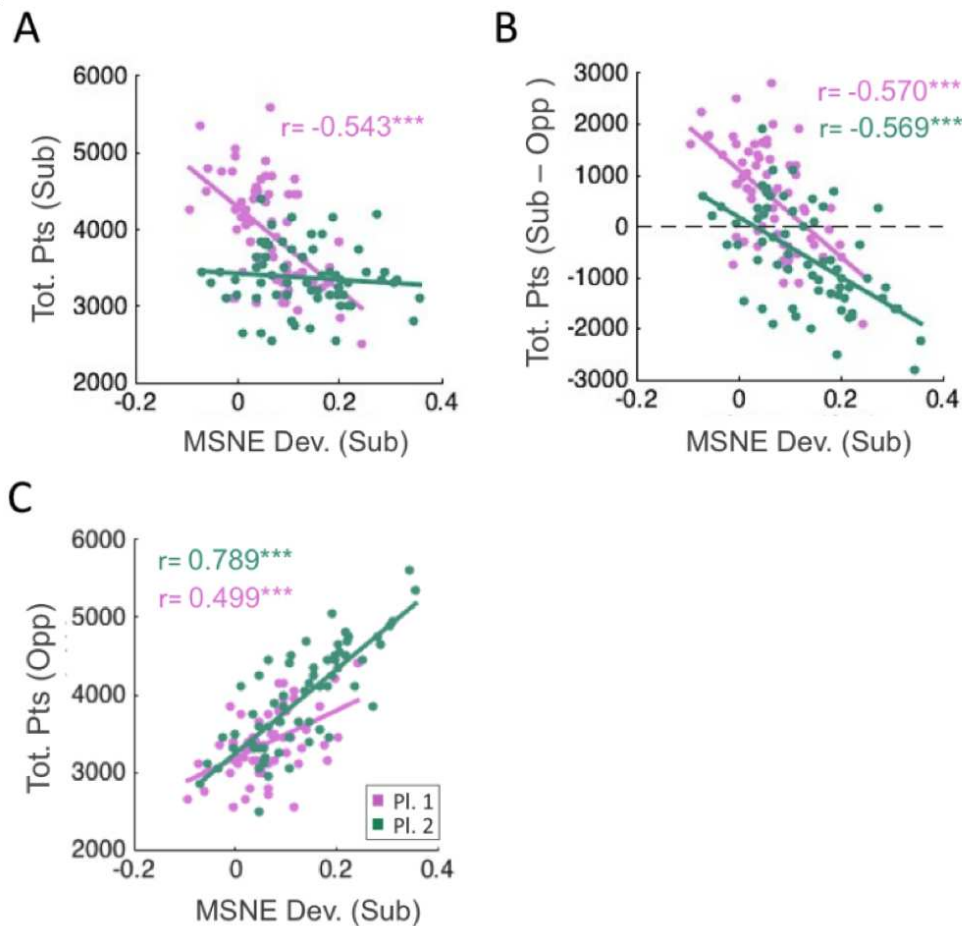
To capture the level of strategic learning of the subjects we first used the computational approach introduced in the preliminary simulation analysis: 3 computational models, corresponding to 3 different levels of strategic sophistication, were fitted individually to each choice series from the two trial block independently (Q-Learning, Fictitious and Influence models). The underlying assumption is that the higher the level of strategic complexity of the model that best fits a subject's behavior, the higher her strategic learning engagement in the interaction. As detailed in the Supplementary Information we also tested additional models to control for the reliability of our computational approach.

## 2) Results

### a) Behavioral results

We first tested our hypothesis that our game settings triggers differences in choice behavior between the 2 roles.

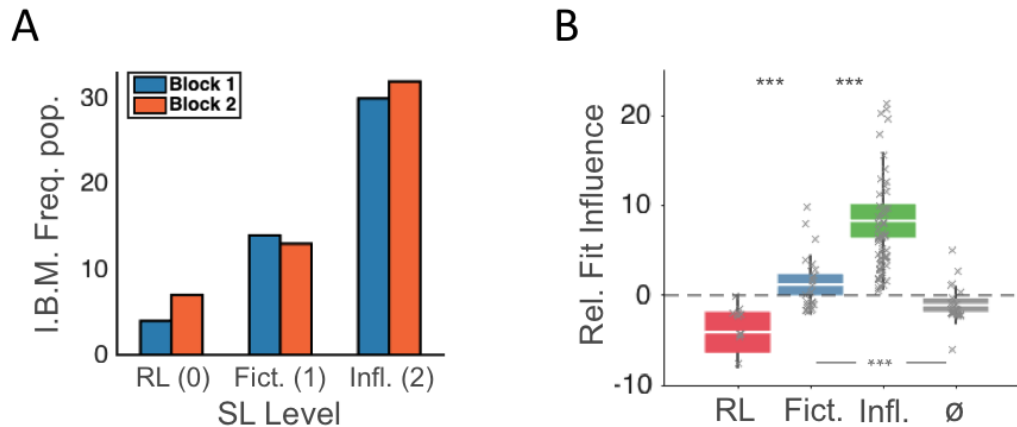
As predicted by our simulation analysis subjects who endorsed the role of Players 1 won more points on average than Players 2 (Block 1, B1:  $F(2,31)= 3.272$   $p= 0.0014$ ,  $t(48.33)= 4.396$   $p<0.0001$  ; Block 2, B2:  $F(2,31)= 2.236$   $p= 0.0282$ ,  $t(54.10)= 3.894$   $p=0.0003$ ), in fact, across the 2 blocks, only 15% of Players 2 won more points than their opponent. The choice behavior of the 2 groups deviated on average from the optimal solution in both blocks (Player 1, P1 - B1:  $P(a) = 0.399(0.065)$ , B2:  $P(a) = 0.391(0.070)$  ; P2 - B1:  $P(A) = 0.482(0.098)$ , B2:  $P(A) = 0.448(0.097)$  ) but Players 2 were the ones who deviated the most from game optimality by choosing the action “A” much more frequently than the mixed strategy equilibrium (MSNE) prescription in comparison to Players 1 (B1  $t(54.09)= 3.935$   $p=0.0002$  unequal variance, B2:  $t(62)= 2.696$   $p=0.009$ ). We thus aimed to test if this difference could explain the difference in performance between the 2 players. As shown on **Fig. 3** Players 2 deviation from MSNE was not correlated to their overall performance like Players 1 (**Fig. 3.A**), but rather to the size of their loss in the interaction (**Fig. 3.B**). In fact the disadvantage in the interaction that was experimentally induced through the structure of the game, lead Players 2 to be constrained to the loss domain, so that the closer their choice proportion was to the MSNE, the less difference in points they had with their opponent. This asymmetry in the interaction seems to have been fully exploited by Players 1 since deviation of Players 2 from the MSNE lead them to perform better than their counterpart did in this situation (**Fig. 3.C - Fig. S1.C**).



**Figure 3.** Model free analysis - Deviation from game optimality affects differently the 2 roles in the strategically asymmetric game, leading to a structural disadvantage of Players 2 (A) The closer to the MSNE the choice distribution of Players 1 is the higher their absolute performance. On the other hand Players 2 choice optimality did not lead to higher absolute performance but to a higher relative performance (B), reducing the gap in points that separate them from their opponent. This structural asymmetry lead Players 1 to fully exploit the disadvantage, their absolute performance increased more with their opponent suboptimality than Players 2 confronted to a suboptimal opponent (C).

Before investigating how the level of strategic learning affects the choice behavior of the two roles, we first tested that our prior assumption that subjects differ in their level of strategic learning was met. Our computational analysis revealed that half of our subjects behavior was best captured by the Influence model (**Fig. 4.A**), while near one third of our population was best fitted by models of lower level of strategic complexity (less than 10% by the reinforcement learning model). Moreover not only the subjects best fitted by the Fictitious model were also better captured by the Influence model in comparison to the

reinforcement model (relative fit of the Influence) (**Fig. 4.B**), but the better a subject's choice behavior was captured by the high SL model, the higher the value of her Influence best fitting parameter  $\lambda$  was (B1:  $r = 0.7534$ ,  $p=6.757e-13$ ; B2 :  $r = 0.7535$ ,  $p=6.7276e-13$ ). Taken together these results reveal that the majority of subjects were engaged in some form of strategic learning throughout a gradient of strategic complexity (SL).



**Figure 4.** Strategic learning heterogeneity captured by our computational approach. Most of the participants engaged in Strategic Learning (SL>0) (A) Individual Best Model (I.B.M.) frequency plot. While at the population level, the Influence model fits the best the population behavior (not shown), at the individual level about half of the subjects were best fitted by high SL and one third by models of lower levels of strategic learning. (B) Population gradient of strategic learning sophistication. The plot represents the average relative fit quality of the Influence model (in comparison to the RL model) for each SL group (I.B.M.). Subjects individually best fitted by higher level of strategic learning model were incrementally better fitted by the Influence.

To maximize the accuracy of our individual characterization and avoid the overestimation of the individual strategic learning level, we conducted an extended computational analysis including additional models. None of the variations of the Reinforcement and Belief-Based models tested improved significantly their fit, thus confirming that most of our subjects indeed engaged in some form of strategic learning (Supplementary Information). In fact our analysis suggests that the SL level might have been under-estimated since more than one third of the subjects previously best fitted by the Influence were better fitted by a 2nd order version of the model (Devaine et al 2014 [24]) (**Fig. S2**). This nevertheless does not affect the superiority of their SL level compared to subjects best fitted by RL or Fictitious models. Although it is important to note that if the relative fit between the 2-Inf and simple reinforcement learning

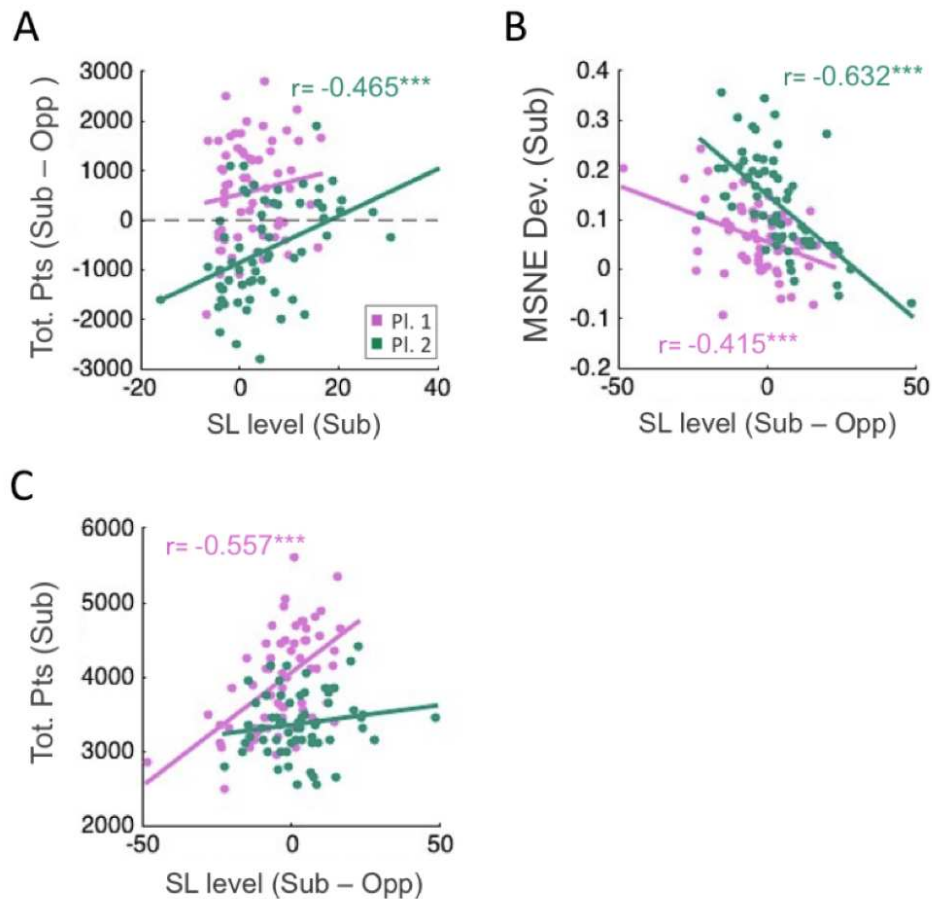
improves the precision of the characterization of the individual SL levels, all the results presented in the following consistently hold when using as SL measure the fit of the Influence relative to the fictitious, or the value of the Influence best fitting parameter  $\lambda$ .

Our computational analysis thus suggests an overall departure from simple reinforcement in repeated competitive interactions, with a population spread across a gradient of strategic sophistication going from value-based (Reinforcement), to low (Belief-Based) and high level (Strategic) level of learning engagement.

Our simulation analysis suggested that the endogenous disadvantage of Players 2 in the game can be overcome by engaging in a higher level of sophistication than the opponent. However the important difference in performance and game optimality observed between the 2 roles in our experiment lead us to hypothesize that, on average, Players 2 did not engage in a higher level of strategic learning than Players 1. Indeed we could not reject the null-hypothesis that the 2 populations of SL came from the same distribution, using as SL measure either their departure from reinforcement towards models of highest strategic complexity ( $D(126)= 0.2031$ ,  $p=0.1250$ ;  $U(126)= 1682$ ,  $Z = 1.7418$   $p= 0.0815$ ) or the weight attributed to 2nd order belief ( $\lambda$ ) ( $D(126)= 0.1719$ ,  $p=0.5809$ ;  $U(126)= 2036$ ,  $Z = 0.0548$ ,  $p= 0.9563$ ), these results hold when the 2 blocks were analyzed separately. The 2 roles did not differ either in how frequently they switched actions from one trial to the next ( $U(126)= 1986.5$ ,  $Z =0.2913$ ,  $p= 0.7708$ ). Our results suggest that the observed disadvantage of players 2 was not due to a difference on average strategic learning sophistication but could rather be caused by a different implication of the SL level in the 2 roles.

We then investigated how the SL level of the 2 Players drove the dynamics of their interaction. We focus first on Players 2 behavior. The level of strategic learning engagement of Players 2 was negatively correlated with deviation from the mixed strategy equilibrium ( $r= -0.6455$ ,  $p= 8.48e-09$ , SL as relative fit of the 2-Inf). Therefore, as suggested by our model free analysis (**Fig. 5.A,B**), their SL level was not correlated directly to the total points won in each block but to the difference in points with their opponent, so that the higher their SL the lower their average relative loss is (**Fig. 5.A**). In fact the higher their SL level was compared to their Player 1 opponent, the closer their action distribution was to the MSNE (**Fig. 5.B**). However this was not enough to overcome the structural disadvantage and increase their absolute performance (**Fig. 5.C**). If Players 2 behavior seems to be constrained by their own SL level, Players 1 behavior presents a quite opposite pattern. Their deviation from MSNE frequency was not directly driven by their SL level ( $r=-0.0396$ ,  $p=0.7561$ ) but by the one of their opponent ( $r=0.49403$ ,  $p=3.34e-05$ ), so that the higher the level of Players 2 the worse their performance was (absolute:  $r= -0.5826$ ,  $p= 4.40e-07$  and relative:  $r= -0.4650$ ,  $p= 0.0001$ ). Since Players 2 who engaged in a higher SL level deviated more from the High Reward action to play closer to the MSNE, they pushed Players 1 to adapt by engaging in a higher SL

level. Indeed, given the structural advantage they have in the game, the better they were at anticipating their opponent's behavior (higher SL level than their opponent), the higher were their relative win ( $r=0.4380$ ,  $p=0.0003$ ) and their absolute performance (**Fig. 5.C**). These results hold when comparing these behavioral measures between high and low SL level (median split) in the 2 roles.



**Figure 5.** Model-based analysis - The Strategic Learning level (SL) of the Players 2 in the game, conditions their capacity to overcome the structural disadvantage of their position in the game. (A) The higher the SL level, measured here by the difference in the fit between the (second order) Influence model compared to the fit of the RL<sup>1</sup>, the more Players 2 reduced their disadvantage compared to their opponent. (B) The higher their SL level compared to their opponent the closer to the MSNE they played. In fact both role converge towards the equilibrium distribution, only Players 2 tend to deviate much more when not engaging in strategic learning. (C) Albeit decreasing the gap in points with their opponent Players 2 could not on average increase their overall performance, constrained by both the structure of the interaction and their own SL level.

<sup>1</sup> Similar results were obtained when running the correlation test analysis with the relative fit of the (first order

Influence. Using the Influence parameter ( $\lambda$ ) values as measure of the SL level or comparing Low and High SL IBM groups of subjects lead conserved the main statistical effects.

---

To capture the effect simultaneously of both the subjects and their opponent SL level on the subject's choice behavior, we ran 3 GLM analyses that confirmed that Players 2's behavior was impacted mainly by their own level of strategic learning sophistication and Players 1 mainly by the SL level of their opponent (**Fig. S3.A.B**).

This dynamic can be further unfolded by looking at the choice accuracy of the subjects. Players 2 who engaged in higher SL level managed to overrule the value-based sub-optimal bias towards the high reward action. Instead the selection of the action A, easily predictable by their opponent in the advantageous situation, was selected more carefully, leading them to switch more often their action from one trial to the other ( $r= 0.3419$ ,  $p=0.0057$ ) and get more frequently the high reward when they chose it (**Fig. S4.C**). Conversely this lead Players 1 to compensate, to avoid deviating more from the optimal play, by engaging in higher strategic learning eventually leading to also increase their accuracy (**Fig. S4.C**).

This overriding of the prime tendency for Players 2 to go for the high reward by engaging in higher level of strategic learning level was also observed from one choice to another. Using a logistic regression analysis we can take a closer look to the series of choices, to investigate how the previous actions impact the next current decision (details provided in the Supplementary Information). This analysis revealed that on average subjects consistently alternated their choices every 2 trials independently of their role (**Fig. S5.A**) but that only Players 2 tended to persist in selecting the action linked to the high reward, taking less into account the opponent's last choice (**Fig. S5.A,B**). And the more Players 2 engaged in strategic learning the more they would alternate their choice (**Fig. S5.C**).

Altogether our analyses suggest that among the subjects endorsing the role of Player 2 in this experiment, only the ones who had a high level of strategic learning sophistication could detach from the game sub-optimal focal point to overcome the structural disadvantage they have in the game interaction. Their opponent, albeit in the easy position, was then forced to adapt and at the end follows the Players 2 to avoid as much as possible to lose their advantage. The leader becomes the follower.

This hypothesis was further backed up by our initial simulation. Indeed running the same type of analysis on our simulation results leads to a very similar dissymmetry in the implication of the SL level between the advantageous and the disadvantageous role (**Fig. S6**).



## **b) Correlation with additional cognitive tasks**

Our main hypothesis was that Players 2's disadvantage in the game would push them to make the effort to use their strategical thinking abilities, so that their performance in the game would be more representative of their cognitive abilities than it would be the case for Players 1. Thus we expected that the heterogeneity in performance among Players 2, as measured in terms of heterogeneity of the SL level of the computational model that best accounted for each subject's behavior, would account more for differences in cognitive capacity. We thus looked at the consistency of their SL level across blocks compared to Players 1. We found that their SL level was significantly more consistent (89% of the subjects were best fitted by the same class - low / high - of SL level models in Block 1 and Block 2) compared to Players 1 (59%, Fisher exact test:  $N = 49$ ,  $p = 0.0269$ ). Also the correlation in SL level across the 2 blocks was significant for Players 2 only (PI.1:  $r = -0.0781$ ,  $p = 0.6710$ ; PI.2:  $r = 0.7171$ ,  $p = 3.88e-06$ ). Moreover no difference on average (or distribution of) SL nor choice behavior (deviation from MSNE, absolute or relative performance, choice accuracy of high reward action, frequency of switch) was found between the 2 blocks for each role. Although our analysis revealed that Players 1 chose faster in the second trial block (PI1:  $t(62) = 2.7102$ ,  $p = 0.0087$ , PI2:  $t(62) = 1.9009$ ,  $p = 0.0620$ ), suggesting some adaptation of Players 1's play in the second block.

Finally we compared the SL level of the subjects and their individual performances in the additional tasks and questionnaires. At the population level only the CRT score (used as a proxy of reasoning ability in the literature) was higher for high SL vs. low SL (median split:  $U(62) = 313.5$ ,  $Z = 2.7678$ ,  $p = 0.0056$ ;  $r = 0.2572$ ,  $p = 0.0402$ ) and in Block 1 only. When comparing the performance in the additional tasks in high vs. low SL level (median split) subjects for each role separately, we found that high SL Players 1 in Block 1 only had a higher CRT score ( $U(30) = 54$ ,  $Z = 2.8891$ ,  $p = 0.0038$ ), performed better in the Raven test ( $U(30) = 62$ ,  $Z = 2.5389$ ,  $p = 0.0111$ ), and were on average more successful in the Tower of London task (ToL :  $t(25) = 2.0675$ ,  $p = 0.04918$ ; ToL (difficult condition: high Goal Hierarchy, high Search Depth) :  $U(25) = 46$ ,  $Z = 2.3271$ ,  $p = 0.0199$ ). No correlation between the performance in any the additional tasks was found with the SL level of Players 2 in any of the 2 blocks. No difference was found between subjects based on the role they endorsed in the game in terms of demographics (Age, salary and education level) nor additional cognitive tasks performances (Working Memory, CRT, Raven, ToL). No difference either in performance in these additional tasks was found between subjects who were consistently fitted by the same SL model in both blocks and the one who switched SL level between the two.

Our results suggest that regardless of the role, the subject's level of strategic sophistication does not depend on the level of the opponent nor the role endorsed in the game but might be related more to different individual cognitive abilities: subjects playing as Player 2 seem to be limited in their propensity to engage in strategic learning preventing them to fully compensate their disadvantage in the game interaction, while the heterogeneity in SL level observed in Players 1, already in a dominant position, might be more driven initially by their executive cognitive abilities (problem-solving, planning). If this is true then Players 2 ability to engage in strategic learning should be correlated to their ability to reason strategically.

During our experimental session, subjects were provided with a second task meant to test specifically the hypothesis that the level of strategic learning engagement in the repeated game matches the ability to play the Nash equilibrium in one-shot games (Griessinger & Coricelli, 2015 [26]). This task was composed of different types of one-shot games played 8 times each, in a random order, with no feedback. In one type of games (Dominant Solvable Other, DSO), higher strategic sophistication was required to form correct belief over the opponent's action and best respond to it but not in the other (Dominant Solvable Self, DSS). The analysis detailed in the Supplementary Information did not allow us to reject our null-hypothesis of an absence of direct mapping between strategic learning and strategic reasoning at the population level. This therefore suggests that different cognitive processes are involved in the engagement in strategic sophisticated play in a repeated game interaction with feedback and in static one-shot games without feedback.

However we found that the more Players 2 reached the N.E. in DSO (requiring higher level of strategic sophistication) the closer to the Nash their frequency of action(a) was in the first Block in the repeated game (all trials B1:  $r=-0.3822$   $p=0.0309$ ; B2:  $r=-0.3078$   $p=0.0865$ ). This correlation, specific to Players 2, was the strongest in the first trials of the all repeated game experiment (B1  $t(1:10)$ :  $r=-0.5647$   $p=7.6e-4$  - not sig. for following bins ; B2  $t(1:10)$ :  $r=-0.1992$   $p=0.2743$  - not sig. for following bins). Using 2 other, more precise, measures of strategic reasoning developed in SI, lead to similar results (SR:  $r=-0.5034$ ,  $p=0.0063$ ; SR':  $r=-0.6853$ ,  $p=0.0068$ ), no correlation was found with the % of NE in DSS.

Taken together these results suggest a transition from static strategic reasoning to on-line computation and update of beliefs over the opponent's behavior when sufficient choice outcomes are observed.

#### D) Exp.3: Human against computerized opponents

To better characterize the interplay of strategic learning sophistication with the strategic asymmetry in game interaction, we conducted a second study in which we controlled for the opponent's behavior by making the subject play against a computer opponent (instructed) and specifically manipulating the SL

level of this opponent. The goal of this experiment was to replicate our initial results and test the specific hypotheses derived from them, that subjects endorsing the disadvantageous role in this strategically asymmetric game were constrained in their choice behavior by their own SL level, while subjects playing in the advantageous position will indeed have the strategic space to adapt, given their SL level, to the behavior of their opponent.

## 1) Material and Methods

### **a) Population**

76 participants (36 male, 40 female; ages 18–30) took part in the experiment. They were student at the University of Trento who had previously joined the Cognitive and Experimental Economics Laboratory (CEEL) recruitment system on a voluntary basis. All participants were right-handed, medication-free, with normal eyesight, no history of neurological disorders. The Ethics Commission of the University of Trento approved the experiment. Informed consent was obtained from each subject before the experiment. Data collection was performed blind to the conditions of the experiment.

### **b) Experimental Design**

The experimental design remained unchanged: participants were randomly assigned to one of the 2 roles of the same game, with the same trial structure and timing, and also played 2 blocks of 100 trials each. Nevertheless, this time they did not play against another randomly picked human participant, but rather played against 2 computerized learning agents: a fictitious play (low SL) and an Influence (High SL), one after the other in a random order. To be fully consistent with the previous experiment, we used as model parameters of the 2 opponents the average best fitting values obtained in the first experiment (details in Supplementary Information).

This design allowed us to test our 2 main hypotheses:

- 1) that the role impacts the average performance and game optimality (Players 1 perform better than Players 2), but not the overall SL distribution of the two groups.
- 2) that Players 1 choice behavior should be impacted by the identity of the opponent with lower performance against the High SL (Influence), compared to Low SL (fictitious). Players 2's behavior on the other hand should only be affected by their own SL level, not the opponent.

Beforehand we simulated once again the experiment by making agents with different SL level play against the 2 algorithms. Our results show the selective effect of the opponent SL level manipulation on

the Players 1 we expect to see in the actual experiment, thus confirming the adequacy of our design (**Fig. S7**). All statistical analyses were performed using Matlab ([www.mathworks.com](http://www.mathworks.com)) with the addition of the Statistical toolbox and other free-download functions. All stimuli and feedbacks were presented using PsychToolBox and appeared on a uniform black background.

## 2) Results

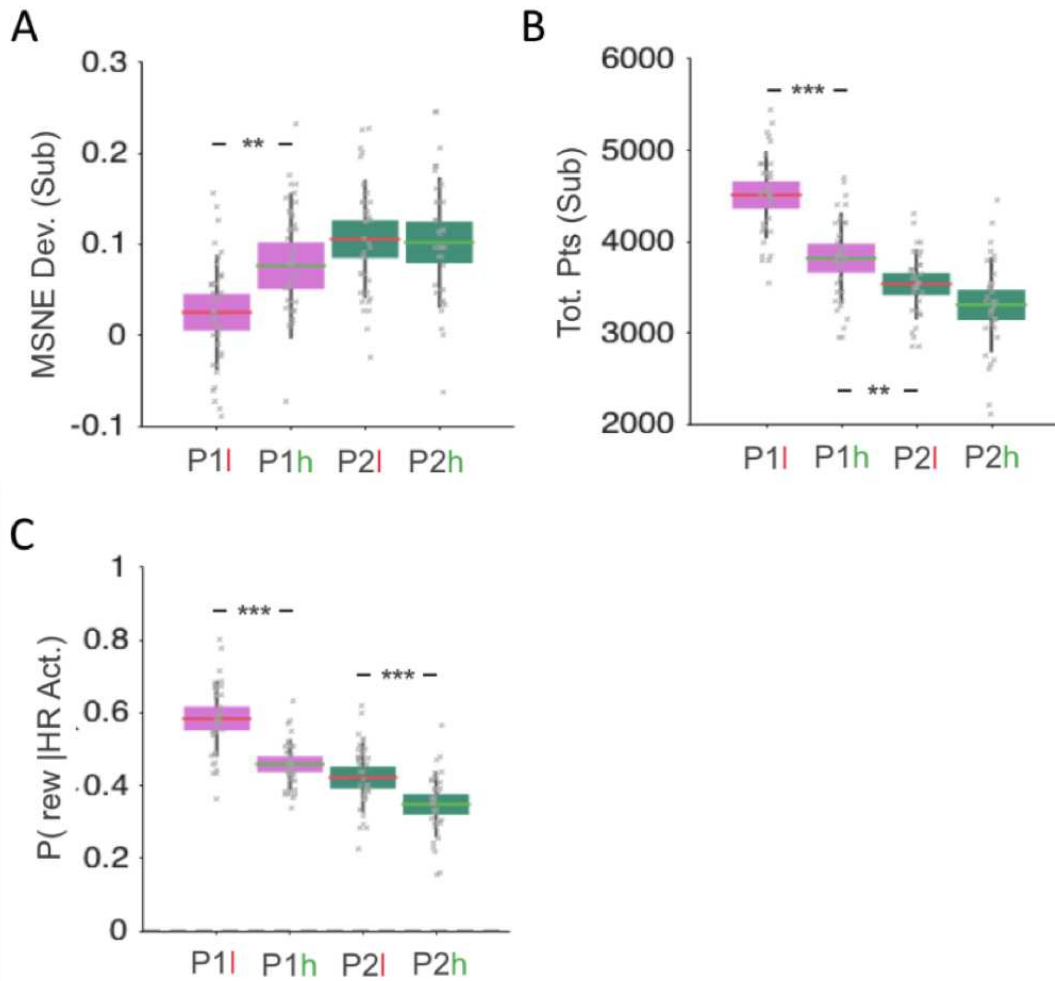
On average, Players 1 won more points ( $t(142) = 8.5298$   $p = 2.7896 \times 10^{-14}$ ) and had a distribution of choices closer to the Mixed Nash Equilibrium ( $t(142) = -4.5144$   $p = 1.322 \times 10^{-5}$  unequal variance) than Players 2.

Our Model-based analysis replicated nicely the distribution and SL gradient across the population observed in Exp.1 (**Fig. S8**). And, as in Exp.1, no difference in SL distribution was found between the 2 roles.

But this time participants were playing against an algorithm, not against another human. And since they played the repeated game in the same experimental conditions as in Experiment 1, we tested if this difference affected their behavior. For each of the 2 roles no significant difference was found in performance (total points, points difference with the opponent) between the two experiments, however a trend towards higher strategic learning engagement when playing against algorithms was observed. When comparing low vs. high SL (median split) Players 1 engaged in strategic learning were found to have a higher SL level in the second experiment (rel. fit 2-Inf:  $U(66) = 233$ ,  $Z = 4.2082$ ,  $p = 2.57 \times 10^{-5}$ ,  $\lambda$  parameter:  $U(66) = 368$ ,  $Z = 2.5495$ ,  $p = 0.0108$  - similar results were obtained when comparing the SL level between subjects best fitted by the Influence models).

No difference in mean SL level (nor distribution) was found between the 2 roles. Running an ANOVA to test if the SL level was modulated by the opponent encountered did not result in any significant effect either. As in experiment 1 Players 2 were most consistently fitted by same SL level models between the 2 opponents than Players 1 ( $P_2 = 0.84$  prop. same low/high SL :  $P_2 = 0.84$ ,  $P_1 = 0.57$  ; Fisher exact test:  $N = 58$ ,  $p = 0.0259$ ).

We next tested our second hypothesis regarding the specific effect of the opponent on the choice behavior of the subject given the role endorsed in the experiment. As shown in **Fig.6**, only Players 1 were affected by the identity of the opponent, exactly as predicted by the simulation (**Fig.5.A.B**, **Fig.S7.A.B**).



**Figure 6.** As hypothesized, only the subjects endorsing the role of Player 1 (pink) in the repeated game were affected in their choice behavior by the SL level of the (computerized) opponent encountered. (A) Players 1 frequency of choice is closer to the MSNE distribution when playing against the low SL opponent compared to the high level. No difference in percentage of deviation from MSNE distribution ( $p(a)=1/3$ ) was found between the 2 opponent's block for Players 2. (B) When opposed to the low SL level opponent, Players 1 won on average more points in total than when playing against the high SL. No difference was found for Players 2. (C) When opposed to the low SL opponent both Players, 1 and 2, were more frequently rewarded when playing the high payoff action (b for player 1 and A for player 2), a proxy for choice accuracy, compared to the high SL level. Effect size was however higher for Player 1 (Cohen's  $d= 1.1909$ ) than Player 2 ( $d= 0.7633$ ).

To refine our analysis, we ran the 3 GLMs we used in Exp. 1, taking into account not only the level of the computerized opponent (low or high) but the SL level of the subjects. Our results replicate strongly the asymmetry found in our first experiment and observed in the average results (**Fig.6**): Subjects playing as

Player 1 have their deviation from MSNE as well as their performance affected by the level of the opponent; Conversely Players 2's choice behavior is modulated only by their own Strategic Learning level (**Fig.6**). In this experiment however, Players 1's behavior seems to have been influenced not only by the level of the opponent but also by their own strategic learning engagement (**Fig.S9**). This effect could be due to the constraint our design added on their opponent's behavior.

At the end of the experiment subjects were provided with an additional task aimed to capture more precisely the working memory capacity of our population (namely 2 and 3-Back tasks - see Supplementary Information for details). We observed a trend towards a higher performance and RT in this task for high SL Players 2 (median split) only when playing against the high SL opponent (% correct in 2-Back:  $t(33) = 2.2047$   $p=0.036166$ , 3-Back :  $t(33)=1.7420$   $p=0.0908$ , albeit a higher % for High SL =75.4(9.5) vs. low SL=69.9(8.9), reaction time 3-Back :  $t(33)=2.2999$   $p=0.027917$ ). Albeit weak, this effect is in line with our results suggesting a cognitive limitation of the subjects playing in the disadvantageous role when confronted to a highly sophisticated learner

## E) Discussion

The present study aimed at testing the prediction that the structure of a repeated game interaction can lead to strategic asymmetry depending on the way it facilitates the engagement in sophisticated learning. More precisely, the hypothesis was that a dissymmetry in the overlap between reward structure and MSNE (even when there is still symmetry in maximum possible payoffs between players) can differently engage human subjects in using sophisticated strategic learning so that their overall performance does not always reflect an individual's general ability of strategic learning or strategic reasoning.

This hypothesis was rooted in research in behavioral economics and cognitive sciences suggesting that humans can use information available about their counterpart to form beliefs over their intentions (Koster-Hale & Saxe, 2013 [3]). Indeed in the case of repeated games, where social interactions are reduced to actions and rewarded feedbacks, payoffs convey informational value about the opponent's behavior and beliefs become analogous to a mapping of action-outcome contingencies, as suggested by the model-based reinforcement learning framework (Doll et al, 2015 [20]). Recent studies suggest that similar brain computations might be involved in the decrease of the uncertainty (increased prediction) over the opponent's next choice allowing one to maximize her overall outcome of the interaction (best response to beliefs) (Lee & Seo, 2016 [27]). Moreover inferential processes might be implicated in the iterative incorporation of the strategic nature of the interaction, not only considering one's behavior but the

interplay of past actions in the history of play, ultimately increasing belief accuracy over an opponent also capable of belief-learning (Griessinger & Coricelli, 2015 [26]; Hyndman et al, 2009 [32]). Nevertheless, important heterogeneity has been observed in the level of engagement in such high-order belief (strategic) learning among individuals (Devaine et al, 2014 [24]; Hampton et al, 2008 [25]).

We thus hypothesized that the reward structure of the interaction might affect subjects differently given their capacity to engage in strategic learning, depending on how much best-response to reward-based and belief-based learning overlap. Based on this prediction we developed a 2x2 strategically asymmetric game where the two roles were meant equal (same payoff distribution and expected payoff at MSNE), but in which inequity arises among individuals with different SL level, from the discrepancy in one role (disadvantaged position) only between the highest reward action (focal point) and the MSNE distribution.

To test this hypothesis, we combined agent simulation and behavioral experiments, unconstrained (human-human interaction) and constrained (human-computer). Our two behavioral experiments lead to the same conclusions, predicted by our simulation analysis. First at the population level, subjects endorsing the disadvantageous position during the repeated game interaction earn significantly less than their opponent, also their choice distribution deviated more from the MSNE prescription. Second the more the disadvantaged participants engaged in strategic learning, the more they overcame the strategic asymmetry, and this effect was even stronger when the opponent did not fully engage in belief-learning. Forming accurate beliefs over their opponent allows these subjects to reduce their disadvantage in total earnings and play closer to the MSNE. Conversely the choice optimality and therefore the absolute performance of the participants playing in the advantaged role was modulated only by the behavior, and ultimately the SL level, of their opponent, but not by their own capacity to engage in strategic learning.

These results provide clear evidence for sophisticated learning in repeated interactions (Lee, 2008 [4]; Shteingart & Loewenstein, 2014 [33]) and the central role of belief accuracy in equilibrium play (Bosworth, 2017 [34]; Crawford et al, 2013 [15]). Moreover our study shows how the reward structure of the repeated game interacts with the observed heterogeneity in belief-learning at the population level by creating a tension between rewards and beliefs. Empirically we show that the high reward attracts maximizing behavior and creates a focal point, which can easily be exploited by a low strategic learning opponent (Coricelli, 2005 [35]) when not aligned with the MSNE prescription. In our stage game the two players had identical expected payoffs, but the fact that for only one of the 2 roles the focal point corresponds to the action that was theoretically optimal to select the most, creates an endogenous asymmetry, strategic in nature which reveals itself throughout the interaction. Previous studies also used a computational approach to capture as precisely as possible the choice behavior of humans in repeated 2x2 games (Hill et al, 2017 [28]; Ho et al, 2007 [36]; Marchiori & Warglien, 2008 [37]). We went a step further and

manipulated the interaction structure to show that the level of the strategic sophistication of individual's learning drives the formation of higher order beliefs and allows them to disengage from the attractions of immediate outcomes and move closer to optimality.

Crucially no correlation was found in any of the 2 experiments between the SL level of the subjects, in any of the two roles, and the one of the opponent. This result replicates the correlation in strategic learning level found in previous studies in which humans were confronted to different opponents also varying in their SL level (Devaine et al, 2014 [24]; Shachat & Swarthout, 2012 [38]). It is worth noting however that recent research suggests that arbitration between model free and model-based learning can be affected by the volatility of the environment (Simon & Daw, 2011 [39]). Indeed in our two experiments most of the subjects did engage in rudimentary form of belief-learning which does not reject the hypothesis that parts of a subject's learning mechanisms may be with low sophistication (model-free).. Also our computational approach was meant to measure the overall level of strategic learning sophistication embedded in individual choice series, and does not allow to track local changes in strategy. In fact, as previously observed (Duersch et al, 2010 [40]; Seo et al, 2014 [41]; Spiliopoulos, 2013 [42]), the SL level of the opponent strongly impacted the behavior of the subjects interacting in the advantageous position. However in our study the influence of these subjects' own SL level on their choice behavior was reported as weaker than the influence of their opponent. One hypothesis for this result is that the sophistication of their beliefs did not condition their behavior, which obviously comes in contradiction with the above-cited literature. An alternative hypothesis is that the accuracy of their beliefs was already sufficient to maximize (up to a certain individual threshold) their earnings and that engaging in higher level of strategic learning did not present a net advantage. We argue that the present study provides the first experimental evidence in favor of the latter explanation. First subjects playing in the advantageous role won on average more points throughout the interaction than their opponent, even when opposed to a high SL computerized agent ( $t(70)=2.7833$ ,  $p=0.0069$ , **Fig.6.B**). Second, albeit good performance, these subjects presented low consistency in SL level across interactions in both experiments. Third their behavior was correlated to planning and problem solving scores captured in additional tasks in the first interaction block only, while in the second block conjunctural evidence of behavioral re-adjustment (faster choices for no change in performance) were observed.

On the other hand disadvantaged subjects behavior were found to be solely conditioned by their level of strategic engagement, not the one of their opponent. They also presented much higher consistency across interactions, and evidence of a correlation between working memory and their SL level was found. Crucially the role endorsed in the repeated game did not seem to impact the SL level of the subjects in both of our experiments.

Altogether our results suggest a dissociation between a strategic learning engagement bounded by



individual cognition for subjects endorsing the disadvantaged position, and what has been proposed to resemble a cost benefits-analysis process (Alaoui & Penta, 2015 [43]) for the subjects ensured to dominate the interaction at lower SL level. This distinction between bounded cognition and bounded rationality in suboptimal play has also been observed in static games by Friedenberg et al (Friedenberg et al, 2016 [44]).

In behavioral game theory the concept of bounded rationality broadly assumes that the capacity of the agents to grasp and use all the required information leading to equilibrium are somehow constrained (Simon, 1991 [45]). In this line a theoretical framework which has accumulated growing support in the past decade has been proposed to explain deviation from optimal choices in static games: level-k models (Crawford et al, 2013 [15]). This class of model relaxes the assumption of full rationality and assumes that players actually best respond to incomplete beliefs varying among individuals in their degree of sophistication (k) over the behavior of their opponent, themselves considered as only capable of a lower order of beliefs ( $k-1$  or  $<1$ , see Camerer et al, 2015 [46]). This hierarchical organisation of beliefs is very close to the computational framework previously developed in (Hampton et al, 2008 [25]).and that we used in this study.

In our approach strategic learning sophistication (SL) is modeled as a hierarchy of different levels of computations: SL0 corresponds to reinforcement learning which computes action values based on the past reward experienced and is agnostic about the choice behavior of the opponent; SL1 is modeled as a fictitious play which best responds to the opponent's probability distribution computed from its past choices; SL2 is modeled as an influence learning process which assumes that the opponent is also learning in a way analog to a fictitious play and thus that its own past actions can have an influence over the action probability of the other (we also included a SL2+ learning rule that considers that the opponent is also learning through influence). Based on this correspondence between the 2 classes of models, we have hypothesized that a direct mapping might exist between the level k in static games and the SL level in repeated interactions (Griessinger & Coricelli, 2015 [26]). We tested this hypothesis but failed to reject a direct correlation between the 2 measures of strategic sophistication (Supplementary Information). However we observed that for the subjects endorsing the disadvantaged position in the repeated game, a correlation was found between their strategic reasoning (level k) and the frequency of play close to the MSNE in the very first trials of the first interaction. This result therefore suggests another type of relationship between the strategic reasoning and strategic learning models of bounded rationality: at first, when beliefs cannot be anchored in enough observations, subjects with a strong incentive to take over the interaction are guided by their ability to reason in an iterative fashion (level k), but when enough experience is accumulated subjects with the capacity to engage in strategic learning form and update beliefs as accurate as possible over their opponent behavior. This hypothesis appears promising to us

since it echoes to other research on the influence of priors in social inference (Chambon et al, 2017 [47]), and therefore calls for proper testing in laboratory.

It is worth noting that another source of suboptimality has been suggested in the behavioral economics literature: heterogeneity in best response. It has been proposed that social preferences for instance could bend utility functions (Fehr & Camerer, 2007 [48]). If our study did not allow to test directly this hypothesis, we still observed a higher strategic engagement in advantaged subjects capable of high strategic learning when confronted to an algorithm compared to another human (social framing effect as in [Devaine]). This result might suggest that social preferences such as altruism or sensitivity to inequity could be reflected in a lower exploitation of their advantageous position when playing with a human counterpart (Silk & House, 2016 [49]).

Altogether our results reveal three possible sources of variability in strategic behavior during repeated game interactions. The first, that we call exogenous, is driven by external factors such as the payoff structure, the salience of the different outcomes of the game, and also the prior knowledge over the opponent. A second source of heterogeneity in game play is endogenous, with differences in social preferences but also motivation (Schmidt et al, 2012 [50]) or sensitivity to rewards (Kim et al, 2015 [51]). Finally a third type of variance emerges from the two previous one, leading to specific dynamics of repeated choice behavior. Indeed even if our experimental setting did not allow further investigation of the phenomenon, it seems clear that the SL level of subjects in the disadvantageous position drove the interaction, while their opponent in the dominant position would simply track and adapt to changes in behavior and ultimately followed them. Leader-follower dynamics have been observed in repeated games (Seip & Grøn, 2016 [52]), however precise understanding of the underlying behavioral forces remain unclear (Sato et al, 2002 [53]). Predicting the learning dynamics in play by tuning the structure of the interaction can help study critical behavior such as strategic teaching (Camerer et al, 2002 [54]).

The present study brings further support to the pertinence of the cognitive neuroscience framework of learning for the analysis of repeated non-cooperative game behavior. Moreover, we advocate in favor of the use of model simulations in the field, that 1) allow to take the most of a normative framework to optimize experimental design and make precise predictions regarding the expected results (Palminteri et al, 2017 [55]) and 2) open the possibility to refine agent based simulation analyses in order to better characterize the interplay between the level of strategic learning and the structure of the game underlying the repeated interaction.

At a broader scope this study points in the direction of a systematic consideration of between-subjects differences and the interaction effect between human variance in learning and the way strategic interactions are constrained. Taking into consideration asymmetric facilitation can help study the

emergence of social hierarchy and strategic dominance in interactions (Qu et al, 2017 [56]), but also better understand how inequity arising from the interaction between the environment and endogenous differences could be reduced in real-life social interactions (Decety & Yoder, 2017 [57]).

## F) References

- [1] Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in cognitive sciences*, 19(2), 65-72.
- [2] Pacherie, E. & Khamassi, M. (2017). Action. In Andler, D., Collins, R. and Tallon-Baudry, C. (Eds) *La cognition*. Paris, France: Gallimard. In press.
- [3] Koster-Hale, J., & Saxe, R. (2013). Theory of mind: a neural prediction problem. *Neuron*, 79(5), 836-848.
- [4] Lee, D. (2008). Game theory and neural basis of social decision making. *Nature neuroscience*, 11(4), 404.
- [5] Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1), 48-49.
- [6] Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.
- [7] Goeree, J. K. et Holt, C. A. (2001). « Ten Little Treasures of Game Theory and Ten Intuitive Contradictions ».
- [8] American Economic Review, vol. 91 no 5 : pp. 1402–1422.
- [9] Fudenberg, D., & Levine, D. K. (2009). Learning and equilibrium. *Annu. Rev. Econ.*, 1(1), 385-420.
- [10] Gauvrit, N., Zenil, H., Soler-Toscano, F., Delahaye, J. P., & Brugger, P. (2017). Human behavioral complexity peaks at age 25. *PLoS computational biology*, 13(4), e1005408.
- [11] Shachat, J. M. (2002). Mixed strategy play and the minimax hypothesis. *Journal of Economic Theory*, 104(1), 189-226.
- [12] Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American economic review*, 848-881.
- [13] Fudenberg, D., & Levine, D. K. (1998). *The theory of learning in games* (Vol. 2). MIT press.
- [14] Camerer, C. F., Ho, T. H., & Chong, J. K. (2004). Behavioural game theory: Thinking, learning and teaching. In *Advances in Understanding Strategic Behaviour* (pp. 120-180). Palgrave Macmillan UK.
- [15] Barros, G. (2010). Herbert A. Simon and the concept of rationality: boundaries and procedures. *Revista de economia política*, 30(3), 455-472.

- [15] Crawford, V. P., Costa-Gomes, M. A., & Iriberri, N. (2013). Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature*, 51(1), 5-62.
- [16] Polonio, L., Di Guida, S., & Coricelli, G. (2015). Strategic sophistication and attention in games: an eye-tracking study. *Games and Economic Behavior*, 94, 80-96.
- [17] Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). A second-person neuroscience in interaction. *Behavioral and brain sciences*, 36(4), 441-462.
- [18] Dennett, D. (1987). *The Intentional Stance* (Cambridge, MA and London).
- [19] Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current opinion in neurobiology*, 22(6), 1075-1081.
- [20] Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature neuroscience*, 18(5), 767-772.
- [21] O'Doherty, J. P. (2014). The problem with value. *Neuroscience & Biobehavioral Reviews*, 43, 259-268.
- [22] Joiner, J., Piva, M., Turrin, C., & Chang, S. W. (2017). Social learning through prediction error in the brain. *npj Science of Learning*, 2(1), 8.
- [23] Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, 15(8), 549-562.
- [24] Devaine, M., Hollard, G., & Daunizeau, J. (2014). The social Bayesian brain: does mentalizing make a difference when we learn?. *PLoS computational biology*, 10(12), e1003992.
- [25] Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences*, 105(18), 6741-6746.
- [26] Griessinger, T., & Coricelli, G. (2015). The neuroeconomics of strategic interaction. *Current Opinion in Behavioral Sciences*, 3, 73-79.
- [27] Lee, D., & Seo, H. (2016). Neural basis of strategic decision making. *Trends in neurosciences*, 39(1), 40-48.
- [28] Hill, C. A., Suzuki, S., Polania, R., Moisa, M., O'Doherty, J. P., & Ruff, C. C. (2017). A causal account of the brain network computations underlying strategic social behavior. *Nature Neuroscience*.
- [29] Beckenkamp, M., Hennig-Schmidt, H., & Maier-Rigaud, F. P. (2007). Cooperation in symmetric and asymmetric prisoner's dilemma games.
- [30] Feldman, M., Kalai, A., & Tennenholtz, M. (2010). Playing Games without Observing Payoffs. In *ICS* (pp. 106-110).
- [31] Friedenber, A., Kets, W., & Kneeland, T. (2016). Bounded Reasoning: Rationality or Cognition.
- [32] Hyndman, K., Terracol, A., & Vaksman, J. (2009). Learning and sophistication in coordination games. *Experimental Economics*, 12(4), 450-472.

- [33] Shteingart, Hanan, and Yonatan Loewenstein. "Reinforcement learning and human behavior." *Current Opinion in Neurobiology* 25 (2014): 93-98.
- [34] Bosworth, S. J. (2017). The importance of higher-order beliefs to successful coordination. *Experimental Economics*, 20(1), 237-258.
- [35] Coricelli, G. (2005). Strategic interaction in iterated zero-sum games. *Homo Oeconomicus*
- [36] Ho, T. H., Camerer, C. F., & Chong, J. K. (2007). Self-tuning experience weighted attraction learning in games. *Journal of Economic Theory*, 133(1), 177-198.
- [37] Marchiori, D., & Warglien, M. (2008). Predicting human interactive learning by regret-driven neural networks. *Science*, 319(5866), 1111-1113.
- [38] Shachat, J., & Swarthout, J. T. (2012). Learning about learning in games through experimental control of strategic interdependence. *Journal of Economic Dynamics and Control*, 36(3), 383-402.
- [39] Simon, D. A., & Daw, N. D. (2011). Environmental statistics and the trade-off between model-based and TD learning in humans. In *Advances in neural information processing systems* (pp. 127-135).
- [40] Duersch, P., Kolb, A., Oechssler, J., & Schipper, B. C. (2010). Rage against the machines: how subjects play against learning algorithms. *Economic Theory*, 43(3), 407-430.
- [41] Seo, H., Cai, X., Donahue, C. H., & Lee, D. (2014). Neural correlates of strategic reasoning during competitive games. *Science*, 346(6207), 340-343.
- [42] Spiliopoulos, L. (2013). Strategic adaptation of humans playing computer algorithms in a repeated constant-sum game. *Autonomous agents and multi-agent systems*, 1-30.
- [43] Alaoui, L., & Penta, A. (2015). Endogenous depth of reasoning. *The Review of Economic Studies*, 83(4), 1297-1333.
- [44] Friedenberg, A., Kets, W., & Kneeland, T. (2016). Bounded Reasoning: Rationality or Cognition.
- [45] Simon, H. A. (1991). Bounded rationality and organizational learning. *Organization science*, 2(1), 125-134.
- [46] Camerer, C. F., Ho, T. H., & Chong, J. K. (2015). A psychological approach to strategic thinking in games. *Current Opinion in Behavioral Sciences*, 3, 157-162.
- [47] Chambon, V., Domenech, P., Jacquet, P. O., Barbalat, G., Bouton, S., Pacherie, E., ... & Farrer, C. (2017). Neural coding of prior expectations in hierarchical intention inference. *Scientific Reports*, 7(1), 1278.
- [48] Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends in cognitive sciences*, 11(10), 419-427.
- [49] Silk, J. B., & House, B. R. (2016). The evolution of altruistic social preferences in human groups. *Phil. Trans. R. Soc. B*, 371(1687), 20150097.
- [50] Schmidt, L., Lebreton, M., Cléry-Melin, M. L., Daunizeau, J., & Pessiglione, M. (2012). Neural mechanisms underlying motivation of mental versus physical effort. *PLoS biology*, 10(2), e1001266.

- [51] Kim, S. H., Yoon, H., Kim, H., & Hamann, S. (2015). Individual differences in sensitivity to reward and punishment and neural activity during reward and avoidance learning. *Social cognitive and affective neuroscience*, 10(9), 1219-1227.
- [52] Seip, K. L., & Grøn, Ø. (2016). Leading the game, losing the competition: identifying leaders and followers in a repeated game. *PloS one*, 11(3), e0150398.
- [53] Sato, Y., Akiyama, E., & Farmer, J. D. (2002). Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences*, 99(7), 4748-4751.
- [54] Camerer, C. F., Ho, T. H., & Chong, J. K. (2002). Sophisticated experience-weighted attraction learning and strategic teaching in repeated games. *Journal of Economic theory*, 104(1), 137-188.
- [55] Palminteri, S., Wyart, V., & Koehler, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*.
- [56] Qu, C., Ligneul, R., Van der Henst, J. B., & Dreher, J. C. (2017). An Integrative Interdisciplinary Perspective on Social Dominance Hierarchies. *Trends in Cognitive Sciences*.
- [57] Decety, J., & Yoder, K. J. (2017). The emerging social neuroscience of justice motivation. *Trends in cognitive sciences*, 21(1), 6-14.

## II - Strategic learning in repeated game interactions: Methodological considerations

In the following section we will further discuss the methodology used in the first three experiments and mention extra analyses that have been conducted on the dataset from Exp.2 but not included in the actual version of the article.

### A) Additional discussion

We would like to discuss the interpretation regarding the use of simple learning models algorithms. We argue that if such models might be useful to approximate the level of strategic sophistication, of departure from RL to engage in beliefs and higher-order inferences, it seems more parsimonious to interpret the information they provide as a “degree of strategic complexity embedded in the subject’s choice series”. This for two main reasons.

The first limit lies under a general concern important to keep in mind regarding the modeling approach used in decision-making neuroscience, and in particular learning models in binary choice tasks. The adequacy of computational models were estimated at the scale of the entire choice series made of binary values. Obviously we insured that, once fitted, the models we used were best tuned with a quite exploitative action selection strategy (inverse softmax temperature,  $\beta$ : exp.2= 2.01(1.30), exp.3= 1.91(1.44)). However, the SL level derived for each individual was still averaged across an entire block. This could prevent the detection of variations in strategic engagement within an interaction block (Schuck et al 2015; Wallin et al, 2017). One evidence for a possible limit of our computational approach is that in our experiment reaction time was highly variable across subjects but quite consistent within-subject, and still no striking correlation between our computational approach and individual variation in choice reaction time was observed. However, recent researches suggest that reaction time is an important piece of information to consider in the framework of game interactions (Gill et al, 2017; Spiliopoulos, 2016). Thus improvement in cognitive modelling is possible in this direction. One direction of future research would be to extend the underlying computational assumptions of our approach to include data from choice reaction time and through the use of multi-objective optimization (see Viejo et al, 2015) to better approximate what could then become a generative model of strategic learning.

This led us to the second limitation, which concerns directly the computational framework we employ. We made the decision to use strategic learning models derived from the field of reinforcement learning

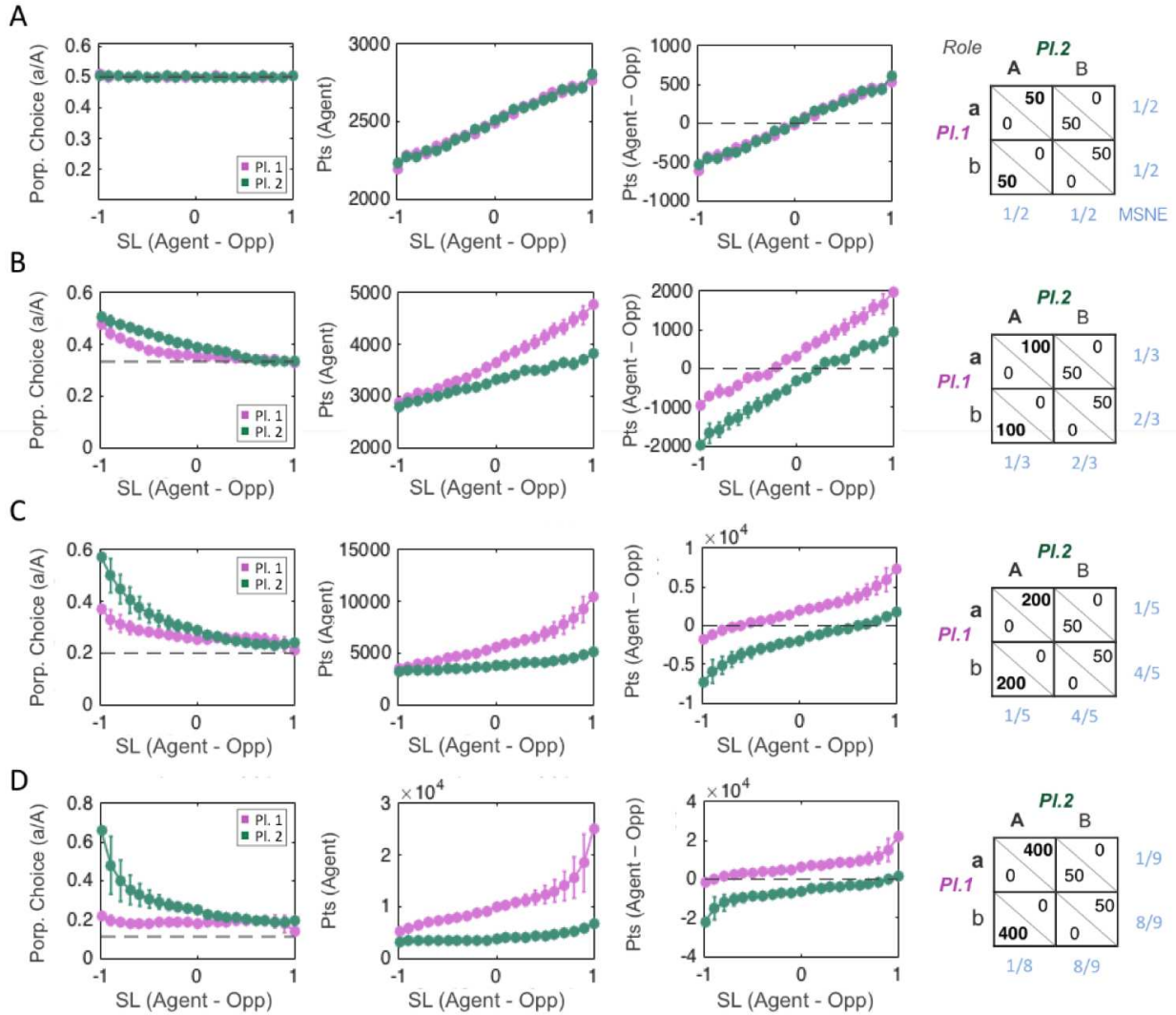
(Hampton et al, 2008). This choice was motivated by two reasons. First, as pointed out previously, regarding the nature of the data we wanted to fit, the use of not-too-heavy computational models seemed the most parsimonious. Indeed the data available for each subject in this kind of social experiment are binary choices repeated 100 times which does not provide tremendous signal-to-noise ratio. Regarding the consistency of our results from simulation prediction to experimental replication, this methodology seems well suited to provide to the experimenter useful information regarding the subject's strategy, thus allowing interpretations regarding the empirical choice dynamics. However more subtle models of strategic learning have been developed in the field. Based on Bayesian computation these models can infer trial by trial the precise statistics of the distributions underlying the opponent's choice allowing better prediction power (Devaine et al, 2010; de Weerd et al, 2010; Yoshida et al, 2010) (see chapter I). This leads to the second reason which motivated the use of models rooted in the RL framework: Our goal was not to uncover the precise computations performed by the human brain, but to capture between-subjects differences in belief-learning. Note that this argument might as well be inverted, since so far mainly computations performed by the reinforcement learning rules have been correlated to neural data in repeated game interactions (Hill et al, 2017; Seo et al, 2017; Zhu et al, 2012). We thus used computational models as a proxy of their level of strategic learning engagement in repeated game interaction. Since a model is by definition always an approximation, and since the simple variations of reinforcement learning rules captured well enough the strategic diversity embedded in the population, it seemed unnecessary to engage in heavy computational work. A third argument in favor of such computational framework, already highlighted in the discussion part of the article, lies in the practicality of such simple, therefore light, strategic models for running many repeated game simulations.

An actually on going work, not presented in this thesis, aims at using this approach to test the predictive power of the strategic learning framework by simulating multiple repeated interactions between agents of different strategic learning levels varying in the underlying game structure (payoff matrix). The first step of this analysis is presented on **Fig.1**. We varied the amplitude of the focal point for the two roles (the payoffs in the game matrix initially equal to 100 pts), while keeping the rest of the payoff equivalent (0, and 50pts). This analysis shows that the higher the SL level of the agent, the closer its choice distribution is to the MSNE in all the variations of the game, from the symmetric version (hide and seek 0, 50pts - **Fig.1.A**) to the highly asymmetric version with focal points of 400 pts (**Fig.1.D**). These results confirmed first that the results presented in the article do not hold for the initial version of the strategic asymmetric game only, but emerge from the strategic asymmetry propriety of the game. These results also suggest that indeed the stronger the focal point for players 2, the higher their SL level has to be to reduce their disadvantage.

The goal of this extended simulation analysis will thus be two-folds: First, to test the generability of the results presented in the article by comparing model predictions to empirical results reported in previous



research in behavioral game theory. Second, to draw a landscape of dyadic inequity in 2x2 repeated game to track how both the amplitude and configuration of the payoffs impact the evolution of relative performance in order to identify equity areas meant to be eventually tested in laboratory.



**Figure 1.** Effect of the focal point amplitude on the simulation results of Influence agents with different values of  $\lambda$  playing the repeated game in each role against each other. The MSNE distribution prescription corresponds to the black dashed line on the left plots. (A-D) Simulation plots represents for each role the effect of the difference between the SL level of the agent and the one of its opponent, on average proportion of choice a/A (left plots), total points accumulated throughout the game interaction (100 repeated choices) (middle plots), and the difference in total points between the agent and its opponent (right plots). The value of the 2 focal points (payoffs associated to action profile (A,a), and (b,A) in the original payoff matrix), varies from 50 pts (reducing to a symmetric hide and seek game - A), 100 pts (initial strategic asymmetric game employed in the article - B), 200 pts (C) and 400 pts (D). The other payoffs remained the same (i.e. 0pts and 50 pts).

[The SL level was modeled by the  $\lambda$  value of the Influence model (varying from 0 to 1, with fixed  $\eta$  to the average values of our empirical distribution [0:0.4]). The simulation was ran 10 times, the results plotted here represent the average across these repetitions. The same effect were observed in the 12 plots when taking as SL measure in our simulation the arbitration parameter ( $\kappa$ ) of the mixture model (Hybrid Influence model) instead.]

---

## B) Additional results Exp.2

In the Supplementary Information 2 (SI2) (**Appendix III**), we develop the details of the extension of the computational analysis presented in the article and that we conducted on the initial dataset (Exp.2). To summarize we ran, among many other models, 2 models that have been developed in the behavioral economics literature (see Chapter I) and that seemed important to test as control. The first model, the Experience Weighted Attraction (EWA), has been developed as an hybrid between reinforcement and belief-based learning, not only updating the action value based on the outcomes experienced through the interaction but also the outcomes that could have been experienced if the agent had chosen differently (Camerer et al, 2002). Such type of fictive computation thus embodies, implicitly, a model of the strategic nature of the interaction and by extent information regarding the choices made by both the agents and their opponent (Hsu et al, 2012). Our model comparison analysis shows that this model did not outperform the belief-based models included in our initial model space (**SI2, Fig.S3.A**). Moreover the computational analysis of the EWA confirms its hybrid nature between RL and belief-based learning.

The second model, the 2-period (weighted) fictitious play (fp2) (Spiliopoulos, 2012), has been developed by Spiliopoulos to extend the original (weighted) fictitious play from (Cheung & Friedman, 1997) which computes the probability that the opponent chooses an action based on its (decaying) frequency of past play. The fp2 model aims at combining pattern detection and fictitious play by tracking the conditional probability of the opponent's choice given the pattern formed by its 2 last choices. The author has shown that in repeated games, subjects were able to detect on average choice patterns 2 trials back (Spiliopoulos, 2013). Moreover, a simulation analysis of agents modeled as fictitious play variations proposed in the economics literature, and interacting with each other in various repeated games, showed that the fp2 model better captures the choice patterns in their opponent choice series than the analog versions (Chu, 2013). Testing this model was therefore a way to control that the strategic complexity captured by the influence models was rooted in the iterative sophistication of the beliefs (ToM like, "I think that you think that I think...", see Hampton et al, 2008; Devaine et al, 2014), and not just a measure of the overall complexity (or autocorrelation) of the individual choice series. The fp2 model did not outperform the influence models, and interestingly stood out as a model orthogonal to the strategic learning gradient

identified by our approach (**SI2, Fig.S3.C,D**).

This result thus suggests that pattern learning could be another strategy implementable in competitive strategic interactions. We are actually running several analyses on our repeated game datasets to further characterize the choice behavior properties that this model captures (not presented in this thesis). Moreover, Chapter IV of this manuscript will present a specific experiment that we designed and ran to assess human subjects' ability to detect patterns in the choice sequence of their opponent.

## - Chapter III -

### **Transfer effect in strategic learning**

#### **(Exp. 4 - extension of Exp.3)**

##### A) Introduction

###### **1) Thesis context**

In the previous chapter we present evidence of an interaction between the capacity of humans to engage in strategic learning emerging and the way the repeated social interaction is structured. The interaction took place in a 2x2 actions space modeled by a competitive game designed to induce specifically strategic asymmetry between the two players (endogenous advantage of one role over the other), by agencing the payoff matrix in order to facilitate only one's engagement in belief-learning, and creating for the opponent a focal point anchoring reward-based learning behavior. We showed that humans differ in their level of engagement in sophisticated belief (i.e. strategic) learning and that the framing effect of the game triggered inequity. Moreover, humans in the disadvantageous position in the repeated game interaction were constrained by their capacity to switch their attention away from the affective value of payoffs and use the information they provide to engage in a complex inferential process over the other's behavior. On the other hand humans in the advantageous position did not engage more in strategic learning but the context of the interaction facilitated the tracking and exploitation of the other's behavior. They also seemed to be less constrained by the interaction structure, and therefore guided more initially by their individual executive abilities to eventually adapt in a cost-benefit tradeoff fashion. What drives adaptation of the subjects endorsing this role remains hypothetical. Further investigation is required to precisely characterize the evolution of subjects' behavior within an interaction block, and between interaction blocks in order to better understand how the recent history of strategic interactions may affect subsequent adaptation of a subject's will to engage in a different strategic learning level.

###### **2) Scientific context**

The level-k framework suggests that sophisticated agents adapt to the level of sophistication of their counterparts in repeated interactive settings (i.e. repeated games). This feature has been directly implemented in the cognitive neuroscience literature by Yoshida et al (2010), who developed a Bayesian

model which infers directly the SL level of their opponent from the history of choice interactions and adjusts its own level to it in order to optimize its behavior. Devaine et al (2014) relaxed the assumption of fully optimal humans able to systematically track the SL level of their opponent and combined considerations about strategic sophistication embodied in the level-k framework to the optimal learning theory of the Bayesian scheme. The authors yet observed that, overall, the level of sophistication of their subjects as captured by their model was correlated across the computerized opponents, similarly to what we found in our own study which is presented in Chapter II. In line with previous experiments studying deviations from RL and engagement in belief-based learning (Seo et al, 2017; Spiliopoulos, 2013b), these results suggest that high SL subjects adapt their behavior to the opponent's past choices but that their SL level is bounded to some individual constraints that remain to be determined.

As detailed in the previous chapter, within the value-based framework, beliefs can be considered as transient representations of the action-outcome contingencies of the task. The advantage of estimating the model generating the outcomes in such environment is that once learned, it allows quick adaptation to changes in the probabilistic structure of the world (Chan et al, 2016; Humphries et al, 2012; Wilson et al, 2014). Recently it has been suggested that previously learned action-outcome contingencies can be recovered in case of environmental redundancy (Koechlin, 2016), making the emphasis on the transferable property of such mental representations (Collins & Franck, 2016).

Vickery et al (2015) showed that when humans play a symmetric game with two identifiable computerized opponents, they differ in the way they relate to their previous choices depending on the identity of the opponent so that they switch from one memorized strategy to the other according to the opponent they face. And even when not explicit, researches suggest that personal features such as task performance, preferences or social status can be learned (Boorman et al, 2013; Devaine et al 2017; Ligneul et al, 2016).

In our study the opponents were anonymized. Therefore only the experience from the previous repeated interaction could be used as a prior by a subject. In fact the degree with which priors over the other's intention influences the behavior during social interactions have been recently linked to the strength of coupling in activity between the rTPJ and the mPFC (Chambon et al, 2017), two regions also implicated in the computation of higher order beliefs in repeated game interactions (Hill et al, 2017). In line with this, we previously found some conjunctural evidence of anchoring effects on choice behavior (strategic thinking, depth of reasoning. Chapter II, Exp 2). However, how previous experience impacts the choice behavior in strategic interaction remains unexplored.

### **3) Hypothesis**

The results presented in Chapter II, suggest that in a situation of payoffs asymmetry subjects in the disadvantageous position were cognitively bounded in their engagement of high level of strategic learning. However, no difference in SL distribution was found when playing against a low vs. a high strategic opponent. An alternative hypothesis lies under the concept of transfer learning: Subjects might transfer some beliefs from the previous interaction onto the next, and thus not (only) being affected by the current opponent's play but (also) by the previous game experience.

The experimental design of the Experiment 3 presented in Chapter II.A, allows us to test specifically the hypothesis of a transfer from one opponent to the other, modulated by the strategic learning level of the subject.

## B) Methods

The study of possible transfer of beliefs within and between interaction blocks requires to subdivide our dataset into subgroups of subjects. Thus to give us the sufficient statistical power to investigate these 2 hypotheses we extended our dataset by running extra experimental sessions in the exact same conditions as Experiment 3 (Chapter II.A). We managed to recruit 58 extra subjects for a total of N=130.

## C) Results

Beforehand we present analyses verifying that increasing the dataset did not impact our previous results. As previously observed no effect was found either of the opponent on the SL level itself<sup>11</sup>, in both roles. In fact all the statistical effects found initially between the strategic learning engagement of the subjects and their choice behavior during the interaction were not only replicated but strengthened when running our analyses on the full dataset (not shown). The extended analysis also showed that for both players the higher their strategic learning level the less often they followed a win-stay lose-shift strategy, as well as the more frequent they switched action from one trial to another (**Fig. 1**), further confirming the pertinence of our computational measures.

---

<sup>11</sup> By default we use as a measure of SL level the relative fit between 2-Infl and Q-learning. As in the previous chapter, when not specified the results presented hold when using the relative fit of the (1-)Influence compared to the fictitious model and the Influence best fitting parameter  $\lambda$  value.

<b>A</b>					
% Switch	Estimate	SE	t	p	
<b>- Player 1</b>					
P1_SL_sub	<b>0.0063</b>	<b>0.0011</b>	<b>5.5836</b>	<b>1.38e-07</b>	
P1_SL_opp	0.0289	0.0162	1.7816	0.0772	
P1_SL_s ~ P1_SL_o	0.0026	0.0015	1.6987	0.0918	
$R^2 = 0.4767$ ; $F(123)=38.26$ , $p = <1.0e-30$				$BIC = -284.30$	
<b>- Player 2</b>					
P2_SL_sub	<b>0.0054</b>	<b>0.0009</b>	<b>6.0143</b>	<b>1.82e-08</b>	
P2_SL_opp	0.0483	0.0183	2.6386	0.0094	
P2_SL_s ~ P2_SL_o	<b>-0.0035</b>	<b>0.0013</b>	<b>-2.6867</b>	<b>0.0082</b>	
$R^2 = 0.2492$ ; $F(68)=13.94$ , $p = 6.61e-8$				$BIC = -273.59$	

<b>B</b>					
% WS-LS	Estimate	SE	t	p	
<b>- Player 1</b>					
P1_SL_sub	0.0018	0.0015	1.1866	0.2376	
P1_SL_opp	<b>-0.0941</b>	<b>0.0220</b>	<b>-4.2757</b>	<b>3.73e-05</b>	
P1_SL_s ~ P1_SL_o	<b>-0.0043</b>	<b>0.0021</b>	<b>-2.0946</b>	<b>0.0382</b>	
$R^2 = 0.2763$ ; $F(123)=16.039$ , $p = 6.83e-09$				$BIC = -205.03$	
<b>- Player 2</b>					
P2_SL_sub	0.0001	0.0011	0.1128	0.9104	
P2_SL_opp	<b>-0.0930</b>	<b>0.0222</b>	<b>-4.1964</b>	<b>5.08e-05</b>	
P2_SL_s ~ P2_SL_o	0.0011	0.0016	0.6747	0.5010	
$R^2 = 0.1680$ ; $F(123)=8.4783$ , $p = 3.57e-05$				$BIC = -223.93$	

**Figure 1.** Effect of the opponent SL on choice behavior - additional GLM results of the distinct influence of the subject SL and the level of the computerized opponent (low, high) on the average frequency of switch and win-stay-lose-shift based choices.

We tested directly if the SL level of the subjects was affected by the opponent. No difference was found for each of the 2 roles. As in our initial experiment we also observed in our extended dataset a higher congruence in SL level across the 2 blocks for Players 2 than Players 1 (% B1-B2 correspondence of low/high SL best fitting models for low/high opp.: Player 1, P1: 0.66, P2: 0.87 ; Fisher exact test:  $N = 102$ ,  $p = 0.0133$ ). As previously shown, the SL level was correlated between the 2 opponent blocks for Players 2 only (only when taking the relative fit of the Influence compared to fictitious  $r=0.4638$ ,  $p=9.9e-05$  or the  $r=0.4515$ ,  $p=1.6e-4$ , but not the full SL gradient captured by the relative fit of 2-Influence compared to Q-learning:  $r=0.2438$ ,  $p=0.0503$ ).

We focused next on the transfer learning hypothesis by investigating how the order, playing against either the low or the high opponent first, affects the choice behavior of the subjects endorsing each role, given their own SL.

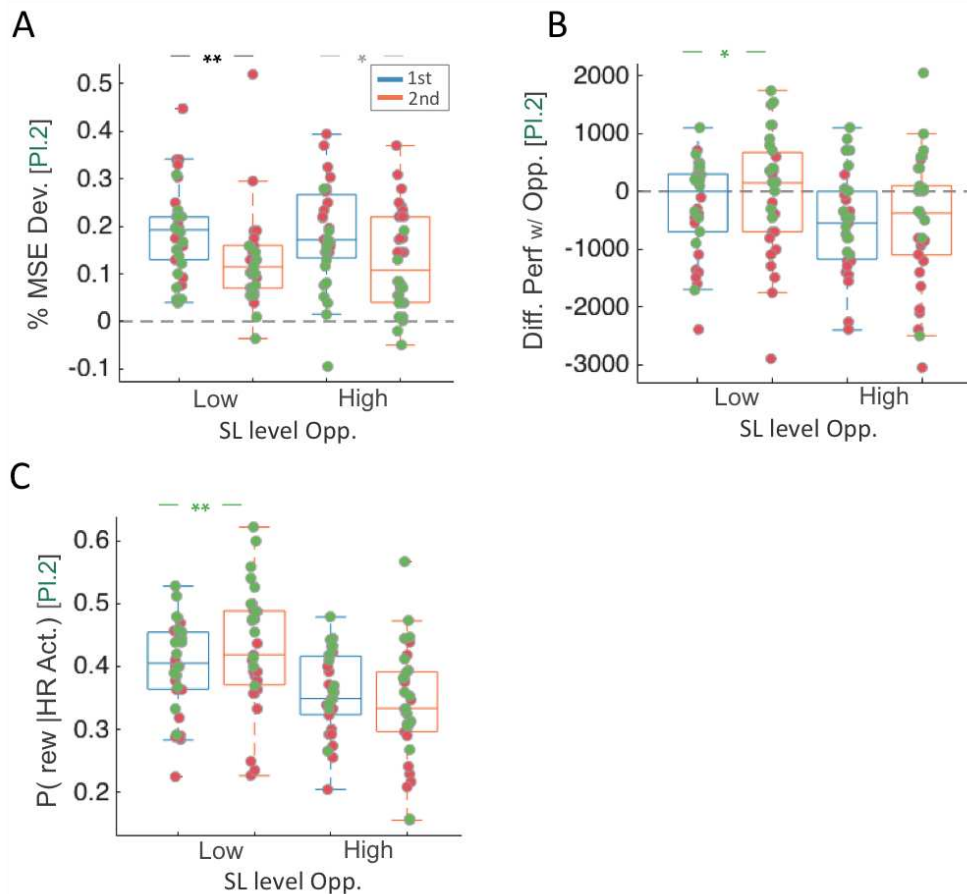
We first focused on Players 1 since the level of sophistication of the opponent impacted their choice behavior (see Fig.5 in Chapter II.A), but not their strategic learning level per se. To test if the difference in performance observed between each opponent was also affected by the order we extended our previous GLM to include this regressor in our analysis while controlling for the previous effects observed. We couldn't find any effect of the order of the opponent encountered on the deviation from MSNE, the absolute, nor relative, performance. However we found a small conjunction effect of both the order and the level of the opponent on their accuracy (**Fig.2.A**).

	<i>Estimate</i>	<i>SE</i>	<i>t</i>	<i>p</i>
<b>A</b>				
Accuracy (%Rew Act. (b/A) )				
- <b>Player 1</b>				
P1_SL_sub	<b>0.0047305</b>	<b>0.0013603</b>	<b>3.4776</b>	<b>0.00070012</b>
SL_opp	<b>-0.16459</b>	<b>0.018857</b>	<b>-8.7284</b>	<b>1.5284e-14</b>
Ord_opp	-0.04383	0.018365	-2.3866	0.018531
P1_SL_sub~ SL_opp	-0.00090875	0.0014661	-0.61984	0.53651
P1_SL_sub~ Ord_opp	0.0013104	0.0014705	0.8911	0.37461
SL_opp~ Ord_opp	<b>0.060942</b>	<b>0.025282</b>	<b>2.4105</b>	<b>0.01741</b>
<i>R2 = 0.5628 ; F(123)=26.39, p = &lt;1.0e-30</i>				<i>BIC = -293.35</i>
<b>B</b>				
% action(a/A)				
- <b>Player 2</b>				
P1_SL_sub	<b>-0.0035</b>	<b>8.96e-4</b>	<b>-3.9588</b>	<b>1.26e-4</b>
SL_opp	-0.0063	-0.014027	-0.4506	0.6531
Ord_opp	-0.0170	0.014432	-1.1776	0.2412
P1_SL_sub~ SL_opp	-0.0016	0.00092879	-1.7016	0.0914
P1_SL_sub~ Ord_opp	-1.47e-4	0.00093027	0.1589	-0.8741
SL_opp~ Ord_opp	<b>0.0388</b>	<b>0.018235</b>	<b>2.1323</b>	<b>0.0350</b>
<i>R2 = 0.5062 ; F(123)=21.03, p = 1.11e-16</i>				<i>BIC = -383.61</i>

**Figure 2.** Impact of the order of the opponent played on the MSNE deviation and accuracy. The extended GLM analysis was ran on the choice series of subjects in each role to dissociate the relative effect of the order of opponent play from the effect of the subject and opponent SL. (1: low first, 2: high first). Here are presented only the results of the models which fitted significantly better than chance the subject's behavior and show a significative effect of the order, i.E. MSNE deviation and accuracy.



When we ran the same analysis on Players 2, we found, in addition to the main effect of their own strategic engagement, a similar interaction influence of the order along with the SL of the opponent on their overall deviation of their choice distribution from MSNE (this effect was small, but consistent across SL measures, improved a bit the overall fit of the GLM and was also observed in ANOVA) (**Fig.2.B**). In fact when comparing the choice behavior data of Players 2 but dividing the data not only by the opponent level played but their order, a dissociation between high and low SL subjects was observed, with an effect marked for the highest half of Players 2 when opposed to the low opponent after having been exposed to a high SL Player 1 (**Fig.3**). This result is therefore coherent with our transfer hypothesis and could be interpreted as a facilitation effect of the sophisticated, and thus more accurate, beliefs formed throughout the previous interaction.



**Figure 3.** Evidence of transfer learning in Players 2. (A) the order of play affected Players 2 deviation from MSNE: subjects who played against the High SL opponent first (N=31) deviated less than the one who were opposed first to the low SL (N=34). (B,C) Improvement of the performance and accuracy when playing against the low SL after having

played against the high, observed only for high (median split) SL players 2 (Green). In green are represented the half of the Players 2 with the higher SL level (N=16), in red with the lowest (N=16).

---

To insure that this order effect was not driven by an increase in strategic learning sophistication when playing first against a high opponent, we tested the alternative hypothesis of an effect of the opponent's order directly on the SL level of Players. Unexpectedly this analysis revealed a higher Strategic engagement of Players 2 compared to Players 1 when opposed to this low SL opponent ( $D(128) = 0.2615$ ,  $p = 0.0188$ ;  $U(128) = 1538$ ,  $Z = 2.6727$ ,  $p = 0.0075$ ). We tested if this new effect was mediated by the order (allowing for interaction effects). We found for Players 2 (weak<sup>12</sup>) evidence of an effect of the opponent's SL and the order. Nevertheless this conjunction effect of the opponent SL and the order on the level of strategic engagement of the subject in the strategically disadvantageous position, was also observed, albeit small, when running an ANOVA on the all dataset, entering the Role as a factor (**Fig. 4.A**). The order did not seem however to directly impact the overall stability of their SL level (prop. same low/high SL IBM: low-high opp.= 0.65, high-low= 0.67) from one opponent to the other. When dividing each condition group (role, order), between the subjects who engaged the less and the one who engaged the most in sophisticated strategic learning we observed that only highly sophisticated Players 2 increased their SL level when playing against the high SL opponent in comparison to the low SL (**Fig. 4.B**). This trend might actually explain the lack of correlation in SL level between blocks previously observed only when using as a measure of the strategic learning engagement the amplitude offered by the entire model space, i.e. the relative fit of the 2-inf compared to the Q-learning.

---

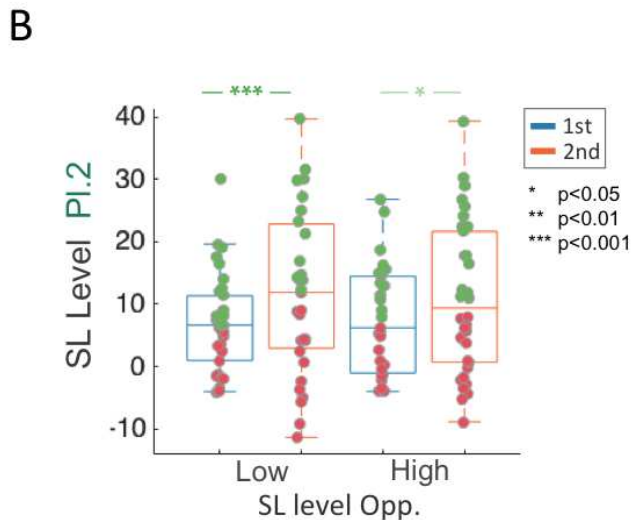
<sup>12</sup> weak effect is defined here as a poor fit of the GLM model,  $R < 0.1$ ; a lack of replicability of the effect across our 3 SL measures, namely the relative fit of the 2-Inf compare to RL, the relative fit of the Influence compared to fictitious and the Influence best fitting parameter  $\lambda$  value. Also using an ANOVA analysis the effect was only marginal.

**A**

[R.Fit 2-Infl.]	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
RoI	664.7	1	664.674	6.94	0.0089
Opp	2.9	1	2.855	0.03	0.8631
Ord	99.1	1	99.102	1.03	0.31
RoI*Opp	0	1	0.016	0	0.9897
RoI*Ord	12.7	1	12.692	0.13	0.7161
Opp*Ord	930.1	1	930.115	9.71	0.002
Error	24230.8	253	95.774		
Total	25927.5	259			

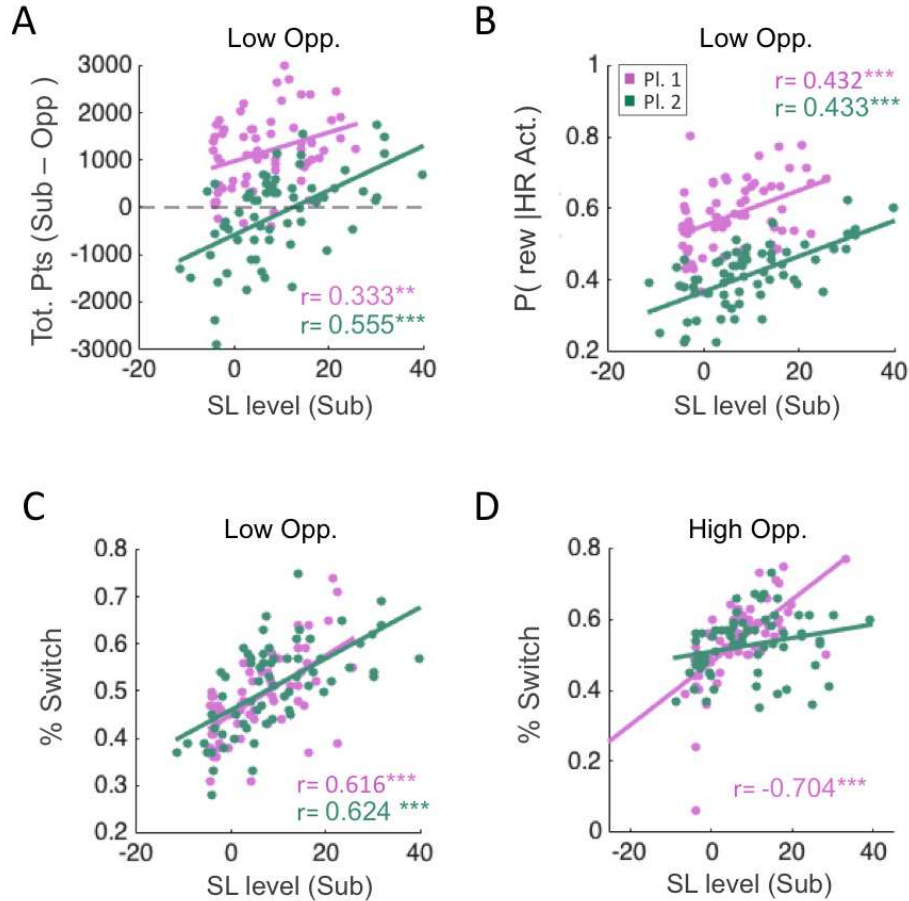
[Best λ Infl.]	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
RoI	0.578	1	0.57798	4.19	0.0416
Opp	0.0629	1	0.06286	0.46	0.5001
Ord	0.0304	1	0.03041	0.22	0.6389
RoI*Opp	0.9284	1	0.9284	6.74	0.01
RoI*Ord	0.0925	1	0.09249	0.67	0.4134
Opp*Ord	1.1704	1	1.17039	8.49	0.0039
Error	34.8692	253	0.13782		
Total	37.6636	259			



**Figure 4.** Relative effect of the order on the SL level of the subjects (A) ANOVA conducted on the SL level of the all population (relative fit of 2-infl and Influence best fitting parameter value) with the role in factor along with the opponent SL and the order of play. (B) Effect of the order on the SL level of Players 2. The SI level is captured here by the relative fit of the 2-infl, the results remained unchanged for the Low opponent with the 2 others measured (Rel.Fit Infl. and Influence parameter), note that the effect on the High opponent does not hold with the other SL measures (shaded green star). Only the half of Players 2 with the higher SL level (green - median split), are affected by the order.

This additional effect suggests that the Players 2 who engaged already in a high strategic learning against the High SL subject remained highly engaged when playing against the low SL opponent. One interpretation is that when they play first against an opponent modelled by low strategic learning algorithm (i.E. fictitious), sophisticated Players 2 managed to overcome their disadvantage and were thus only rationally bounded. However when confronted first to a high SL Influence model, their performance was constrained by individual, probably cognitive, limits (only the one able to fully engage in strategic learning reduced their disadvantage).

Indeed Players 2 with a higher strategic learning boundary, when confronted to the low SL opponent, improved their behavior (**Fig. 5.A,B**) and even managed to increase their absolute performance ( $r=0.4318$   $p=3.0e-4$ ). According to our computational approach, they achieve to do so against the low SL opponent, by tracking the interplay between their own behavior and the one of their opponent, leading them to deviate from the focal point to play closer to the MSNE and to choose more accurately the high reward action. Additional evidence of exploitation of the low SL opponent behavior can be observed in Players 2 play: the higher their SL level the higher their accuracy and the higher frequency of switch. And conversely not only both were impaired against higher SL opponent, (**Fig. 5.C,D**), but a correlation can then be observed between their SL level and their choice reaction time, with an increase after a loss and a decrease after a win (win:  $r= 0.335$ ,  $p= 0.006$  ; loss:  $r= 0.315$ ,  $p= 0.011$ ).



**Figure 5.** Correlation between the SL level of the subjects in both role and selected choice behavior. (A,B) correlation with performance and accuracy when opposed to low SL opponent resp. (C,D) Difference in correlation with frequency of switch between the 2 types of opponent observed for Players 2 only.

These results seem congruent with the hypothesis that Players 2 due to their disadvantageous position in the game are constrained to their own cognitive capacity, only reached by high SL subjects in this task when playing against a high SL opponent.

To further challenge this hypothesis we analysed the additional Working Memory task that the subjects had to perform at the end of this experiment. This time, working memory capacity was measured using a N-Back task. We confirmed the trend observed in our second experiment: both 2-back and 3-back performance correlated to the SL level of the subjects endorsing the role of Players 2 (2-Back:  $r=0.3025$   $p=5.19e-4$  ; 3-Back:  $r=0.2971$   $p=6.58e-4$ ), no correlation was found between working memory capacity and

SL level in Players 1. But more interestingly, this time we could look at this correlation for each type of opponent separately, and we found that the correlation was only driven by the block where the Players 2 were opposed to the high SL opponent (2-Back:  $r= 0.3980$ ,  $p=0.0011$ , 3-Back:  $r= 0.4120$ ,  $p=7.18e-4$ ). Players 1 on the other hand, already in an advantageous position, did not seem to be directly affected by the order of play, as if the sophisticated ones only needed to adapt to the opponent. Their choice behavior was thus not directly affected by the opponent (**Fig. 5.C,D**), and no difference in choice reaction time after a win or a loss was observed in both blocks.

## D) Discussion

The previous chapter (II) presented evidence for the predictive power of the value-based decision making framework developed in neuroeconomics applied to repeated game interactions. Our previous results (Chapter II) suggested that the behavior of the subjects interacting in the disadvantageous position was driven by their own level of strategic learning, which was itself not directly impacted by the level of sophistication of the computerized opponent encountered. Based on recent research in the field (see Wikenheiser et al, 2016 [17]), we exploited the experimental design of the Experiment 3 (Chapter II) and ran a follow up study by including additional subjects to our dataset, to test the specific hypothesis that humans who are capable of engaging in higher-order belief (strategic) learning would transfer their sophisticated beliefs from one opponent to the next in order to improve their performance.

The results presented here consolidated our previous results but did not allow the rejection of our null hypothesis of an absence of transfer of sophisticated beliefs across the interaction blocks. Our analysis revealed that the high level of strategic sophistication of the opponent first encountered directly enhanced the level of engagement in high order belief-learning of the subjects playing in the disadvantageous role (i.e. Players 2).

The transfer effect we observed in the present (extended) dataset appeared to be specific to the strategic learners (subjects capable to engage in high strategic learning level), who endorsed the disadvantaged role. This result seems to argue in favor of a dissociation between bounded cognition and bounded rationality in equilibrium play as suggested by Friedenberg et al (2017) and discussed in chapter II. Indeed, according to this theory, there are two possible reasons why not all participants reach equilibrium, and *a fortiori* in repeated interactions, engage in higher strategic learning. First the incentive to engage in such costly cognitive process is not high enough (according to an individual trade off threshold), as it seems to be the case when subjects endorse the dominant position in the strategically asymmetric

interaction since this position ensures an outcome superior to the one of their opponent for minimal effort (bounded rationality). Second the observed heterogeneity also reflects between-subject differences in cognitive capacities, preventing some participants to fully engage in strategic learning even if it was maladaptive not to do so (bounded cognition). The results presented here then refine this apparent dichotomy as they suggest that even when the position is disadvantageous, a cost-benefit tradeoff arises for subjects capable of high strategic learning.

This interpretation lies under the assumption that the engagement in (high order) belief-based learning requires higher mental effort, a hypothesis at the heart of the bounded rationality framework (Polonioli, 2016). Interestingly a recent review article proposed that cognitive control is at the origin of such trade-off between the cost of computation and its expected benefits (Shenhav et al, 2017). Some authors indeed showed a correlation between the cognitive control engagement and the switch between model-free and model-based learning (Otto et al, 2014). The idea that cognitive control might subserve the engagement in model-based learning has been evoked as a potential link within the value-based decision making framework between motivation and cognitive control through incentives (Botvinick & Braver, 2015). And recently authors presented compelling evidence that the level of cognitive control and the balance between cost and expected reward could indeed drive the engagement in model-based RL (Deserno et al, 2015; Kool et al, 2017). A posteriori, we could have interpreted the correlation found by Yoshida and colleagues (Yoshida et al, 2010) between the adjusted level of strategic learning (from the direct estimated sophistication of the opponent computed by their fully optimal model) and the dlPFC, as a first evidence of cognitive control implication in the level of strategic learning during a repeated game interaction (Duverne & Koechlin, 2017).

An alternative hypothesis would be that playing against a strategically sophisticated learner emphasizes the strategic nature of the interaction and somehow facilitates the formation of higher-order beliefs. To our knowledge this hypothesis is not rooted in one pre-existing theoretical framework. However, the literature on value-based decision making offers several ways to explore possible interpretations. First, it has been suggested that attention can be driven by the value attached to stimuli through model-free learning (Le Pelley, et al, 2016; Preciado et al, 2017). One possibility is thus that in our task the saliency of the focal point is modulated by the subjective value eventually attached to it, and that once the subjects in the disadvantageous position formed a high-order representation of the strategic interaction, the attention paid to the high payoff / suboptimal action eventually decreased, facilitating their strategic learning engagement in the next block. A second interpretation lies on the social effect of the interaction itself. Authors have shown that when engaged in a situation involving a counterpart the other's actions tend to influence our own decision-making process even when it is not optimal to do so (Apps & Ramnani, 2017; Suzuki et al, 2016; Wittmann, et al, 2016). In strategic interactions, higher order beliefs are modeled as simulations embodying the interplay of the past behavior of both the subject and her opponent. Thus

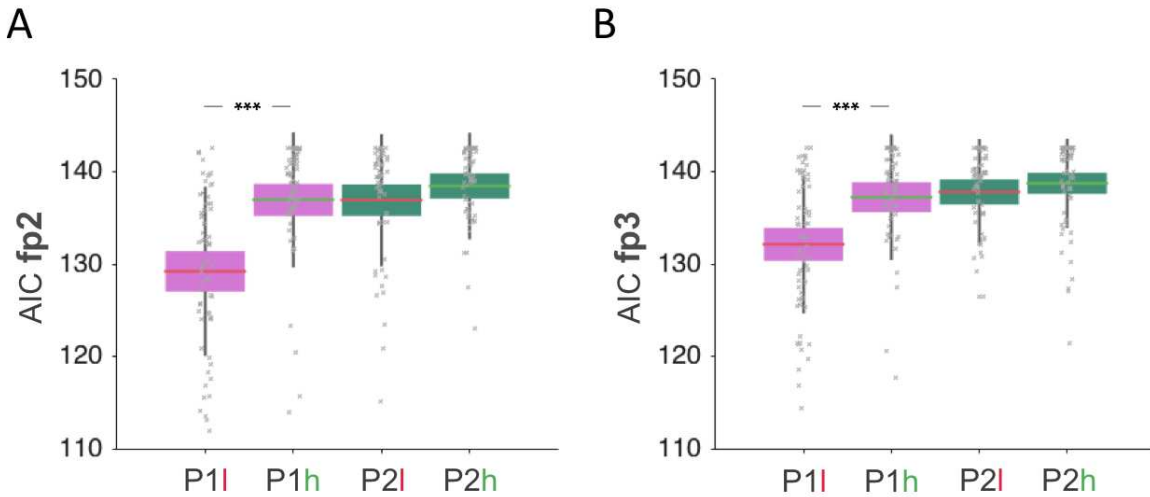
perspective taking facilitation could enhance such representations (Nicolle et al, 2012). This effect could also be amplified by the joint agentivity emerging from dynamic social settings (Bolt & Loehr, 2017), so that realizing that our own actions influence the opponent might improve the accuracy of the learned generative model (Di Costa et al, 2017). Manipulating the noise in the action selection (i.e. the inverse temperature parameter of the softmax function in Equation X) of the belief-based learning algorithm generating the opponent's choice could impact the perception of joint agency, and help disentangle these two interpretations.

We argue that these two hypotheses, of respectively endogeneous and exogeneous effects in transfer learning, could be tested experimentally. One way to do it would be to manipulate the strategic asymmetry. This can be done using the simulation process presented in the previous chapter (Exp.1 in chapter II.A, chapter II.B) to increase or decrease the strength of the focal point by carefully changing the payoff values of the game matrix.

Finally, we wanted to emphasize a main difference found between the 2 experiments, when participants were playing against a humans vs. a computerized algorithm: the higher the SL level of Players 1 (advantaged role) the more they deviated from the MSNE when playing against the low SL opponent ( $r=0.42802$ ,  $p=3.75e-4$ ), an effect that seems to be specific to this condition (GLM  $F(126)= 8.66$ ;  $p= 2.88e-05$ ;  $sub\_SL$ :  $t=2.5492$ ,  $p= 0.0119$ ;  $opp\_SL$ :  $t=2.9391$ ,  $p=0.0039$ , no interaction effect). Since performance and accuracy increased with a subject's SL level, we hypothesized that this core difference was driven by an exploitation of the Low SL computerized opponent. One possibility would be that these subjects managed to exploit the regularity emerging from the fictitious play behavior of their opponent. We investigated this hypothesis by comparing the quality of fit of the two pattern versions of the fictitious play developed by Spiliopoulos (2013a), *fp2* and *fp3* (Chapter II, Appendix 2) in each opponent block for each player role. The pattern fictitious model *fp2* tracks how many times two (or three for *fp3*) temporally consecutive sequences of actions have been observed, and then computes from the estimated (weighted) frequency the conditional probability of an action being played based on her last two (or three) past actions. As in Exp.1 (Chapter II), at the population level, the fit of the Influence model was higher compared to the *fp2* and *fp3* models (AIC comparison both opponents: *fp2*,  $U(128)= 3696$ ,  $z=7.8404$ ,  $p=4.4897e-15$  ; *fp3*,  $U(518)= 17379$ ,  $z=9.5848$ ,  $p=9.2603e-22$  ; Similar results were obtained for each opponent separately). When comparing the fit of the two models in each opponent block for each player, we observed that the models better captured the choice behavior of players 1 when playing against the low SL opponent only (**Fig.6**). However even in this condition, the Influence model still best fitted their overall behavior (Players 1 low SL opponent: AIC(*fp2*),  $U(128)= 1318$ ,  $z=3.6971$ ,  $p=2.1807e-04$  ; AIC(*fp3*),  $U(128)= 943$ ,  $z=5.4432$ ,  $p=5.2325e-08$ ). These results thus suggest that players 1 might have used statistical redundancies emerging from the behavior of the low SL opponent in order to exploit its strategy and maximize their



earning, even if it meant deviating from the MSNE distribution (i.e. from mutual best response, to unilateral best response).



**Figure 6.** Better fit of the pattern fictitious models only on Players 1 against low SL opponent. (A) Fit of the two choices pattern fictitious (fp2) in each opponent condition for each role. (B) Fit of the three choices pattern fictitious (fp3) in each condition.

- Chapter IV -  
**From strategic learning to Pattern detection  
in competitive interactions  
(Exp. 5, 6)**

The experimental work presented in this chapter has been done in collaboration with Larsen Tobias.

**I - Pattern detection in strategic dyadic interactions (Exp.5)**

A) Introduction

**1) Thesis context**

In our previous experiments (Exp.1-4) we found that most of our subjects engage in some form of strategic learning during competitive interactions. Subjects, however, compute the information provided by the history of play at different depths, or levels of sophistication. This results in the formation of different orders of beliefs leading subjects to anticipate differently the opponent's next action and ultimately ending in observable heterogeneity in best responses. Through several forms of repeated game experiments conducted in laboratory, we showed that such between-subject differences in strategic engagement can explain part of the variance observed in choice sub-optimality. We also presented evidence that the payoff structure of the game can drive different behaviors among individuals depending on the way it facilitates higher engagement in strategic learning and their propensity to do so. Finally, our results suggest that human learning sophistication operates upon the implicit computation of a cost-benefit ratio of their cognitive engagement into the social interaction.

Altogether the results presented in the first part of my PhD work (Chapter II and III) lead to the conclusion that humans are able to consider their opponent's behavior in their own learning process and thus engage in belief-based learning. However individuals vary in the sophistication of their inferences over the strategic implication of the opponent's behavior (high order belief-based learning).

Nevertheless, similar to the vast majority of the existing literature on human strategic learning and on computational models of human learning and decision-making abilities, this work was based on a

statistical frequentist approach. More precisely, the models for different levels of strategic learning (SL) consisted in estimating the frequency with which the opponent selected one option over another, assuming that subjects are just estimating average frequencies of choices over a certain window of trials, rather than detecting finer choice regularities in the behavior of the opponent. Moreover, these models also consider that the opponent is himself doing similar frequentist estimations. These models thus use these different estimations to choose the most rewarding option under the assumption that the opponent will continue its choices with the same learning strategy. However, the behavioral data acquired during the first part of this PhD work suggest that subjects may be doing more than simply estimating choice frequencies (Exp. 4, Chapter III). In some cases, the subjects seem to be able to identify regular patterns in the sequence of the opponent's choices. For instance, if the opponent performs the following sequence: Left Right Right Left Right Right, there is more information in this behavior than simply considering that the opponent selects Left 33% of the time and Right 67% of the time. There is a certain pattern or structure to extract from this sequence so that one can more precisely predict the next move, rather than simply considering that Right is the most likely next option. In this particular example, the actual most probable move from the opponent after this sequence is Left, even if its overall frequency is lower than that of Right. Thus if subjects are able to extract and respond to such patterns, then their behavior may sometimes deviate from what can be captured by models adopting a pure frequentist approach, while still performing a high level of strategic thinking and learning.

Here the goal was thus to more systematically investigate the question whether individuals are able to detect repetitiveness in their opponent's choices, and form beliefs over such choice patterns to maximize their final outcome. We moreover explore how this type of social learning interacts with the strategic learning engagement to form accurate beliefs over their opponent's (choice) behavior.

## **2) Scientific context**

The pattern learning hypothesis has been very rarely explored in the field of game theory (Sonsino, 1997), until recently. In a recent in-depth work (Spiliopoulos, 2012, 2013a), Spiliopoulos hypothesized that in repeated (competitive) games, humans can exploit regularities in the choice series of their opponent in order to improve their beliefs' accuracy and maximize their final outcome. To test this hypothesis he extended the fictitious play model introduced previously to allow for the computation of the conditional probability of choice given the past 2 choices of the opponent. This way the model (*fpn*) tracks the *n* choices patterns in her behavior, update the probability (decayed frequency) associated to the *n!* patterns, and best respond to it. The *fp2* model has been shown to outperform classic fictitious play in

human-human repeated game interactions (Spiliopoulos, 2013a). If it seems clear that humans are not effective randomizers (Gauvrit et al, 2017), the model fitting approach used by Spiliopoulos however does not ensure that the subject's opponent did actually display obvious, tractable and exploitable patterns.

Actually the additional computational results presented at the end of our chapter II (“A) Additional discussion”) suggest that, in a repeated 2x2 competitive game, humans engage in iterative inference over their opponent's behavior instead of computing the probability of their opponent's choice conditionally on the pattern emerging from her history of play. On the other hand, the results presented in chapter III show that high SL subjects playing in the advantageous position and confronted to a computerized fictitious play opponent (low SL) might exploit regularities in its play to improve their choice accuracy and maximize their final earnings (even if this implies deviating from the MSNE choice distribution).

Thus, to our knowledge, no strong evidence is available showing that humans can indeed learn from temporal sequences (patterns) in her opponent's choice series during a repeated game interaction. Nevertheless, human's ability to detect statistical redundancies in the environment has been extensively explored in psychology and recently in cognitive neuroscience (Schwarb & Schumacher, 2012). Authors consistently showed that humans detect deterministic (implicit) patterns (increased prediction accuracy, decreased reaction time) in tasks using temporal sequences of identifiable stimuli (Baker et al, 2014; Schwarb & Schumacher, 2012), and this in a quasi-optimal fashion (Meyniel et al, 2016; Yu & Cohen, 2009). Moreover it has been shown that humans tend to respond inappropriately to local statistical redundancy (emerging patterns) even in purely random sequences (Hahn & Warren, 2009; Oskarsson et al, 2009). For instance in probabilistic decision tasks, such as the classical two-arm bandit task, humans can respond to irrelevant regularities in outcome temporal sequences, leading to overall deviation from optimal behavior which consists in this case in exploiting the option leading to the highest expected payoff (Hanaki et al, 2017). Ultimately Lahav et al (2009) showed that humans can even best respond to action-specific patterns in a similar choice task, a repeated game “against nature” where the (hidden) transition probability between each of the 3 actions was predetermined.

We propose that this probabilistic pattern generation approach used in Lahav et al, transposed to strategic games, can offer a controlled setting to test the hypothesis that humans are able to detect patterns in their opponent's choice series. This way we can really measure the participant's accuracy to detect the induced statistical choice redundancies and best respond to the learned patterns while making sure that these patterns 1) do not depend on the subjects' behavior (and thus remain constant across subjects), and 2) generate overall choice frequencies that follow MSNE distribution.

An inherent drawback of this experimental strategy is that, by using such fixed pattern-driven strategy, the

opponent does not adapt to the behavior of the subject and thus does not maximize payoff *per se*. This pitfall is inevitable if we want to control for pattern detection in strategic setting where the outcome of one's choice depends directly on the decision simultaneously made by the opponent. We thus developed a variation of this algorithm, closer to the *fjn* model of Spiliopoulos in the sense that it generates probabilistic patterns directly over the subject's past history of play. To further control for the maximization effect we also included another computerized opponent which follows deterministic patterns this time, and switches its predetermined choice sequence when exploited by the subject.

### **3) Hypothesis**

We hypothesized that in a repeated game setting where the computerized opponent would implicitly use a pattern-driven strategy, subjects would manage to learn this feature from the opponent's choice history, in order to use this type of information to better predict their opponent's next choice and thus maximize their total earnings in the game interaction. We expect subjects to vary in their pattern learning engagement, and predict that their ability to do so is a stable and transferable cognitive trait across opponents, that is, an individual subject should show the same propensity to identify and respond to patterns when confronted with different opponents.

A secondary hypothesis we formulate is that it will be more difficult for subjects to detect that opponents track and use patterns embedded in their own choice series -- a higher order pattern strategy corresponding to some sort of Influence model based on pattern detection rather than simply choice frequencies. We expect subjects who are better at doing so to also be better at engaging in higher levels of strategic learning.

We developed a novel task from the framework of repeated 2x2 game interaction used in computational neuroscience. In this experiment subjects had to interact, in a random order, with 4 different computerized opponents in an interactive setting. The four opponents gradually varied in the strategy implemented to interact with the participants, going from probabilistic choice patterns blind to the subject's choices, to a maximization strategy adapting on-line to their opponent's behavior.

## **B) Methods**

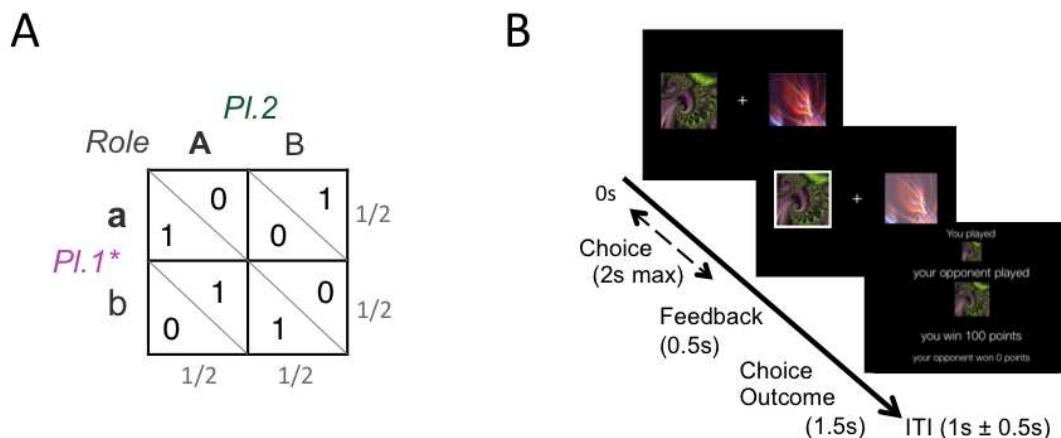
### **1) Participants**

67 participants (31 male, 36 female; ages 18–30) took part in the experiment. All participants were right-handed, medication-free, with normal eyesight, no history of neurological disorders. The Ethics Commission of the University of Trento approved the experiment. Informed consent was obtained from each subject before the experiment. Data collection was performed blind to the conditions of the experiment.

## 2) Experimental design and task

The experiment consisted in 4 blocks of 102 repetitions of a 2x2 hide and seek game, each block against a different computerized opponent.

Points earned at each trial were accumulated through each block and summed up to determine the final payoff, which would ultimately be converted to euros according to a predetermined rule. At each trial the two game actions, represented by randomly assigned colored fractals, were presented for 3s to each player. The choice was made by pressing the corresponding button (left or right). 4s after the trial onset, players were provided with the outcome feedback of their choice for 2s. During outcome feedback the two fractals chosen respectively by the opponent and the subject were displayed along with a sentence indicating if they won or lost, and the corresponding points for each player (“you” vs. “your opponent”) highlighted (**Fig.1.B**). The payoff matrix of the game is presented in **Fig.1.A**. Subjects always played in the Role 1, so that the rule was simply: “in order to win, try to select the same fractal as the opponent at the same trial”.



**Figure 1.** Task Design. (A) Symmetric game. Each option (a/A and b/B on the figure corresponded to two different fractals in the experiment, randomly picked at the beginning of the experiment). The participants always played as

Player 1. (B) At each trial participants had to choose between two fractals, randomly drawn and assigned at the beginning of the experiment. Once the choice was made, the choice outcome was displayed presenting both the participant and her opponent's choice along with the points earned in this trial. In case the choice was not made in time, the trial was considered as missed and lead the participant to automatically lose (0 pts). Since participants endorsed the Player 1 role, the rule was the same for everyone: "in order to win, try to choose the same fractal as your opponent at the same trial".

---

Each of the four opponents used a different strategy of adversarial play in order to maximize their own pay-off at the expense of the subject. Three of them were following pattern rules, among which two were probabilistic (O,S) and one deterministic (D). The fourth opponent used a belief-learning rule (modeled by a weighted Fictitious play algorithm) tracking and updating at each trial the subject's probability of choice from her past play (**Fig.2.A**).

The two probabilistic opponents differ in their level of implementation of the pattern rule.

First, Opponent O (for "Opponent-based") followed a probabilistic rule so that the choice made in the last two trials by the algorithm determined its own next choice with a certain probability. Two types of probability of choice following a pattern combination were implemented, either low (0.7) or high (0.9). This allows us to test specifically if subjects actually learn the full two-choice patterns, or rather simply use a two-back learning strategy pairing the choice at t-2 with a rough probability of choice. The probabilistic nature of this algorithm ensured that 1) no meta-pattern emerged from the choice series of the opponent leading to confusion about the number of past choices constituting a pattern, 2) the choice patterns were less obvious, and 3) the overall opponent's behavior looked realistic in the strategic sense to the eye of the subject. Probabilities associated to each pattern combination were chosen so that the overall frequency of each action available to the opponent matches the action distribution prescribed by the Mixed Nash Equilibrium Strategy (MSNE:  $p(A)=0.5$ ) (**Fig.2.A**).

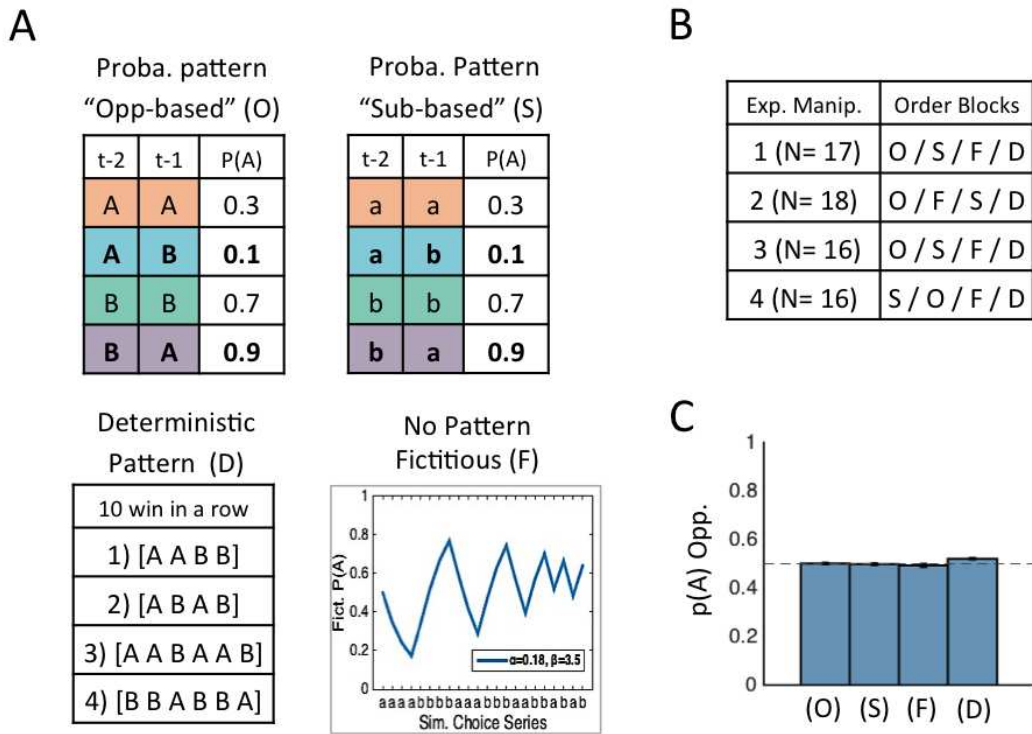
Second, Opponent S (for "Subject-based") follows the same probabilistic rule associated to each combination of past two choices as Opponent O, except that the choices pattern determining the opponent's next choice are the one of the subject, not its own. Thus the subject's own behavior triggered directly, with a fixed probability, the next action of the opponent. This opponent S was made to be a maximizing version of the Opponent O. This algorithm could thus be seen as a intermediate between a pattern-driven opponent (S) and a fictitious play, in the sense that it tracks regularities in the subject's choices (to not estimate probability distribution over her two actions, but relative frequency of two last actions patterns - much like *fp2* (Spiliopoulos, 2012), however the opponent does not best respond to it but follow a probabilistic pattern ("if the subject played "a" at t-2 and "a" at t-1, then select "B" with

$p=0.7$ ). In that sense detecting patterns in this block requires a higher level of strategic awareness, so that not only the subject must realize that Opponent S follows patterns, but also that her own choices are affecting its behavior and thus that the patterns could be determined by their own behavior.

Third, Opponent D (for “Deterministic”) was meant as a control for pattern detection ability, and was simply an algorithm following a fixed, determined, two-choice pattern. In order to maximize the information provided by this block we implemented a rule so that the type of fixed patterns changed once the subject had learned the current pattern successfully (the threshold was fixed at 10 win in a row), this way we could compute different measures quantifying the subject’s ability to learn non-realistic, fixed, choice patterns. This deterministic pattern was always played in the last block to avoid any confound with their performance in the three other blocks of interaction, whose order was randomized across subjects (**Fig.1.B**). To control for the adequacy of our algorithmic design, we insured that the choice behavior of the four opponents indeed followed the MSNE distribution (**Fig.1.C**). No correlation was found between the participants overall performance (total points) and the proportion of choice “A” in any of the four opponents.

This design was optimized from pilot studies (**Fig.S1**) in two ways. First, it seemed that subjects were better at discriminating the high probability pattern from the low one when they were paired to the pattern corresponding to two different choices in the last two trials (AB and BA - **Fig.1.A**). Second, pilot results suggest that subjects were better at detecting patterns in the opponent’s behavior when the rule was to choose the same action as the opponent to win (**Fig.S1.D**).



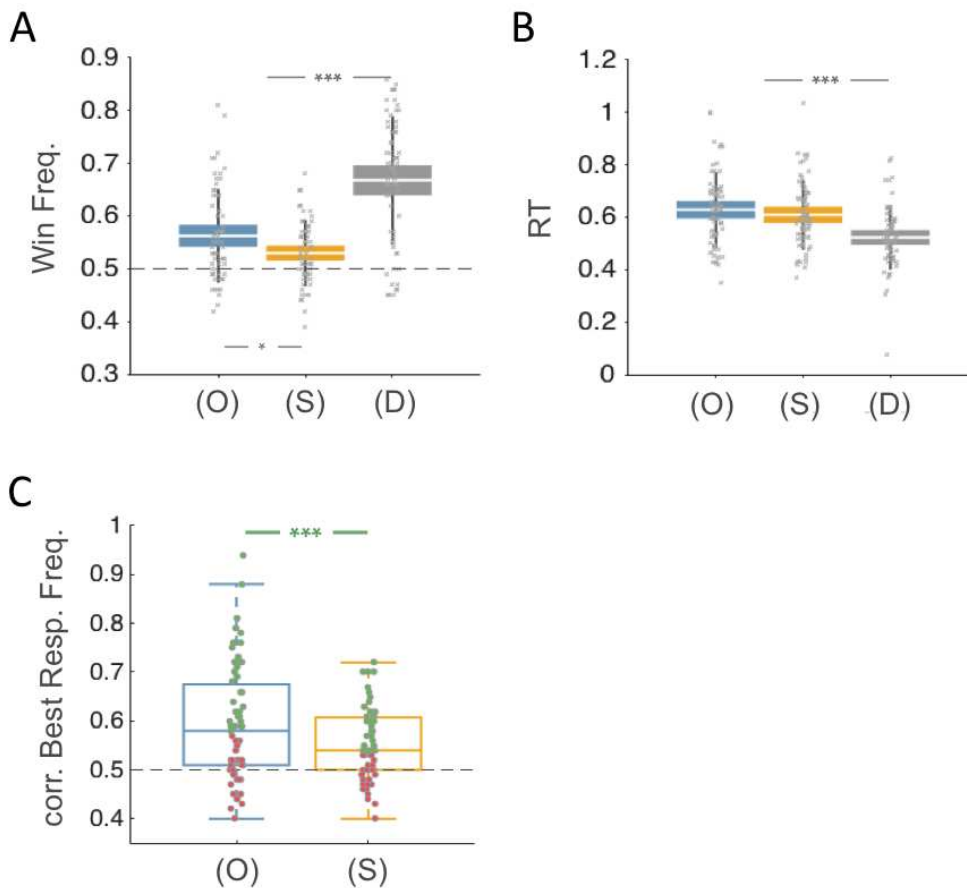


**Figure 2.** Experimental Setting. (A) Four computerized opponents divided into four blocks of 102 trials, each one facing a different opponent: three pattern opponents, of which two were probabilistic, and one fictitious-playing opponent. (B) Experimental manipulation where the order of the four pattern blocks with different opponents varied. (C) On average the choice behavior of each of the four computerized opponents followed the MSNE distribution.

## C) Results

We first look at the overall performance in each pattern condition to test our main hypothesis that subjects are performing on average better than chance in the pattern blocks, and are better in the "Opp-based" (O) than in the "Sub-based" (S) blocks. On average subjects performed better than chance (won more point than randomly drawn) in all the 3 pattern blocks (**Fig.3.A**). They were much better but also faster in the deterministic pattern condition than in the 2 probabilistic versions, thus confirming the control nature of this block (**Fig.3.A.B**). However we found only a slight difference in percentage of win between the opponent O and the opponent S ( $U(132)=1.9668, p=0.0492$ ). In these 2 blocks, the game was probabilistic, meaning that in some trials some players might have detected the pattern in the opponent

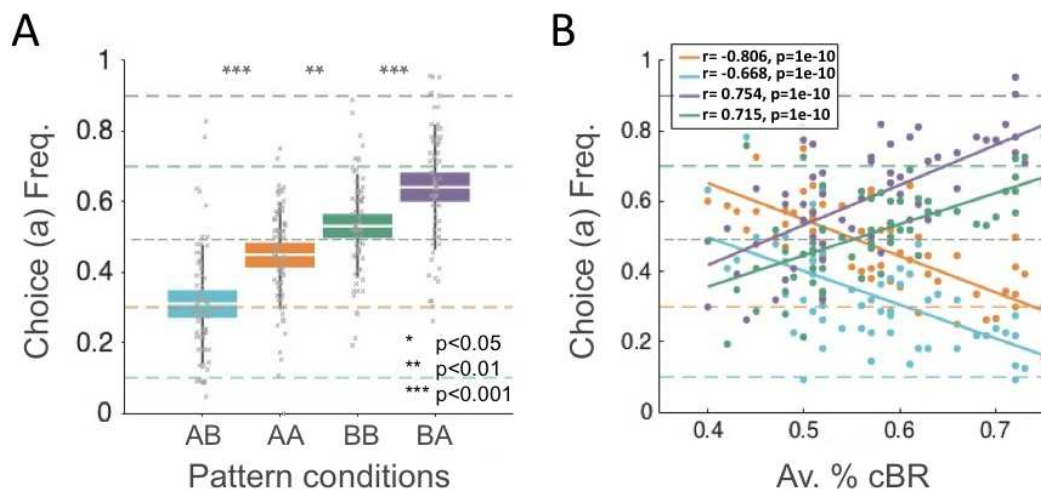
(or their own) past 2 choices and responded accurately to it, but the opponent could have played otherwise, leading the subject to lose despite her correct best response. These trials, thereafter called unpredictable trials (in opposition to the predictable trials where a subject who has accurately learned the pattern contingencies and best respond to it would win), represented around 20% (on average across the 4 pattern conditions) of the total block trials. We thus defined individual performances as the frequency of correct best-response in a block, i.e. the amount of action selected that corresponded to the opponent's most probable action, independently of the actual agent's choice. Taking this measure of performance however erased the small difference in percentage of win found previously between the 2 blocks (**Fig.3.C**). But when plotting the population distribution the heterogeneity among the subjects became obvious: some subjects failed to learn the patterns and performed around chance level, while some managed to learn it and perform much better than chance. When we then considered the half of the subjects who performed the best in each of the 2 blocks (**Fig.3.C, green data points**), their average percentage of best response was much higher when the choice patterns were in the opponent past choices (O), than in their own past choice (S) ( $U(746)=3.4386$ ,  $p=5.85e-04$ ).



**Figure 3.** Performance comparison across pattern blocks. (A) Subjects on average did win slightly more in the “opponent-based” (O) compared to the subject-based (S) probabilistic pattern block. However they won significantly more in the block where the pattern in the computerized opponent choice was deterministic compared to the 2 blocks. (B) Similarly subjects were on average faster in the Deterministic pattern block, while no difference between the 2 versions of the probabilistic pattern was found. (C) No difference in correct Best Response frequency was found between the 2 blocks O and S at the population level. When splitting the subjects between low (red) vs. high (green) correct Best response rate, a strong difference in performance appears between the 2 blocks, with an advantage in the block where the patterns in the opponent’s behavior was driven by its own past choices.

To insure that the subjects in the “Opp-based” block actually learned the patterns in their opponent’s past choices we looked closely at the distribution of best-response frequency across the 4 pattern conditions within the interaction block.

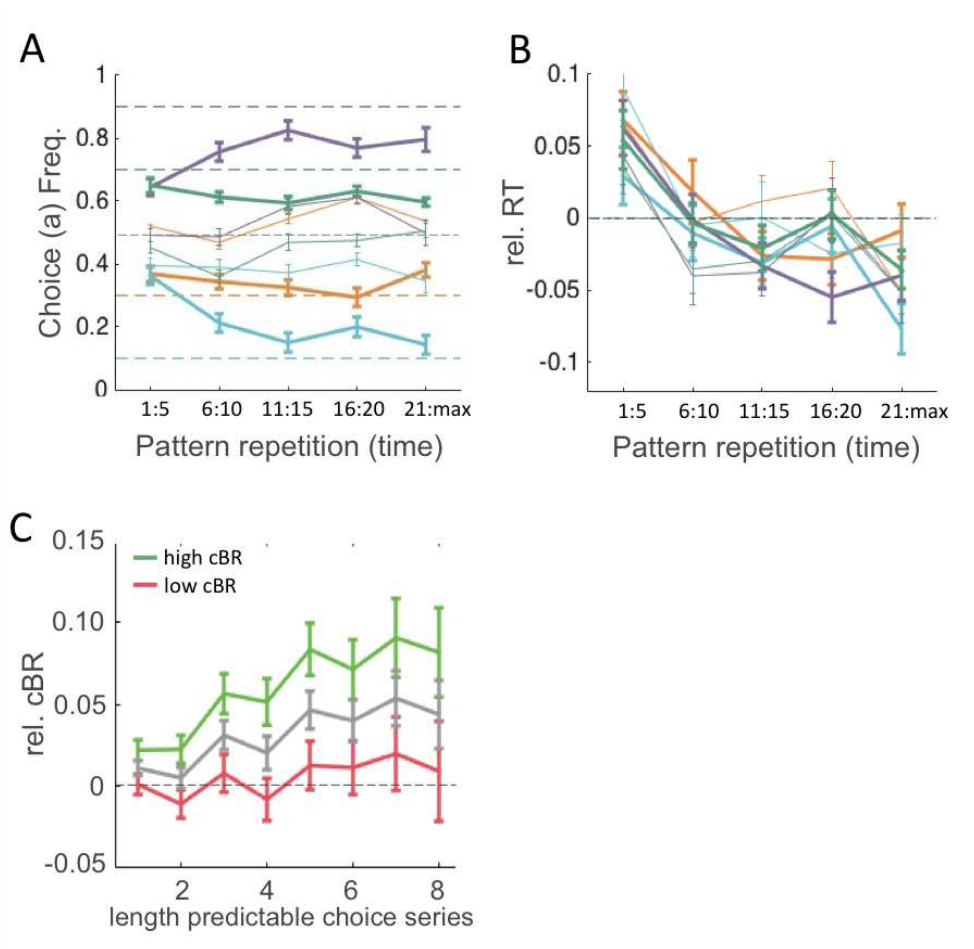
On average subjects’ choice distribution matched the probabilities associated to each of the 4 combinations of the 2-choices patterns, they best responded significantly more to high probability patterns than low probability patterns. They also selected the appropriate action to the probability patterns (0.7) more than chance (**Fig.4.A**). The variability in performance previously observed, also reveals the between-subjects variance in pattern discrimination. Indeed the more subjects best-responded on average (across all pattern conditions) the better they were at discriminating between the 4 patterns, and this in a symmetric fashion (all patterns were learned equally) (**Fig.4.B**). In addition, as observed before, we clearly observe a distinction between subjects who managed to detect the patterns and the others (**Fig.4.B** - 59% of the subjects above/below 0.55% of correct best response).



**Figure 4.** Subjects choice probability matches in best response the probabilistic patterns in the opponent's choice behavior when driven by its own past choices (opponent-based block) (A) On average, the higher the probability of the opponent selecting action "A" given the pattern displayed by its past choices, the more subjects selected correctly their action "a". Frequency of choice in the low probability conditions (AA/BB) were significantly different from chance level (AA:  $t = -2.5579$ ,  $p = 0.0128$  ; BB:  $t = 2.0487$ ,  $p = 0.0445$ ) (B) When sorting our population by average Best response rate (on the all block trials, independently of the pattern condition), we observed a strong asymmetric increase in their appropriate best response across the all pattern conditions, with the lower Best responders playing at random level independently of the current pattern condition and the higher Best responder's choices matching accurately the probabilities of the opponent given the appropriate pattern condition.

---

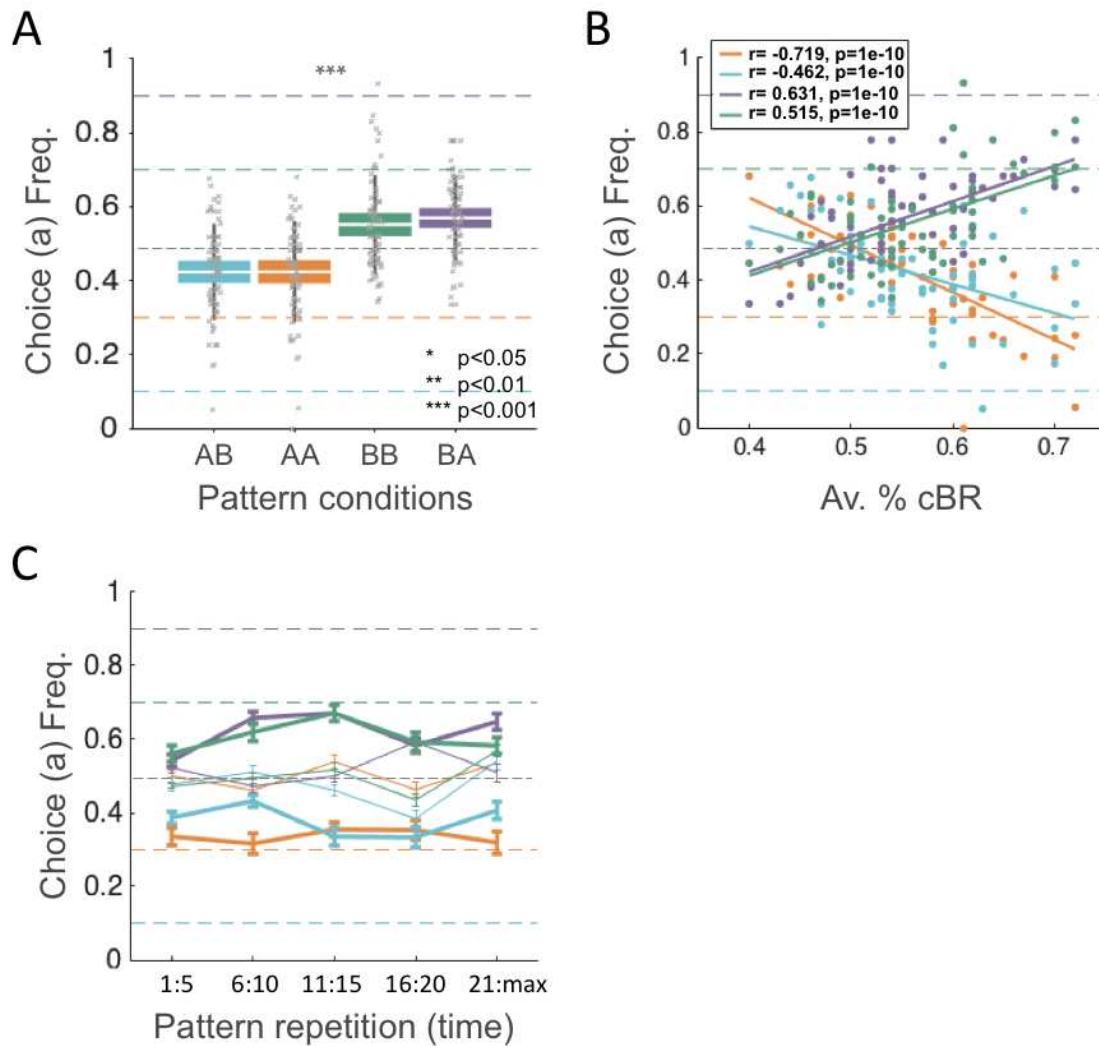
We went a step further to investigate if this heterogeneity in best response to probabilistic pattern was the product of learning or an intrinsic pattern detection ability. Looking at the evolution of best response rate through pattern repetition for the best performers revealed clear attributes of classic learning curves: increased best response and decreased reaction time with exposure for each pattern (**Fig. 5**). This concomitant effect of increase accuracy and decrease reaction time replicate extensive literature on sequence learning (Meyniel et al, 2016; Schwarb & Schumacher, 2012). In fact when we divide our population between portion of our subjects who learned the pattern and best responded better than chance and the one who did not (median split on average best response, but similar results are obtained when taking the subjects who best respond on average more that 55% of the time - Fig.4B), we observe in the best responders a decrease in their choice reaction time after a predictable trial ( $t(28) = -2.7922$ ,  $p = 0.0093$ ) and, in comparison, a increased reaction time after non predictable trials ( $t(30) = 2.3908$ ,  $p = 0.0232$ ). Since the propensity to detect patterns varied gradually at the population level (Fig.4.B), we looked at the linear correlation between the overall percentage of best response and the difference in choice reaction time at the individual level in the trials following a predictable action from the opponent vs. a non predictable one, and found a significant effect ( $r = -0.3491$   $p = 0.0038$ ). Moreover evidence was found that the more trials in a row are predictable, the higher the best response of the subject. This result thus suggest that unpredictable trials, and thus the probabilistic nature of the generated patterns, slow down pattern learning. Note that a similar effect of predictability was found for relative RT: the more the predictable trials are presented in a row, the lower their choice reaction time in these trials compared to their average reaction time across the all block (not shown).



**Figure 5.** Subjects manage to learn over time the probabilistic patterns in the opponent's choice behavior when driven by its own past choices (opponent-based block) (A) Choice learning curve of each opponent's choice pattern for the high best responders (Bold line - median split) compared to the low Best responders (thin lines). (B) Learning trace in relative reaction time, which decrease as a function of the number of experience of each pattern. Since average reaction time was strongly correlated within-subjects, we plotted the relative reaction time which corresponds to the difference between the choice reaction time at the current trial and the subject's average reaction time at the block level. (C) Exposure to long deterministic patterns increases the probability of best response compared to average performance for pattern learners. The more the subject was exposed to predictable trials in a row, the higher her performance for high Best responders (median split) only.

We then wanted to check if the difference in performance with the "Sub-based" block was indeed triggered by the additional need for a perspective switch to strategic awareness as designed.

We ran the same analysis we conducted on the “Opp-based” block on the the data of the other probabilistic pattern block, the “Sub-based”. This time subjects best responded better than chance to all patterns, however they failed to distinguish between low and high probability patterns. This was true even for the best responders (**Fig.6.A,B**). Similarly best responders did not show clear learning curves as in the “Opp-pattern” block, suggesting that they were constrained in their ability to accurately detect and learn patterns (**Fig.6.C**).



**Figure 6.** Subjects manage to differentiate between high and low probabilistic patterns only when the opponent’s choice is driven by the subject past choices (subject-based block), but fail to learn the precise probabilities. A) On average, the subjects selected more the action “a” for both low and high probability (that the opponent selects action “A” given the pattern displayed by the subjects past choices), but did not differentiate between the low and high pattern probability. Frequency of choice in all probability conditions were significantly different from chance level. B)

When sorting our population by average Best response rate (on the all block trials, independently of the pattern condition), we observed an asymmetric increase in their appropriate best response across the main pattern conditions, with the lower Best responders playing at random level independently of the current pattern condition and the higher best responder's choices matching the probabilities of the opponent, however they did not either respond differently in the low and high pattern conditions. C) Choice learning curve of each opponent's choice pattern for the high Best responders (Bold line - median split) compared to the low Best responders (thin lines). The former group (high) best responded in a similar fashion across repetitions for both high and low pattern conditions.

---

To better understand the difference in performance between the 2 probabilistic blocks, and what exactly in the opponent's choice behavior of the "Sub-based" block triggered such impairment in pattern learning, we should focus on one inherent feature of our design. In our task, the better a subject is, i.e. the more she best responds to the opponent's patterns, the more the pattern in the opponent's choice transposes to her own choice. Indeed the rule of the game being "choose the same choice as the opponent to win", the more they win, the more similar their choice series becomes to the one of the opponent, and thus the more confusion there might be regarding whose past behavior is actually predicting the choice of the opponent with some probability, and thus the more the 2 probabilistic pattern blocks become alike.

This feature is observable by looking at the ratio between correct best response and incorrect best response rate. As described previously correct best response rate is the percentage of trials where subjects best responded appropriately, i.e. to the opponent's past 2 choices in the "opp-based" block, and her own past 2 choices in the "sub-based" block). Conversely incorrect best response is the frequency of choices driven by the incorrect pattern, i.e. own's past 2 choices in "opp-based" block and the opponent's past 2 choices in "sub-based" blocks. As shown on **Fig.7.A**, the more subjects best responded correctly the higher the frequency of incorrect best-response in the opp-based block. This was also true, albeit less strongly, in the sub-based block ( $r=0.4528, p=1.2e-04$ ).

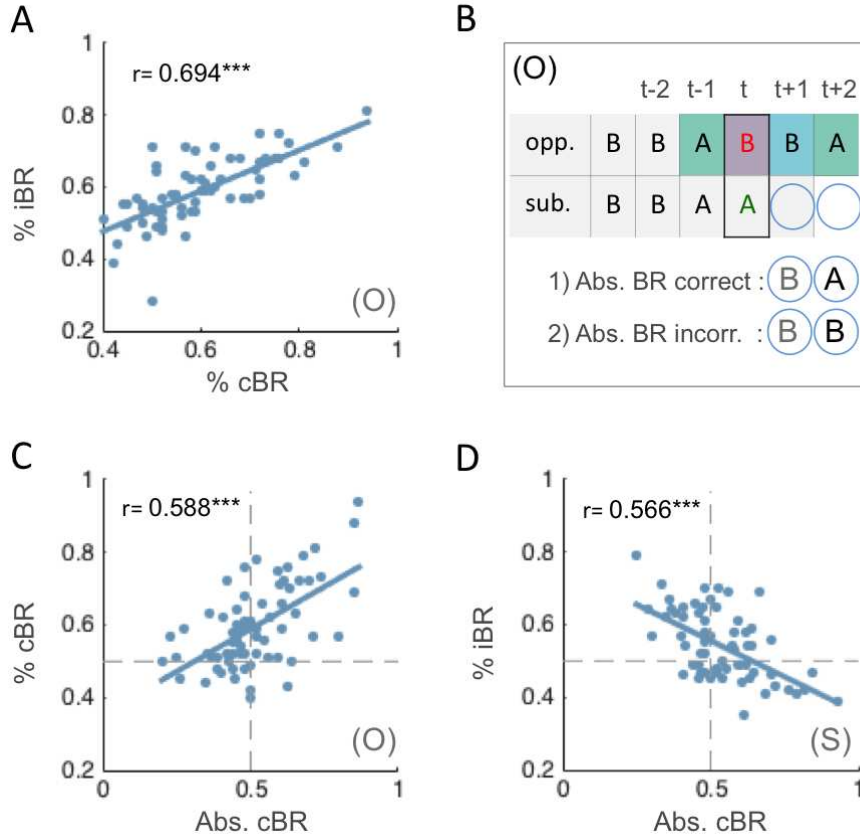
A way to overcome this confound it to not consider the trials where the subject chose the same action as their opponent in the last 2 trials (2 win in a row), since their best response in the next trial could be driven by the correct as well as the incorrect best response. We thus looked at the trials where the subjects lost, either because they did not detect the correct pattern or they did recognize the correct pattern but the opponent choice was in fact part of the 20% of unpredictable trials triggered by the probabilistic nature of their pattern-driven behavior. In the situation where the subject did not choose the same fractal as the opponent on trial (t), but then best responded in the next trial (t+1), the choice at trial (t+2) provides a clearer information about the correct best response (**Fig.7.B**). We thus consider patterns of these trials to compute an alternative way to discriminate correct best responses from incorrect best responses that we called "absolute best response". Absolute best response corresponds to the ratio between the frequency

of correct best response and incorrect best response given that on the previous trial they chose correctly but not 2 trials before. In the “opp-based” block, this refined measure did correlate with the initial correct Best-response (**Fig.7.C**), but not with the incorrect best response rate (not shown). Conversely in the “sub-based” block, this measure correlated negatively with the incorrect Best-response rate (**Fig.7.D**), but not the correct one (not shown). Taken together these results suggest that the better performance in the “Sub-based” block, was not due to an increased recognition of the pattern embedded in the subjects past choices, but a decrease in wrong pattern detection.

Note that this measure strongly correlates to the simple ratio between our initial correct Best-response and the incorrect Best response rate ((S):  $r= 0.8215$ ,  $p= 1.65e-17$  ; (O):  $r=0.7650$ ,  $p=4.8e-14$ ), and thus leads to similar results as presented on (**Fig.7.C,D**).

Another way to capture this ratio between correct vs. incorrect best response would be to only consider the unpredictable trials (noisy choices) of the opponent followed by a win of the subject and run the same computation. However not only the 1/10 of the trials unpredictable are not similarly distributed among the subjects, the one followed by a win represent only 10% of the block trials compared to our “absolute” best-response measure which represents 1/4 of the choices (note that despite these constraints the 2 measures are correlated (O):  $r=0.4282$ ,  $p=3e-4$ ; (S):  $r= 0.2441$ ,  $p= 0.0465$ ).





**Figure 7.** To disentangle the frequency of correct Best-response from the amount of incorrect best-response we computed an alternative measure, the absolute best-response rate. (A) Our design by its nature produces a crucial confound, the more often a subject best-responds to the opponent's pattern driven behavior, the more often her choice can be mistaken for an incorrect best-response, i.e. a best-response to the pattern embodied in her own past choice. (B) Absolute best-response gets rid of confound trials by focusing only on the trials following a mismatch in choices between the subject and her opponent, this way correct best-response can be dissociated to incorrect best-response. (C,D) The absolute best-response measures gives a more accurate distinction between correct and incorrect best-response in our task. Indeed, the higher the frequency of correct absolute best-response the higher the frequency of overall best-response in the opp-based block, while the opposite was observed in the sub-based block.

We then looked at the deterministic pattern condition to test if the ability to detect pattern when those remained fixed in the opponent's behavior is linked to the ability to detect probabilistic patterns.

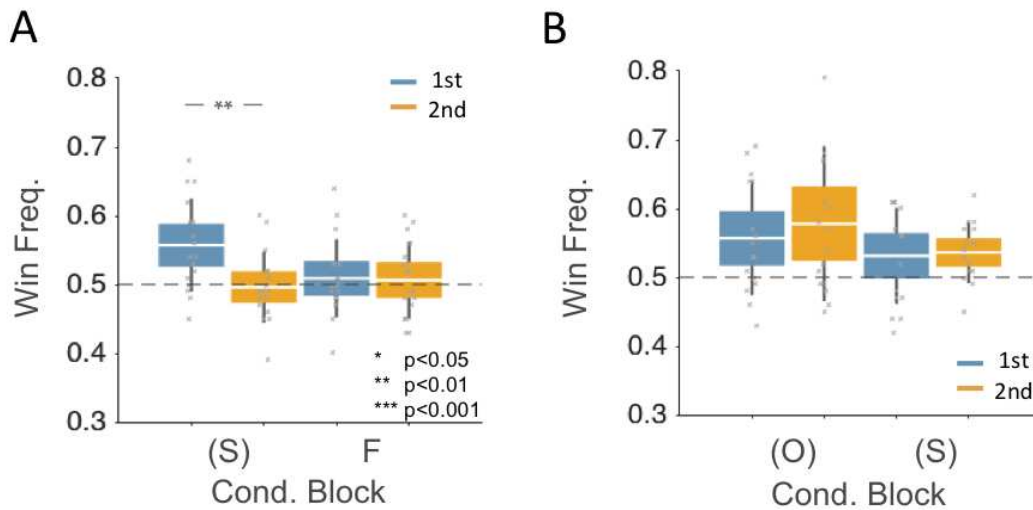
From our deterministic pattern block, 3 measures could be extracted: 1) the overall performance, 2) the amount of pattern passed (amount of times the subject did perform correctly 10 trials in a row), and 3) the number of trials needed to pass the first pattern which was meant to be a deterministic version of the

probabilistic pattern used in the 2 other blocks.

We found correlations between the first 2 measures and the best-response rate in both probabilistic pattern blocks ([1] (O):  $r=0.3969$ ,  $p= 8.8e-04$ , (S):  $r=0.4621$   $p= 8.3e-05$  ; [2] (O)  $r=0.2914$   $p= 0.0167$ , (S)  $r= 0.3236$   $p= 0.0076$ ), but not with the third one. None of the 3 measures correlated with the absolute best-response rate.

Finally we tested if subjects' performance in the non-pattern block against the fictitious opponent was correlated to the pattern learning ability. No correlation was found between the individual ability to beat the fictitious and their best response rate (nor absolute) in the 2 probabilistic pattern blocks. No correlation was found either with any of the 3 performance measures in the deterministic (D) block.

Additionally our design allowed us to test some transfer effect between opponent (**Fig.2.B**). We found that subjects who encountered first the fictitious opponent, thus non pattern-driven, had a reduced performance when playing next against the subject-based opponent (**Fig.8.A**). Given our previous results, this effect could thus be interpreted as a reduced incorrect best response from the subjects who first experienced an interaction against a maximizing opponent, which required to engage in higher strategic learning in order to win on average better than chance.



**Figure 8.** Performance of subject decreases in the subject-based block after playing against the non-pattern driven (fictitious) opponent. (A) Decrease in % of win (similar effect in Best response) when opposed to subject-based block after playing against the fictitious (N=17) compared to when encountered before (N=18). (B) no order effect on subjects who encountered the 2 probabilistic pattern block in shuffle order (N=32)

## D) Discussion

Research in psychology and in economics suggest that humans can detect patterns in their opponent's behavior even when randomizing (Dyson et al, 2016; Oskarsson et al, 2009). We aimed to test specifically the hypothesis that, in competitive repeated games, humans can not only detect but best respond to generated patterns in their opponent's choice series. We designed a novel experiment that was meant to be a controlled version of previous work by Spiliopoulos. The author suggested that participant's behavior in human-human interaction could be well fitted by a learning model that tracks and best responds to statistical redundancies in the opponent's choice series (patterns) (*fn* model in Spiliopoulos, 2012), although our own data (presented in this thesis, chapter II, III) suggest that subjects are in fact better fitted by a Influence model that computes higher order beliefs through iterative inference (level-k analog).

In the present experiment, participants interacted with 4 computerized algorithms in a pseudo random order through a repeated symmetric competitive game (hide and seek, see Devaine et al, 2014 for a strategic learning analysis of the game).

First, we showed that subjects were on average able to learn two-choice patterns in the opponent's behavior when they were generated probabilistically from its past choices ("opp-based" opponent). This result is congruent with Spiliopoulos who shows that on average humans can detect patterns embedded in the last two choices of the opponent (Spiliopoulos, 2013a). The learning behavior displayed in the repeated game interaction with a computerized opponent follows characteristics of classical pattern learning: accuracy increases and reaction time decreases with time and repetition (but also higher reaction time after an unexpected break in the pattern); overall suggesting improved choice prediction and thus learning. We even observed a trend towards a facilitation effect over patterns composed by two same opponent choices in a row, conveying twice more information as two different choices, as predicted by Meyniel et al (2016).

Second, we observed important inter-individual variability in subjects' ability to learn probabilistic patterns in their opponent's behavior, with a bit more than half of the population who performed gradually better than chance level. Looking back at sequence learning studies, it seems that important heterogeneity is often observed empirically, with a part of the sampled population who never detected the statistical redundancies (Meyniel et al, 2016; Sonsino & Sirota, 2016). However, no systematic distinction is made in the analysis. Moreover, our data show evidence of a consistency in the individual capacity to learn patterns when interacting in the probabilistic condition (non-maximizing algorithm) and in the deterministic

condition (maximizing algorithm, i.e. switching pattern when exploited). Interestingly, it seems that even the best responders in our task did not fully exploit the learned patterns. Such suboptimal behavior has been consistently observed in the decision-making literature, an effect called probability matching [Sugrue et al 2004 Science; Niv et al 2006 comment in Nat Neurosci on Morris et al 2006). Authors have suggested that the observed tendency of the participants to explore instead of exploit the fixed probabilistic nature of their choice environment can be, at least in part, explained by the propensity to look for temporal patterns (Baker et al, 2014). It has been recently proposed that a combination between reinforcement learning and pattern detection could actually better capture this effect (Gaissmaier & Schooler, 2008; da Silva et al, 2017). This tendency to explore might explain the matching effect in our task. As illustrated by the increased best response rate with the number of predictable (not noisy) trials in a row, and the difference in (relative) reaction time between a predictable and a non-predictable trial, prediction error might have impacted the subject's optimal performance. Recent research suggest that in a perceptual decision tasks with stable transition probabilities, such non-predicted trials could be considered as non-relevant and ignored (Filipowicz et al, 2016). Nevertheless, when patterns of stimuli are tractable such unpredictable event is experienced as a prediction error (Stefanics et al, 2011). In repeated games such unpredictable trials could trigger engagement in belief-learning, as each trial might be considered as informative to accurately predict the opponent's next choice.

One could then ask if, in social interactions, the statistical redundancies in the opponent's choices are actually considered by the subjects as probabilistic patterns, or whether they assume that their opponent is rather trying to maximize his own payoff and thus engage in some sort of belief-learning allowing for pattern detection (as suggested by Spiliopoulos' results).

Our design was meant to shed light on this question. We added a "sub-based" condition, in which the opponent was employing probabilistic patterns, as in the "opp-based" block, but directly on the participant's behavior, thus requiring a higher level of strategic sophistication to detect it. We showed that subjects actually failed to track such higher order patterns. Indeed the best responses displayed by subjects performing non-randomly in this block were actually the byproduct of the underlying choice patterns: these subjects appeared to have focused primarily on their opponent's behavior and thus to have ended up learning the wrong generative model underlying the emergence of choice patterns. The choice series of the two players being partially correlated in such dyadic interaction setting, such misattribution of belief led them to a higher than chance performance. The difficulty to engage in such high-order pattern was already suggested by Stöttinger et al (2014) who showed that in a repeated competitive game, participants failed to best respond to a computerized opponent generating probabilistic behavior conditionally on the individual's previous choice, and this even after being primed through training.

Moreover, we did not find any correlation (positive or negative) between pattern detection ability and the

performance in the fictitious block, where the opponent was only maximizing and which thus required to engage in higher-order belief-learning. We did observe however that participants who played first against this opponent failed to detect patterns in the “sub-based” condition, and thus performed at chance level since no misattribution occurred. This priming effect is interesting because it challenges the null-hypothesis according to which pattern learning was not correlated with strategic learning. We could thus not reject this null-hypothesis in this study. However, two interpretations are possible: either subjects in this experimental group (fictitious opponent before “sub-based”) were impaired in their general ability to detect and best respond to patterns, or they were more aware of the strategic dependency (opponent tracking patterns in the subject’s choice), but failed to engage in higher order pattern learning. To disentangle the two hypotheses we are planning to run a follow-up experiment with two extra experimental manipulations (not presented in the manuscript) where a group of subjects would be playing against the “opp-based” opponent after having interacted with the fictitious: in one condition they would encounter first the fictitious, in another they would have encountered the “sub-based” first to control for priors. In the next section, we will address directly the question of the interaction between pattern-learning and strategic learning, using a different, more suited, experimental paradigm.

## **II - Pattern detection and Strategic sophistication in Rock-Paper-Scissor (Exp.6)**

### **A) Introduction**

#### **1) Thesis context**

In our previous experiment we found heterogeneity in the human’s ability to detect and best respond to statistical regularities hidden in the opponent’s choice series. We observed a gradient of behavior going from subjects missing the patterns in the opponent’s choice and playing at chance level, to individuals who were successfully learning the patterns in their opponent’s behavior to anticipate their next action and best respond accordingly.

We also found evidence for a priming effect from the previous interaction block on the subject’s ability to detect and best respond to patterns in the next block. Subjects who were first exposed to a strategic learner, an opponent maximizing its behavior by tracking the subject’s past choices, missed the patterns when interacting against a pattern-driven opponent. These results thus cannot lead to reject the null-hypothesis of a common individual propensity to engage in both types of learning.

To take into account this possibility we formulated the following alternative hypothesis: two different processes are implicated in pattern and strategic learning; therefore humans are constrained in their ability to effectively combine the two during a (competitive) repeated game interaction.

This hypothesis thus suggests that engaging in one type of learning might prevent the engagement in the other type of learning during strategic interactions. Along with the results presented in Chapter III, the previous study suggests that priming from a previously experienced strategic interaction might drive sophisticated players to either track and learn statistical contingencies in their opponent's behavior or engage in sophisticated computation over their strategic intentions. If not rejected, this hypothesis would extend our conception of strategic sophistication and refine our understanding on how the two learning systems can interact in strategic repeated settings. Note that this hypothesis is different from Spiliopoulos' hypothesis that pattern detection is imbricated within strategic learning so that a fictitious play (for instance) can rely on estimating both choice proportions and choice patterns in the opponent's behavior.

## **2) Scientific context**

Statistical learning, and in particular pattern detection seems to be a human ability shared across sensory modalities and domains (Aslin & Newport, 2012). Faced with inconsistent data on the cognitive cost of such cognitive process, some authors have suggested that while detection of statistical regularities can be quasi-automatic (Kimura et al, 2010), learning the latent structure of the environment in order to adjust a goal-directed behavior might be computationally costly (Collins, 2017; Sun et al, 2015; Unsworth & Engle, 2005). Based on our previous study, we thus hypothesized that human's capacity to track and learn patterns in the opponent's behavior might prevent the iterative computation of high-order beliefs and engage in heavy strategic learning.

It has been shown that human's mental representation of the detected patterns can be influenced by priming (Schuur et al, 2013). And in a competitive game (Rock-Scissor-Paper), Stöttinger et al (2014) showed that prior over a probabilistic opponent's strategy (bias towards one action) facilitates the adaptation to change in the opponent's behavior (bias towards another action). Moreover Filipowicz et al (2014) showed that participants learn better the deterministic pattern embedded in a computerized opponent in a RSP paper game, when the framing of the task facilitates the pattern detection (i.e. when the choice patterns matched the location pattern on the screen). Such facilitative transfer of previously learned patterns have been identified experimentally in a more general social domain, like sport (Broadbent et al, 2017; Loffing et al, 2015).

We therefore hypothesized that interacting with an opponent displaying tractable patterns might facilitate

pattern detection in the following interaction, and that conversely interacting with a maximizing opponent adapting trial-by-trial to the subject's behavior might impair such learning process.

### **3) Hypothesis**

We designed a task divided in three blocks in which participants will interact with three different computerized opponents in a repeated version of the Rock-Scissor-Paper game. During first block participants will compete against either (1) an opponent following MSNE distribution, and partially randomizing while not following a deterministic rule producing tractable patterns, or (2) an opponent maximizing through fictitious play by tracking and best responding to their behavior. This first block will serve as a priming condition and the second block will be the test block. In the test block all the participants will be confronted to a hybrid version of the two opponents, maximizing through belief-learning in parallel with a fixed rule producing deterministic patterns. This test block will thus allow us to observe the priming effect of each type of learning on the behavior displayed during the following interaction. Following our hypothesis, we expect that in this block subjects primed by the pattern-only opponent will best respond mainly to the deterministic patterns embedded in its choice series and not to its maximizing trials. Conversely, we expect the priming of the fictitious block to drive them to focus less on the temporal patterns and engage more in iterative belief inference. In the following, we present a preliminary analysis of the data from the experiment we conducted in order to test this double dissociation prediction.

## **B) Methods**

58 participants (30 male, 28 female; ages 18–25) took part in the experiment. All participants were right-handed, medication-free, with normal eyesight, no history of neurological disorders. The Ethics Commission of the University of Trento approved the experiment. Informed consent was obtained from each subject before the experiment. Data collection was performed blind to the conditions of the experiment.

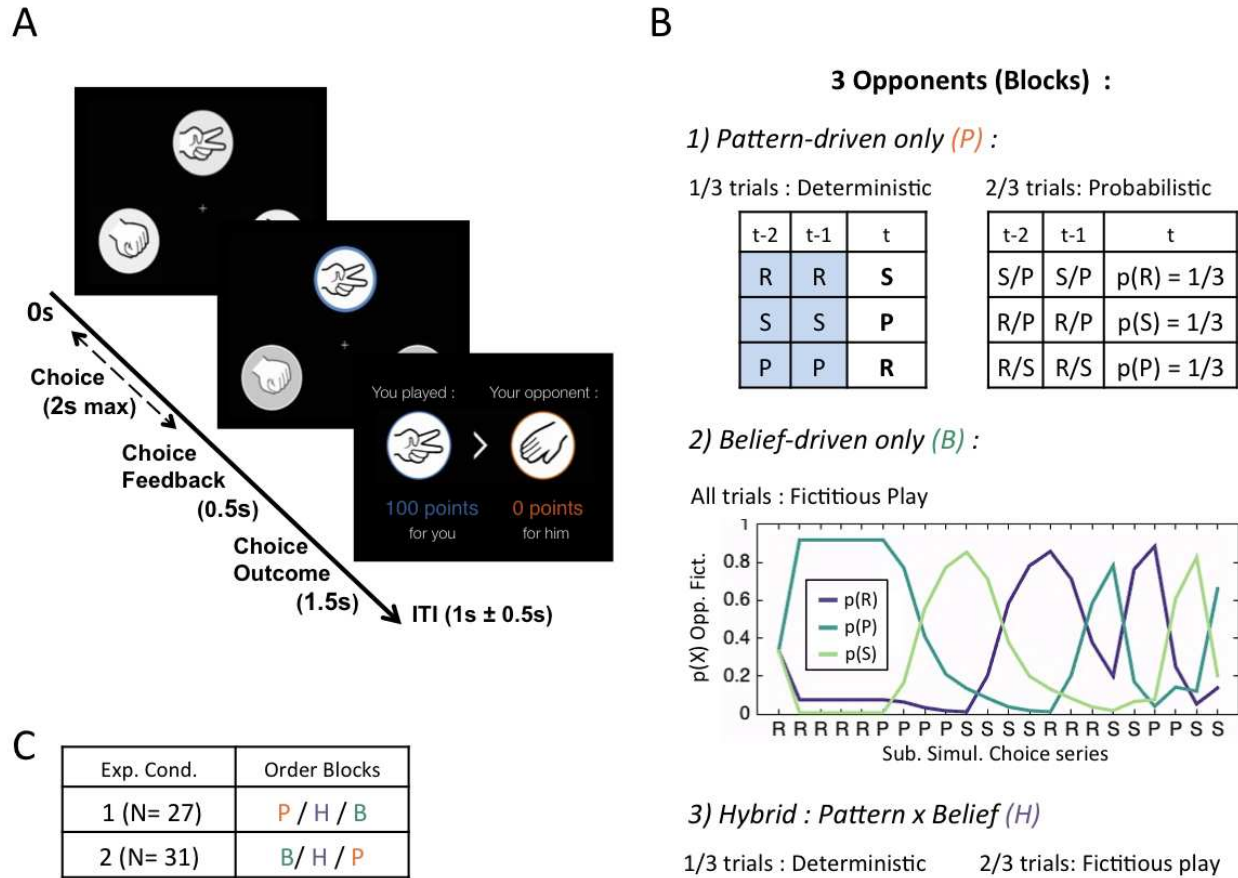
The experiment consisted of 3 blocks of a repeated competitive interaction game against 3 different computerized opponents. The interaction consisted of a 3x3 game known as Rock-Paper-Scissor, repeated 200 times and rewarded from 0 to 100 points depending on the choice combination at each trial. The rule was the following: Rock beats Scissor, Scissor beats Paper, and Paper beats Rock. 100 points

were won if the option chosen beats the opponent's symbol, 50 points for a draw (same symbol selected) and 0 points in case of a loss. Points earned at each trial were accumulated through each block and summed up to determine their final payoff which would ultimately be converted to euros according to a predetermined rule.

At each trial the three game actions, represented by three different hand sign symbols, were presented for 2s (subjects were initially instructed and trained to insure they were familiar with the choice settings) (**Fig.9.A**). The choice was made by pressing the corresponding button (left, up or right). 2.5s after the trial onset, both players were simultaneously provided with the outcome feedback of their choice for 1.5s. During outcome feedback, the two chosen hand sign symbols (one chose by the opponent, one by the subject) were displayed along with an indication of who won, and a sentence indicating their points underneath.

The subjects were playing against 3 computerized opponents in 3 interaction blocks (**Fig.9.B**). The Pattern-driven opponent (P) was programmed to play each action randomly except when the combination of the past 2 choices matched one of the 3 target patterns, in these trials the choice was predetermined. The Belief-driven opponent (B) was modeled by a (weighted) fictitious play algorithm, computing at each trial the weighted frequency of the subjects' past choices and best responding to it. As opposed to (P), the opponent (B) forms beliefs over the subject's behavior to maximize its play. The third opponent was a hybrid (H) between these two algorithms, best responding to the participant choice (weighted) frequency by following the fictitious learning rule at each trials except when the past 2 choices matched the predetermined patterns used by the (P) algorithm. Note that unlike the Pattern-driven only (P) opponent, the occurrence of the predetermined pattern choices in the (H) choice series varies given the subject's choice series. To make sure that this won't be an issue we simulated different agents using classic strategies of play. None of these strategies lead to an over- or under-representation of the deterministic choices in the opponent's choice series (**Supplementary Information 3 (Appendix IV) - Fig. S2**).





**Figure 9.** Experimental Design. (A) Trial Structure: At each trial the three game actions, represented by the corresponding hand sign symbol (subjects were previously instructed and trained), were presented randomly in each of the 3 positions on the screen, for 2s. The choice was made by pressing the corresponding button (left, up or right). 2.5s after the trial onset, both players were simultaneously provided with the outcome feedback of their choice for 1.5s. (B) Subjects played 3 interaction blocks against 3 different opponents. The pattern-driven opponent plays the Mixed Strategy Nash Equilibrium ( $p(\text{choice})=1/3$ ) except when its 2 last choices were the same, in which case the next choice was predetermined. The Belief-driven opponent was modelled by a weighted fictitious, best responding in all trials to its estimation of the subject's action probability computed from the last past choices. The hybrid opponent, a mixture between the first two algorithms, maximizing through a fictitious play algorithms in all trials, except when its 2 last choices were the same in which case the next choice was predetermined.

The interaction against the Hybrid opponent was meant to be the test block. We therefore manipulated the opponent against which the subjects would play before the opponent block, by randomizing the order of the 2 "pure" opponent blocks (**Fig.9.C**). We hypothesized that the subjects' behavior in the Hybrid block

would be affected by the opponent encountered first: playing against the Belief-driven only (B) would prime them to focus on the non-pattern trials (belief-driven) trials; while playing against the Pattern only (P) opponent would orientate them towards pattern-learning. We expected that this attentional manipulation would drive the strategic players to engage either in strategic learning, or in pattern-learning. This would in turn translate into different performances in each type of trials; non-pattern trials and pattern trials i.e. belief-driven choices and predetermined choices, respectively.

## C) Results

We first looked at the overall performance between each opponent condition. We found that on average, performance (quantified as the average points accumulated across each block) was above chance level in each block (**Fig.2.A**). When comparing how subjects performed in the 2 “pure” conditions, respectively pattern-driven only (P) and belief-driven only (B) blocks, we found that the average performance (outcome) was higher in the latter condition. However this difference in performance between these 2 blocks could be explained solely by the fact that in the former block (P) 2/3 of the trials cannot be predicted by the subjects (the opponent chooses randomly) and thus their performance is somehow constrained between 0.5 and 0.67 (max Expect.Perf =  $(0.5 \cdot 2/3) + (1 \cdot 1/3)$ ). To take this structural difference into account we computed subjects' performance in the (P) block as their average outcome in the predictable trials (opponent's choices following a predictable pattern) only. By doing so, we observed that in this block subjects performed better in the predictable trials compared to the unpredictable trials (random choices) ( $t(84)=4.8832$ ,  $p=5e-06$ , individual ratio of performance between the predictable vs. unpredictable trials compared to 0:  $t(57)=5.2023$ ,  $p=2.78e-06$ ). However no difference in reaction time was found between the two types of trials. Ultimately, considering the performance in (P) as the average outcome in the predictable trials only did make the difference in performance between the 2 blocks P and B fade away. No difference on average choice reaction time was found between these two blocks.

The same effect was observed when comparing the average performance between the 2 pattern blocks (P vs. H): when looking at the overall performance a difference was found (**Fig.10.A**), but it disappeared when considering the performance in the pattern-predictable trials only (the same lack of difference was found when considering the best response to pattern rates instead of average outcome in the 2 blocks). In the Hybrid opponent (test) block, both pattern and non-pattern (belief-driven) trials lead to better than chance performance (pattern:  $t(57)=5.1824$ ,  $p=3e-06$ ; non-pattern:  $t(57)=5.7113$ ,  $p=4.25e-07$ ). However unlike

the pattern-only block (P), no difference on average performance was found between pattern and non-pattern trials in the Hybrid opponent ( $t(87.3) = 1.8640$ ,  $p = 0.0657$  - indivi. ratio performance pattern vs. non pattern trials to 0:  $t(57) = 1.8014$ ,  $p = 0.0769$ ), thus suggesting equivalent performance in pattern and non-pattern trials. These results thus suggest that subjects were on average able to not only best respond to the deterministic patterns but also to the remaining belief-driven trials in which the opponent best responded.

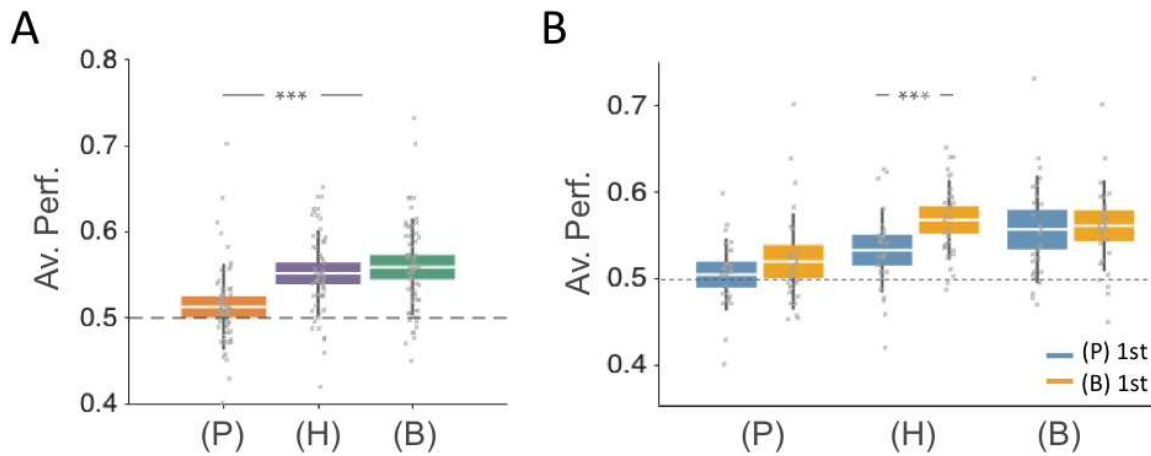
We investigated the hypothesis that the better subjects were at best responding to patterns (predictable trials) in the pattern only (P) block, the better they were at doing so in the hybrid block (H) too. We found a correlation in best response selection between the 2 blocks when measuring pattern best response as their average performance in the predictable trials only ( $r = 0.4222$ ,  $p = 0.001$ ). Similarly we found that that the higher subjects in earn during their interaction with belief-driven opponent (block B), the higher their performed in the non-pattern (maximized) trials in the block H ( $r = 0.4222$ ,  $p = 0.001$ ).

We then tested our hypothesis of a competition between pattern learning and strategic learning. No correlation was found between the performance in pattern trials and non-pattern trials of the Hybrid block ( $r = 0.0848$ ,  $p = 0.5268$ ), whereas we would have expected to find a negative correlation between the two. In fact half ( $N = 31$ ) of our subjects had an average performance higher than 0.5 in both types of trials (while  $N = 11$  had a performance  $> 0.5$  in only pattern trials,  $N = 11$  in only non-pattern trials, and  $N = 1$  in any of the 2 types of trials). We then computed for each subject the difference in performance (points) between the second half of the trials compared the the first half of the trials, this allows us to approximate learning since a positive score meant improvement in performance across the interaction. We did not observed any significant correlation either in this learning measure between the pattern trials and the maximized trials ( $r = 0.03093$ ,  $p = 0.0182$ ).

Finally we looked at the correlation between the difference in performance in pattern vs. non pattern trials in the Hybrid block, and the average performance when interacting against the 2 other opponents. No correlation was found with the performance in the belief-driven block (B) ( $r = -0.1386$ ,  $p = 0.2994$ ) and only a weak correlation was observed with the performance in the predictable (pattern) trials of the pattern-only Block (P) ( $r = 0.2645$ ,  $p = 0.0448$ ).

To further investigate the interaction between the pattern learning and strategic learning engagement we looked at the effect of priming our design was initially meant to trigger.

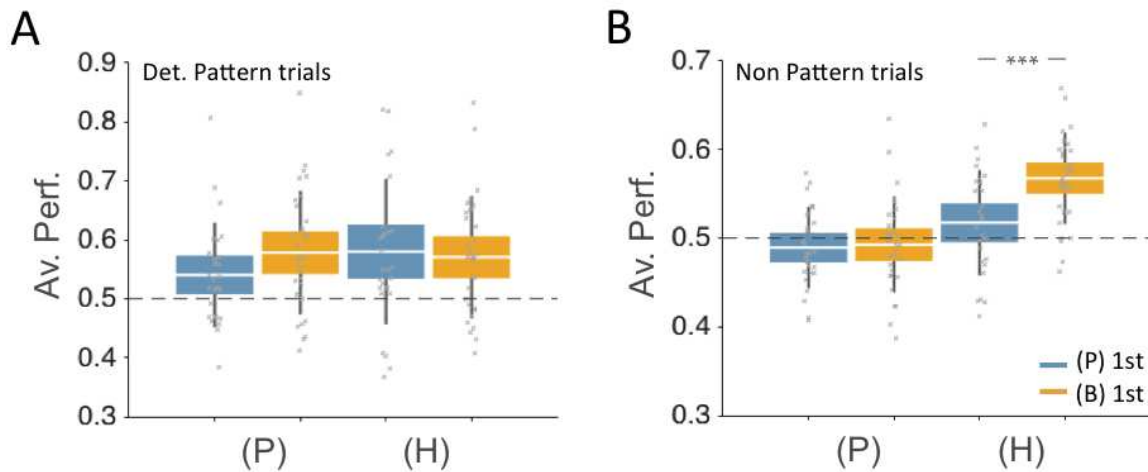
First we observed a difference in overall performance in the hybrid (H) block depending on the opponent played first, with a higher performance for the subjects exposed first to the belief-driven only opponent (**Fig.10.B** ; no effect of the order of play was observed in the pattern only (P) nor belief-driven only (B) blocks)



**Figure 10.** Subjects performed better in the deterministic pattern block when the non-pattern trials were maximized by the opponent (A) and this effect was mainly driven by the subjects who were playing against the belief-driven (fictitious) opponent first (B). Performance here is the translation from average points across the all block to percentage, so that 1 corresponds to an average of 100 pts, 50pts thus being the chance level (i.e. 0.5).

Based on our previous results showing that no difference in pattern detection could be observed between the 2 pattern blocks, we hypothesized that this difference in performance conditional on the type of opponent encountered first was driven by an enhanced ability to focus on belief-driven trials and thus a priming effect of the belief-driven opponent over the hybrid block.

Our data indeed suggests that playing against the belief-driven only (B) opponent first does not affect the performance in the pattern trials in the hybrid block (H) (**Fig.11.A**) but enhanced the performance in the belief-driven remaining trials (**Fig.11.B**).



**Figure 11.** Splitting the trials in the 2 pattern blocks between the non predictable and the predictable (from deterministic patterns) reveals that subjects did not respond better to the deterministic pattern in the “H” block, when exposed to the belief-driven opponen “B” first (A), but accurately best responded to the belief-driven part of the hybrid opponent (B)

If this hypothesis was correct then we should find a correlation between subject’s ability to engage in strategic learning (taking their overall performance as proxy) and their performance in the non-pattern trials of the hybrid block only for subjects who were exposed first to the belief-driven only (B) block. As expected a significant correlation could be observed between the 2 blocks for the subjects who played first against the belief-driven ( $r = 0.5449$ ,  $p = 0.0015$ ), and not for the other group ( $r = 0.1385$ ,  $p = 0.4909$ ). Note that the correlation of the performance in pattern trials only between the 2 pattern blocks holds what ever opponent was encountered first ((P) first:  $r = 0.4411$ ,  $p = 0.0213$  ; (B) first:  $r = 0.4516$ ,  $p = 0.0108$ ).

## D) Discussion

We hypothesized that, during a competitive repeated game, a human subject’s ability to detect and best respond to patterns embodied in the choice series of the opponent is in competition with her ability to engage in strategic learning, so that the triggering of one process prevents the other one and vice versa. We designed a task divided in three blocks. In the priming block subjects were either interacting with a partially randomizing (MSNE) opponent following a deterministic rule producing patterns in one third of the trials, or with a belief-driven (weighted fictitious play) algorithm. We then tested the

priming effect of participants' interaction in a test block where the opponent followed patterns in one third of the trials but best responded in a belief-driven fashion (fictitious) in the remaining trials (a combination of the 2 "priming" opponents). We hypothesized that pattern priming would enhance pattern learning in the test block and prevent strategic learning, while fictitious priming would have the opposite influence, preventing pattern learning.

The preliminary analysis of the data did not reveal the hypothesized double dissociation between the two types of learning in the test block. Indeed, no difference in pattern learning was found when subjects were primed by interacting with either one of the two opponents. However, we found that when playing first against the fictitious (belief-driven) opponent, their performance in the test block improved. Looking more closely at this effect revealed that subjects on average did not learn the patterns less in the test block, but responded more accurately to the belief-driven choices of the hybrid opponent. In fact, our data do not seem to show evidence suggesting that a competition between the two types of learning takes place in our task, solely a facilitative effect of being primed by a strategic opponent.

Two results are interesting here. First, subjects' best response rate to patterns in the test block did not decrease when primed by the interaction against the fictitious opponent. However, their performance in the remaining trials (choices where the opponent was belief-driven, i.e. following a fictitious strategy) was improved. Moreover, this facilitative effect depended on their ability to override the fictitious play when encountered in the first interaction block. This suggests that subjects able to engage in higher order belief-learning were able to also keep track of the statistical regularities (deterministic patterns) emerging from their opponent's choice behavior.

Second, subjects on average did not improve their best response to deterministic patterns. In fact, their average performance remained quite low in the pattern-only block. Indeed, on average they did perform a bit higher than chance (albeit not significantly), by winning only 56.1( $\pm$ 9.7)% of the maximum reward and, when looking at the actual percentage of correct best response, only 39.5(1.4)% of the predictable trials. Moreover, this average score did not improve over time since the comparison of the average best response between the first and the fourth quartile of the predictable trials shows no significant difference ( $t(114)=-0.3617$ ,  $p=0.7182$ ), even within subjects (difference fourth - first quartile compared to 0:  $t(57)=0.4264$ ,  $p=0.6714$ ). This result is puzzling since we previously observed that pattern learning can take place in a competitive repeated game (Exp.5).

A more fine-grained analysis investigating the dynamics of the interaction as well as between-subjects differences is required. For now we can only speculate from the striking difference found with our previous experiment that this effect might be triggered by the nature of the patterns, here deterministic and static across the whole interaction block. Indeed, in our previous experiment either patterns were probabilistic --

and thus considered as statistical regularities which could be learned but not as direct sequence learning --, or patterns were deterministic but changed across the interaction as subjects were able to exploit them. Duffy et al. (2016) observed that subjects engaged in a repeated 2x2 game against an opponent following deterministic patterns (repetition of the same 3-action sequence) did perform a bit better than chance but were far from fully exploiting these patterns. Interestingly, the subjects under high cognitive (working memory) load were better at it than the one under low cognitive load.

An alternative hypothesis would be that the noise surrounding the deterministic patterns, and accounting for two thirds of the trials, impaired the pattern learning of the subjects. In other words, subjects who managed to detect some statistical regularities and performed higher than chance might have tried to engage in high order belief learning to learn from the random trials as well. This could have thus prevented them from best responding in the predictable (pattern) trials. One way to test this hypothesis with the current dataset would be to then compare the difference in behavior in these random trials in subjects who played first the fictitious in comparison to the group who played it at the beginning of the experiment.

Taken together these two main results suggest that humans are able to combine information from the two sources of information that were manipulated in this experiment, namely the statistical regularities of the opponent's behavior (pattern detection) and the correlation between the opponent's behavior and her own past choices (high order belief).

Much refined analyses remain to be performed on this dataset. Nevertheless, we can already rule out our initial hypothesis and consider that (some) subjects can simultaneously engage in both pattern and strategic learning.

In an additional study, Spiliopoulos et al (2013b) controlled for one side of the repeated interaction by making participants play against *fpn* (2,3) algorithms, he showed that subjects managed to adapt their behavior from one opponent to the other. In the light of the results obtained in this experiment, this suggests that human might actually be able to combine pattern detection and engagement in strategic learning. If true, this then raises the question of the nature of the interaction between pattern and strategic learning and the factors influencing such highly sophisticated learning processes. Better characterizing this interaction could help better capturing inter-individual variability in (competitive) repeated game interactions.

### **III- Conclusion of Experiments 5 and 6**

Designing an experiment directly testing this hypothesis we have shown that subjects who interact first against a maximizing opponent engaging in belief-learning (fictitious play), have seen their performance improved against an opponent employing in parallel a deterministic rule displaying tractable choice patterns. This result sheds light on the finding of our previous experiment (presented in section A of this chapter), in which we observed that the participants who played against a fictitious opponent first did, rightly, not learn patterns in their opponents behavior when the algorithm was exploiting patterns in their own choice history. Together the results of these two experiments (A,B) suggest that interacting first against a strategic agent allows to improve one's belief accuracy by making her fully consider the strategic nature of the interaction.

This hypothesis could be tested directly by fitting the learning models varying in the order of belief inference operated over the opponent's choice behavior (Chapter I and II of this thesis). We make the prediction that subjects interacting through a 2x2 symmetric game against the "Opp-based" (exp.1 - A) opponent, and the one playing RSP in the Hybrid block (exp.2 - B) would be found to engage more in strategic learning (Influence model) compared to a simple RL or even belief-based such as Fictitious. Moreover we hypothesize that such subjects might actually be better fitted by the Influence model compared to the *fp2* model of Spiliopoulos which explicitly forms beliefs over statistical redundancies (patterns) in the opponent's history of play. This analysis thus represents a logical next step for the analysis of the present data.

One striking difference observed between these two experiments, and already outlined in the discussion of section B, is that subjects failed to really learn patterns, or at least to correctly best respond to it, in the RSP repeated game where the opponent was employing a deterministic pattern strategy. We proposed several hypotheses and suggested ways to test them in the next step of our data analysis. However, one more conceptual hypothesis might lie under the nature of the algorithm used to generate patterns in this task. Indeed, one might argue that this opponent, that we called pattern-driven, is actually implementing a specific rule, which thus produces identifiable sequences in its choice history. In the previous experiment however this "rule" was probabilistic, generating patterns that were more statistical redundancies than fixed sequences of choices.

While human's ability to infer structure from statistical regularities (statistical or structure learning) has been shown to encompass a panel of cognitive functions (Goldstein et al, 2010), our capacity to detect fixed sequences, or chunks, have been highlighted in psychology as we were proved to be particularly skilled when it comes to recognize implicit, deterministic patterns ( Du & Clark, 2017; Fonollosa et al 2015). Recently authors have suggested that such chunk detection might probe statistical learning (Daltrozzo & Conway, 2014; Jiménez, 2008), notably in goal-directed settings where cognitive control is



required ( Deroost et al, 2012; Jones & McLaren, 2009).

Authors have proposed such human's ability to learn underlying structure of their environment to subserve value-based decision making (Abrahamse et al, 2010; Nakahara, H., & Hikosaka, 2012). Indeed in model-based learning, knowledge of the task structure is required, and thus refining our mental map of the environment is key for adapted behavior in a changing world (Doll et al, 2015; Green et al, 2010). Accordingly, recent research in cognitive neuroscience suggest that the orbital part of the prefrontal cortex (OFC), region previously shown to be implicated in the state representation subserving model-based learning (Schuck et al, 2016; Wilson et al, 2014), might encode the mapping of the task by integrating signals from the hippocampus (Wikenheiser & Schoenbaum, 2016). Moreover authors have suggested that encoding task-sets, chunks of pre-encoded rules or strategies associating stimuli to action and actions to outcome, might be at the root of flexible learning (Collins et al, 2013; Donoso et al, 2014). Independently Abrahamse et al (2010) suggested that “the representations that are most relevant (and thus most active) for current purposes (on the basis of task set and/or task context) ultimately determine the nature of sequence learning”.

Together we suggest that pattern detection in repeated games might trigger learning of the rule, or task-set strategies employed by the opponent, in order to refine the mental representation of the structure of the strategic interaction. If this is true then pattern-learning could then subserves strategic learning to improve belief accuracy, as suggested by the results presented in this chapter.

However as in statistical learning, such encoding of the task structure have been shown to be computationally costly and to require important cognitive control (Collins et al, 2013, 2017; Jones & McLaren, 2009). Investigating this new hypothesis might ultimately help us to better understand how the two learning systems outlined here interact and vary among individuals. We believe that this line of inquiry will allow us to improve our understanding of the observed individual deviations from optimal play and characterize the behavioral dynamics emerging from repeated interactions (Wang et al, 2014).

## - Chapter V -

### General Discussion

In this PhD work, we aimed to investigate the cognitive mechanisms underlying human strategic learning in repeated (competitive) game interactions. The work presented in the manuscript was divided into three parts.

In chapter II (Exp.1-3) we studied how the level of engagement in strategic learning, allowing for the formation of higher-order beliefs over the opponent's play, interacted with the structure of the strategic interaction. We showed that the engagement in strategic learning drives the formation of more accurate beliefs and eventually leads to an overall convergence towards game optimality embodied in the concept of (Mixed Strategy) Nash Equilibrium. Moreover, we demonstrated that the heterogeneity in choice behavior usually observed in repeated strategic interactions can be captured by considering the interplay between the (payoff) structure of the interaction and the individual ability to engage in strategic learning. We also observed that the strategic sophistication behavior displayed by the opponent does not influence the engagement in strategic learning, but that this engagement is endogenously driven.

In chapter III (Exp.4) we showed that the level of strategic learning sophistication of the previously encountered opponent can increase the engagement in higher-order belief-based learning of humans (cognitively) capable of doing so, i.e. strategic players when they are in a situation of strategic disadvantage. We also provided evidence that highly strategic individuals in an advantageous situation in the interaction can exploit a belief-based opponent, by taking advantage of the predictability of her behavior.

In chapter IV (Exp.5,6) we showed that humans are indeed capable of learning statistical redundancies (i.e. patterns) in the choice behavior of their opponent, despite important heterogeneity in their behavior., We also found that subjects' capacity to detect and learn patterns in the opponent's past play was not correlated to the engagement in strategic learning. Moreover, our preliminary data suggest that the participants displaying pattern learning were not impaired in their strategic learning capacity when an opponent displayed both rule-based and strategic behavior, suggesting that humans can combine the information from both types of learning to improve the accuracy of their beliefs over the opponent's behavior.

In the following, we will (A) first discuss this experimental work from a methodological and conceptual point of view and discuss the implications of the results obtained with respect to the existing literature; (B)

We will then discuss directions for future research. Finally we will briefly conclude on the limits and implications of this PhD work.

## A) Conclusion of the experimental studies (Exp1-6)

### 1) On the methodological approach

At first sight, one can highlight a general observation regarding the experimental work conducted in this thesis: the computational approach developed and tested in cognitive neuroscience proved here once again a useful tool to make sense of the heterogeneity of the repeated choice data obtained in laboratory. We found that such a modelling strategy could be particularly useful in three ways:

1) Computational learning models can be fitted to individual choice series, to identify different types of (averaged) behavior (Exp. 2-4). By using different types of models varying in their strategic sophistication (order of beliefs generated), we were able to capture between-subject differences in behavior and to establish that individuals can vary in their learning strategy. We thus used computational modelling to build an abstract measure of the level of strategic learning engagement, as discussed in Chapter II.B.1. This strategy appeared to be fruitful as we showed that the level of strategic learning we computed for each subjects (SL level) fitted both our theoretical and simulation predictions, and eventually lead to correlational results replicated across experiments,

2) Computational models can be also used to simulate *ex ante* the expected behavioral data in a given experimental setting such as a repeated game interaction (Exp. 1). Simulating how a computerized agent, modelled by a given learning rule with given parameters, would behave in a game might help refine the experimental design to improve the quality of the experimental data eventually obtained and enhance the statistical power of the following analyses of the dataset (Palminteri et al, 2017).

3) Computational modelling provides us with an additional experimental approach, giving the possibility to manipulate one side of the interaction in repeated games, and to control for instance the action-outcome contingencies that a player can learn (Exp. 3-6). As presented in this thesis, this approach could be used to test how the strategic learning sophistication of the opponent (Exp.3,4) or the statistical redundancies in its choice series affect human performance or learning (Exp.5,6). It is worth noting that this experimental strategy has also been recently advocated in behavioral game theory by Spiliopoulos (2013b). Actually, the use of computerized agents can be pushed even further, by inducing online

*individual-specific behavioral manipulation*. In Exp.3, typically, the computerized agent generates its choices through a computational learning model that uses the information provided by either its own choice (Q-learning), the choices of the participant with which it interacts (Fictitious), or both (Influence). The algorithm generates its choice at each trial based on the value of the hidden variables the model updates dynamically (trial-by-trial), and given the parameter values entered by the experimenter. Thus at each trial the agent generates a map of the unobservable (hidden variables) that embodies an information about the current interaction, and by extension about the participant's behavior. The experimenter could thus develop a (meta) algorithm which would use these values generated (indirectly) by the choices of the participant to change the parameter values of the model, to drive online learning strategy adaptation depending of the behavior generated by the human encountered. To our knowledge this computational strategy has not been yet implemented in an experimental setting<sup>1</sup>.

A second general observation regarding the approach we employed in this PhD work can be made: game theory provides a useful framework to study learning in (dyadic) social interactions, and this at three levels: experimental, analytical and conceptual. Experimentally, games, which model strategic interaction in a minimal fashion, are similar to a Markov decision process (MDP), (with an exit probability as the number of games (trials) that will be played is unknown), in which players are provided with a set of actions and a known state (matrix cell) structure with deterministic payoffs associated to each, only the transition function is unknown as it depends upon the choice behavior of the opponent. As humans interact directly with another human (or with a human-like algorithm), it thus provides a simple social setting to study the learning process underlying human behavior in dynamic interactions.

At the analytical level, the value of such a model of social interaction, is that it comes with a strong theoretical framework which provides a benchmark to study human behavior. Indeed, Game theory formulates (mathematical) solution concepts, such as the Mixed Strategy Nash Equilibrium (MSNE), which prescribes an (mutually) optimal behavioral strategy. This allows to study human behavior and its departures from such game optimality. In Exp.2-4, we started with the assumption that humans differ in their belief accuracy over their opponent behavior, as MSNE is achieved through mutual best response. By first analysing how subjects' aggregated choices differed from the prescribed choice distributions (Exp.2), we were able to formulate the hypothesis that the relative performance (in comparison with the one of the opponent) was more informative of the accuracy of the behavior in the game than absolute performance (total points earned in a block). In Exp.4, the benchmark of MSNE provided us with the insight that the more players in the advantageous role engaged in high-order learning (SL level) against a low SL (belief-based) algorithm, the more they departed from the MSNE distribution, while still managing

---

<sup>1</sup> Still we would like to mention that this approach is inspired by Geana & Niv (2014), who used the inter-block time interval in a (non-social) armed-bandit learning task to (very basically) optimize a computational model on the choice series the participant generated in the first block to tune the parameters of the task in the second block.

to increase their final earnings. This result was the only striking difference with the initial (human-human) version of the task (Exp.2), and lead us to infer that subjects might have detected and used statistical redundancies in their opponent's past play to improve the accuracy of their belief and exploit their opponent's choice behavior beyond the mutually optimal distribution (MSNE). Thus, taking the MSNE as a reference point for belief accuracy provided us with a guide for behavioral analyses, computational modelling adequacy, between-subject comparison, and simulation prediction.

In fact, the theoretical framework in games turned out to have been useful at a third level, from a conceptual viewpoint. Indeed, research in behavioral game-theory proposed that the mutual knowledge of rationality (mutual best response) premise of the MSNE concept should be relaxed (Chapter I section III.A) in order to take into account the empirical departure from equilibrium systematically observed in human choice. Following this line, a class of model has been proposed based on the idea of bounded rationality. The level-k/CH models (Crawford et al, 2013) posit that players differ in their level of (iterative) strategic sophistication when reasoning in one-shot games, and empirical data suggest that such models effectively capture the choice departure from MSNE. The idea behind this conceptual framework is that humans are cognitively constrained (bounded), or at least that humans differ in their propensity to exert cognitive control in order to engage in higher strategic reasoning levels.

Based on this theory, we proposed that the engagement in strategic learning (SL) might as well be constrained. Indeed, previous studies suggested that humans can use past outcomes and track the past actions of the opponent in order to adapt their following decision (Zhu et al, 2012), but also that they can take into account their own influence over the interaction (Hampton et al, 2008). This hypothesis lead us to test different models of various SL level (from Q-Learning, up to the Influence model developed by Hampton et al, 2008). These models embodied different hypotheses regarding the type of learning and the level of (strategic) sophistication of the beliefs generated. Our data (Exp1-4) seem to fit this conceptual framework of bounded rationality as dissociating between different levels of strategic engagement during a repeated interaction allowed to capture part of the variance in the population, but also to better understand the consistent departure from equilibrium observed in the behavioral game theory literature.

Crucially, our results showed that engaging in higher-order learning lead to more accurate beliefs, and thus to higher performance in a competitive game interaction. A second concept derived from the MSNE prescription is the idea of randomization. MSNE assumes that, to be unpredictable, players should randomize over their action-set following the prescribed (mutually optimal) probability distribution. However, empirical data in both behavioral game theory and neuroeconomics shows that humans fail to fully randomize and unconsciously display patterns or statistical redundancies in their behavior (such as over-alternation) (Camerer, 2003). This lead us to specifically investigate in Exp.5 the ability of humans to detect and exploit (learn) the probabilistic patterns embodied in the choice behavior of a (computerized)

opponent, and to study how (if) this information could be implemented in higher-order beliefs formed through strategic learning (Exp.6).

## **2) Implication of the present work for the related literature**

We first showed that in repeated strategic interactions humans vary in their level of strategic learning sophistication, from reinforcement learning to higher-order belief-based learning in which not only the individual best responds to the beliefs formed over their opponent's choice distribution (belief-based learning), but also implements higher-order beliefs that take into consideration how much they believe their own past choices may have influenced their opponent's action distribution (strategic learning). The computation of such higher-order beliefs has been shown to be subserved by the medial prefrontal cortex (mPFC) which implements the influence-related signals encoded by mentalizing-related brain areas such as the (right) tempo-parietal junction (rTPJ) (Hampton et al, 2008; Hill et al, 2017).

In regard to the cognitive neuroscience literature, the computation of beliefs over the action-outcome contingencies of a choice environment is computationally costly but allows a more flexible type of learning (model-based) that increases the adaptability of the choice behavior (Doll et al, 2012; Khamassi & Humphries, 2012). In non-social decision-making, the overall performance is a well-established measure of accuracy of beliefs over the choice environments, as animals maximize their subjective value (see introduction I.A). In the social interactions however the notion of belief accuracy over someone else's intention is difficult to estimate as the actual action-outcome contingencies emerging in the behavior of an individual are dictated by internal states and beliefs, thus unknown to the experimenter (Zaki & Ochsner, 2011). Competitive strategic interactions modelled as games simplify this endogenous problem as the goal of the opponent is obvious: maximizing as much as possible her outcomes. The common knowledge of the payoff structure of the game clears potential uncertainty about the belief that the opponent aims to maximize (best response). Accordingly, we found (Exp.2) that individuals endorsing a delicate position in a competitive game (strategically disadvantageous role, in which the action linked to the highest reward does not align with the MSNE prescription) start with this belief (prior) as they seem to engage in (iterative) strategic reasoning in the absence of past experience to be learned from.

Moreover, strategic games provide a theoretical benchmark to estimate belief accuracy, departure from optimal belief, as discussed in the previous section of this chapter. Competitive game interactions thus allow to quantify a measure of belief accuracy and provide an ideal framework to test how model-based learning might take place. If the MSNE provides a relatively independent and formal measure of (mutual) optimality, controlling the opponent's choice behavior through the use of computerized agents informs about the specific accuracy of the behavior adopted but also about the optimal (normative) behavior (learning strategy) to employ.

By their structure, repeated game interactions thus permit to fill the conceptual gap with the non-social learning literature. Behavioral game theory proposed learning strategies close to the ones developed in cognitive neuroscience, from model-free reinforcement learning to model-based learning. The latter has been implemented either through direct action-outcome learning (similar to a simple Bayesian generative model learner, with negligible priors) (Fictitious play, Fudenberg & Levine, 1998), or through the use of a heuristic like the action (expected) value update from the counterfactual outcome (EWA, Camerer et al, 2002). The crucial component of influence in learning is an add-on specific to social learning as the non-social environment is usually not directly affected by our own choices, or not to the point where this component should be considered as highly predictive of the action-outcome contingencies in the world. Similarly to the generative models developed in the Bayesian framework applied to non-social decision-making, and which track and update hidden hyperparameters of the task structure such as its fluctuation in volatility, Devaine et al (2014) proposed a Bayesian influence model built upon the concept of hierarchy of beliefs proposed by the level-k/CH models. Their study showed that even if humans are better modelled on average as highly strategic learners, taking into account the influence of their own behavior on the one of the opponent, and beyond, individuals still differ in their engagement in high-order belief learning in a symmetric game in which perspective taking is facilitated. The Influence model developed by Hampton et al (2008) offers a simpler way to capture strategic learning engagement and higher-order belief formation, while allowing for comparison with the belief-based and reinforcement models developed in the game theoretical literature.

Using game-theoretical manipulations along with model simulations (agent-agent) and empirical analyses of human-human and human-agent interactions (Exp1-4), we provided extensive evidence that the level of strategic learning, similarly to the level-k/CH, allows the formation of more accurate beliefs over the choice behavior of the opponent, thus converging towards theoretical predictions in a competitive repeated game interaction. This result matches the prediction made by Camerer (2003) according to whom MSNE can be seen as an equilibrium in beliefs in which players do not need to randomize, as long as other players cannot guess what they will do. The level-k/CH model posits that humans iterative (strategic) reasoning is somehow bounded, leading to different orders of belief in the population. Studies in behavioral game theory have suggested that the level of strategic sophistication in one-shot games might be cognitively bounded (Camerer et al, 2002), and that executive functions such as working-memory of logical reasoning might constrain the formation of higher-order beliefs (Carpenter et al, 2013). However, the question of what constraints the level of strategic learning and the formation of higher-order beliefs remained unanswered. In their symmetric repeated game, Devaine et al (2014) failed to link subjects' strategic learning level with other cognitive functions, and here we failed (Exp.2) to find a correlation between the strategic sophistication level in one-shot games and the strategic learning level in the repeated game. We found that the working-memory capacity of humans playing a competitive game

in a position requiring to form higher-order beliefs to overcome their endogenous (strategic) disadvantage correlated to their level of strategic learning engagement. Conversely, the SL level displayed by their opponent in the advantageous position correlated in the first interaction block with various measures of reasoning abilities. In fact, taken together the results presented in Exp.2-4 seem to suggest that the boundary leading heterogeneity in strategic learning level in the population might be a *rationaly cognitive* one, in the sense that given an individual trade-off, humans might engage in a cost-benefit analysis driving their engagement in high-order belief formation through strategic learning. In fact, such an interpretation of bounded rational models has been proposed in one-shot games too (Alaoui & Penta, 2015). This hypothesis needs to be tested specifically. However, it is worth noting that it fits the current literature on model-based learning in non-social choice environments as well, which suggests that the level of engagement in action-outcome learning might come at a computational cost requiring a high level of cognitive control (Otto et al, 2014). Recently, authors in the field proposed that such a trade-off between model-free and model-based reinforcement learning might be arbitrated in a cost-benefit fashion (Kool et al, 2017).

Besides, our data also provide insights for the leader-follower dynamics often observed in repeated competitive games (Frey & Goldstone; Seip & Grøn, 2016). Indeed, we showed that the *strategic asymmetry* of the game lead, for each position (advantaged or disadvantaged), to very distinct influences of the level of strategic learning engagement: the former choice dynamic was mainly influenced by the consequences of her own past decisions on the strategic interaction, while the latter in a teaching position was driven mainly by the behavior of her opponent.

In line with the hierarchy of beliefs proposed by the level-k/CH model, our results (Exp.1) indeed suggest that forming accurate beliefs over the other's behavior requires to be capable of a higher-order of strategic learning. However, our data (Exp.4) also show that players capable of engaging in strategic learning and playing in an advantageous position against a computerized agent of lower SL level (belief-based learning strategy), might have been able to exploit the statistical redundancies from the behavior it displayed in order to best respond to this new belief, and to deviate from the mutually optimum frequency (MSNE) to increase their earnings.

We then aimed to test if actually humans can detect statistical redundancies when controlling specifically for the patterns embodied in the opponent choice series (Exp.5,6). Our results show that humans can learn and best respond to the patterns when they are deterministic, as already highlighted by Sonsino (1997) two decades ago, who suggested that players must recognize the repeated pattern if *it has been repeated successively with no interruptions a large enough number of times*. In behavioral game theory, Spiliopoulos (2012) tackled this question and proposed to extend a belief-based model computing the



probability (weighted frequency) of the opponent's actions conditionally on the last couple of choices she selected. He showed that such pattern fictitious strategy fitted better human's choice data. However, we argued in the present manuscript that such a strategy did not ensure that humans could actually detect statistical redundancies, as in his task the opponents were human that might not have displayed any strong patterns within their choice series. And indeed, we showed that in such human-human competitive interactions, a model incorporating a parameter of influence in belief learning and leading to a higher-order strategic behavior fitted better the choice data than such pattern belief-based model (Exp.2).

We thus tested (Exp.5) this pattern-learning hypothesis directly by manipulating one side of the interaction using computerized agents, allowing us to observe that humans can indeed detect and learn probabilistic patterns hidden in the opponent choice series. Nevertheless, we found that important between-subjects heterogeneity could be observed. More intriguing, when specific sequences of deterministic choices were generated by an opponent randomizing over the MSNE distribution in a competitive game (Exp.6), participants appeared to be bad pattern learners (or good detectors, but bad exploiters), as actually no strong evidence of pattern learning could be found.

One possible hypothesis is that strategic competitive interactions elicit beliefs that the opponent must respond to some beliefs over the strategic nature of the game. As mentioned in the introduction, recent studies in psychology (Jara-Ettinger et al, 2016) suggest that humans, and this even at an early age, start with the prior over another person's intentions that her behavior is driven by the maximization of the outcome of her actions. The results presented in Exp.2-4 indeed suggest that most of individuals engage in some sort of belief-based learning over their opponent's behavior. However, to our knowledge, if the research in cognitive neuroscience suggests that humans can combine model-based and rule-based behavior in non-social decisions tasks requiring to do so (Donoso et al, 2014), the question of the human's ability to form beliefs about rule-based behavior, such as the one used by the computerized opponent in our experiment, remains open (but see the recent work by Velez-Ginorio et al (in prep.) going in this direction). However, the rule-based behavior employed by these computerized agents to generate experimentally tractable choice patterns was not directed towards a specific goal but should have been considered as systematic biases emerging from a (strong) inability to randomize. As mentioned at the end of Chapter IV, more work is required to investigate under which conditions pattern learning might take place in strategic social interactions and how this type of learning interact with higher-order belief-learning.

Finally, we showed that the behavior of the opponent has a limited impact on human's own overall learning strategy in (competitive) repeated game interactions. In Exp.2-3 we could not find evidence that the level of strategic learning sophistication (SL level) of the opponent, either human (Exp.2) or

computerized (Exp.3), influenced the SL level of the subjects. However, we observed that the SL level of the subjects endorsing the disadvantageous role (Players 2) was quite stable across opponents in both experiments, but also that their SL level seemed to be correlated to their working memory capacity suggesting that their strategic learning engagement was driven by endogenous inter-individual differences. Moreover, the transfer effect observed in Exp.4 was limited to the subjects who already engaged in a high-level of strategic learning.

In the group of participants who endorsed the advantageous role in the game, the SL level did not correlate strongly across opponents, and in Exp.2 this measure of sophisticated learning engagement correlated with reasoning abilities as measured in additional tasks in block 1. Still, no systematic increase or decrease in their sophisticated learning engagement could be predicted from the level of their opponent (either previously or currently encountered) in the second block.

Similarly, in Exp.3 as in Exp.4 the (manipulated) SL of their opponent in previous or current interaction did not influence in a coherent way their strategic learning level despite the low congruency in SL level also observed between the two blocks. While Hampton et al (2008; Hill et al, 2017) did not investigate this question, Devaine et al (2014) coherently found a correlation in strategic learning level across the computerized opponents (with different SL levels). These results are thus congruent with the idea of an improvement in behavioral accuracy through the formation of higher-order beliefs (Camerer & Ho, 1999). Indeed, implementing the Influence factor in the learning of the action-outcome contingencies in the behavior of the opponent should not prevent the learning of a simpler generative model.

Similarly, in Exp.4-5 we did not find any evidence of an impact of the type of pattern-driven (rule-based) opponent on the subject's learning strategy. In Exp.4 we found that the ability of the subjects to learn the statistical redundancies in the choice behavior of a probabilistic pattern opponent (O) were also able (on average) to learn the patterns in the deterministic (albeit adaptive) opponent (D). In Exp.5 the performance in the pattern-driven trials in the test block (H) was correlated to the performance in the Pattern-driven-only block (P), while the performance in the Belief-driven trials of block (H) was correlated to the performance in the Belief-driven-only block (B). Taken together these results thus suggest that our experimental work failed to trigger important (quantifiable) changes in the learning process (strategy) in which humans engaged throughout repeated game interaction, leading us to conclude that such a learning engagement is mainly driven by endogenous characteristics and not by the behavior displayed by their opponent.

## B) Future research directions

At the end of chapter IV we suggested that strategic individuals might combine the two types of information regarding the behavior of the opponent, from learning the statistical redundancies in her past choices and forming iterative high-order (influence) beliefs through strategic learning.

Recently, a neuroscience study (FitzGerald et al, 2017) showed that adding sequence representation of state to the generative model of a Bayesian learner (thus adding one dimension to the MDP representation of the choice environment) better fitted the behavior of participants performing a probabilistic learning tasks, in which action-outcome probabilities were switched (reversals) at regular intervals. Using fMRI analysis, they showed that the ability to take into consideration the state sequence was linked to higher grey matter density in the Brodmann area 10 of the PFC, while the length of the sequence correlated to the lateral prefrontal cortex activity. Interestingly, the Brodmann area 10 has been recently proposed, in an extensive review on human and non-human primates (Mansouri et al, 2017), to subserve the management of competing goals in decision-making.

Thus, it does not seem implausible that forming more accurate action-outcome contingencies in a strategic repeated interaction might require to alternate between the information related to the time-related belief over the statistical redundancies in the opponent's choice series, and the strategic information provided by forming high-order beliefs implementing the trial-by-trial influence of one's own actions.

It is indeed possible that humans can alternate between two different types of learning during a strategic interaction. Wan et al (2015), for instance, showed that the activity of dmPFC (ACC) matches the relative weight given to the value-related computations associated to two different types of strategies (one related to attack moves, the other to defense moves) in Shogi (Japanese chess).

The mPFC has been even related to spontaneous strategy switches in a complex non-social task (Schuck et al, 2015), thus suggesting that during repeated games changes in learning strategy might occur. This is problematic since the methodology presented in this thesis requires to consider the entire choice series of an individual to estimate (approximate) her (averaged) level of strategic learning (see related discussion at the end of Chapter II, B.1). One possibility would be to manipulate on-line the behavior of the computerized opponent to make it switch between either a more oriented rule-based behavior, or a belief-based (adaptive) learning strategy either at fixed intervals of trials, or once exploited by the participant in a dynamic fashion (See section A.1 of the present discussion).

It is worth noting that the interdependency of the behaviors in (dyadic) social interactions, and specifically in the strategic ones, makes these kinds of manipulations very tricky as a change in the opponent's behavior might also trigger a change in the strategy of a (strategic) participant. This covariation in behavior is what makes the concept of equilibrium so delicate, since a strategic individual should find, for instance, a balance between exploiting (best responding) the statistical regularities in the opponent's

behavior while still making sure that this best response does not induce any change in these learned patterns. In other words, in a repeated competitive game the goal is to override the opponent's behavior in a way that she does not realize it.

We argue that the question of influence in high-order belief-learning remains too weakly characterized. In the Influence model of Hampton, as in the sophisticated Bayesian version of Devaine, the influence component is a statistical one. This means that while it provides a useful tool to analyse choice data, its conceptual power is restrained. Indeed, no difference is made between the actual influence of one's actions on the choices of the opponent (covariation), and the beliefs that one's behavior influences the other. This *representation of our own influence* relates to the concept of agency.

Agency, or the awareness of being in control of our own actions and responsible for their outcome (Haggard & Tsakiris, 2009), has been recently found to be indeed implicated in social interactions. Based on Pacherie's (2012) theory of joint action and agency, Bolt & Loehr (2017) indeed showed experimentally that when performing joint (motor) actions<sup>2</sup>, a feeling of joint agency (shared control over the observable outcome of the interaction) emerged when individuals could successfully predict the other's action.

As predictability seems to play a role in the feeling of agency, we propose that the belief of one's influence over the opponent's choices might as well be triggered by manipulating the level of joint predictability. One might imagine an experiment in which participants would interact through a repeated (competitive game) interaction with a computerized opponent (as in Exp. 3) which would be modelled as a (weighted) fictitious model with different levels of determinism. We can do this by manipulating the exploration parameter of the action selection process ( $\beta$  in the softmax, see chapter II) at regular intervals throughout the interaction. This way, the model would switch (either at regular intervals or in an adaptative way) between MSNE randomized play, and a fully exploitative fictitious that generates a choice behavior almost entirely predictable from the participants' own previous choices (depending on the decay parameter of the model, see appendix 2)<sup>3</sup>. If humans in such experimental setting indeed report higher joint agency when their influence over the opponent's choices is made more evident, then the question becomes: how does one's (acknowledged) influence in a strategic interaction impacts her strategic learning engagement? In other words, can joint agency promote the formation of higher-order beliefs?

Related questions follow such as the question of the arbitration between the strength of influence-driven

---

<sup>2</sup> Participants were asked to perform motor actions to produce self-paced tones conjointly with another participant (not visible), and report their feeling of agency over the tones (self, other, joint). The other participant was in fact a computer that produced either predictable tones or unpredictable ones.

<sup>3</sup> It is worth noting however that the two learning strategies (noisy, exploratory / predictable, exploitative) should be clearly distinguishable, as contamination effects of one condition onto the other might happen. Caruana et al (2017) indeed showed that the belief of agency might sometimes override the actual correlation in behavior, i.e. actual influence.

beliefs and the pattern learning strategy, or the teaching behavior in leader-follower dynamics of play. Interestingly, Xue et al, (2013) showed that the agency over choices increased the risky-behavior of participants performing a decision-making task in which the outcomes were actually randomly generated, and activated specifically more the lateral and medial PFC.

The question of agency in learning is closely related to the topic of causal learning. Infants for instance seem to not simply learn the probabilistic associations between world events, but could compute the hidden causal structure generating them (Gopnik & Wellman, 2012), and this through the interaction with others (Legare et al, 2017). Studies in computational neuroscience also started to tackle the question of causal learning (Jocham et al, 2016). Very recently Morris et al (2017) suggested that humans could learn action-outcome contingencies through causal learning, and that the mPFC played an important role in encoding these associations using prediction error.

Another appeal to this line of inquiry is the conceptual relationship between agency and regret, as the latter, the experience of regret, depends on the former, since it takes place only when we are aware (believe) that by choosing another action we could have obtained a different outcome. Regret and counterfactual learning share common themes (Coricelli & Rustichini, 2010) and can thus present another way to extend this hypothesis to strategic learning.

## C) General conclusion

During this PhD work we aimed at improving our understanding of the cognitive mechanisms at play during (dyadic) social interactions. To do so, we combined the computational modelling and the conceptual approaches of cognitive neuroscience, to the theoretical and experimental insights developed in behavioral game theory.

We decided to focus on behavior for two main reasons:

- 1) From a cognitive neuroscience perspective we believe that choice behavior in strategic interactions provides a rich source of information, mostly embodied in the observed between and within-subject variance, that has not yet been fully exploited, but also in related behavioral measures such as reaction times. We argue that a better understanding of the behavior can lead to a better characterization of the cognitive mechanisms and computations performed by the brain (see the recent manifesto by Krakauer et al, 2017).
- 2) From a behavioral game-theory standpoint, improving the understanding of the cognitive mechanisms

at play in (repeated) games through experimental manipulations and computational modelling, and studying to which extent the experimentally observed behavior relates to the normative framework, might allow us to improve its predictive power (see the review by Konovalov & Krajbich, 2016).

Thus we believe that refining our understanding of how the two fields of research interact in laboratory might lead to the development of a dedicated neuroeconomics of strategic interaction.

- Appendix I -

Griessinger, T., & Coricelli, G. (2015). The neuroeconomics of strategic interaction. *Current Opinion in Behavioral Sciences*, 3, 73-79.

# The neuroeconomics of strategic interaction

Thibaud Griessinger<sup>1</sup> and Giorgio Coricelli<sup>2</sup>

We describe here the theoretical, behavioral and neural bases of strategic interaction — multiagent situations where the outcome of one's choice depends on the actions of others. Predicting others' actions requires strategic thinking, thus thinking about what the others might think and believe. Game theory provides a canonical model of strategic thinking implicit in the notion of equilibrium and common knowledge of rationality. Behavioral evidence shows departures from equilibrium play and suggests different models of strategic thinking based on bounded rationality. We report neural evidence in support of non-equilibrium models of strategic thinking. These models suggest a cognitive-hierarchy theory of brain and behavior, according to which people use different levels of strategic thinking that are associated with specific neural computations.

## Addresses

<sup>1</sup> Département d'Etudes Cognitives, Laboratoire de Neurosciences Cognitives, Inserm unit 960, Ecole Normale Supérieure - PSL Research University, Paris 75005, France

<sup>2</sup> Department of Economics, University of Southern California, 3620 S. Vermont Ave., Los Angeles, CA 90089, United States

Corresponding author: Coricelli, Giorgio ([giorgio.coricelli@usc.edu](mailto:giorgio.coricelli@usc.edu))

**Current Opinion in Behavioral Sciences** 2015, **3**:73–79

This review comes from a themed issue on **Social behavior**

Edited by **Molly J Crockett** and **Amy Cuddy**

For a complete overview see the [Issue](#) and the [Editorial](#)

Available online 7th February 2015

<http://dx.doi.org/10.1016/j.cobeha.2015.01.012>

2352-1546/© 2015 Elsevier Ltd. All rights reserved.

## Introduction

Everyday social interactions affect our individual decisions. What makes information relative to others relevant for our own decisional process, and how it is dynamically incorporated in our valuation system remain open questions in neuroscience. The specific case of strategic interactions where the outcome of one's action depends directly on the other's behavior narrows down social interactions to situations where each agent should take into consideration not only her own actions, but also the actions of the others [1,2]. Game theory (GT) prescribes

precise theoretical solutions for optimal behavior embodied in the premises of rationality (and the notion of equilibrium), and provides a benchmark for the analysis of its behavioral departures. The emergence of bounded rationality models [3–5] and behavioral game theory provide a theoretical framework to unravel the neural roots of strategic reasoning, and shed light on the decision-making mechanisms involved in social interaction. Within this framework, we review work on the neural substrates of equilibrium and nonequilibrium play, and we identify a network associated with strategic thinking in interactive games. In addition, we investigate the interplay between uncertainty and belief inference in repeated interactions with a network related with strategic thinking, thus identifying, respectively, neural substrates of strategic uncertainty and strategic learning.

## Theoretical background of strategic interaction

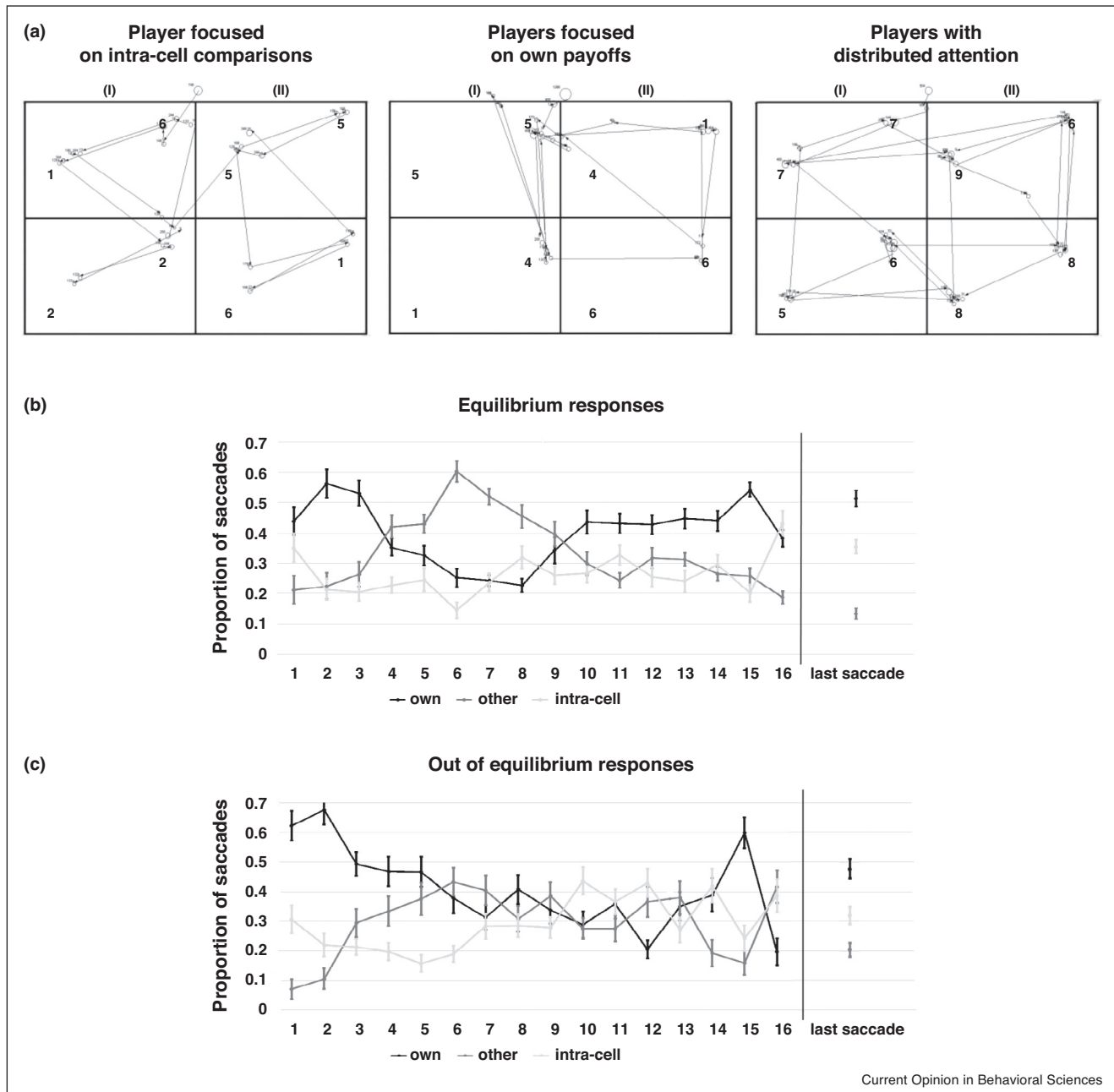
Game theory models strategic interactions as games representing decisions between agents where one's payoffs depend on the other's actions, therefore extending the model of individual decision making to the understanding of the interactions in multi-agents situations. Solution concepts provide an answer about which action profile will result from playing a game. Nash equilibrium [6], for instance, prescribes an (optimal) action profile by which no player can increase her payoff by changing her action given the other players' (optimal) actions. Players are assumed to select strategies that maximize their utility over the payoffs of the game. The choice of the strategies is based on their beliefs about what other players will do. At equilibrium beliefs are correct. Nash equilibrium implies that players' are certain and accurate about the strategies of the others, indeed the equilibrium is an equilibrium in beliefs that assumes rational expectation and mutual knowledge of beliefs, and thus mutual rationality.

## Do people (think and) play at equilibrium?

Equilibrium reasoning (i.e., rationality-based inference) can be cognitively extremely demanding and eventually implausible. Several experimental and empirical studies show behavioral responses that deviate from the prescription of standard game theory, and report extensive evidence of non-equilibrium play [7,8]. From the basic assumptions of strategic reasoning in standard game theory, there are two main departures suggested by behavioral game theory and



Figure 1



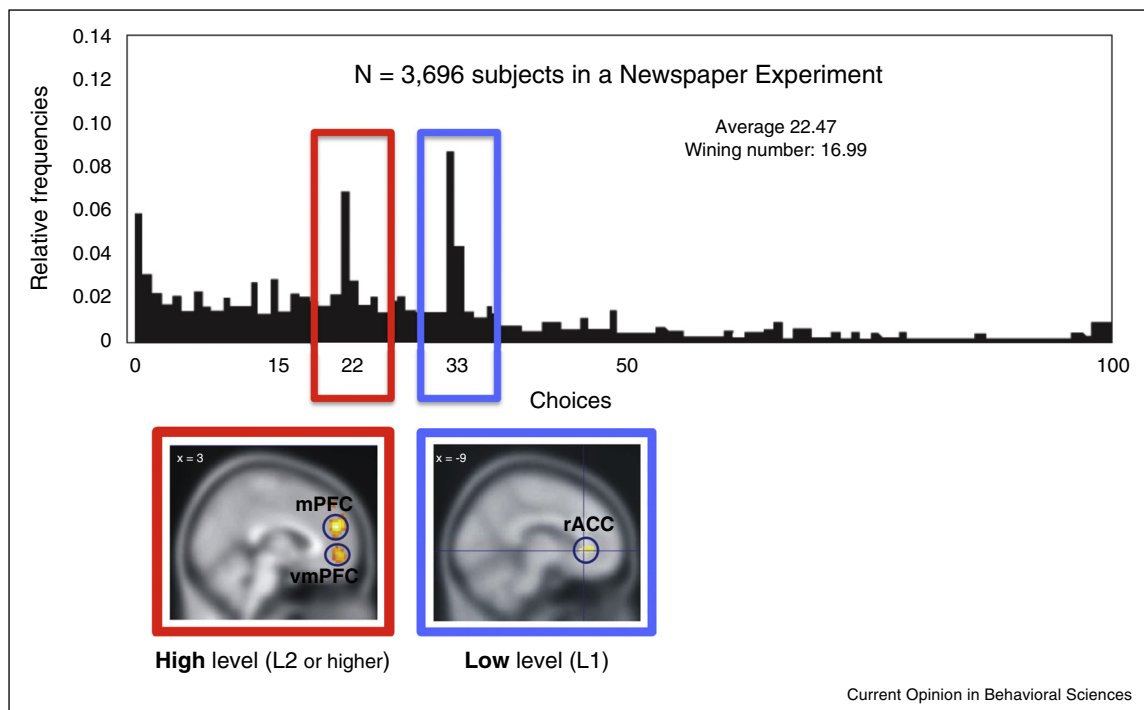
Equilibrium and non-equilibrium play in normal form games: an eye-tracking study [15\*]. The authors used eye-tracking to measure the dynamic patterns of visual information acquisition in two players normal form games (i.e. represented in matrix form). Panel (A) shows the pattern of saccades (i.e. eye movements) performed by 3 typical participants who played as column players (i.e., can take action I or II and have their own payoffs on the top right corner of each cell of the game matrix) in the experiment: (left panel) shows data from a participant who focused her attention on own and other player's payoffs within each cell of the matrix; (center) focused on own payoffs (i.e., systematically neglected the payoffs of the other player); and (right) players with distributed attention (strategic player). Lines indicate the saccades and the circles the fixation location. Panels (B) and (C) show the first 16 saccades (mean and standard error) in a group of players clustered as distributed attention (strategic players, i.e., level 2 in CH model), divided by equilibrium responses, Panel B: shows that participants started looking at their own payoffs, then they evaluated the payoffs of their counterpart, and finally, they chose their best response when re-evaluating their own payoffs; and out of equilibrium responses, Panel C: showing an undefined temporal pattern of visual analysis. On the right side of each Panel are reported the proportion of own, other and intra-cell saccades at the time of choice (last saccade). These data show how deviation from a distinctive and well-characterized pattern of visual information acquisition determines out of equilibrium behavior. Adapted from [15\*].

## ABSTRACT

Social interactions rely on our ability to learn and adjust on the spot to the other's behavior. Strategic games provide a useful framework to study the cognitive processes involved in the representation of the other's intentions and their translation into the most adapted actions. In the last decades, the growing field of behavioral economics provided evidence of a systematic departure of human's behavior from the optimal prescription formulated by game theory. Based on recent advances in cognitive sciences, we hypothesized that characterizing the source of heterogeneity in behavior might provide key insights to understand the boundaries over human social learning, and therefore deviation from mutually beneficial interactions.

We first address the question of the interplay between the game environment and the heterogeneity in formation of high-order beliefs over the opponent's behavior through strategic learning. We show that in a competitive repeated interaction, the payoff structure of the underlying game can influence the engagement in strategically sophisticated learning and explain deviation from game optimality (equilibrium). Our data suggest that participants in a disadvantaged role are constrained in their learning sophistication, and thus in the overcoming of their position, by their own cognitive capacities. Their opponents, albeit advantaged, still need to engage in strategically sophisticated learning but to follow and adjust their behavior in order to maximize their earnings. This study provides the first evidence of the key implication of strategic learning heterogeneity in equilibrium departure and provide insight to explain the emergence of a leader-follower dynamics of choice. In addition our results suggest that a cost-benefit analysis might drive the engagement of strategic players in a more sophisticated learning process. In a second step, we investigated the hypothesis that the depth of strategic learning is not the only factor in play to grasp the other's mind during competitive interaction, but that the capacity to detect and exploit patterns in her behavior is also important. We found that not only subjects were able to detect patterns in the opponent's behavior, but that the capacity to do so was not correlated to a lower engagement in sophisticated strategic learning, therefore suggesting that humans can combine information from both types of learning to improve belief accuracy during social decision making.

Figure 2



Neural substrates of High (Level 2 or higher, mPFC) vs. low (Level 1, rACC) level of strategic reasoning in the Beauty Contest game. In the experimental game, participants choose a number between 0 and 100. The winner is the person whose number is closest to  $2/3$  times the average of all chosen numbers. Level- $k$  model (iterated best response) predicts that a naïve player (level 0) chooses randomly. A level 1 (low level) player thinks of others as level 0 reasoning and chooses  $33 (=2/3 \times 50)$ , where 50 is the average of randomly chosen numbers from 0 to 100. A more sophisticated player (level 2, high level) supposes that everybody thinks like a level 1 player and therefore he or she chooses  $22 (= (2/3)^2 \times 50)$ . Zero is the equilibrium solution of the game.

Adapted from [48,22].

bounded rationality: the first is about equilibration of beliefs (i.e. the assumption of correct beliefs about the behavior of the others), the second is about errors in the choice process. In what follows we discuss two leading non-equilibrium models of strategic thinking and we report evidence about their neural substrates: (1) Quantal Response Equilibrium (QRE); and (2) level- $k$  and Cognitive Hierarchy (CH) models.<sup>3</sup>

### Non-equilibrium models of strategic thinking

#### Noisy and stochastic choice: Quantal Response Equilibrium

Quantal Response Equilibrium [9] belongs to a class of bounded rationality models that relaxes the assumption of best response and considers errors in choices, keeping the assumption of (statistically) accurate beliefs and equilibrium responses. In interactive settings, a small amount of noise can have large effect, and QRE models that incorporate stochastic elements in the analysis of interactive decisions can explain ‘anomalous’ behavior (i.e. deviation

from rationality) in several experimental games. According to QRE models individuals are more likely to select better than worse actions, but they are often unable to select the very best one. QRE theory has several features in common with findings in recent neuroeconomics literature on noisy and stochastic choice [10]. It has been recently suggested that QRE can be reduced to a form of bounded accumulation models [11,12], a class of models that has been proven relevant to capture under a common theoretical framework stochasticity in value-based decision, reaction time, and visual fixation [13,14]. In a recent paper Polonio *et al.* [15\*] observe that equilibrium play in normal form games corresponds to a distinctive and well characterized attentional pattern (in terms of transitions in visual information acquisition between own and other player’s payoffs), and any deviation generates non-equilibrium responses (Figure 1). This suggests how limited attention or noise in the decision process could lead to out of equilibrium behavior.

### A cognitive hierarchy theory of brain and behavior

Level- $k$  models [16,17] and Cognitive Hierarchy models (CH, [18,19]) maintain the rational assumption of best

<sup>3</sup> Additional models are  $k$ -rationalizability and finitely iterated dominance ([45,46], for an extensive review of non-equilibrium strategic thinking see [47\*\*]).

response to beliefs, but relax the assumption of ‘correct’ beliefs (and rational expectation about beliefs). This class of models considers the presence of heterogeneous players in terms of a hierarchy or level of strategic sophistication: level 0 players, are strategically naïve (e.g. they play randomly, or do not fully consider the incentives of the game), while higher levels iteratively best respond (i.e. respond optimally) to a distribution (Poisson for CH, and as k-1 for Level-k models) of lower levels players (e.g. L1 best respond to L0, L2 best respond to a distribution of L1 and L0, and so on). Limited strategic thinking (usually 0–2 steps of iteration) is due to limited cognition (limited recursive thinking, limited memory, etc. [20]) and personality characteristics such as overconfidence [21]. According to this model, high-level reasoners (L2 or higher) expect the others to behave strategically, whereas low-level reasoners (L1) choose based on the expectation that others will choose randomly.

Coricelli and Nagel (2009, [22]) ran a fMRI version of the ‘beauty contest game’ — a game suitable for investigating whether and how a player’s mental process incorporates the behavior of the other players in his strategic reasoning [23]. In their fMRI study Coricelli and Nagel found enhanced brain activity in the medial prefrontal cortex (mPFC), rostral anterior cingulate (rACC), superior temporal sulcus (STS) and bilateral temporo-parietal

junction (TPJ) when subjects made choices facing human opponents rather than a computer (that chose randomly) in the beauty contest game. This network is often associated to Theory of Mind (ToM) or mentalizing, thus the ability to attribute mental states and beliefs to others [24–27].


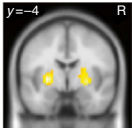

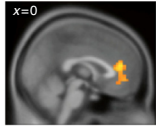

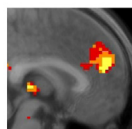
**Pattern of neural activity related with recursive thinking**

When Coricelli and Nagel (2009, [22]) analyzed separately L2 and L1 subjects, they found the activity in the medial prefrontal cortex to be stronger in subjects classified as L2 (Figure 2). Similar result was recently found in Bhatt *et al.* [28]fMRI results [22] show additional brain activities related to L2 versus L1 reasoning in the lateral orbitofrontal cortex and the dorsolateral prefrontal cortex, areas likely related to performance monitoring and cognitive control [29–31]. This suggests that a complex cognitive process subserves the higher level of strategic reasoning; consistent with the hypothesis that L2 or higher levels imply recursivity (reasoning about reasoning others) and the fact that a strategic player considers the impact of his or her own behavior on the behavior of the others.

**Strategic learning**

A critical aspect of strategic interactions lies in its time-dependent dynamic. Learning is functional in beliefs formation and in shaping social preferences (i.e. reputation,

Figure 3

Learning mechanisms	Levels of strategic thinking	Neural correlates
<p><b>Reinforcement learning (RL)</b></p> $V_{t+1}^a = V_t^a + \eta \delta_t$ <p>(Sutton &amp; Barto, 1998)</p>	<p><b>Level zero k=0</b></p> 	<p>Striatal activity:</p> $\delta(t) = r(t) - V_a(t)$ <p>RL Prediction error</p>  <p>(Zhu et al, 2012)</p>
<p><b>Fictitious Play learning (FL)</b></p> $P_{t+1}^* = P_t^* + \eta \delta_t^p$ <p>(Fudenberg et al, 1998)</p>	<p><b>Low level k=1</b></p> 	<p>rACC</p> $\delta_t^p = P_t - P_t^*$ <p>Belief Prediction error</p>  <p>(Zhu et al, 2012)</p>
<p><b>Influence learning (IL)</b></p> $P_{t+1}^* = P_t^* + \eta_1 \delta_t^p + \eta_2 \lambda_t^p$ <p>(Hampton et al, 2008)</p>	<p><b>High level k=2</b></p> 	<p>mPFC</p>  <p>(Hampton et al, 2008)</p>

Current Opinion in Behavioral Sciences

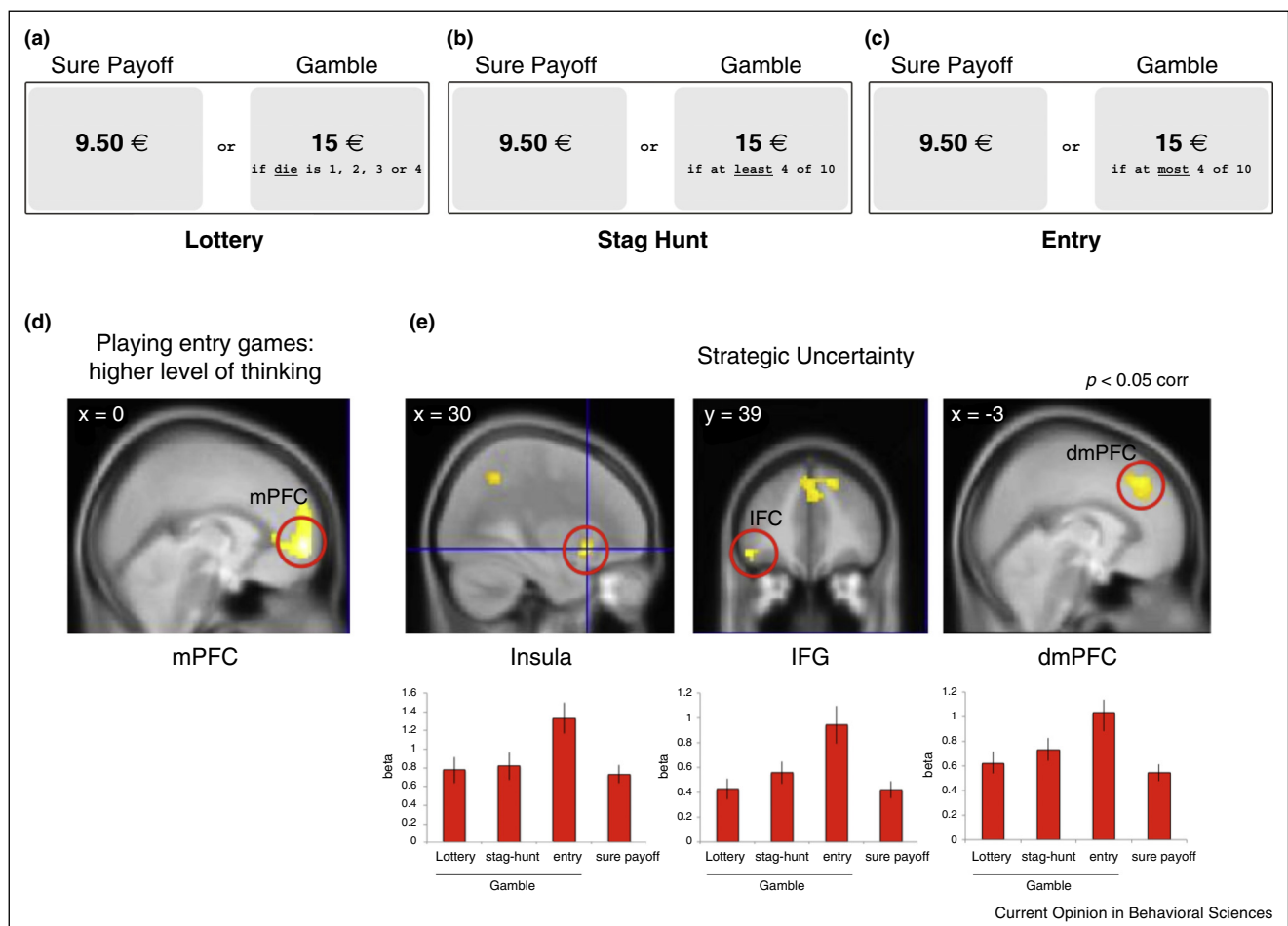
Thinking and Learning: computational and neural correlation between strategic thinking and learning. Level zero of strategic thinking can be associated with RL algorithms (Sutton and Barto [49]), low level (level 1) of thinking can be associated with Fictitious play algorithms (Fudenberg and Levine [50]) and high level of thinking (Level 2 or higher) can be associated with Influence learning algorithms. Adapted from [35\*\*,39].

reciprocity etc.). Several studies tackled the dynamic update of belief under a Bayesian framework [32\*,33]. Yoshida *et al.* (2010, [34]) first investigated the neural substrates of (optimal) dynamic beliefs formation in a repeated game in terms of estimates of the opponent's level of strategic sophistication and beliefs uncertainty. The results of this study show that the mPFC plays a role in encoding the uncertainty of inference of the strategy of the (computerized) opponents and the dLPFC is associated with the level of sophistication implemented by the subjects.

An alternative approach, in line with level-k and CH models, has been proposed by Hampton *et al.* [35\*\*] using learning models incorporating different levels of recursive information integration in repeated strategic games. Interestingly they show that the mPFC, found to

support high level of strategic thinking [22], also implements the individual propensity to dynamically incorporate in their learning process a representation of the opponent's adaptive behavior (captured in their model by a parameter of influence, i.e. how one's own actions had influenced the behavior of the other). The role of the mPFC has been found in other studies on social learning [36–38] as representing other's action-reward and action-outcome contingencies. In addition, the implication of the rACC, found to be active for a lower level of strategic thinking [22] has also been shown in repeated strategic games as associated to a lower level of sophisticated learning. Zhu *et al.* [39] showed for instance that during repeated strategic interactions the activity of the rACC correlates with the estimated departure from reinforcement learning (RL) to (first order) belief leaning. Seo *et al.* [40] recently showed

Figure 4



Strategic Uncertainty: how strategic thinking modulates the perception of risk in games. Participants played lotteries (A), stag hunt games (B), with a sure payoff choice (e.g. 9.50) and an uncertain choice (gamble) in which a player who chooses it gets 15 if at least 4 out all 10 players (including her) choose the gamble and otherwise she gets 0, and entry games (C) with a sure payoff choice and an uncertain choice in which a player who chooses it receives 15 if at most 4 out of all 10 players (including her) choose it and zero otherwise. (D) mPFC activity correlates with choices in the entry games only (thus reflecting higher level of strategic thinking). (E) neural network associated with Strategic Uncertainty (SU, SU entry > SU stag hunt = risk). Adapted from [41\*\*].

that the equivalent area in monkeys encodes the amount of switch from RL, a function of the ability of the computerized opponent to exploit its ongoing learning strategy. All together these results may suggest a cognitive hierarchy of strategic learning mechanisms rooted in a similar level of recursive information integration (see Figure 3).

### The role of mPFC in the interplay between deliberation (i.e. degrees of strategic thinking) and strategic uncertainty

Strategic uncertainty arises when the outcome of one's choice depends on other people's actions, and thus is the result of strategic interaction. Nagel *et al.* [41\*\*] investigated how this kind of uncertainty is related to exogenous individual risk and to degrees of strategic thinking (i.e. deliberation). The authors used fMRI to measure the neural correlates of uncertainty in lotteries (i.e. choice under risk) and two kinds of coordination games (i.e. strategic uncertainty), the stag hunt game, where participants have incentives to coordinate on the same action, and the entry game that incentivizes coordination on opposite actions. Solving the former requires low and the latter high degrees of strategic reasoning (of the kind 'I think that you think that I think etc.'). The results of this study (see Figure 4) demonstrate that a common brain network composed of the thalamus, dorsal medial prefrontal cortex, inferior frontal gyrus and anterior insula (commonly associated to individual risk [42]) is engaged by both individual and social contexts for the resolution of uncertainty. The activity in this network is similar in lotteries and in stag hunt games, but is higher in the entry games. They also found enhanced mPFC activity in the entry games, where more level of strategic thinking is required. Thus, the pattern of activity in the medial prefrontal cortex reflects the interaction between degrees of strategic thinking and uncertainty in interactive games: more deliberation correlates with higher strategic uncertainty.

### Conclusions and directions for future work

We can hypothesize that degrees of knowledge of the others and of the context, ranging from certainty to uncertainty, and the different levels of recursive reasoning (depths of reasoning: i.e. the player's mental processing that incorporates the thinking process of others in strategic reasoning), are crucial factors in the definition of the brain circuits needed to solve strategic interactive situations. The brain data reviewed here provide substantial support for a cognitive hierarchy model of strategic thinking. A higher level is associated with recursive thinking, which is the realization that others can also produce any thought process that we produce, while a lower level reflects self-referential thinking. Different portions of the prefrontal cortex clearly distinguish high-versus-low levels of strategic thinking, and naïve

versus sophisticated learning, thus encoding the complexity underlying human social behavior.

We believe that several lines of theoretical research could provide additional relevant insights and tools for the understanding of the neural basis of strategic interaction. Examples are concepts from epistemic game theory (EGT, [43]) and from Global games (GG, [44]). EGT studies the behavioral implications of different notions of rationality and mutual beliefs. EGT can provide important insights into the definition of types of players in terms of beliefs about the structure of the game and the strategies of other players in the game and others' beliefs, i.e. hierarchies of beliefs). The theory of GG relaxes the assumption of common knowledge and assumes that elements of the game (such as payoffs) are observed with a small amount of noise and that in an *ex ante* stage of the game any payoff is possible (global games). The assumption is that each player observes a private signal over the course of the payoffs. The result is a unique equilibrium in games with small amounts of noise.

### Conflict of interest statement

Nothing declared.

### References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
  - of outstanding interest
1. Lee D: **Game theory and neural basis of social decision making.** *Nat Neurosci* 2008, **11**:404-409.
  2. Bhatt M, Camerer CF: **The cognitive neuroscience of strategic thinking.** In *Handbook of social neuroscience*. Edited by Decety J, Cacioppo JT. Oxford Library of Psychology; 2011:949-960.
  3. Simon HA: **A behavioral model of rational choice.** *Quart J Econ* 1955, **69**:99-118.
  4. Rubinstein A: *Modeling bounded rationality*. MIT Press; 1998.
  5. Gigerenzer G, Selten R (Eds): *Bounded rationality: the adaptive toolbox*. MIT Press; 2002.
  6. Nash J: **Equilibrium points in n-person games.** *Proc Natl Acad Sci* 1950, **36**:48-49.
  7. Costa-Gomes M, Crawford V, Broseta B: **Cognition behavior in normal-form games: an experimental study.** *Econometrica* 2001, **69**:1193-1235.
  8. Camerer CF (Ed): *Behavioral game theory: experiments on strategic interaction*. Princeton University Press; 2003.
  9. McKelvey R, Palfrey T: **Quantal response equilibria for normal form games.** *Games Econ Behav* 1995, **10**:6-38.
  10. Haile PA, Hortaçsu A, Kosenok G: **On the empirical content of quantal response equilibrium.** *Am Econ Rev* 2008, **98**:180-200.
  11. Webb R: *Dynamic Constraints on the Distribution of Stochastic Choice: Drift Diffusion Implies Random Utility*. Working Paper. 2013;. Available at SSRN 2226018.
  12. Dickhaut J, Rustichini A, Smith V: **A neuroeconomic theory of the decision process.** *Proc Natl Acad Sci* 2009, **106**:22145-22150.
  13. Reutskaja E, Nagel R, Camerer CF, Rangel A: **Search dynamics in consumer choice under time pressure: an eye-tracking study.** *Am Econ Rev* 2011, **101**:900-926.



14. Krajbich I, Armel C, Rangel A: **Visual fixations and the computation and comparison of value in simple choice.** *Nat Neurosci* 2010, **13**:1292-1298.
15. Polonio L, Di Guida S, Coricelli G: *Strategic sophistication and attention in games: an eye-tracking study.* ECARES Working Papers; 2014.
- The authors use eye-tracking to classify the participants according to their visual pattern of information acquisition in four classes of two players normal form games, in which reaching the equilibrium play necessitates to incorporate different levels of information about the other's payoffs. They show individually heterogeneous-but stable-patterns of visual information acquisition based on subjective levels of strategic sophistication and social preferences.
16. Nagel R: **Unraveling in guessing games: an experimental study.** *Am Econ Rev* 1995, **85**:1313-1326.
17. Stahl DO, Wilson PW: **On players' models of other players: theory and experimental evidence.** *Games Econ Behav* 1995, **10**:218-254.
18. Camerer CF, Ho TH, Chong JK: **A cognitive hierarchy model of games.** *Quart J Econ* 2004, **119**:861-898.
19. Ho TH, Camerer CF, Weigelt K: **Iterated dominance and iterated best response in experimental p-beauty contests.** *Am Econ Rev* 1998:947-969.
20. Carpenter J, Graham M, Wolf J: **Cognitive ability and strategic sophistication.** *Games Econ Behav* 2013, **80**:115-130.
21. Camerer CF, Lovaglio D: **Overconfidence and excess entry: an experimental approach.** *Am Econ Rev* 1999, **89**:306-318.
22. Coricelli G, Nagel R: **Neural correlates of depth of strategic reasoning in medial prefrontal cortex.** *Proc Natl Acad Sci* 2009, **106**:9163-9168.
23. Fletcher PC, Happé F, Frith U, Backer SC, Dolan RJ: **Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension.** *Cognition* 1995, **57**:109-128.
24. Gallagher HL, Happé F, Brunswick N, Fletcher PC, Frith U, Frith CD: **Reading the mind in cartoons and stories: an fMRI study of theory of mind in verbal and nonverbal tasks.** *Neuropsychologia* 2000, **38**:11-21.
25. McCabe K, Houser D, Ryan L, Smith V, Trouard T: **A functional imaging study of cooperation in two-person reciprocal exchange.** *Proc Natl Acad Sci* 2001, **98**:11832-11835.
26. Bird CM, Castelli F, Malik O, Firth U, Husain M: **The impact of extensive medial frontal lobe damage on Theory of Mind and cognition.** *Brain* 2004, **127**:914-928.
27. Amodio DM, Frith CD: **Meeting of minds: the medial frontal cortex and social cognition.** *Nat Rev Neurosci* 2006, **7**:268-277.
28. Bhatt MA, Lohrenz T, Camerer CF, Montague PR: **Neural signatures of strategic types in a two-person bargaining game.** *Proc Natl Acad Sci* 2010, **107**:19720-19725.
29. Koechlin E, Summerfield C: **An information theoretical approach to prefrontal executive function.** *Trends Cogn Sci* 2007, **11**:229-235.
30. Dixon ML, Christoff K: **The lateral prefrontal cortex and complex value-based learning and decision making.** *Neurosci Biobehav Rev* 2014, **45**:9-18.
31. Shenhav A, Botvinick MM, Cohen JD: **The expected value of control — an integrative theory of anterior cingulate cortex function.** *Neuron* 2013, **79**:217-240.
32. Yoshida W, Dolan RJ, Friston KJ: **Game theory of mind.** *PLoS Comput Biol* 2008, **4**:e1000254.
- The authors propose that the expected value of a given action in a dynamic coordination game depends directly on subjects' belief over the other's sophistication level. The Bayesian learning model first estimates at each trial the level of strategic thinking of the opponent and then adjusts the subject's level of strategic thinking from which an optimal action is selected.
33. Devaine M, Hollard G, Daunizeau J: **The social Bayesian brain: does mentalizing make a difference when we learn?** *PLoS Comput Biol* 2014, **10**:e1003992.
34. Yoshida W, Seymour B, Friston KJ, Dolan RJ: **Neural mechanisms of belief inference during cooperative games.** *J Neurosci* 2010, **30**:10744-10751.
35. Hampton AN, Bossaerts P, O'Doherty JP: **Neural correlates of mentalizing-related computations during strategic interactions in humans.** *Proc Natl Acad Sci* 2008, **105**:6741-6746.
- Using a repeated zero-sum inspection game they show that their influence model fits significantly better than the average choice data of their subjects compared to a simple fictitious play and a reinforcement learning model. The fictitious play first infers from the frequencies of the opponent's past choices the probability of choosing one action or another, and then decides so as to maximize the action's consequent expected reward.
36. Nicolle A, Klein-Flügge MC, Hunt LT, Vlaev I, Dolan RJ, Behrens TEJ: **An agent independent axis for executed and modeled choice in medial prefrontal cortex.** *Neuron* 2012, **75**:1114-1121.
37. Suzuki S, Harasawa N, Ueno K, Gardner JL, Ichinohe N, Masahiko H, Cheng K, Nakahara H: **Learning to simulate others' decisions.** *Neuron* 2012, **74**:1125-1137.
38. Seid-Fatemi A, Tobler PN: **Efficient learning mechanisms hold in the social domain and are implemented in the medial prefrontal cortex.** *Social Cogn Affective Neurosci* 2014, **10**:1093.
39. Zhu L, Mathewson KE, Hsu M: **Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning.** *Proc Natl Acad Sci* 2012, **109**:1419-1424.
40. Seo H, Cai X, Donahue CH, Lee D: **Neural correlates of strategic reasoning during competitive games.** *Science* 2014, **346**:340-343.
41. Nagel R, Brovelli A, Heinemann F, Coricelli G: *Neural correlates of risk and strategic uncertainty.* (forthcoming) 2014.
- The findings of this study suggest a cognitive-hierarchy theory of brain and behavior, according to which different levels of strategic thinking contribute to the resolution of the uncertainty underlying social interactions.
42. Mohr PN, Biele G, Heekeren HR: **Neural processing of risk.** *J Neurosci* 2010, **30**:6613-6619.
43. Aumann R, Brandenburger A: **Epistemic conditions for Nash equilibrium.** *Econometrica* 1995, **63**:1161-1180.
44. Carlsson H, Van Damme E: **Global games and equilibrium selection.** *Econometrica* 1993, **61**:989-1018.
45. Bernheim BD: **Rationalizable strategic behavior.** *Econometrica* 1984, **52**:1007-1028.
46. Pearce D: **Rationalizable strategic behavior and the problem of perfection.** *Econometrica* 1984, **52**:1029-1050.
47. Crawford VP, Costa-Gomes MA, Iriberry N: **Structural models of nonequilibrium strategic thinking, theory, evidence, and applications.** *J Econ Literature* 2013, **51**:5-62.
- This article provides a clear and detailed theoretical and conceptual review of the main Non-Equilibrium models in Game Theory along with recent evidence in favor of level-k models.
48. Bosch-Domènech A, Montalvo JG, Nagel R, Satorra A: **One, Two, (Three), Infinity ...: Newspaper and Lab Beauty-Contest Experiments.** *Am Econ Rev* 2002, **92**:1687-1701.
49. Sutton RS, Barto AG: *Reinforcement learning: an introduction.* Cambridge, MA: MIT Press; 1998, .
50. Fudenberg D, Levine DK: *The theory of learning in games.* Cambridge, MA: MIT Press; 1998, .

## - Appendix II -

---

### SUPPLEMENTARY INFORMATION 1

---

Griessinger Thibaud, Khamassi Mehdi\*, Coricelli Giorgio\*. The interplay of learning sophistication and strategic asymmetry. About to be submitted.

#### I - Supplementary Figures

#### II - Simulation Analysis

- 1- Computational models
- 2- Simulation procedure

#### III - Computational Analysis

- 1- Model Space Extension
- 2- Optimization Process
- 3- Subject-level Model comparison

#### IV - Additional Cognitive Tasks

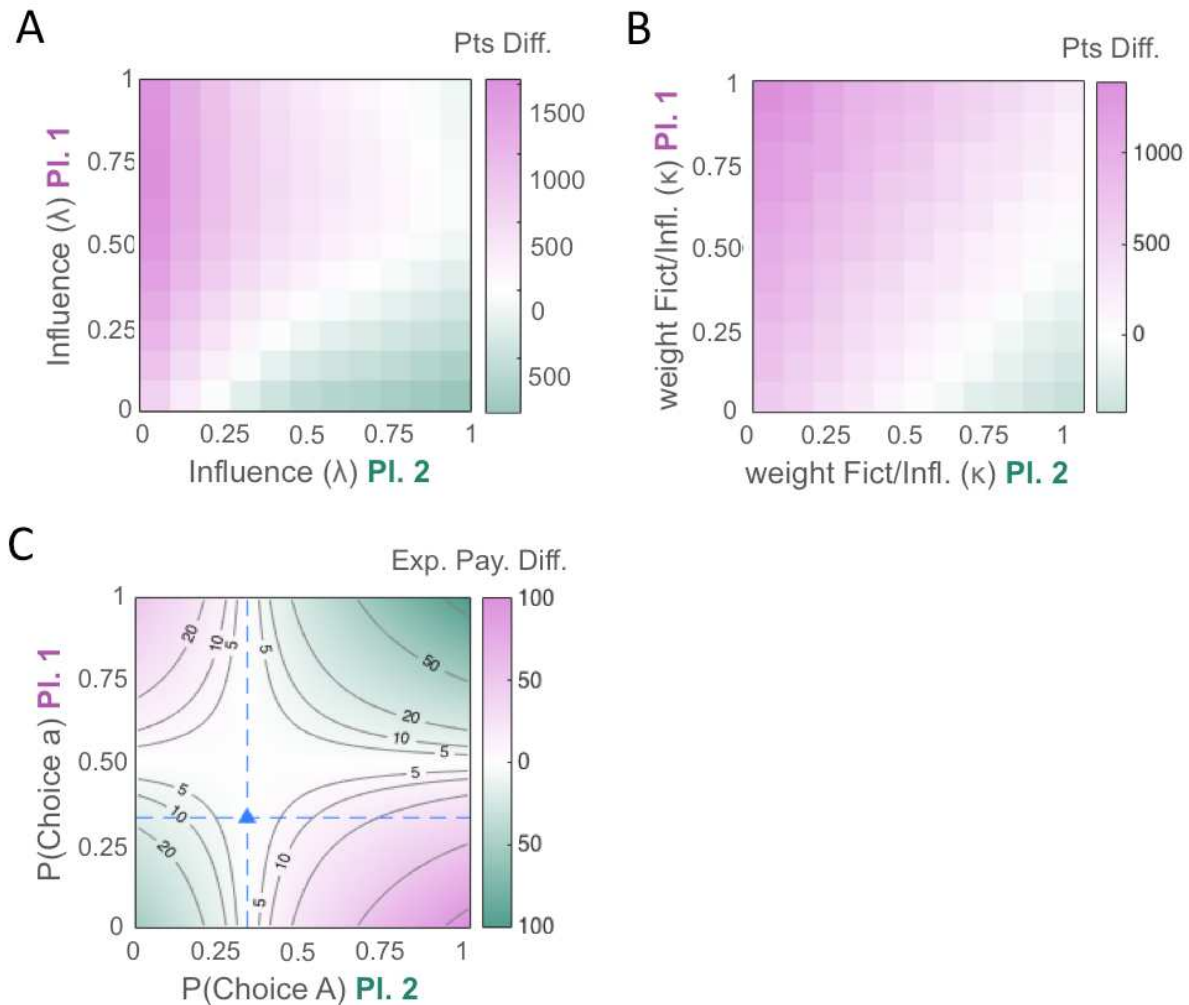
#### V - One-shot games experiment

- 1- Methods
- 2- Results

#### VI - References

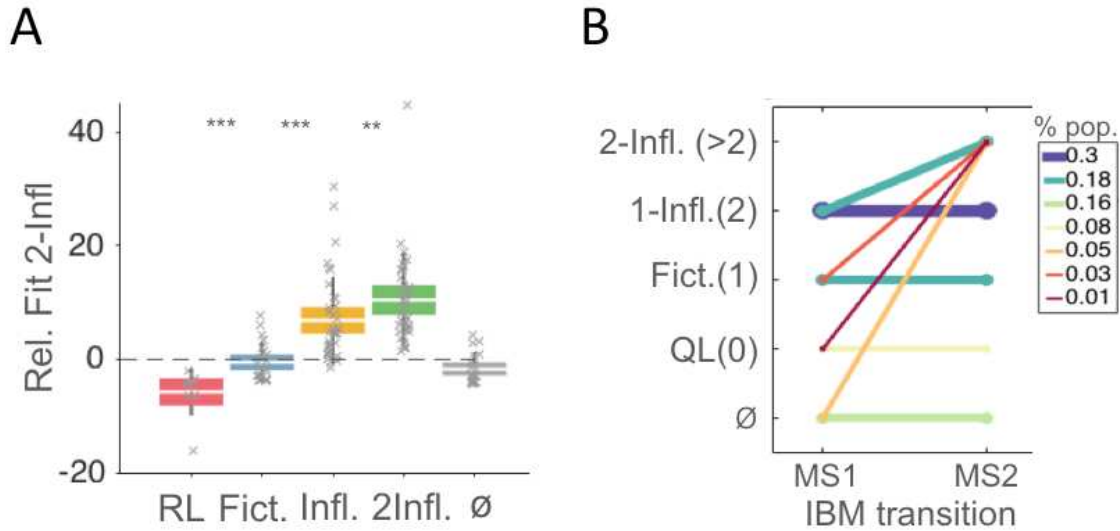
#### **I - Supplementary Figures**





**Figure S1. Strategic asymmetry of the game replicated in simulation with different measures of individual Strategic Learning.**

(A) Agents were modelled by the Influence model varying in the values of their parameter  $\lambda$  from 0 (low SL level - fictitious) to 1 (high SL level - full influence), as well as their  $\eta$  ([0:1]). The heatmap were obtained by averaging across the whole range of  $\eta$  values, for each  $\lambda$  combination. (B) Agents were then modelled by a mixture model (Hybrid Influence model) including an arbitration parameter ( $\kappa$ ) controlling for the relative weight given to the 1st and 2nd order update of the opponent's action probability (i.e. low  $\kappa$  = Fictitious vs. high  $\kappa$  = Influence) in the final action value computation. Heatmaps represent the difference between the 2 players in their total payoffs obtained for every combination of their respective  $\kappa$  parameter values, averaged across all instances of  $\eta_1$ ,  $\eta_2$ , and  $\lambda$  ([0:1]). (For both A and B analyses, each game play was averaged across 100 simulations - see text for details). (C) Difference in theoretical expected payoff between the 2 players as a function of their respective probability of choice. If theoretically the expected utility computation assumes independence between trials, this plot suggests that, in order to reduce their disadvantage in this game, players 2 will have to deviate from the focal point (i.e. remind what is the focal point) to converge towards the probability of choice prescribed by the Mixed Strategy Nash Equilibrium (in blue).



**Figure S2. Strategic learning heterogeneity refined by our additional computational analysis. Introducing an extension of the Influence (i.e., the 2-Inf model) allows our model space to cover higher SL levels.**

(A) Population (extended) gradient of strategic learning sophistication. Subjects now best fitted by the 2-Inf. model have a higher relative fit compared to the 3 other SL groups (IBM), thus capturing more of the population continuum of SL level. The subjects individually best fitted by the 2-Inf present also higher  $\omega$  (2nd order influence parameter) ( $U(62) = 212$ ,  $Z = 5.0885$ ,  $p = 3.6087e-7$ ) and slightly higher 1st order  $\lambda$  parameter ( $U(62) = 860$ ,  $Z = 1.9163$ ,  $p = 0.0553$ ) than the group best fitted by the Influence. (B) 37% of the subjects initially best fitted by the Influence are now best fitted by the 2-Inf. Model. Adding the 2-Inf. in the model comparison improved the overall fit by capturing higher SL level, initially constrained to the SL level of the Influence (not shown). MS1/MS2: Model Spaces 1 and 2. IBM: Individual Best Model.

<b>A</b>					
% action(a/A)	Estimate	SE	t	p	
<b>- Player 1</b>					
P1_SL_sub	0.0005	0.0015	0.3339	0.7395	
P1_SL_opp	<b>0.0034</b>	<b>0.0008</b>	<b>4.2911</b>	<b>6.58e-05</b>	
P1_SL_s ~ P1_SL_o	3.65e-05	0.0001	0.2456	0.8068	
$R^2 = 0.2481 ; F(60)=6.60, p = 6.01e-4$				$BIC = -165.54$	
<b>- Player 2</b>					
P2_SL_sub	<b>-0.0125</b>	<b>0.0010</b>	<b>-6.0833</b>	<b>8.91e-08</b>	
P2_SL_opp	0.0024	0.0002	1.2356	0.2214	
P2_SL_s ~ P2_SL_o	-5.31e-05	0.0002	-0.2826	0.7784	
$R^2 = 0.4323 ; F(60)=15.23, p = 1.76e-10$				$BIC = -135.55$	

<b>B</b>					
Performance (Pts)	Estimate	SE	t	p	
<b>- Player 1</b>					
P1_SL_sub	7.8518	14.486	0.5423	0.5898	
P1_SL_opp	<b>-38.9540</b>	<b>7.3843</b>	<b>-5.2753</b>	<b>1.91e-06</b>	
P1_SL_s ~ P1_SL_o	0.3818	1.3874	0.2752	0.7841	
$R^2 = 0.34669 ; F(60)=10.613, p = 1.08e-5$				$BIC = 1004.6$	
<b>- Player 2</b>					
P2_SL_sub	9.2498	5.5253	-6.0833	0.1406	
P2_SL_opp	3.4612	10.8390	0.3193	0.7506	
P2_SL_s ~ P2_SL_o	-0.5596	1.0381	-0.5390	0.5918	
$R^2 = 0.4323 ; F(60)=0.7778, p = 0.5109$				$BIC = 967.48$	
<b>Rel. Perf. (Diff Pts)</b>					
<b>- Player 2</b>					
P2_SL_sub	<b>47.2014</b>	<b>12.0350</b>	<b>3.9222</b>	<b>2.28e-04</b>	
P2_SL_opp	-4.3906	23.609	-0.1850	0.8531	
P2_SL_s ~ P2_SL_o	-0.9414	2.2614	0.4163	0.6786	
$R^2 = 0.2210 ; F(60)=5.6744, p = 0.0017$				$BIC = 1067.1$	

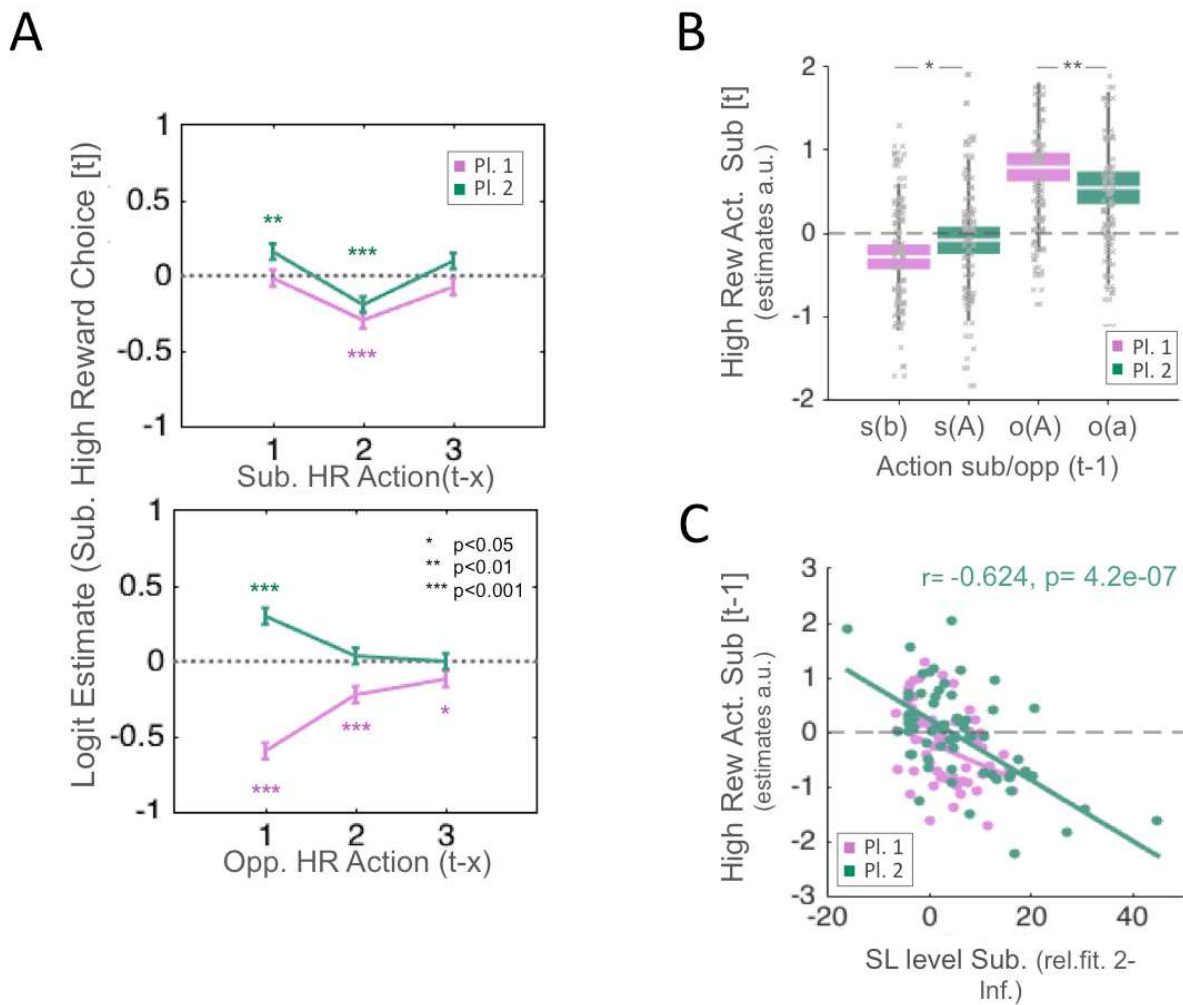
  

<b>C</b>					
Accuracy (%Rew Act. (b/A) )	Estimate	SE	t	p	
<b>- Player 1</b>					
P1_SL_sub	0.0018	0.0024	0.7583	0.4512	
P1_SL_opp	<b>-0.0073</b>	<b>0.0012</b>	<b>-5.8626</b>	<b>2.08e-07</b>	
P1_SL_s ~ P1_SL_o	8.18e-5	2.35e-4	0.3473	0.7295	
$R^2 = 0.4007 ; F(60)=13.374, p = 8.6448e-7$				$BIC = -106.6$	
<b>- Player 2</b>					
P2_SL_sub	<b>0.0049</b>	0.0012	<b>3.8936</b>	<b>2.51e-04</b>	
P2_SL_opp	9.99e-4	0.0024	0.4026	0.6886	
P2_SL_s ~ P1_SL_o	-1.05e-4	2.38e-4	-0.4412	0.6606	
$R^2 = 0.20612 ; F(60)=5.1926, p = 0.0029$				$BIC = -105.5$	

**Figure S3. GLM results of the distinct influence of the subject SL level and the one of the opponent on her own choice behavior.**

(A) Players 2's deviation from MSNE was influenced only by their own SL level. Conversely Players 1's choice probability was influenced solely by the SL level of their opponent SL not their own. (B) Players 2 had to engage in

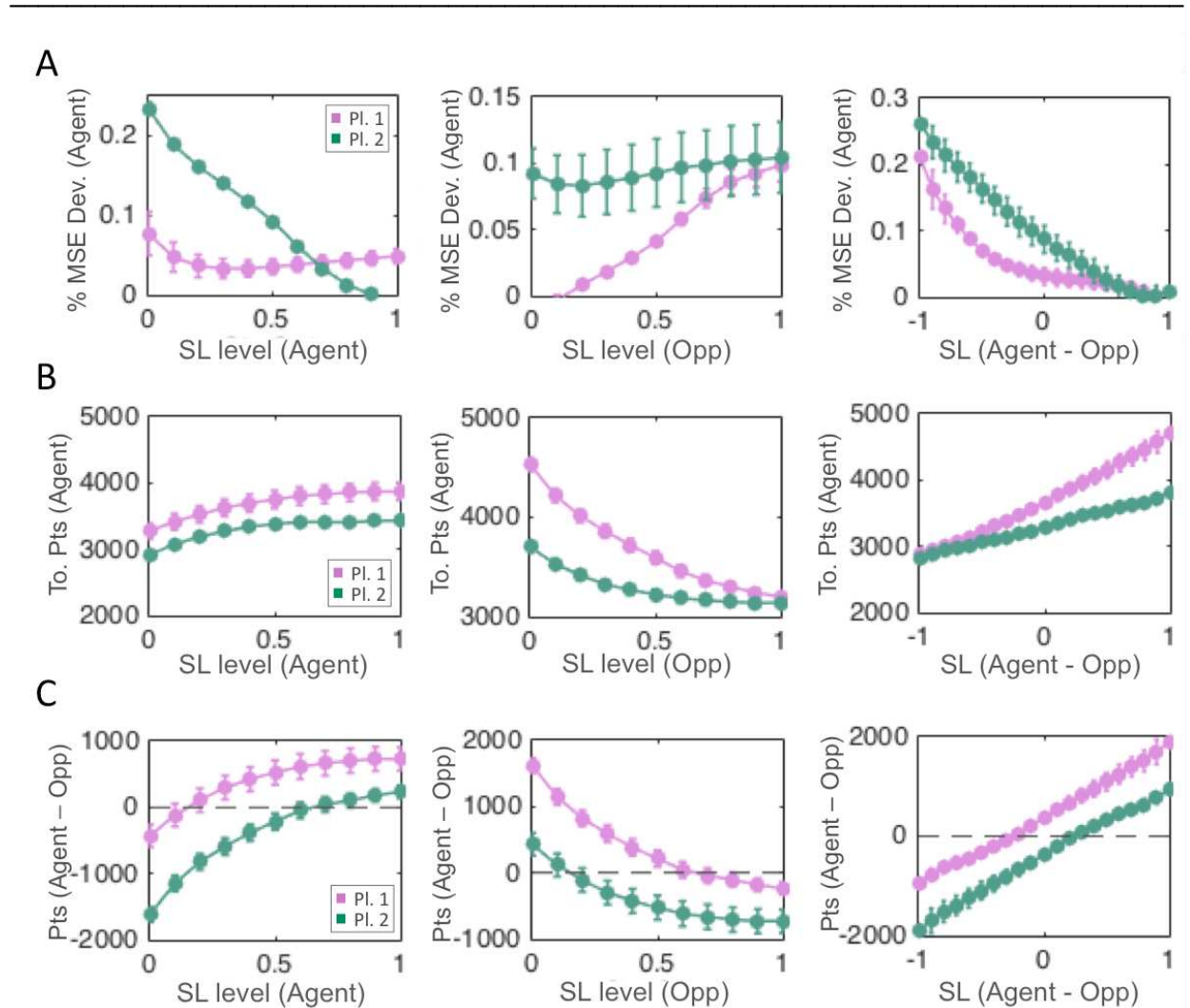
higher level of strategic learning in order to reduce their disadvantage, relative performance, but this is not sufficient to increase their absolute performance, while Players 1's higher performance is solely affected (reduced) by their opponent's behavior (SL level), suggesting a hierarchical leader-follower dynamics. (C) This dynamics is confirmed by the similar asymmetry between the 2 roles observed in choice accuracy (percentage of trials where the selection of the subject's action linked to the highest payoff is actually rewarded).



**Figure S4. Logistic regression capturing the average effect of the actions observed (own and opponent) in the past 3 trials on the actual high reward action choice of the subjects, for each role separately.**

(A) Graphic representation of the results of the logistic regression ran on the 2 populations of subjects endorsing each role. Players 2 tend to perseverate in choosing the action leading to the highest reward (Sub. HR Action(t-1)), but not Players 1. Player 1 take into account the opponent's past choices up to 2 trials back, Players 2 only the previous choice. These results suggest that Players 2 are more focused on the choice of the high reward action and

take less into consideration the opponent's history of play. (B) We then compared the weight (estimate value) of the previous high reward choice of the subject or the opponent on the current choice, by running the logistic regression on each subject independently. Players 2 alternate less than player 1 their high reward choice from one trial to the next. Moreover, this choice is also less affected by the opponent's last choice. (C) Correlation analysis of the individual estimates of the past high reward choice of the subjects endorsing each role suggests that for Players 2 the higher their SL level, the more they tend to alternate the choice of the high reward action (A). (similar results were obtained by using the relative fit of the Influence compared to fictitious, or the Influence best fitting parameter value  $\lambda$ )



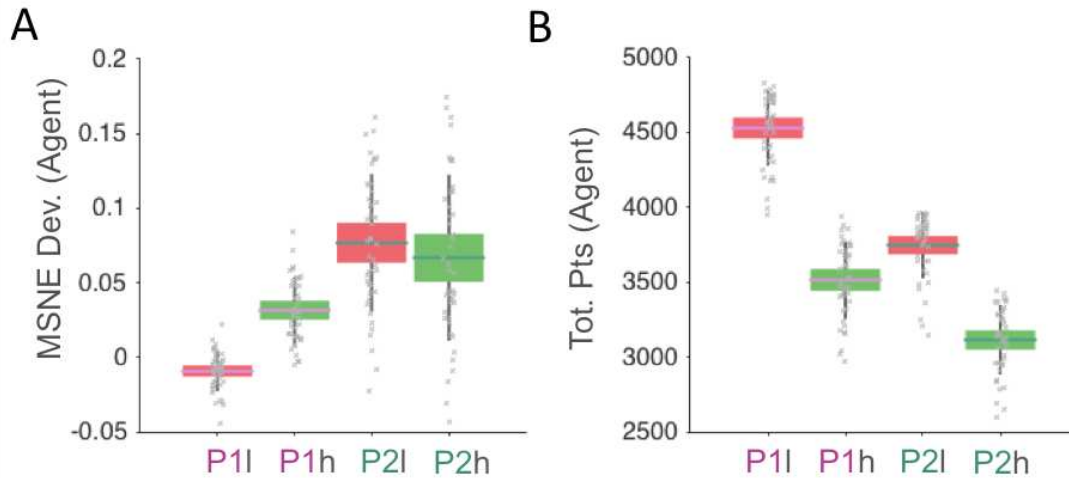
**Figure S5. Simulation results of Influence agents with different values of  $\lambda$  playing the repeated game in each role against each other.**

The SL level was modeled by the  $\lambda$  value of the Influence model (varying from 0 to 1, with fixed  $\eta$  to the average values of our empirical distribution [0:0.4]). The same effect were observed in the 9 plots when taking as SL measure

in our simulation the arbitration parameter ( $\kappa$ ) of the mixture model (Hybrid Influence model) instead. Simulation plots represents for each role the effect of the SL level, own, opponent's or difference between the 2, on deviation of action probability distribution from MSNE (A), total points accumulated throughout the game interaction (100 repeated choices) (B), and the difference in total points between the agent and its opponent (C).

---

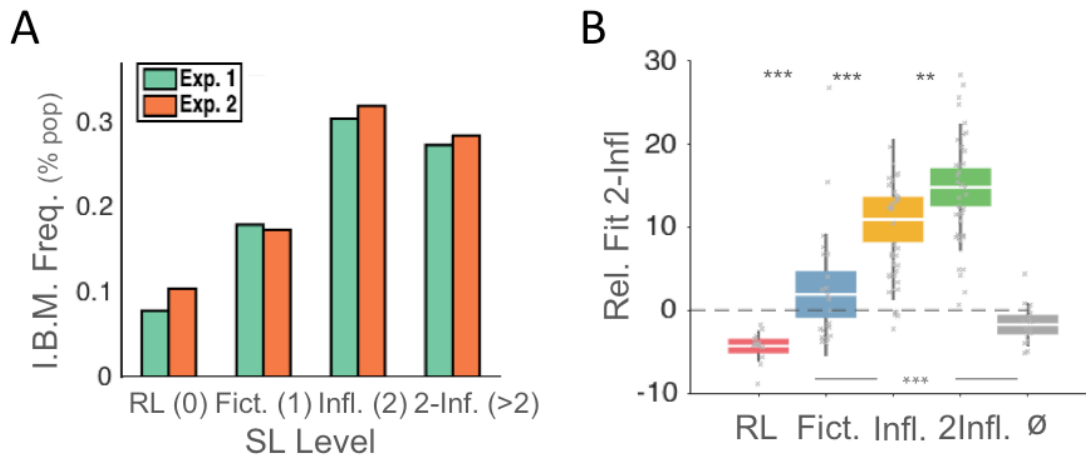
---



**Figure S6. Results expected in Exp. 2, as obtained with model simulation.**

Simulation was performed by making Agents vary in their SL level, and play the repeated game in each role against the 2 fixed algorithms. The SL level was modeled here as the  $\lambda$  value of the Influence model (varying from 0 to 1, with fixed  $\eta$  to the average values of our empirical distribution [0:0.4]). The same patterns of results were observed when taking as SL measure in our simulation the arbitration parameter ( $\kappa$ ) of the mixture model (Hybrid Influence model) instead.

---



**Figure S7. SL level model-based analysis of Exp. 2 replicates the results of Exp. 1.**

(A) Histogram of the frequency of fitted individual best models (I.B.M. - BMS: AIC+LRT) in the 2 populations (with  $N(\text{exp.1})=64$  and  $N(\text{exp.2})=72$ ) (B) Population gradient of strategic learning sophistication in Exp.2. As in Exp.1, subjects best fitted by higher SL models have on average a higher relative fit ( $\text{fit}(2\text{-Infl.}) - \text{fit}(\text{RL})$ ) than the groups of subjects best fitted by models of lower SL level.



<b>A</b>					
% action(a/A)	Estimate	SE	t	p	
<b>- Player 1</b>					
P1_SL_sub	<b>0.0054</b>	<b>0.0020</b>	<b>2.8838</b>	<b>0.0052</b>	
P1_SL_opp	<b>0.0743</b>	<b>0.0277</b>	<b>2.6842</b>	<b>0.0091</b>	
P1_SL_s ~ P1_SL_o	-0.0012	0.0025	-0.4558	0.6500	
$R^2 = 0.2731 ; F(68)=8.52, p = 7.04e-5$				$BIC = -117.57$	
<b>- Player 2</b>					
P2_SL_sub	<b>-0.0050</b>	<b>0.0011</b>	<b>-4.6258</b>	<b>1.73e-05</b>	
P2_SL_opp	-0.0010	0.0232	-0.0431	0.9657	
P2_SL_s ~ P2_SL_o	-0.0023	0.0017	-1.3999	0.18473	
$R^2 = 0.4342 ; F(68)=17.40, p = 1.74e-8$				$BIC = -152.74$	

<b>B</b>					
Performance (Pts)	Estimate	SE	t	p	
<b>- Player 1</b>					
P1_SL_sub	<b>24.241</b>	<b>8.1641</b>	<b>2.9692</b>	<b>0.0041</b>	
P1_SL_opp	<b>-729.82</b>	<b>120.79</b>	<b>-6.0421</b>	<b>7.18e-8</b>	
P1_SL_s ~ P1_SL_o	-1.6412	11.088	-0.1480	0.8828	
$R^2 = 0.4906 ; F(68)=21.827, p = 5.20e-10$				$BIC = 1089.2$	
<b>- Player 2</b>					
P2_SL_sub	<b>13.042</b>	<b>5.9539</b>	<b>2.1906</b>	<b>0.0319</b>	
P2_SL_opp	-200.34	127.82	-1.5674	0.1217	
P2_SL_s ~ P2_SL_o	0.5017	9.3736	0.0535	0.9575	
$R^2 = 0.1667 ; F(68)=4.5342, p = 0.0059$				$BIC = 1087.9$	
<b>Rel. Perf. (Diff Pts)</b>					
<b>- Player 2</b>					
P2_SL_sub	<b>47.620</b>	<b>12.308</b>	<b>3.4628</b>	<b>9.30e-04</b>	
P2_SL_opp	-407.77	264.22	-1.5433	0.1274	
P2_SL_s ~ P2_SL_o	8.7787	19.377	0.4536	0.6516	
$R^2 = 0.2957 ; F(68)=9.5155, p = 2.49e-5$				$BIC = 1192.1$	

<b>C</b>					
Accuracy (%Rew Act. (b/A) )	Estimate	SE	t	p	
<b>- Player 1</b>					
P1_SL_sub	<b>0.0055</b>	<b>0.0014</b>	<b>3.8475</b>	<b>2.66e-04</b>	
P1_SL_opp	<b>-0.1217</b>	<b>0.0214</b>	<b>-5.6984</b>	<b>2.83e-07</b>	
P1_SL_s ~ P1_SL_o	-0.0022	0.0020	-1.1001	0.2751	
$R^2 = 0.5127 ; F(68)=23.845, p = 1.17e-10$				$BIC = -154.9$	
<b>- Player 2</b>					
P2_SL_sub	<b>0.0052</b>	0.0011	<b>4.9795</b>	<b>4.62e-06</b>	
P2_SL_opp	-0.0483	0.0227	-2.1306	0.0367	
P2_SL_s ~ P1_SL_o	-0.0016	0.0016	-0.9909	0.3252	
$R^2 = 0.4245 ; F(68)=16.718, p = 3.07e-8$				$BIC = -155.8$	

**Figure S8. Exp 2 - GLM results of the distinct influence of the subject SL and the level of the computerized opponent (low, high) on her behavior.** (A,B) As in Exp. 1, Players 2's deviation from MSNE and relative performance were influenced only by their own SL level, while Players 1's choice and performance (absolute and relative, not shown) are impacted by the SL level of their opponent. Two additional effects are observed in this



experiment: Players 2's absolute performance is also, albeit marginally, directly impacted by the SL level of the opponent encountered; and in Players 1, their own SL level seems to condition their action probability and, slightly, their performance (absolute and relative : Est: 32.45 (15.8),  $t(68)=2.0467$ ,  $p= 0.0445$ ). (C) In this more controlled experiment, Players 1's accuracy was not only impacted by the level of the opponent but also independently by their own SL level. Players 2's accuracy was still modulated only by their own SL level.

---

## II - Simulation Analysis

### 1- Computational models:

The 3 following models (learning rules) approximating 3 different levels of strategic sophistication (none, low, high), and forming the initial model space (MS1) used in our computational analysis, were used to simulate the playing behavior of agents interacting with each other:

a- Learning rules :

(in the following 50 points = 1 point unit (a.u) to simplify the equations, see Hampton et al, 2008)

- Reinforcement Learning (SL level 0)

Reinforcement learning is modelled as a Q-learning algorithm with a single state.

At trial  $t$  the chosen option value  $Q_c(t)$  (the option being either action  $a$  or  $b$  for Player 1 and  $A$  or  $B$  for Player 2 - see payoff matrix on Fig.1.A main text) is updated with the following learning rule:

$$Q_c(t) = Q_c(t-1) + (\alpha_1 \times \delta_c(t)) \quad (1.1)$$

where  $\alpha_1$  is the learning rate for the chosen option.  $\delta_c$  is a (reward) prediction error term calculated from  $R_c$  the reward (points) received as the outcome of the chosen action:

$$\delta_c(t) = R_c(t) - Q_c(t) \quad (1.2)$$

- Fictitious Model (SL level 1) (Hampton et al, 2008 [1])

The agent infers the probability that the opponent will choose one action or another, and then decides so as to maximize the action's consequent expected reward (best response). Reminder: Player 1 can choose between action  $a$  and action  $b$ ; Player 2 can choose between action  $A$  and action  $B$ ; the Pay-off matrix showing the outcome for each combination of these two choices is shown in Figure 1A of the main article. From the point of view of a given agent, the opponent's probability  $P_{A/a}^*$  of choosing an action ( $a/A$ ) (with  $P_{B/b}^* = 1 - P_{A/a}^*$ ) is dynamically inferred by tracking the history of the actions that the opponent makes ( $P_{A/a}^*$  was initiated to 0.5 at  $t=0$ ):

Example from Player 1's perspective, the probability that Player 2 selects the action  $A$

$$P_A^*(t) = P_A^*(t-1) + (\eta \times \delta_A(t)) \quad (1.3)$$

where  $\eta$  is the learning rate for the chosen option.  $\delta_A$  is a (action) prediction error term calculated from C the choice observed at trial t (1 if opponent chose action A, 0 otherwise):

$$\delta_A(t) = C(t) - P_A^*(t-1) \quad (1.4)$$

The payoff matrix (in point units) is thus used to convert the probability  $P_{A/a}^*$  of the opponent choosing action (A/a) into the expected value (Q) of each action. For player 1:

$$Q_a(t) = 1 - P_A^*(t) \quad Q_b(t) = 2 \times P_A^*(t) \quad (1.5)$$

And for player 2 :

$$Q_A(t) = 2 \times P_a^*(t) \quad Q_B(t) = 1 - P_a^*(t) \quad (1.6)$$

#### - 1-Influence (SL level 2) (Hampton et al, 2008 [1])

The agent considers that the opponent is playing according to a fictitious play strategy, and thus infers the probability the opponent computes over its own actions ( $P_{a/A}^{**}$ ) and how this influences her action probability ( $P_{A/a}^*$ ). As in fictitious play, the agent then uses this probability to decide so as to maximize its action's consequent expected reward. The opponent's probability  $P_{A/a}^*$  of choosing an action is dynamically inferred by the agent by tracking the last action the opponent makes, and its own past choice. The learning rate of the opponent's fictitious play thus becomes the influence parameter  $\lambda$ , which captures the weight of the agent's actions on her choice behavior (action probability), while the agent's own learning rate is embedded in the parameter  $\eta$ .

For Player 1 the probability of the opponent playing action A at trial t ( $P_A^*(t)$ ) is thus obtained through :

$$P_A^*(t) = P_A^*(t-1) + (\eta \times \delta_A(t)) + (3 \times \lambda \times (P_A^*(t-1) \times (1 - P_A^*(t-1)) \times (C_a - P_a^{**}(t)))) \quad (1.7)$$

with

$$P_a^{**}(t) = 1/3 - (1/3 \times \beta \times (\log((1 - P_A^*(t-1))/P_A^*(t-1)))) \quad \text{and} \quad \delta_A(t) = C_A - P_A^*(t-1) \quad (1.8)$$

And for Player 2:

$$P_a^*(t) = P_a^*(t-1) + (\eta \times \delta_a(t)) - (3 \times \lambda \times (P_a^*(t-1) \times (1 - P_a^*(t-1)) \times (C_a - P_A^{**}(t)))) \quad (1.9)$$

with

$$P_A^{**}(t) = 1/3 + (1/3 \times \beta \times (\log((1 - P_a^*(t-1))/P_a^*(t-1)))) \quad \text{and} \quad \delta_a(t) = C_a - P_a^*(t-1) \quad (1.10)$$

The computed opponent's choice probability  $P_{A/a}^*(t)$  is then used to estimate at each trial t the expected value of each action of the agent :

$$\text{Player 1 : } Q_a(t) = 1 - P_A^*(t) \quad Q_b(t) = 2 \times P_A^*(t) \quad (1.11)$$

$$\text{Player 2 : } Q_A(t) = 2 \times P_a^*(t) \quad Q_B(t) = 1 - P_a^*(t) \quad (1.12)$$

b- Action selection :

Best response is made noisy (probabilistic) through a logistic (softmax) function. For instance for player 1:

$$P_a(t) = 1/(1 + \exp(\beta \times (Q_b(t-1) - Q_a(t-1)))) \quad (1.13)$$

c- Hybrid extension:

We added an additional model to this initial model space to directly control for the between-subject balance between fictitious play and Influence. This hybrid model is meant to dissociate the weight of the implementation of the own past play on the computation of the opponent's action probability in the influence ( $\lambda$  parameter), and the relative weight of the first and second order beliefs (fictitious, vs. influence). This model is very close to the Influence agent varying in Influence parameter but insures an additional control for our simulation analysis.

The "hybrid Influence" model has 2 learning rules, 2 modules operating in parallel (here illustrated for Player 1):

A fictitious module (F) updating the estimated probability that the opponent chooses A:

$$\delta_{A(F)}(t) = C(t) - P_{A(F)}^*(t-1) \quad (1.14)$$

$$P_{A(F)}^*(t) = P_A^*(t-1) + (\eta_1 \times \delta_{A(F)}(t)) \quad (1.15)$$

And a Influence module (I) updating the same probability at a higher order:

$$\delta_{A(I)}(t) = C(t) - P_{A(I)}^*(t-1) \quad (1.16)$$

$$P_{A(I)}^*(t) = P_{A(I)}^*(t-1) + (\eta_2 \times \delta_{A(I)}(t)) + (3 \times \lambda \times (P_{A(I)}^*(t-1) \times (1 - P_{A(I)}^*(t-1)) \times (C_a - P_{a(I)}^{**}(t)))) \quad (1.17)$$

with

$$P_{a(I)}^{**}(t) = 1/3 - (1/3 \times \beta \times (\log((1 - P_{A(I)}^*(t-1))/P_{A(I)}^*(t-1)))) \quad (1.18)$$

The 2 Q-values are then computed based on a weighted mixture of the 2 probabilities computed by the 2 modules.

This weight thus corresponds to an arbitration, a fifth, parameter ( $\kappa$ ):

$$Q_a(t) = ((1 - \kappa) \times (1 - P_{A(F)}^*(t)) + (\kappa \times (1 - P_{A(I)}^*(t))) \quad (1.19)$$

and

$$Q_a(t) = ((1 - \kappa) \times (2 \times P_{A(F)}^*(t)) + (\kappa \times (2 \times P_{A(I)}^*(t))) \quad (1.20)$$

Conceptually,  $\kappa$  represents the propensity to engage in second order belief (Influence) compared to first order (fictitious), while  $\lambda$  represents more the weight of the Influence in this second order belief (how much the opponent takes into account the subject's past choices in her own choice behavior).

## 2- Simulation procedure :

Each simulation followed the same procedure. We made models play against each others, one endorsing each role, for 100 simulated repetitions of the stage game, thus representing one simulated game interaction. We defined range values for each of the free parameters of each of the 2 models playing against each other (typically for parameters ranging from 0 to 1, 11 uniformly distributed values), and simulated all parameter combinations of play. Each simulated game interaction with a specific combination of free parameters was itself simulated 100 times to reduce potential noise in our data.

We then computed for each game simulation, with each combination of parameters and each combination of models playing against each other, the key behavioral measures such as the total payoffs accumulated by each player throughout the interaction, or action distribution. Finally, the behavior measures of all repetitions of each interaction simulation across all combinations of parameter values of no interest (not representing the SL level per se) were

averaged. The results were thus plotted as heatmaps representing how on average each behavioral measure varies with the value of the parameters of interest (**Fig 1.B, Fig S1.A,B**).

### III - Exp. 1: Computational Analysis

#### 1- Model Space Extension (MS2):

We extended the initial model space MS1 with additional models used in the strategic learning literature of repeated games:

- Reinforcement Learning variations and extensions :

Generalized RL (Khamassi et al, 2015 [2]):

$$Q_c(t) = Q_c(t-1) + (\alpha_1 \times \delta_c(t)) \text{ with } \delta_c(t) = R_c(t) - Q_c(t) \quad (1.21)$$

And the counterfactual update:

$$Q_u(t) = Q_u(t-1) + ((1 - \kappa) \times (Q_u(0) - Q_u(t))) \quad (1.22)$$

where  $\kappa$  is the forgetting rate [0:1] and  $Q_u(0)$  is the initial Q-value of the unchosen action, which corresponds to the expected value of this action computed from the payoff matrix using  $P_{Ala}^* = 0.5$ .

CounterFactual RL (Palminteri et al, 2015 [3]):

$$Q_c(t) = Q_c(t-1) + (\alpha_1 \times \delta_c(t)) \text{ with } \delta_c(t) = R_c(t) - Q_c(t) \quad (1.23)$$

And the counterfactual update, since both rewards (factual from chosen action, counterfactual from the one left unchosen) are observable when the payoff matrix is displayed at the time of the outcome:

$$Q_u(t) = Q_u(t-1) + (\alpha_2 \times \delta_u(t)) \text{ with } \delta_u(t) = R_u(t) - Q_u(t) \quad (1.24)$$

- Weighted Fictitious play (Cheung and Friedman, 1997 [4]):

The standard weighted fictitious model (Cheung and Friedman, 1997 [4]) does not take a reinforcement form with a prediction error but simply computes the probability of the opponent's action from a weighted average of frequency of its past choices, the steep of the exponential decay being controlled by the parameter  $\eta$ :

$$P_{A'}^*(t) = (C_{A'}(t) + \sum_{x=1}^{t-1} (\eta^x \times C_{A'}(t-x))) / (1 + \sum_{x=1}^{t-1} \eta^x) \quad (1.25)$$

- Influence Extension (2-Influence) (Devaine et al, 2014 [5]):

The agent considers that the opponent is playing using herself an influence learning model, and thus infers the probability the opponent computes over its own actions ( $P_{a/A}^{**}$ ) and how this influences her action probability ( $P_{Ala}^*$ ). As in fictitious and influence models, the agent then uses this probability to decide which action should be performed in order to maximize its action's consequent expected reward. The opponent's probability  $P_{Ala}^*$  of choosing an action

is dynamically inferred by tracking the last action the opponent makes and taking into consideration how her own past action (weighted by the influence parameter  $\lambda$ ) influences this calculus, as well as the agent's own past action. The influence parameter of the opponent's influence model thus becomes the influence second-order parameter  $\omega$ , which also captures the weight of the agent's actions on the opponent's choice behavior (action probability).

For Player 1 the probability  $P_A^*(t)$  of the opponent is obtained through :

$$P_A^*(t) = P_A^*(t-1) + (\eta \times \delta_A(t)) + (3 \times \lambda \times ((P_A^*(t-1) \times (1 - P_A^*(t-1)) \times (C_a - P_a^{**}(t))) - (3 \times \omega \times P_a^{**}(t) \times (1 - P_a^{**}(t)))))) \quad (1.26)$$

with

$$P_a^{**}(t) = 1/3 - (1/3 \times \beta \times (\log((1 - P_A^*(t-1))/P_A^*(t-1)))) \quad \text{and} \quad \delta_A(t) = C_a - P_A^*(t-1) \quad (1.27)$$

And for Player 2:

$$P_a^*(t) = P_a^*(t-1) + (\eta \times \delta_a(t)) - (3 \times \lambda \times ((P_a^*(t-1) \times (1 - P_a^*(t-1)) \times (C_a - P_A^{**}(t))) + (3 \times \omega \times P_A^{**}(t) \times (1 - P_A^{**}(t)))))) \quad (1.28)$$

with

$$P_A^{**}(t) = 1/3 + (1/3 \times \beta \times (\log((1 - P_a^*(t-1))/P_a^*(t-1)))) \quad \text{and} \quad \delta_a(t) = C_a - P_a^*(t-1) \quad (1.29)$$

The expected value of each action are then calculated from the estimated probability of the opponent's action  $P_{A/a}^*$  :

$$\text{Player 1 : } Q_a(t) = 1 - P_A^*(t) \quad Q_b(t) = 2 \times P_A^*(t) \quad (1.30)$$

$$\text{Player 2 : } Q_A(t) = 2 \times P_a^*(t) \quad Q_B(t) = 1 - P_a^*(t) \quad (1.31)$$

## 2- Optimization Process:

### a- Optimization procedure

Each model was fitted to subjects' choices using log-likelihood maximization with a slice sampling procedure (Bishop, 2006 [6], Drugowitsch et al, 2016 [7]). A slice sampler "samples" the parameter space and constructs Markovian "chains" of samples in which the frequency of each set of parameters is proportional to the likelihood function. This method has a high computational cost but presents advantages for high-dimensional parameters space. It allows for checking a posteriori the shape of each parameter posterior distribution. We ensured that samples were independent enough so that the parameters' average estimate was reliable. The slice sampler has a few parameters which we tuned empirically. We initialized our chain at random positions within the parameter space, and used 3 chains of 1 million samples each. We performed a last gradient ascent from the best sample in order to fine-tune the parameter optimization. We found that this method leads to better optimization results than random parameter sampling or grid search, gradient descent (fmincon initialized with multiple starting points in matlab), or a combination of the two (not shown). With this fitting procedure, we were able to identify, for each model the free parameter set that best fitted subjects' choices.

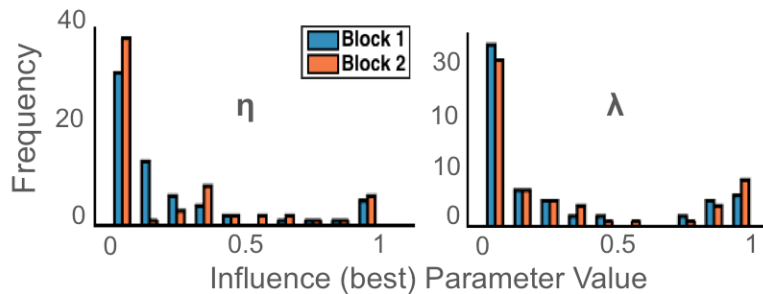
b- Population-level Model comparison (Bayesian Model Selection)

We compared the 3 models of our initial model space at the population level. To do so, we compared for each block the difference on average across all subjects in terms of LLH, AIC, BIC but also Log of posterior probability (LPP) and Exceedance probability (EP) (both computed using the VBA toolbox elaborated by Daunizeau et al, 2014 [8]). The Influence model fitted on average the behavior significantly better than the 2 other models for all of our criteria (Fig.S9).

A

	Block 1			Block 2		
	Q-Learning	Fictitious	Influence	Q-Learning	Fictitious	Influence
LLH	64.42 (5.39)	64.55 (5.42)	<b>62.02 (5.67)</b>	64.24 (5.93)	64.74 (5.64)	<b>62.38 (5.71)</b>
AIC	134.85 (10.78)	133.1 (10.84)	<b>130.04 (11.33)</b>	134.48 (11.86)	133.47 (11.28)	<b>130.77 (11.42)</b>
BIC	142.66 (10.78)	<b>138.3 (10.84)</b>	<b>137.85 (11.33)</b>	142.29 (11.85)	<b>138.68 (11.28)</b>	<b>138.58 (11.42)</b>
LPP	66.22 (5.27)	66.34 (5.31)	<b>63.92 (5.36)</b>	66.05 (5.80)	66.53 (5.56)	<b>64.22 (5.51)</b>
EP	0	0	<b>1</b>	0	0	<b>1</b>

B



**Figure S9. Exp.1, optimization results at the population level.** The Influence model fits best the choice behavior of our population of subjects but inter-individual heterogeneity is observed. (A) Bayesian Model Selection (BMS) with Model as fixed-effect, i.e. a single model best describes our population of subjects. Criteria included in the BMS: Log Likelihood (LLH), Akaike information criterion (AIC), Bayesian information criterion (BIC), Log of posterior probability (LPP) and Exceedance probability (EP). Best model indicated in bold. IGrey shades show the models whose population values are significantly different from the two other models ( $p < 0.01$ ). For the BIC score, the Influence model does not fit the population statistically better than the Fictitious, but instead both the fictitious and the influence (in bold) fit (significantly) better the subject's behavior compared to the Q-learning (thus, in grey) in the 2 blocks. The values between parentheses represent the (population) standard deviation of each selection criterion value reported in this table. (B) Distribution of the best parameter values of the Influence model obtained through the optimization process.

---

### 3- Subject-level Model comparison (extended model space)

#### a- Q-Learning models:

To increase the power of the standard Q-learning, we tested a Generalized version of the algorithm (see supplementary section II.1). The Generalized version did not fit significantly better the choice data of the subjects than the standard version (AIC:  $U(254)= 7031$ ,  $Z= 1.9591$ ,  $p= 0.0501$ , this result holds when running the comparison on each of the two blocks separately). Across the 2 blocks, 49% of the subjects were fitted better than chance (LRT,  $p<0.05$ ) by the standard Q-learning against 43% by the Generalized version and no difference in fit was found for those subjects (AIC:  $U(254)= 1868$ ,  $Z= 0.7283$ ,  $p = 0.4664$ , this result holds when running the comparison on each of the two blocks separately). Indeed, the average value of the best fitting (counterfactual decay) parameter of the Generalized Q-learning was quite low (0.70,  $std=0.35$ ), while no difference in Beta or Alpha was observed between the 2 versions). We therefore included the standard version of the Q-learning algorithm in our main Model Space (MS1).

#### b- Fictitious models:

We then controlled for the adequacy of the version of the fictitious model we used (from Hampton et al, 2008). To do so, we compared its performance to the weighted Fictitious model described in supplementary section II.1 (2 parameters each). No difference in fit was observed at the population level between the 2 models ( $U(254)= 8467$ ,  $Z= 0.4634$ ,  $p = 0.6431$ , **nor difference in fit difference between blocks:  $U(126)= 2308$ ,  $Z= 1.2367$ ,  $p= 0.2162$** ), accordingly across the 2 blocks 48% of the subjects were fitted better than chance (LRT,  $p<0.05$ ) by the weighted-fictitious for 51% by the TD version ( $\eta$  fictitious= 0.38(0.37),  $\eta$  weighted fictitious = 0.58(0.38)).

## IV - Additional Cognitive Tasks

### A) Reasoning tasks

#### 1- CRT

The Cognitive Reflexion Task (CRT) has been elaborated by Frederick, 2005 [9]. The task consists of three short questions that can be answered in less than 3 minutes. The three items of the CRT are designed such that the intuitive response is incorrect, but can be correctly reconsidered through some deliberation. In this sense, the CRT measures cognitive reactivity or impulsiveness, respondents' automatic response versus more elaborate and deliberative thinking. The three questions have an obvious incorrect answer that can be easily corrected upon minimal reflection. Those who arrive at the correct answers are less impulsive and more likely to engage in reflective

thinking. In this sense, the CRT can be viewed as a combination of cognitive capacity and the disposition for judgement and decision-making. Albeit very simple, the CRT has been proven to be quite robust [10].

## 2- Raven's Test

We used the Raven (Advanced Progressive Matrices) test to measure efficient problem-solving and abstract reasoning (Raven et al., 1998 [11]). The task consists of a series of pattern-matching tasks that do not require mathematical or verbal reasoning abilities. 30 different test items were presented on the screen, in each test item participants were asked to identify the missing element that completes a pattern of shapes. The patterns are presented in the form of a 3x3 matrix, and possible matching (missing) shapes were presented below with a number from 1 to 8, that had to be entered and validated before jumping to the next problem (**Fig.S10.A**). Performance at Raven's test is usually used as a non-verbal estimate of fluid intelligence [12]

### B) Working Memory tasks

In Exp. 2, we provided subjects with the Digit Span verbal test, the 7th item of the Wechsler adult intelligence scale, WAIS-III (Wechsler, 1997 [12]). The task consists of two parts, in which a series of digits with increasing length (2 to 9) are provided, and participants asked to repeat the 3 - 9 digits forward (first part) and 2 - 9 digits backwards (second part). The score represents the limit of performance reached by the participant in both parts (maximum length of digit retrieved). The task thus measures short-term memory but also attention and concentration.

In Exp. 3, we developed a computerized version of the N-Back test, an experimental paradigm widely used in cognitive neuroscience to assess working-memory capacities (Owen et al, 2005 [13]; Blain et al, 2016 [14]). The task was divided into two blocks: first a series of 100 2-Back trials, and then a series of 100 3-Back trials. At each trial, a random letter was presented in the middle of the screen and participants were instructed to respond by pressing the left or the right arrow on the keyboard (key-response association was counterbalanced across participants), indicating yes if the current letter was the same as the letter presented N trials before, or no otherwise (**Fig.S10.B**).

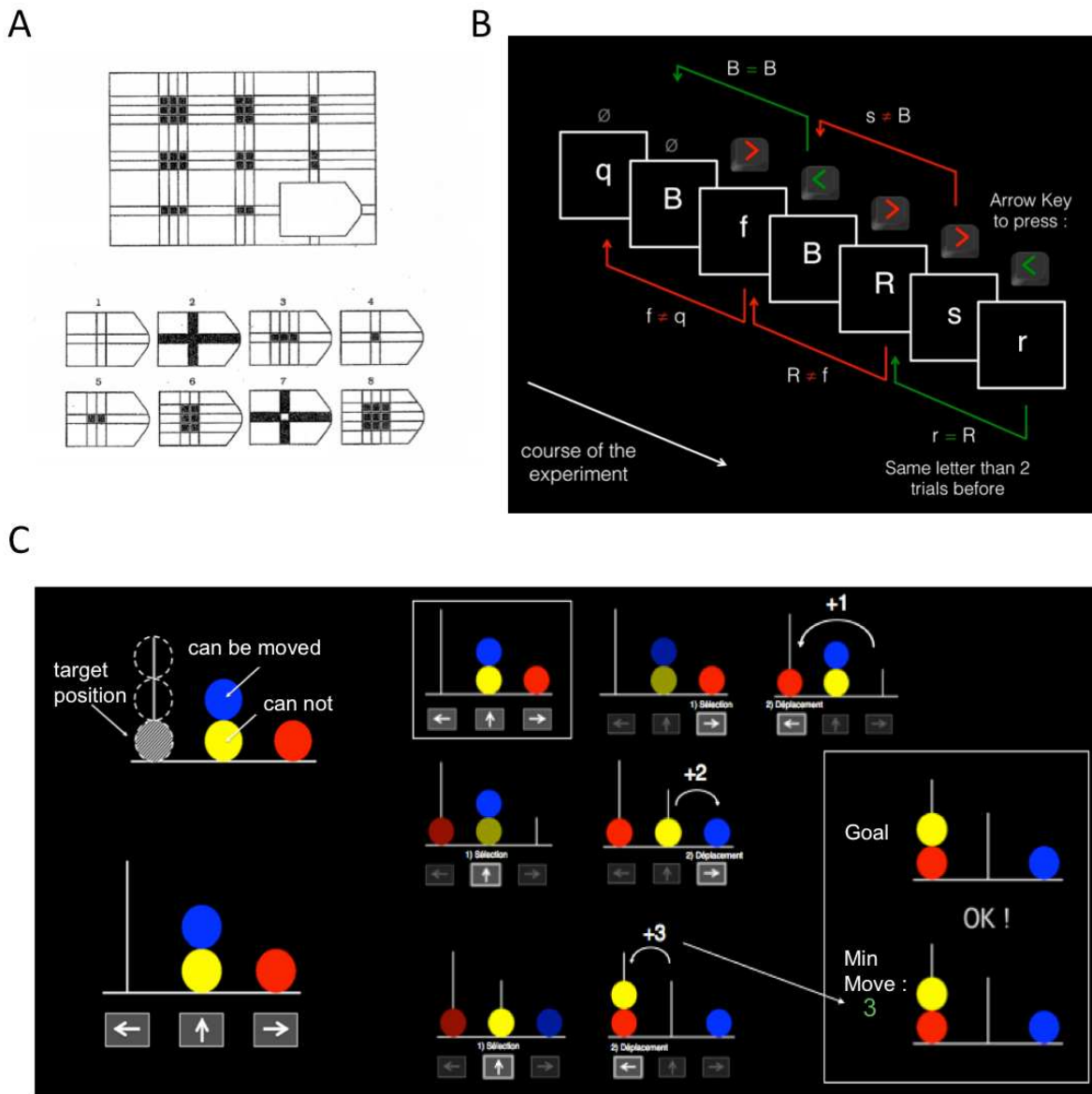
### C) Tower of London

The tower of London task is a task developed in neuropsychology in order to estimate planning abilities in humans. We adapted the psychometric version validated by Kaller et al 2012 [15]. The task consisted of three differently colored balls placed on three vertical rods of different heights that may hold at maximum either one, two or three balls, respectively. Start state and goal state were presented in the lower and upper parts of the screen, respectively. Subjects were asked to transform the start state to match the goal state while following three rules: (a) only one ball can be moved at a time; (b) a ball may not be moved if another ball is already on top of it; and (c) three balls can be accommodated at the tallest peg on the left, two balls at the peg in the middle, and one ball at the smallest peg on the right. The computer program did not allow rule-incongruent moves. (**Fig.10.C**). Two parameters are manipulated in a factorial design in this version of the task: Search depth (number of intermediate moves before the first ball can be placed into its goal position) and Goal hierarchy (ambiguity of information on subgoal ordering, i.e. the degree to which the sequence of final goal moves can be derived from the configuration of the goal state). The 32 problems



presented to the participants also varied, orthogonally to the two main parameters, in the minimum number of moves required to transform the respective start state into the goal state (**Fig.10.C**).




All the additional tasks (except for the digit span verbal test) were presented using PsychToolBox (Brainard, 1997 [16]) and appeared on a uniform black background.



**Figure S10.** Illustration of the Raven's task (A), the N-Back test (B), and the Tower of London © used as additional cognitive tasks.

## V - One-shot games experiment

A

Obs. Level of Strategic Learning	Exp. Level of Strategic Reasoning
<b>No SL</b> Reinforcement (RL)	<b>No SR</b> k-level =0 
<b>Low SL</b> Deviation from RL	<b>Low SR</b> k-level =1 
<b>High SL</b> Higher order Prob. Learning	<b>High SR</b> k-level >1 

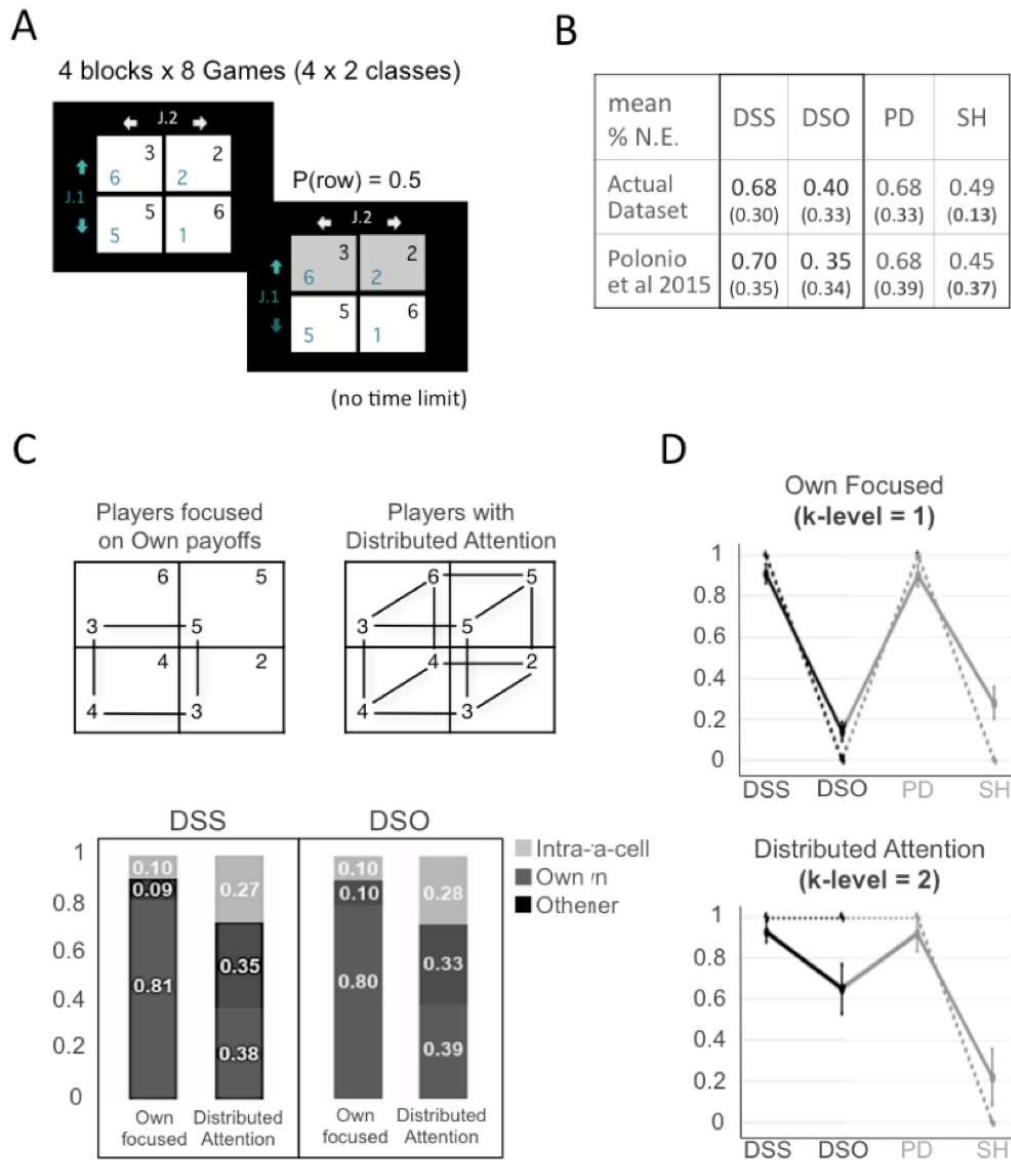
B

32 one-shot games in 4 blocks

4 classes of 8 games each :

		A	B			A	B
a		6,3	2,2	a		3,6	5,5
b		5,5	1,6	b		4,4	3,2
<b>DSS</b>				<b>DSO</b>			
		A	B			A	B
a		6,1	2,2	a		4,1	5,5
b		5,5	1,6	b		6,6	1,4
<b>PD</b>				<b>SH</b>			

**Figure S11. Endogeneous hypothesis of strategic learning sophistication (adapted from Griessinger et Coricelli, 2015).** (A) Hypothesized correspondence between individual strategic learning sophistication in the repeated game captured by our computational approach from the game behavior in the main task and the expected strategic reasoning ability measured in the secondary task. Each model dynamically differed in how much individuals incorporated the information relative to the opponent throughout the repeated interactions, and therefore in their level of strategic sophistication. (B) Secondary task design. Subjects were provided with 32 one-shot games without direct feedback (4 classes of 8 games each: Dominant Solvable Self, Dominant Solvable Other, Prisoner's Dilemma and Stag Hunt - adapted from Polonio et al, 2015 [17]) shuffled into 4 blocks of trials.



**Figure S12. Secondary task structure and rationale.** (A) 32 one-shot games shuffled in 4 blocks of 8 trials (2 games of each class in random order). In each game, subjects had to choose between one of the two options knowing that each trial could be selected for their final payoff. In half of the trials subjects were playing as row player, choosing between up and down; in the other half as column player, choosing between left and right. (B) Behavioral results of secondary task (% of Nash Equilibrium reached in each static game) replicate the results from Polonio et al 2015 at the population level. (C-D) (Adapted from Polonio et al, 2015 [16]): the strategic reasoning level (k-level) corresponds to specific information acquisition patterns. (C) Saccade patterns reveal different information processing in 2x2 static games: some subjects consistently focused on their own payoffs (own focused) while other focused on both their own and the opponent's payoffs (distributed attention) (top: black lines represent schematic saccades patterns). The proportion of these information processing types is independent of the payoff matrix (bottom). (D)

“Own focused” subjects were able to reach equilibrium play in DSS but not in DSO (low SR ; k-level = 1). On the other hand, “distributed attention” subjects reached equilibrium in both DSS and DSO, therefore playing as a high SR level (k-level = 2).

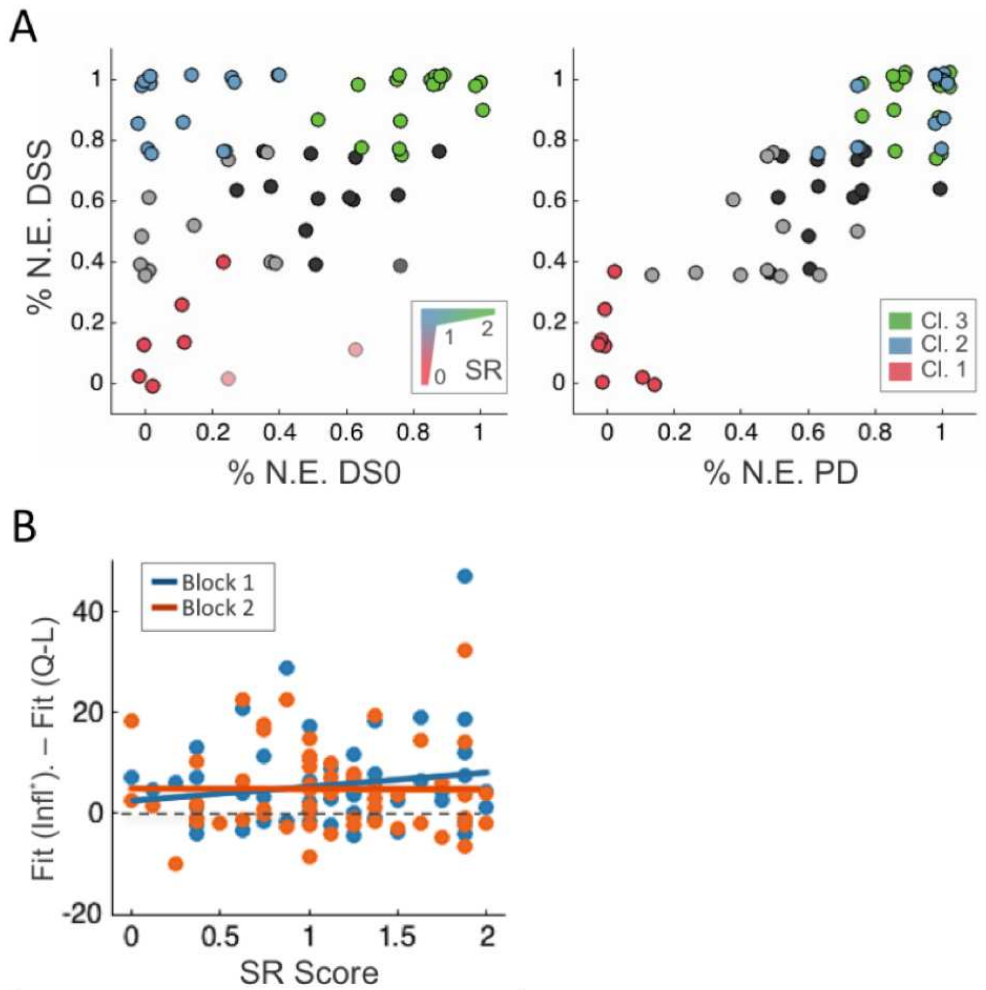
---

## 1- Methods:

The 32 games (8 in each class) were randomly but equally divided into four blocks in which subjects had to play either as row or column players. No choice feedback was provided at the end of each trial, but subjects were instructed that each trial could be the one selected at the end to determine their final payoff.

## 2- Results:

In this task, subjects were asked to make choices without direct feedback in a series of static games borrowed from Polonio et al (2015 [17]) (**Fig. S11.B; Fig. S12.A**). Performance was measured by the percentage of optimal choices in a game-theoretic sense (action leading to the Nash Equilibrium - NE) in each of the 4 classes of games for each subject. Performance of our subjects in this task replicates results from Polonio et al quite accurately (**Fig. S12.B**). No difference in percentage of N.E. nor reaction time was found between row and column play blocks (**N.E.: U(126)= 2096, Z= 0.2274, p = 0.8201, RT : U(126)= 2099, Z= 0.2407, p = 0.8098**), nor between each of the 4 blocks (**N.E.: F(3, 252)=0.05, p= 0.9866**; same results when ANOVA for 4 games separately, **RT: F(3, 252)=2.38, p= 0.0699**; same results when ANOVA for DSS, DSO and PD, but not SH (faster through blocks): **F(3, 252)=5 p= 0.0022**), nor even between games with low and high payoff amplitude within each class (**NE: U(126)= 1897, Z= 0.7258, p = 0.7258**), **RT: U(126)= 2394, Z= 1.6465, p = 0.0997**). To next investigate the congruence of the subjects' behavior across the 4 games, we conducted a multivariate cluster analysis to regroup subjects according to their performance in the 4 games. The best fitting model was the one with 5 clusters, represented in (**Fig. S13.A**) and which corresponds to 5 different coherent types of subjects across the 4 games. Indeed, the more subjects reached the Nash Equilibrium Dominant solvable games, (i.E. the DSS and DSO games, see next paragraph), the more they were able to choose the optimal action in the prisoner's dilemma game (these cluster results replicate Polonio et al). This result also replicates eye-tracking results from Polonio et al showing that reaching the NE in all of these 3 games requires to compute the information relative to both own payoffs and the other's payoffs in a dynamic strategic fashion. Note that subjects' behavior did not differ much between our 5 clusters in the Stag Hunt game (**Fig. S14.C**), congruent with the smaller variance observed in our data in comparison to Polonio et al (**Fig. S12.B**).



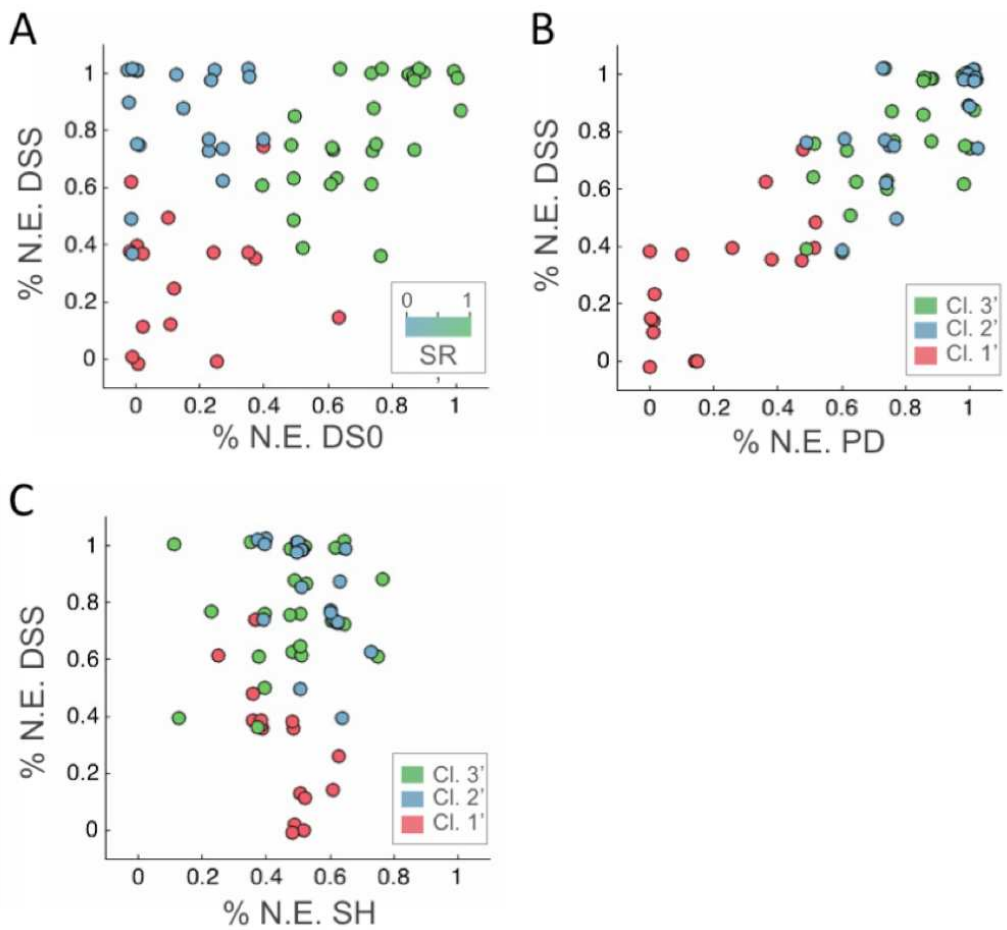
**Figure S13. Choice behavior in the secondary task are consistent but no direct matching with the level of Strategic Learning in the main task was found.** (A) Each dot represents for each subject the percentage of N.E. reach in each game (left: DSS and DSO, right: DSS and PD). The 5 colors represent the subjects in our 5 clusters across all of the 4 games, in blue and green our clusters of interest for our analysis. From 0 to 2 the SR score, gradient color matches the cluster colors. (B) Correlation between the difference in individual fit of the (high order) Influence model compared to the fit of the Q-learning in each block of the repeated game and the SR score. Correlation not significant for either block.

To test our hypothesis (Griessinger & Coricelli, 2015 [18]) (**Fig. S11.A**) that SL is not intrinsic, we focused on performances of subjects in the DSS games, in which reaching the Nash Equilibrium required to best respond to their own payoffs only, and in the DSO games in which choosing optimally required to also best respond to the other's payoffs. We first compared performances in the repeated game task between our 2 clusters of interest: subjects reaching most of the time the NE in DSS but not DSO (low strategic reasoning, SR) and subjects reaching NE in both

DSS and DSO (high SR) (namely Cluster 2 and 3 on **Fig. 13.A**). As expected, no difference in average points in the repeated game was observed between low and high SR subjects in either block (**Bl.1:  $U(29) = 82, Z = 1.451, p = 0.146$ ; Bl.2:  $U(29) = 100, Z = 0.735, p = 0.462$** ), nor in reaction time (**Bl.1:  $U(29) = 254, Z = 0.695, p = 0.487$ ; Bl.2:  $U(29) = 108, Z = 0.417, p = 0.677$** ).

We then tested directly our hypothesis on our entire population of subjects by comparing our measures of SL engagement in the repeated game for each group of low and high SR subjects. No difference was found between the SR groups in block 1 nor in block 2 in the quality of fit of the Influence model compared to the Q-learning model (**Bl.1:  $U(29) = 129, Z = 0.377, p = 0.706$ ; Bl.2:  $U(29) = 100, Z = 0.734, p = 0.463$** ), or the individual value of the Influence parameter (**Bl.1:  $U(29) = 120, Z = 0.019, p = 0.984$ ; Bl.2:  $U(29) = 106, Z = 0.496, p = 0.619$** ). Consistently no difference in any of the additional cognitive tasks was found between low and high SR subjects (i.e. age, Raven, CRT, WM, Tower of London score).

Another way of measuring the Strategic Reasoning ability in the static game task consists in simply computing a continuous SR score that corresponds to the sum of their performance in both DSS and DSO games, taking values from 0 to 2 (**Fig. 13.A** – colored gradient represented on the left plot legend). Indeed, the population behavior distribution makes subjects with such SR score below 1 display low or no strategic reasoning in the task, while subject above 1 display higher strategic behavior. No correlation was neither observed between this SR measure and the SL level in the repeated game, as measured by the quality of fit of the Influence model (or higher order influence) compared to the Q-learning model (**Bl.1 :  $r = 0.163, p = 0.226$ ; Bl.2 :  $r = -0.011, p = 0.933$** ) (**Fig. 13.B**), nor the individual Influence parameter value (**Bl.1 :  $r = -0.06, p = 0.661$ ; Bl.2 :  $r = -0.01, p = 0.929$** ) in either block. This was also true when we took as Strategic Reasoning measure only the performance in DSO (SR' score) for subjects in SR Cluster 2' and 3' (**Fig. S14.A**).



**Figure S14. Choice behavior in the secondary task are consistent across the 4 games.** (A-C) In colors are represented the subjects in each cluster when forcing the cluster routine to 3 clusters. Same conventions as Figure S13. Results are coherent with initial cluster analysis. No difference in percentage of N.E. in the SH game is observed between the 3 clusters.

Thus, while our subjects' choice behavior displayed in the secondary task replicates previous results using similar static games, we could not reject our null-hypothesis of an absence of direct mapping at the population level between the individual propensity to reason strategically in static games and the individual engagement in strategic learning during the repeated game interaction.

## VI - References

- [1] Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences*, 105(18), 6741-6746.
- [2] Khamassi, M., Quilodran, R., Enel, P., Dominey, P. F., & Procyk, E. (2015). Behavioral regulation and the modulation of information coding in the lateral prefrontal and cingulate cortex. *Cerebral cortex*, 25(9), 3197-3218.
- [3] Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature communications*, 6.
- [4] Cheung, Y. W., & Friedman, D. (1997). Individual learning in normal form games: Some laboratory results. *Games and Economic Behavior*, 19(1), 46-76.
- [5] Devaine, M., Hollard, G., & Daunizeau, J. (2014). The social Bayesian brain: does mentalizing make a difference when we learn?. *PLoS computational biology*, 10(12), e1003992.
- [6] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- [7] Drugowitsch, J., Wyart, V., Devauchelle, A. D., & Koechlin, E. (2016). Computational precision of mental inference as critical source of human choice suboptimality. *Neuron*, 92(6), 1398-1411.
- [8] Daunizeau, J., Adam, V., & Rigoux, L. (2014). VBA: A Probabilistic Treatment of Nonlinear Models for Neurobiological and Behavioural Data. *PLoS Comput Biol*, 10(1), e1003441.
- [9] Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, 19(4), 25-42.
- [10] Bialek, M., & Pennycook, G. (2017). The Cognitive Reflection Test is robust to multiple exposures. *Behavior Research Methods*, 1-7.
- [11] Raven, J. C., & John Hugh Court. (1998). *Raven's progressive matrices and vocabulary scales*. Oxford, UK: Oxford Psychologists Press.
- [12] Wechsler, D. (1997). *WAIS-III: Wechsler adult intelligence scale*. Psychological Corporation.
- [13] Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human brain mapping*, 25(1), 46-59.
- [14] Blain, B., Hollard, G., & Pessiglione, M. (2016). Neural mechanisms underlying the impact of daylong cognitive work on economic decisions. *Proceedings of the National Academy of Sciences*, 113(25), 6967-6972.
- [15] Kaller, C. P., Unterrainer, J. M., & Stahl, C. (2012). Assessing planning ability with the Tower of London task: psychometric properties of a structurally balanced problem set. *Psychological assessment*, 24(1), 46.
- [16] Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial vision*, 10, 433-436.
- [17] Polonio, L., Di Guida, S., & Coricelli, G. (2015). Strategic sophistication and attention in games: an eye-tracking study. *Games and Economic Behavior*, 94, 80-96.
- [18] Griessinger, T., & Coricelli, G. (2015). The neuroeconomics of strategic interaction. *Current Opinion in Behavioral Sciences*, 3, 73-79.



# - Appendix III -

## SUPPLEMENTARY INFORMATION 2

Additional analyses on the dataset from Exp.2 in:

Griessinger Thibaud, Khamassi Mehdi\*, Coricelli Giorgio\*. The interplay of learning sophistication and strategic asymmetry. About to be submitted.

### I- Computational Analysis - Model Space Extension 2

A- Model-space extension 2:

In this part we develop the details of the additional computational analysis we conducted on this dataset.

As presented in the article we extended the initial model space with additional versions of the following models :

- Q-Learning: the classic version was extended with a Generalized (Barraclough et al., 2004; Ito & Doya, 2009; Khamassi et al, 2014) and a Counter-factual version (Palminteri et al 2015)
- Fictitious: the version of Hampton was extended with the original weighted-fictitious (Cheung and Friedman, 1997)
- Influence: the Influence from Hampton et al (2008) was extended with the 2-Influence (Devaine et al, 2014)

Here we detail additional models related to the strategic learning literature in repeated games:

- EWA:

The original EWA from (Camerer et al, 1999, 2002) is considered as a hybrid between a RL and a Belief-Based model. Among its 3 parameters  $\delta$  is the more important, it represents the relative weight between foregone payoffs and actual payoffs in the update of the action's value. This parameter controls for the arbitrage between the RL and the Belief-Based component. The concept of Belief learning in the model is embodied in the foregone payoffs whereby counter-factual actions are reinforced, in addition to the factual Reinforcement of the chosen action by the payoff actually received at each trial. The hybrid model thus reduces to the RL model when  $\delta_i = 0$  , and the belief learning model when  $\delta_i = 1$  such as:

$$Q_c(t) = 1/N(t-1) \times ((\varphi \times N(t) \times Q_c(t-1)) + R_c(t)) \quad (1.1)$$

$$Q_u(t) = 1/N(t-1) \times ((\varphi \times N(t) \times Q_u(t-1)) + (\delta \times R_u(t))) \quad (1.2)$$

$$\text{With } N(t) = (\rho \times N(t-1)) + 1 \quad (1.3)$$

$\phi$  represents the belief about the speed of adaptation of the opponent, a small  $\phi$  means the agent believe that her opponent depreciates past values faster.

$N(t)$  captures the strength of past experience, i.E. the number of observation of past experience relative to one period of current experience, and impacts directly the update of the action value. If initiated at high value, it plays as a Bayesian prior, we thus set  $N(0)$  to 0.

A player with a low  $N(t)$  puts little weight on past “attractions”; a player with a huge  $N(t)$  is barely affected by immediate experience. the  $\rho$  parameter is considered to control for the influence of the out-of-game prior beliefs during the first trials of the interaction.

Note that a fourth parameter was originally entered in the models,  $\kappa$  which represented the discount rate for  $N(t)$ . When  $\kappa$  is large the effect of the prior beliefs will fade quickly.

- EWA variations (Zhu et al, 2012, Hampton et al, 2008):

Zhu et al developed a TD version of the EWA to allow the model to update the reward predictions through a prediction error therefore separating the reward prediction from the prediction error. Due to the equivalence observed empirically they considered that  $\phi = \delta$ , their model therefore includes only 2 parameters, the prior and the weight of the forgone payoff in the counterfactual update:

$$Q_c(t) = Q_c(t) + (1/N(t) \times (R_c(t) - Q_c(t-1))) \quad (1.4)$$

$$Q_u(t) = Q_u(t) + (1/N(t) \times ((\delta \times R_u(t)) - Q_c(t-1))) \quad (1.5)$$

$$\text{With } N(t) = (\rho \times N(t-1)) + 1 \quad (1.3)$$

Note that In their study Hampton et al (2008) also tested a simpler version of the EWA, with no decay or discount rate and where  $R_1$  is the reward obtained had action  $a$  been chosen (Fictitious learning), and  $R_2$  is the reward given that action  $a$  was chosen - zero otherwise (Reinforcement Learning).

$$Q_c(t) = ((1 - \alpha_1) \times Q_c(t-1)) + (\alpha_1 ((\delta \times R_u(t)) + ((1 - \delta) \times R_c(t)))) \quad (1.6)$$

- Fictitious play (pattern) variation (Spiliopoulos, 2012; 2013a)

The standard weighted fictitious by Cheung & Friedman (1997) computes the probability of the opponent’s action from a weighted average of frequency of its past choices, the steep of the exponential decay being controlled by the parameter  $\eta$ :

$$P_A^*(t) = (C_A(t) + \sum_{x=1}^{t-1} (\eta^x \times C_A(t-x))) / (1 + \sum_{x=1}^{t-1} \eta^x) \quad (1.7)$$

From this (first order) weighted fictitious model, Spiliopoulos developed a belief based model based on pattern recognition. In his fictitious play extension, named 2-p fictitious, the probability distribution over the opponent’s choice that will lead the action selection in the next trial is conditional over her last 2 choices instead only the choice made in the current trial such as:

$$P_{AA}^*(t) = (C_{AA}(t) + \sum_{x=1}^{t-1} (\eta^x \times C_{AA}(t-x))) / (1 + \sum_{x=1}^{t-1} \eta^x) \quad (1.8)$$

With  $p_{AA}(t)$  being the probability that the opponent (playing as player 2 in this case), chooses the action  $A$  after she played the action  $A$  in the previous trial. This way the fp2 algorithm tracks *how many times two temporally*

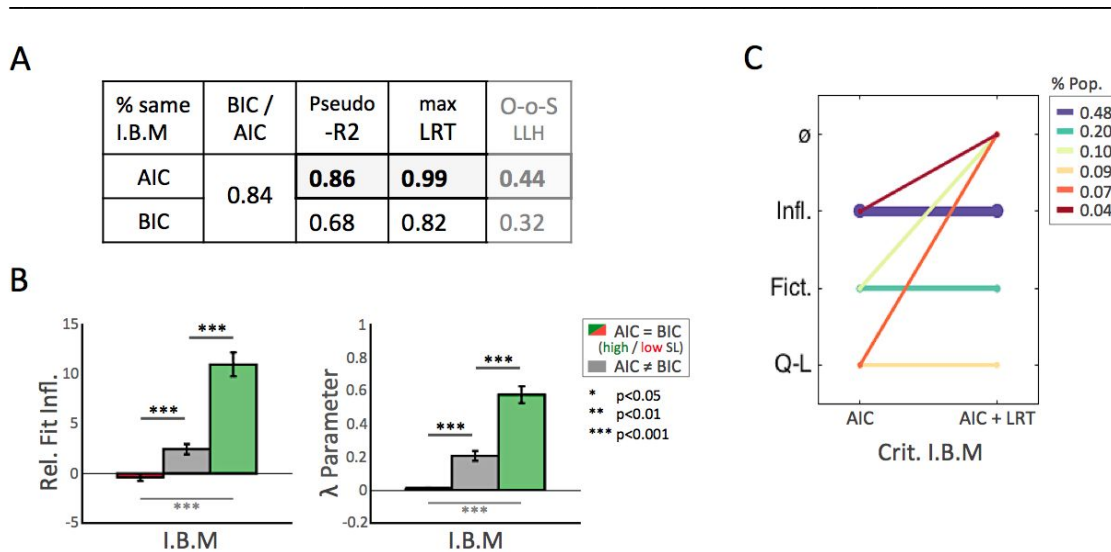
consecutive sequences of actions have been observed and then by conditioning the probability of an action being played on the previous action chosen by the opponent.

B- Bayesian Model Selection (BMS) - details:

Our goal here was to characterize for each subject the model, from the target Model Space that fitted the best their choice behavior in the repeated game interaction.

For each subjects we considered a model as the Individual Best Model (I.B.M.) if it fitted better than chance her choice series (Likelihood ratio test) and better than the other model included in the model space (lower AIC).

We initially considered 5 different model selection criteria: LLH, AIC, BIC, Pseudo-R2 (Camerer et al, 1999; Daw, 2011) and the max-LRT score (Daw, 2011). As presented on **Fig.S1.C** AIC presents the highest congruence between all the BMS criteria used. Note that since we hypothesized that the opponent encountered at each block might affect the model best fitting the individual behavior we did not include the Out-of-Sample LLH (ref., i.E. the LLH of the model when fitted to the choice series of one of the 2 block trial with the best parameter set obtained through optimization on the other block), but the AIC was also more coherent with this criterion.



**Figure S1.** Congruence between the different criteria of quality of fit used for the model selection. AIC > BIC

A post-hoc test was performed to confirm that the I.B.M. selection based on AIC (+LRT) was the most appropriate to our computational analysis. As presented on **Fig.S1.A** for 0.16% of the choice series (across the 2 block trials) AIC and BIC did not lead to the selection of the same I.B.M., the former considering those datasets as best fitted by the Influence model, while the latter mainly lead to the selection of the Fictitious play (for 88%, Q-Learning otherwise). We compared the relative fit of the Influence (individual difference in AIC between the Influence and the Q-learning

models) and the value of its  $\lambda$  parameter for the choice series that were considered by both AIC and BIC as best fitted either by the Belief-Based models (Q-Learning, Fictitious) (Low SL, Red ) or by the Strategic model (Influence, High SL, Green) to the choice series which were best fitted by the Influence according to the AIC and by the Belief-based models according to the BIC (Grey - **Fig.S1.B**).

This result further confirms that considering these choice series as best fitted by the Influence makes more sense, and therefore confirms the adequacy of the AIC criterion in our analysis. Thus using the AIC criterion instead of the BIC increased the discrepancy between the Low and High strategic learning subjects and the heterogeneity of the 2 groups while not changing the congruency of our measures of SL level nor the results presented in the main text.

As mentioned previously we included a prerequisite to the model selection procedure: for a model in the model space to be considered as (individually) best fitting a subject's choice behavior in a repeated game series (block), it had to fit statistically better than chance (i.E. a reduced model that would predict with a probability 0.5 the subject's choice at each trial). We thus performed for each model (each subject, each choice series) first a Log Likelihood ratio test taking into account the degree of freedom (number of free parameters in the model), and then considered as I.B.M. the model with the lowest AIC among those which fitted significantly better than chance ( $p < 0.01$ , Daw, 2011). Only few choice series did not pass this additional criterion (**Fig.S2.C**)

#### C- Model Comparison of the extended model space 2:

##### 1- Within-model comparison:

###### - Fictitious fp2:

We then tested how the 2-pattern version of the fictitious (fp2) fitted the dataset. The fp2 model did not fit significantly better the choice data of the subjects than the fictitious model (**AIC=135.45(7.52), U(254)=7377, Z= 1.3750, p=0.1691**), and fit worse the population data than the Influence model (**AIC: U(254)=10828, t=4.4492, p= 1.7383e-08**). Finally we compared the relative fit (individual difference between the fit of the target model compared to the fit of the Q-learning) of the fp2 and the Influence model. The Influence model had a higher relative fit than the fp2 (**U(254)=5275, z=8.4999e-07 1.7383e-08**). These three results hold for the comparison within each block (separately) and using BIC as criteria for goodness of fit. Similar results were obtained when using the 3-pattern version (fp3) from Spiliopoulos (2012)

###### - EWA:

We compared the simplified TD version of the EWA developed by Zhu et al (2012) to the original version of the EWA from Camerer which contains an extra parameter. The two models lead to equivalent fit (no difference in AIC nor BIC). In fact the fit of the two models were highly correlated (**correlation in AIC: B= 1.0050, r= 0.9896, p < 0.00001, in relative fit with Q-L : B= 0.9618, r= 0.9126, p < 0.00001**), as well as the (best) values of  $\delta$  obtained through the optimization process (**B= 0.7153, r= 0.7486, p < 0.00001**). We thus chose to consider the TD-version of the EWA from Zhu et al to allow better comparison with Q-Learning versions, but also for its more parsimonious form. We thus

compared the TD-EWA to the one-parameter version developed by Ho et al, 2007, the Self Tunique EWA, but no improvement in fit was found ( $U(254)= 9193$ ,  $Z= 1.6890$ ,  $p= 0.0912$ ).

## 2- Between-model comparison:

We first compared the Belief-based mixture embodied in the EWA with the Fictitious model which computed directly the choice probability of the opponent based on her past choices. The assumption behind the EWA is that the more subjects consider the counterfactual reward in their learning process the more they form belief over the opponent's behavior and best respond to it. However if the relative fit of the two models correlated with each other significantly it was not as strong as the correlation with the Counter-factual Q-learning (which updates the counter-factual Q-value at a certain rate - weight) (**rel fit TD-EWA vs. Fict.:  $r= 0.4463$ ,  $p< 0.0001$ , CF-QL vs. Fict:  $r= 0.7764$ ,  $p< 0.0001$** ) and no correlation was found between the propensity to consider the counterfactual reward in the update process (embodied in the  $\delta$  parameter) and the relative fit of the fictitious model ( $r= 0.0872$   $p= 0.3280$ ).

Moreover the weight attributed to the choice of the opponent in the update of the belief over her choice probability at each trial (i.E.  $\eta$  parameter) did not correlate either with the weight on counterfactual reward in EWA ( $r= 0.1034$   $p= 0.2455$ ) but with the counterfactual learning rate ( $\alpha_c$ ) in the CF-Q-learning ( $r= 0.7107$ ,  $p< 0.0001$ ). Similar results were obtained when comparing the full version of the EWA and Fictitious and Counter-factual Q-learning.

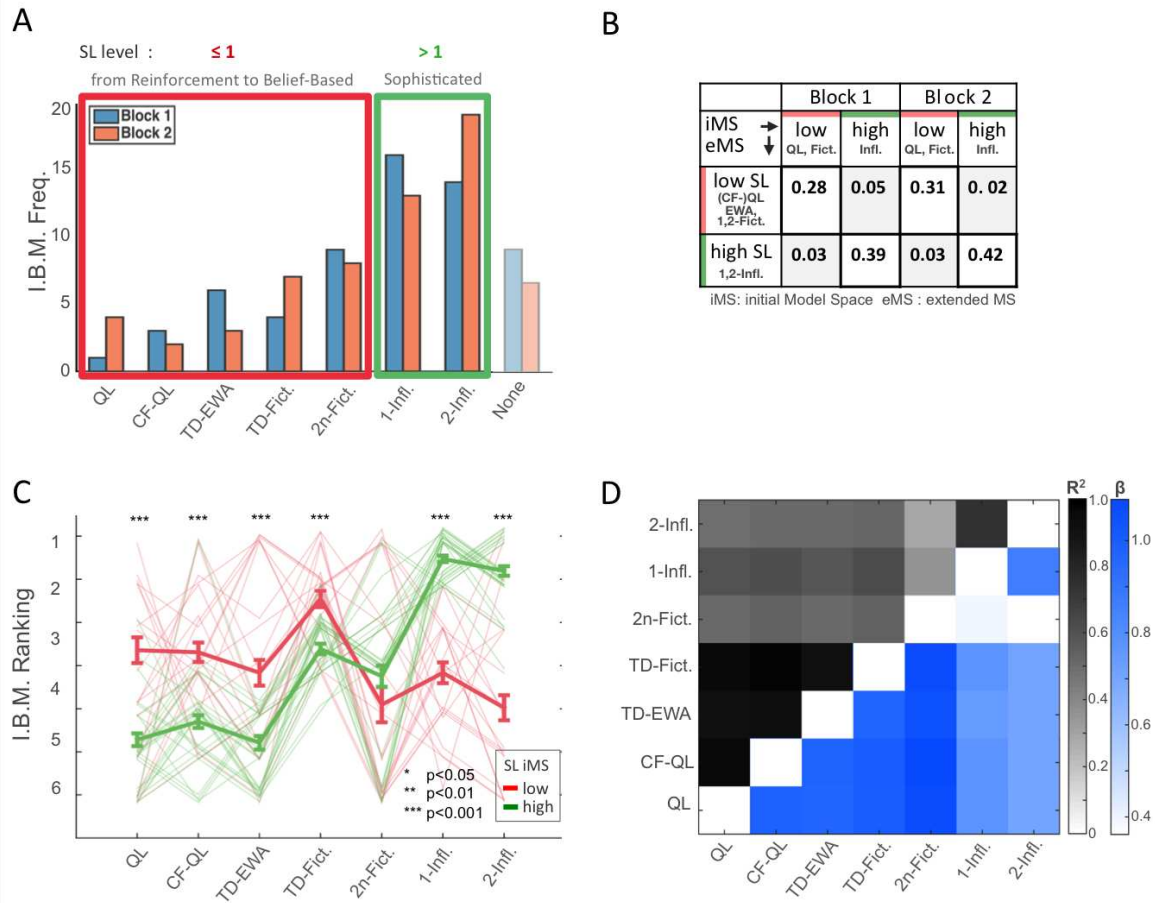
One way to explain why the Counter-factual Q-learning behaves closely to a Belief-Based model that the EWA is to look at their computational differences. The two models could update the unchosen option value but in 2 different ways, the first updates the corresponding Q-value (when  $\delta$  tends towards 1) to a certain degree (individual parameter) influencing the weight of the outcome received, the other modulates directly the updating rate of the unchosen Q-value. Thus the Counter-factual QL is the one adjusting its belief regarding the pertinence of the counter-factual update.

We also ran 2 versions of a WS-LS (Win Stay - Lose Shift) model, a deterministic version ( $\beta$  in softmax fixed at 10 to exploit Q-valued computed - see Devaine et al, 2014) and a probabilistic version ( $\beta$  estimated through optimization). Both models did not fit the population better than any of the models included in the 3 model spaces tested for the model comparison. The correlation between the WS-LS fit and the relative fit of the influence (or the fictitious) was not significant (and no significant difference in fit was found between low and high SL subjects for both versions of the WS-LS model - results not shown). This result is coherent with the observation that very few subjects were best fitted by the Q-learning algorithms in the 2 blocks and confirms that our subjects incorporated the information relative to the opponent in their choice process.

## 3- Between-subject comparison:

Once the optimization process launched on all the models included in the extended model space, we selected for each subject the I.B.M. following the procedure described previously. **Fig.S2.A** represents the frequency of subjects individually best fitted by each model.

Most of the subjects previously best fitted by the Q-learning or the fictitious play models are now best fitted by models of lower strategic sophistication, while subjects previously best fitted by the Influence model are best fitted by one of the two versions of the Influence models (1st or 2nd order) (**Fig.S2.B**).



**Figure S2.** Extended computational modeling analysis (A) Individual Best Model (I.B.M.) frequency plot of extended Model Space (EMS). At the individual level, the sophisticated models (in green) best fits about half of the subjects, while the remaining were best fitted by lower levels of strategic learning models (red). (B) Subjects individually best fitted by low SL level models (Q-L, Fictitious) in the Initial - reduced - Model Space (IMS) are consistently best fitted by the low Strategic Learning (SL) models of the Extended Model Space. Subjects best fitted by the high SL model (Influence) of the IMS are coherently best fitted by the Higher SL models of the EMS.

Dividing our population in low vs. high strategic learners based on the model that best fitted their choice behavior requires that the models considered as low level of strategic sophistication and the one identified as high level cluster together in terms of quality of fit. This was confirmed by the the results of the correlations in relative fit between each

model of our extended model space. The analysis ran on our whole dataset is summarized on **(Fig.S2.D)** As hypothesized the correlation analysis confirms that the quality of fit of the 2 influence models correlate with each other, as well as the Belief-Based models but the 2 categories of models do not correlate with each other (black frames). When plotting the average best model ranking of our low vs. high level of strategic learning subjects we observe the expected double dissociation between Belief-based better ranked in terms of quality of fit compared to the Influence models for low SL, and vice versa **(Fig.S2.C)** - similar highly significant dissociation was obtained when clustering our subjects in 2 equal groups given the values of their best Influence parameter ( $\lambda$ ), low close to 0 and high close to 1).

#### 4- Q-learning variations 2:

We also tested 4 additional variation of Q-Learning but none of these learning rules provided a fit of sufficient quality to be considered in our extended Model Space.

#### 2-States Q-Learning models with Belief-Based heuristic:

The initial RL model being equivalent to a Q-Learning with only one state. We tested Q-Learning algorithm considering 2 states (s), each corresponding to one action available to the opponent ( $p(s=1) = P$ ).

$$Q_{s,c}(t) = Q_{s,c}(t-1) + (\alpha_1 \times \delta_{s,c}(t)) \text{ with } \delta_{s,c}(t) = R_c(t) - Q_{s,c}(t)$$

To update the action values in the accurate state (Q(s,a)) the model thus has to know in which of the 2 states he's in, i.E. to use heuristic to predict which choice the opponent will make at each trial to choose the action with the higher value in the corresponding state selected. Here learning on Q-values, not on the opponent strategy since heuristics are used.

At each trial the model selects the state, based on a simple heuristic, and they update the corresponding Q-value once the outcome is displayed.

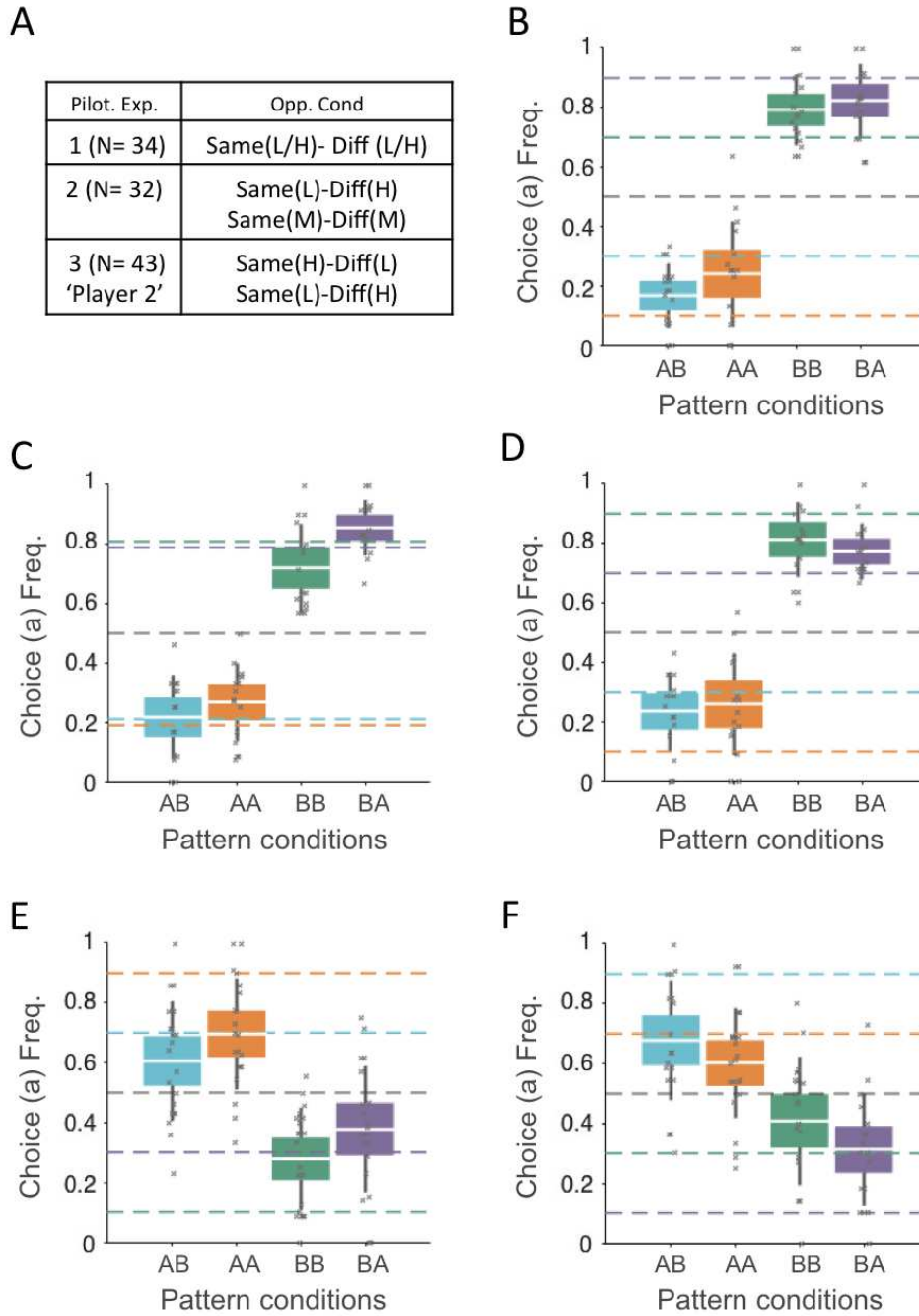
4 heuristics were tested (from Player 1 perspective):

- 1- "Previous": state s corresponds to the action previously selected by the opponent
- 2- "Highest": state s corresponds to the state leading to the highest Q-value
- 3- "Frequent" : state s corresponds to the action the most frequently selected by the opponent in all the past trials of the same block
- 4- "WS-LS" : state s corresponds a deterministic Win-Stay Lose Shift strategy apply to the opponent, so that if the action previously chosen by the opponent lead him to not lose ( $R_t^O > 0$ ), then the state s is the same, if the action selected by the opponent lead her to lose ( $R_t^O = 0$ ) then s corresponds to the other action.

## - Appendix IV -

**SUPPLEMENTARY INFORMATION 4**

**I- Supplementary Figures**

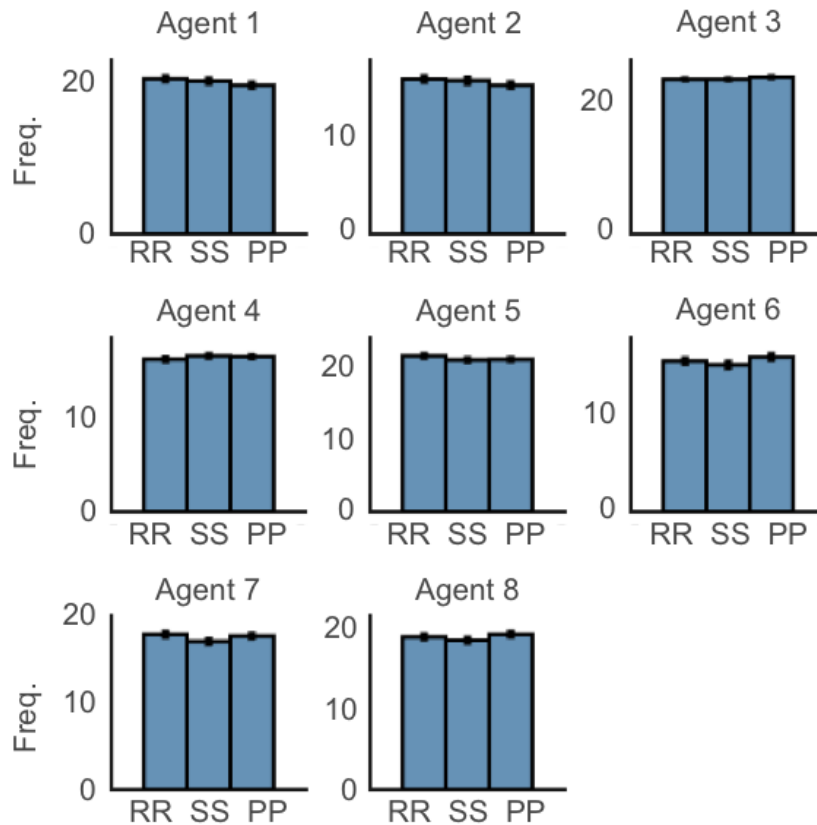


**Figure S1.** Overview of the pattern learning in the “opponent-based” block - additional experiments (pilot studies - second half of the interaction block). (A) Summary of the 3 pilot studies. Colored dashed lines represent the



probability of the opponent to choose the action "A" (B,C,D) or "B" (E,F), in each pattern condition. The black dashed line indicates chance level ( $p(a)=0.5$ ) The probability associated to each pattern condition (either "same" for patterns composed by the two same choices, i.e AA and BB, or "different", when two different choices) could either be low (0.7, "L"), medium (0.8,"M") or high (0.9,"H"). Pilots 1 and 2 only vary in the probability associated to each of the four pattern conditions, Pilot 3 also differs in the rule of the game (role endorsed by the player) . Subjects' choice probability in the last 50 trials matches in best response the probabilistic patterns in the opponent's choice behavior especially when the opponent chose differently in the last 2 trials (B) Initial experiment (Pilot 1). The probabilities associated to each pattern were chosen to avoid two-back learning: the two patterns labeled as "same" or the two labeled as "different" were associated to different probabilities. Altogether, these results consistently show a higher, albeit non significant, tendency of subjects to best respond to patterns composed by two different choices in the last two trials (i.e. AB and BA). (C,D) This tendency was confirmed in Pilot 2 when pairing all the four pattern conditions to the same probability (0.8): subjects best responded to the "different" rather than to the "same" pattern conditions. (plot C) We also tested a version of the probabilistic pattern task (plot D) in which the 2 combination in the "same" and "different" pattern conditions were each paired to a different probability, respectively low (0.7) and high (0.9) probability. In this setting, subjects managed to discriminate between low and high probability patterns. (E,F) Finally we tested in Pilot 3 a version of the task were subjects played as Player2 and not as Player1, thus changing the rule to "in order to win, try to select the opposite fractal as your opponent at the same trial". On average subjects seem to be still able to discriminate between each pattern condition but performed worse than when playing as Player 1 (for exact same probabilistic pattern condition with opposite role/rule - E vs. D : all block :  $t(43.6)=2.1514$ ,  $p=0.0370$ , last 50 trials:  $t(38.23)=2.3278$ ,  $p= 0.0253$ ).

---



**Figure S2.** Control of the suitability of the Hybrid Algorithm for the opponent (H). 8 agents were simulated (100 times) playing the repeated RPS game for 200 trials as Player 1 against different simulated opponents. Frequency indicates the number of times each pattern emerged from the interaction. Each agent implemented a different strategy : (1) MSNE play ( $p(R,P,S) = 1/3$ ), (2) Optimal ( $P(cBR) = 1$  in pattern trials, random in non-pattern trials), (3) Fictitious play (same parameter values as Maximize (F) opponent), (4) Hybrid opponent (same parameter values as (H)). We then tested different heuristics, among them : (5) Win-Stay/Lose-Shift (6) Best response to Opponent's best response to agent's choice at  $(t-1)$ , (7) Alternate between options (selects randomly one of the 2 options not chosen last trial), (8) MSNE except if two same choices in a row, then shift to randomly one of the 2 other options.

## II - Computational Modelling (computerized opponents)

- Fictitious play Opponent (F):

This opponent was model by a standard weighted fictitious (Cheung & Friedman, 1997) which simply computes the probability of the subject's action from a weighted average of frequency of her past choices, the steep of the exponential decay being controlled by the parameter  $\eta$  :

$$P_a^*(t) = (C_a(t) + \sum_{x=1}^{t-1} (\eta^x \times C_a(t-x))) / (1 + \sum_{x=1}^{t-1} \eta^x)$$

The estimated probability of the subject's action is then transposed into Q-values given the payoff matrix of the game for each opponent action (illustrated here for Exp.5) :

$$Q_A(t) = 1 - P_a^*(t) \quad Q_B(t) = P_a^*(t)$$

Which are then converted to action selection through a softmax function :

$$P_A(t) = 1 / (1 + \exp(\beta \times (Q_B(t-1) - Q_A(t-1)))) \text{ and } P_B(t) = 1 - P_A(t)$$

The fictitious play model has thus 2 free parameters: the inverse temperature  $\beta$  controlling for the noise ratio in the action selection of the opponent (arbitration between greedy exploitation and noisy exploration of the Q-value associated to each action) and the decay parameter  $\eta$  controlling for the size of the memory. For Exp. 5, we fixed the parameters at  $\beta = 5$  and  $\eta = 0.5$ , to obtain the learning curve displayed on **Fig.2.A** (bottom right plot) presenting a credible tradeoff between maximizing learning adaptation and probabilistic action selection. For Exp. 6, we fixed the parameters at  $\beta = 5$  and  $\eta = 0.55$ , to obtain the learning curve displayed on **Fig.9.B** (middle plot).

## - References -

- Abe, H., & Lee, D. (2011). Distributed coding of actual and hypothetical outcomes in the orbital and dorsolateral prefrontal cortex. *Neuron*, 70(4), 731-741.
- Abrahamse, E. L., Jiménez, L., Verwey, W. B., & Clegg, B. A. (2010). Representing serial action and perception. *Psychonomic bulletin & review*, 17(5), 603-623.
- Alaoui, L., & Penta, A. (2015). Endogenous depth of reasoning. *The Review of Economic Studies*, 83(4), 1297-1333.
- Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature reviews. Neuroscience*, 7(4), 268.
- Apperly, I. A. (2008). Beyond simulation–theory and theory–theory: why social cognitive neuroscience should use its own concepts to study “Theory of Mind”. *Cognition*, 107(1), 266-283.
- Apps, M. A. J., & Ramnani, N. (2017). Contributions of the Medial Prefrontal Cortex to Social Influence in Economic Decision-Making. *Cerebral Cortex*, 27(9), 4635-4648.
- Apps, M. A., & Ramnani, N. (2014). The anterior cingulate gyrus signals the net value of others' rewards. *Journal of Neuroscience*, 34(18), 6190-6200.
- Apps, M. A., Rushworth, M. F., & Chang, S. W. (2016). The anterior cingulate gyrus and social cognition: tracking the motivation of others. *Neuron*, 90(4), 692-707.
- Aslin, R. N., & Newport, E. L. (2012). Statistical learning: From acquiring specific items to forming general rules. *Current directions in psychological science*, 21(3), 170-176.
- Bahlmann, J., Aarts, E., & D'Esposito, M. (2015). Influence of motivation on control hierarchy in the human frontal cortex. *Journal of Neuroscience*, 35(7), 3207-3217.
- Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1, 0064.
- Baker, R., Dexter, M., Hardwicke, T. E., Goldstone, A., & Kourtzi, Z. (2014). Learning to predict: Exposure to temporal sequences facilitates prediction of future events. *Vision research*, 99, 124-133.
- Balleine, B. W., & O'doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, 35(1), 48.
- Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind”? *Cognition*, 21(1), 37-46.

- Barraclough, D. J., Conroy, M. L., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature neuroscience*, 7(4).
- Battalio, R., Samuelson, L. et Van Huyck, J. (2001). « Optimization Incentives and Coordination Failure in Laboratory Stag Hunt Games ». *Econometrica*, vol. 69 no 3 : pp. 749–64.
- Bault, N., Joffily, M., Rustichini, A., & Coricelli, G. (2011). Medial prefrontal cortex and striatum mediate the influence of social comparison on the decision process. *Proceedings of the national Academy of sciences*, 108(38), 16044-16049.
- Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. *Nature*, 456(7219), 245.
- Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature neuroscience*, 10(9).
- Bhatt, M. A., Lohrenz, T., Camerer, C. F., & Montague, P. R. (2010). Neural signatures of strategic types in a two-person bargaining game. *Proceedings of the National Academy of Sciences*, 107(46), 19720-19725
- Bhatt, M., & Camerer, C. F. (2005). Self-referential thinking and equilibrium as states of mind in games: fMRI evidence. *Games and economic Behavior*, 52(2), 424-459.
- Biele, G., Rieskamp, J., Krugel, L. K., & Heekeren, H. R. (2011). The neural basis of following advice. *PLoS biology*, 9(6), e1001089.
- Bolt, N. K., & Loehr, J. D. (2017). The predictability of a partner's actions modulates the sense of joint agency. *Cognition*, 161, 60-65.
- Bond, R., & Smith, P. B. (1996). Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological bulletin*, 119(1), 111.
- Boorman, E. D., O'Doherty, J. P., Adolphs, R., & Rangel, A. (2013). The behavioral and neural mechanisms underlying the tracking of expertise. *Neuron*, 80(6), 1558-1571
- Bosch-Domenech, A., Montalvo, J. G., Nagel, R., & Satorra, A. (2002). One, two,(three), infinity,...: Newspaper and lab beauty-contest experiments. *The American Economic Review*, 92(5), 1687-1701.
- Botvinick, M., & Braver, T. (2015). Motivation and cognitive control: from behavior to neural mechanism. *Annual Review of Psychology*, 66.
- Broadbent, D. P., Ford, P. R., O'Hara, D. A., Williams, A. M., & Causer, J. (2017). The effect of a sequential structure of practice for the training of perceptual-cognitive skills in tennis. *PloS one*, 12(3), e0174311.
- Brown, G. W. (1951). Iterative solution of games by fictitious play. *Activity analysis of production and allocation*, 13(1), 374-376.

- Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences*, 107(32), 14431-14436.
- Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.
- Camerer, C. F. (2008). Neuroeconomics: opening the gray box. *Neuron*, 60(3), 416-419.
- Camerer, C. F., Ho, T. H., & Chong, J. K. (2002). Sophisticated experience-weighted attraction learning and strategic teaching in repeated games. *Journal of Economic theory*, 104(1), 137-188.
- Camerer, C. F., Ho, T. H., & Chong, J. K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3), 861-898.
- Camerer, C. F., Ho, T. H., & Chong, J. K. (2015). A psychological approach to strategic thinking in games. *Current Opinion in Behavioral Sciences*, 3, 157-162.
- Camerer, C. F., Loewenstein, G., & Prelec, D. (2004). Neuroeconomics: Why economics needs brains. *The Scandinavian Journal of Economics*, 106(3), 555-579.
- Camerer, C. F., Loewenstein, G., & Rabin, M. (Eds.). (2011). *Advances in behavioral economics*. Princeton University Press.
- Camerer, C., & Hua Ho, T. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, 67(4), 827-874.
- Camerer, Colin F., Teck-Hua Ho, and Juin Kuan Chong. "Behavioural game theory: Thinking, learning and teaching." *Advances in Understanding Strategic Behaviour*. Palgrave Macmillan UK, 2004. 120-180.
- Carpenter, J., Graham, M., & Wolf, J. (2013). Cognitive ability and strategic sophistication. *Games and Economic Behavior*, 80, 115-130.
- Caruana, N., Spirou, D., & Brock, J. (2017). Human agency beliefs influence behaviour during virtual social interactions. *PeerJ*, 5, e3819.
- Chambon, V., Domenech, P., Jacquet, P. O., Barbalat, G., Bouton, S., Pacherie, E., ... & Farrer, C. (2017). Neural coding of prior expectations in hierarchical intention inference. *Scientific Reports*, 7(1), 1278.
- Chan, S. C., Niv, Y., & Norman, K. A. (2016). A probability distribution over latent causes, in the orbitofrontal cortex. *Journal of Neuroscience*, 36(30), 7817-7828.
- Chase, H. W., Kumar, P., Eickhoff, S. B., & Dombrovski, A. Y. (2015). Reinforcement learning models and their neural correlates: an activation likelihood estimation meta-analysis. *Cognitive, affective, & behavioral neuroscience*, 15(2), 435-459.

Chater, N., Felin, T., Funder, D. C., Gigerenzer, G., Koenderink, J. J., Krueger, J. I., ... & Stanovich, K. E. (2017). Mind, rationality, and cognition: An interdisciplinary debate. *Psychonomic Bulletin & Review*, 1-34.

Cheung, Y. W., & Friedman, D. (1997). Individual learning in normal form games: Some laboratory results. *Games and Economic Behavior*, 19(1), 46-76.

Christopoulos, G. I., & King-Casas, B. (2015). With you or against you: Social orientation dependent learning signals guide actions made for others. *Neuroimage*, 104, 326-335.

Chu, Ronald. *Extending Fictitious Play with Pattern Recognition*. MS thesis. 2013.

Chung, D., Christopoulos, G. I., King-Casas, B., Ball, S. B., & Chiu, P. H. (2015). Social signals of safety and risk confer utility and have asymmetric effects on observers' choices. *Nature neuroscience*, 18(6), 912-916.

Collins, A. G. (2017). The cost of structure learning. *Journal of Cognitive Neuroscience*.

Collins, A. G. E., & Frank, M. J. (2016). Neural signature of hierarchically structured expectations predicts clustering and transfer of rule sets in reinforcement learning. *Cognition*, 152, 160-169.

Collins, A. G., & Frank, M. J. (2013). Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological review*, 120(1), 190.

Collins, A., & Koechlin, E. (2012). Reasoning, learning, and creativity: frontal lobe function and human decision-making. *PLoS biology*, 10(3), e1001293.

Cooper, J. C., Dunne, S., Furey, T., & O'Doherty, J. P. (2012). Human dorsal striatum encodes prediction errors during observational learning of instrumental actions. *Journal of cognitive neuroscience*, 24(1), 106-118.

Coricelli, G., & Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, 106(23), 9163-9168.

Coricelli, G., & Rustichini, A. (2010). Counterfactual thinking and emotions: regret and envy learning. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1538), 241-247.

Crawford, V. P., Costa-Gomes, M. A., & Iriberri, N. (2013). Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature*, 51(1), 5-62.

Crawford, Vincent P., Miguel A. Costa-Gomes, and Nagore Iriberri. "Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications." *Journal of Economic Literature* 51.1 (2013): 5-62.

da Silva, C. F., Victorino, C. G., Caticha, N., & Baldo, M. V. C. (2017). Exploration and recency as the main proximate causes of probability matching: a reinforcement learning analysis. *bioRxiv*, 104752.

Daltrozzo, J., & Conway, C. M. (2014). Neurocognitive mechanisms of statistical-sequential learning: what do event-related potentials tell us?. *Frontiers in human neuroscience*, 8.

- Davey, C. G., Allen, N. B., Harrison, B. J., Dwyer, D. B., & Yücel, M. (2010). Being liked activates primary reward and midline self-related brain regions. *Human brain mapping*, *31*(4), 660-668.
- Daw, N. D. (2011). Trial-by-trial data analysis using computational models. *Decision making, affect, and learning: Attention and performance XXIII*, *23*, 3-38.
- Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204-1215.
- Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, *8*(12), 1704-1711.
- Deroost, N., Vandebossche, J., Zeischka, P., Coomans, D., & Soetens, E. (2012). Cognitive control: a role for implicit learning?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(5), 1243.
- Deserno, L., Huys, Q. J., Boehme, R., Buchert, R., Heinze, H. J., Grace, A. A., ... & Schlagenhauf, F. (2015). Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proceedings of the National Academy of Sciences*, *112*(5), 1595-1600.
- Devaine, M., & Daunizeau, J. (2017). Learning about and from others' prudence, impatience or laziness: The computational bases of attitude alignment. *PLoS computational biology*, *13*(3), e1005422.
- Devaine, M., Hollard, G., & Daunizeau, J. (2014). The social Bayesian brain: does mentalizing make a difference when we learn?. *PLoS computational biology*, *10*(12), e1003992.
- de Weerd, H., Diepgrond, D., & Verbrugge, R. (2017). Estimating the use of higher-order theory of mind using computational agents. *The BE Journal of Theoretical Economics*.
- Di Costa, S., Théro, H., Chambon, V., & Haggard, P. (2017). Try and try again: Post-error boost of an implicit measure of agency. *The Quarterly Journal of Experimental Psychology*, (just-accepted), 1-28.
- Di Paolo, E., & De Jaegher, H. (2012). The interactive brain hypothesis. *Frontiers in human neuroscience*, *6*.
- Diaconescu, A. O., Litvak, V., Mathys, C., Kasper, L., Friston, K. J., & Stephan, K. E. (2017). A computational hierarchy in human cortex. *arXiv preprint arXiv:1709.02323*.
- Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 67-78.
- Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, *22*(1), 1-18.
- Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312-325.



- Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature neuroscience*, 18(5), 767-772.
- Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current opinion in neurobiology*, 22(6), 1075-1081.
- Dollé, L., Sheynikhovich, D., Girard, B., Chavarriaga, R., & Guillot, A. (2010). Path planning versus cue responding: a bio-inspired model of switching between navigation strategies. *Biological cybernetics*, 103(4), 299-317.
- Donoso, M., Collins, A. G., & Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science*, 344(6191), 1481-1486.
- Doya, K., Samejima, K., Katagiri, K. I., & Kawato, M. (2002). Multiple model-based reinforcement learning. *Neural computation*, 14(6), 1347-1369.
- Du, Y., & Clark, J. E. (2017). New insights into statistical learning and chunk learning in implicit sequence acquisition. *Psychonomic bulletin & review*, 24(4), 1225-1233.
- Duffy, S., Naddeo, J. J., Owens, D., & Smith, J. (2016). Cognitive load and mixed strategies: On brains and minimax.
- Dunbar, R. I. (2002). The social brain hypothesis. *Foundations in social neuroscience*, 5(71), 69.
- Duverne, S., & Koechlin, E. (2017). Rewards and Cognitive Control in the Human Prefrontal Cortex. *Cerebral Cortex*, 27(10), 5024-5039.
- Dyson, B. J., Wilbiks, J. M. P., Sandhu, R., Papanicolaou, G., & Lintag, J. (2016). Negative outcomes evoke cyclic irrational decisions in Rock, Paper, Scissors. *Scientific reports*, 6.
- Eppinger, B., Walter, M., Heekeren, H. R., & Li, S. C. (2013). Of goals and habits: age-related and individual differences in goal-directed decision-making. *Frontiers in neuroscience*, 7.
- Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American economic review*, 848-881.
- Erev, I., & Roth, A. E. (2014). Maximization, learning, and economic behavior. *Proceedings of the National Academy of Sciences*, 111(Supplement 3), 10818-10825.
- Filipowicz, A., Anderson, B., & Danckert, J. (2014). Learning what from where: Effects of spatial regularity on nonspatial sequence learning and updating. *The Quarterly Journal of Experimental Psychology*, 67(7), 1447-1456.
- Filipowicz, A., Valadao, D., Anderson, B., & Danckert, J. (2016). Rejecting Outliers: Surprising Changes Do Not Always Improve Belief Updating.
- Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, 11(2), 77-83.

- FitzGerald, T. H., Hämmerer, D., Friston, K. J., Li, S. C., & Dolan, R. J. (2017). Sequential inference as a mode of cognition and its correlates in fronto-parietal and hippocampal brain regions. *PLoS computational biology*, *13*(5), e1005418.
- Fonollosa, J., Neftci, E., & Rabinovich, M. (2015). Learning of chunking sequences in cognition and behavior. *PLoS computational biology*, *11*(11), e1004592.
- Forstmann, B. U., Wagenmakers, E. J., Eichele, T., Brown, S., & Serences, J. T. (2011). Reciprocal relations between cognitive neuroscience and formal cognitive models: opposites attract?. *Trends in cognitive sciences*, *15*(6), 272-279.
- Frank, Michael J. "Linking across levels of computation in model-based cognitive neuroscience." *An introduction to model-based cognitive neuroscience*. Springer New York, 2015. 159-177.
- Frey, S., & Goldstone, R. L. (2013). Cyclic game dynamics driven by iterated reasoning. *PloS one*, *8*(2), e56416.
- Friedenberg, A., Kets, W., & Kneeland, T. (2016). Bounded Reasoning: Rationality or Cognition.
- Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual review of psychology*, *63*, 287-313.
- Frith, U., & Frith, C. (2010). The social brain: allowing humans to boldly go where no other species has been. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *365*(1537), 165-176.
- Fudenberg, D., & Levine, D. K. (1998). *The theory of learning in games* (Vol. 2). MIT press.
- Fudenberg, D., & Levine, D. K. (2009). Learning and equilibrium. *Annu. Rev. Econ.*, *1*(1), 385-420.
- Fumagalli, R. (2016). Economics, Psychology, and the Unity of the Decision Sciences. *Philosophy of the Social Sciences*, *46*(2), 103-128.
- Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition*, *109*(3), 416-422.
- Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, *2*(12), 493-501.
- Garvert, M. M., Moutoussis, M., Kurth-Nelson, Z., Behrens, T. E., & Dolan, R. J. (2015). Learning-induced plasticity in medial prefrontal cortex predicts preference malleability. *Neuron*, *85*(2), 418-428.
- Gauvrit, N., Zenil, H., Soler-Toscano, F., Delahaye, J. P., & Brugger, P. (2017). Human behavioral complexity peaks at age 25. *PLoS computational biology*, *13*(4), e1005408.

Geana, A., & Niv, Y. (2014). Causal model comparison shows that human representation learning is not Bayesian. In *Cold Spring Harbor symposia on quantitative biology* (Vol. 79, pp. 161-168). Cold Spring Harbor Laboratory Press.

Gershman SJ, Gerstenberg T, Baker CL, Cushman FA (2016) Plans, Habits, and Theory of Mind. *PLoS ONE*11(9): e0162246

Gershman, S. J. (2017). Dopamine, Inference, and Uncertainty. *bioRxiv*, 149849.

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current opinion in neurobiology*, 20(2), 251-256.

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, 143(1), 182.

Gershman, S. J., Zhou, J., & Komers, C. (2017). Imaginative Reinforcement Learning: Computational Principles and Neural Mechanisms. *Journal of Cognitive Neuroscience*.

Gęsiarz, F., & Crockett, M. J. (2015). Goal-directed, habitual and Pavlovian prosocial behavior. *Frontiers in behavioral neuroscience*, 9.

Gill, D., & Prowse, V. L. (2012). Cognitive ability and learning to play equilibrium: A level-k analysis. *Analysis*.

Gill, D., & Prowse, V. L. (2017). Using response times to measure strategic complexity and the value of thinking in games.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4), 585-595.

Glimcher, P. W., Dorris, M. C., & Bayer, H. M. (2005). Physiological utility theory and the neuroeconomics of choice. *Games and economic behavior*, 52(2), 213-256.

Goeree, J. K. & Holt, C. A. (2001). « Ten Little Treasures of Game Theory and Ten Intuitive Contradictions ». *American Economic Review*, vol. 91 no 5 : pp. 1402–1422.

Goldstein, M. H., Waterfall, H. R., Lotem, A., Halpern, J. Y., Schwade, J. A., Onnis, L., & Edelman, S. (2010). General cognitive principles for learning structure in time and space. *Trends in cognitive sciences*, 14(6), 249-258.

Gopnik, A., & Wellman, H. M. (1994). 10 The theory theory. *Mapping the mind: Domain specificity in cognition and culture*, 257.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, 138(6), 1085.

- Green, C. S., Benson, C., Kersten, D., & Schrater, P. (2010). Alterations in choice behavior by manipulations of world model. *Proceedings of the national academy of sciences*, *107*(37), 16401-16406.
- Griessinger, T., & Coricelli, G. (2015). The neuroeconomics of strategic interaction. *Current Opinion in Behavioral Sciences*, *3*, 73-79.
- Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): comment on Bowers and Davis (2012).
- Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nature Neuroscience*, *18*(9), 1233-1235.
- Haggard, P., & Tsakiris, M. (2009). The experience of agency: Feelings, judgments, and responsibility. *Current Directions in Psychological Science*, *18*(4), 242-246.
- Hahn, U., & Warren, P. A. (2009). Perceptions of randomness: why three heads are better than four. *Psychological review*, *116*(2), 454.
- Hampton, A. N., Bossaerts, P., & O'doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience*, *26*(32), 8360-8367.
- Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences*, *105*(18), 6741-6746.
- Hanaki, N., Kirman, A., & Pezanis-Christou, P. (2017). Observational and Reinforcement Pattern-Learning: An Exploratory Study.
- Hare, T. A., Camerer, C. F., Knoepfle, D. T., O'Doherty, J. P., & Rangel, A. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *Journal of Neuroscience*, *30*(2), 583-590.
- Hari, R., Henriksson, L., Malinen, S., & Parkkonen, L. (2015). Centrality of social interaction in human brain function. *Neuron*, *88*(1), 181-193.
- Hertz, U., Palminteri, S., Brunetti, S., Olesen, C., Frith, C., & Bahrami, B. (2017). Neural Computations Underpinning The Strategic Management Of Influence In Advice Giving. bioRxiv, 121947.
- Heyes, C. M., & Galef Jr, B. G. (Eds.). (1996). *Social learning in animals: the roots of culture*. Elsevier.
- Hill, C. A., Suzuki, S., Polania, R., Moisa, M., O'Doherty, J. P., & Ruff, C. C. (2017). A causal account of the brain network computations underlying strategic social behavior. *Nature Neuroscience*.
- Ho, T. H., Camerer, C. F., & Chong, J. K. (2007). Self-tuning experience weighted attraction learning in games. *Journal of Economic Theory*, *133*(1), 177-198.

- Ho, T. H., Camerer, C., & Weigelt, K. (1998). Iterated dominance and iterated best response in experimental "p-beauty contests". *The American Economic Review*, 88(4), 947-969.
- Hsu, C. T., Sims, T., & Chakrabarti, B. (2017). How mimicry influences the neural correlates of reward: An fMRI study. *Neuropsychologia*.
- Hsu, M., & Zhu, L. (2012). Learning in games: neural computations underlying strategic learning. *Recherches économiques de Louvain*, 78(3), 47-72.
- Hughes, B. L., Zaki, J., & Ambady, N. (2017). Motivation alters impression formation and related neural systems. *Social cognitive and affective neuroscience*, 12(1), 49-60
- Humphries, M. D., Khamassi, M., & Gurney, K. (2012). Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Frontiers in neuroscience*, 6.
- Ito, M., & Doya, K. (2009). Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *Journal of Neuroscience*, 29(31), 9861-9874.
- Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: computational principles underlying commonsense psychology. *Trends in cognitive sciences*, 20(8), 589-604.
- Jenkins, A. C., & Mitchell, J. P. (2011). Medial prefrontal cortex subserves diverse forms of self-reflection. *Social neuroscience*, 6(3), 211-218.
- Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, 168, 46-64
- Jiménez, L. (2008). Taking patterns for chunks: is there any evidence of chunk learning in continuous serial reaction-time tasks?. *Psychological Research*, 72(4), 387-396.
- Jocham, G., Brodersen, K. H., Constantinescu, A. O., Kahn, M. C., Ianni, A. M., Walton, M. E., ... & Behrens, T. E. (2016). Reward-guided learning with and without causal attribution. *Neuron*, 90(1), 177-190.
- Joiner, J., Piva, M., Turrin, C., & Chang, S. W. (2017). Social learning through prediction error in the brain. *npj Science of Learning*, 2(1), 8.
- Jones, F. W., & McLaren, I. P. (2009). Human sequence learning under incidental and intentional conditions. *Journal of Experimental Psychology: Animal Behavior Processes*, 35(4), 538.
- Jones, M., & 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(4), 169-188.

Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: understanding the relations between judgments of competence and warmth. *Journal of personality and social psychology*, 89(6), 899.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the econometric society*, 263-291.

Kanske, P., Böckler, A., Trautwein, F. M., Parianen Lesemann, F. H., & Singer, T. (2016). Are strong empathizers better mentalizers? Evidence for independence and interaction between the routes of social cognition. *Social cognitive and affective neuroscience*, 11(9), 1383-1392.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS computational biology*, 7(5), e1002055.

Khamassi, M., Girard, B., Clodic, A., Devin, S., Renaudo, E., Pacherie, E., Alami, R. and Chatila, R. (2016). Integration of Action, Joint Action and Learning in Robot Cognitive Architectures. *Intellectica*. Vol 2016/1 No 65 Pages 169-203.

Khamassi, M., & Humphries, M. D. (2012). Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies. *Frontiers in behavioral neuroscience*, 6.

Khamassi, M., Enel, P., Dominey, P. F., & Procyk, E. (2013). Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. *Prog Brain Res*, 202, 441-464.

Khamassi, M., Quilodran, R., Enel, P., Dominey, P. F., & Procyk, E. (2014). Behavioral regulation and the modulation of information coding in the lateral prefrontal and cingulate cortex. *Cerebral cortex*, 25(9), 3197-3218.

Khamassi, M., Wilson, C., Rothé, R., Quilodran, R., Dominey, P. F., & Procyk, E. (2011). Meta-learning, cognitive control, and physiological interactions between medial and lateral prefrontal cortex. *Neural basis of motivational and cognitive control*, 351-370.

Killcross, S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral cortex*, 13(4), 400-408.

Kimura, M., Schröger, E., Czigler, I., & Ohira, H. (2010). Human visual system automatically encodes sequential regularities of discrete events. *Journal of Cognitive Neuroscience*, 22(6), 1124-1139.

Koechlin, E. (2016). Prefrontal executive function and adaptive behavior in complex environments. *Current opinion in neurobiology*, 37, 1-6.

Konovalov, A., & Krajbich, I. (2016). Over a decade of neuroeconomics: What have we learned?. *Organizational Research Methods*, 1094428116644502.

Kool, W., Cushman, F. A.\*, & Gershman, S. J.\* (in press). Competition and cooperation between multiple reinforcement learning systems. In R. W. Morris, A. M. Bornstein, & A. Shenhav (Eds.), *Understanding Goal-Directed Decision Making: Computations and Circuits*. Elsevier.

- Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science*, 28(9), 1321-1333.
- Koster-Hale, J., & Saxe, R. (2013). Theory of mind: a neural prediction problem. *Neuron*, 79(5), 836-848.
- Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences*, 110(14), 5648-5653.
- Krakauer, J. W., Ghazanfar, A. A., Gomez-Marín, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist Bias. *Neuron*, 93(3), 480-490.
- Kuzmanovic, B., & Rigoux, L. (2017). Valence-Dependent Belief Updating: Computational Validation. *Frontiers in Psychology*, 8
- Lahav, Y. (2009). Behavioral pattern learning models for decision making in games. *Journal of Pattern Recognition Research*, 4(1), 133-151.
- Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: An integrative review.
- Lebreton, M., Jorge, S., Michel, V., Thirion, B., & Pessiglione, M. (2009). An automatic valuation system in the human brain: evidence from functional neuroimaging. *Neuron*, 64(3), 431-439.
- Lebreton, M., Kawa, S., d'Arc, B. F., Daunizeau, J., & Pessiglione, M. (2012). Your goal is mine: unraveling mimetic desires in the human brain. *Journal of Neuroscience*, 32(21), 7146-7157.
- Lee, D. (2008). Game theory and neural basis of social decision making. *Nature neuroscience*, 11(4), 404.
- Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3), 687-699.
- Lee, V. K., & Harris, L. T. (2013). How social cognition can inform social decision making. *Frontiers in neuroscience*, 7.
- Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, 1, 0067.
- Legare, C. H., Sobel, D. M., & Callanan, M. (2017). Causal learning is collaborative: Examining explanation and exploration in social contexts. *Psychonomic Bulletin & Review*, 1-7.
- Levy, D. J., & Glimcher, P. W. (2012). The root of all value: a neural common currency for choice. *Current opinion in neurobiology*, 22(6), 1027-1038.
- Ligneul, R., Obeso, I., Ruff, C. C., & Dreher, J. C. (2016). Dynamical Representation of Dominance Relationships in the Human Rostromedial Prefrontal Cortex. *Current Biology*, 26(23), 3107-3115.

- Loffing, F., Stern, R., & Hagemann, N. (2015). Pattern-induced expectation bias in visual anticipation of action outcomes. *Acta psychologica*, 161, 45-53.
- Luce, R. D. (1977). The choice axiom after twenty years. *Journal of mathematical psychology*, 15(3), 215-233.
- Ma, N., Baetens, K., Vandekerckhove, M., Kestemont, J., Fias, W., & Van Overwalle, F. (2013). Traits are represented in the medial prefrontal cortex: an fMRI adaptation study. *Social Cognitive and Affective Neuroscience*, 9(8), 1185-1192.
- Ma, N., Wang, S., Yang, Q., Feng, T., & Van Overwalle, F. (2016). The neural representation of competence traits: An fMRI study. *Scientific reports*, 6, 39609.
- Mahy, C. E., Moses, L. J., & Pfeifer, J. H. (2014). How and where: Theory-of-mind in the brain. *Developmental cognitive neuroscience*, 9, 68-81.
- Mansouri, F. A., Koechlin, E., Rosa, M. G., & Buckley, M. J. (2017). Managing competing goals—a key role for the frontopolar cortex. *Nature reviews. Neuroscience*.
- McKelvey, R. D., & Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1), 6-38.
- McNamee, D., Rangel, A., & O'doherty, J. P. (2013). Category-dependent and category-independent goal-value codes in human ventromedial prefrontal cortex. *Nature neuroscience*, 16(4), 479-485.
- Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2012). The neural dynamics of updating person impressions. *Social cognitive and affective neuroscience*, 8(6), 623-631.
- Meyniel, F., Maheu, M., & Dehaene, S. (2016). Human inferences about sequences: A minimal transition probability model. *PLoS computational biology*, 12(12), e1005260.
- Mill, J. S. (1901). *Utilitarianism*. Longmans, Green and Company.
- Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, 50(4), 655-663.
- Mobbs, D., Yu, R., Meyer, M., Passamonti, L., Seymour, B., Calder, A. J., ... & Dalgleish, T. (2009). A key role for similarity in vicarious reward. *Science*, 324(5929), 900-900.
- Molenberghs, P., Johnson, H., Henry, J. D., & Mattingley, J. B. (2016). Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews*, 65, 276-291.
- Morris, R. W., Dezfouli, A., Griffiths, K. R., Le Pelley, M. E., & Balleine, B. W. (2017). The algorithmic neuroanatomy of action-outcome learning. *bioRxiv*, 137851.
- Nagel, R. (1995). Unraveling in guessing games: An experimental study. *The American Economic Review*, 85(5), 1313-1326.



- Nakahara, H., & Hikosaka, O. (2012). Learning to represent reward structure: A key to adapting to complex environments. *Neuroscience research*, 74(3), 177-183.
- Nicolle, A., Klein-Flügge, M. C., Hunt, L. T., Vlaev, I., Dolan, R. J., & Behrens, T. E. (2012). An agent independent axis for executed and modeled choice in medial prefrontal cortex. *Neuron*, 75(6), 1114-1121.
- Nicolle, A., Symmonds, M., & Dolan, R. J. (2011). Optimistic biases in observational learning of value. *Cognition*, 119(3), 394-402.
- Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, 35(21), 8145-8157.
- Nyarko, Y. et Schotter, A. (2002). « An Experimental Study of Belief Learning Using Elicited Beliefs ». *Econometrica*, vol. 70 no 3 : pp. 971–1005.
- O'Doherty, J. P. (2014). The problem with value. *Neuroscience & Biobehavioral Reviews*, 43, 259-268.
- O'Doherty, J. P., Lee, S. W., & McNamee, D. (2015). The structure of reinforcement-learning mechanisms in the human brain. *Current Opinion in Behavioral Sciences*, 1, 94-100.
- O'Reilly, J. X., Jbabdi, S., & Behrens, T. E. (2012). How can a Bayesian approach inform neuroscience?. *European Journal of Neuroscience*, 35(7), 1169-1179.
- O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, reward, and decision making. *Annual review of psychology*, 68, 73-100.
- O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of sciences*, 1104(1), 35-53.
- Oskarsson, A. T., Van Boven, L., McClelland, G. H., & Hastie, R. (2009). What's next? Judging sequences of binary events. *Psychological bulletin*, 135(2), 262.
- Otto, A. R., Skatova, A., Madlon-Kay, S., & Daw, N. D. (2014). Cognitive control predicts use of model-based reinforcement learning. *Journal of cognitive neuroscience*.
- Pacherie, E. (2012). The phenomenology of joint action: Self-agency vs. joint-agency. In A. Seemann (Ed.), *Joint attention: New developments* (pp. 343–389). Cambridge, MA: MIT Press .
- Padoa-Schioppa, C. (2008). The syllogism of neuro-economics. *Economics & Philosophy*, 24(3), 449-457.
- Padoa-Schioppa, C. (2011). Neurobiology of economic choice: a good-based model. *Annual review of neuroscience*, 34, 333-359.
- Padoa-Schioppa, C. (2013). Neuronal origins of choice variability in economic decisions. *Neuron*, 80(5), 1322-1336.

- Padoa-Schioppa, C., & Schoenbaum, G. (2015). Dialogue on economic choice, learning theory, and neuronal representations. *Current opinion in behavioral sciences*, 5, 16-23.
- Palminteri S, Lefebvre G, Kilford EJ, Blakemore S-J (2017) Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS Computational Biology* 13(8): e1005684.
- Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature communications*, 6.
- Palminteri, S., Wyart, V., & Koechlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*.
- Patel, D., Fleming, S. M., & Kilner, J. M. (2012, December). Inferring subjective states through the observation of actions. In *Proc. R. Soc. B* (Vol. 279, No. 1748, pp. 4853-4860). The Royal Society.
- Pezzulo, G., Donnarumma, F., & Dindo, H. (2013). Human sensorimotor communication: A theory of signaling in online social interactions. *PloS one*, 8(11), e79876.
- Pezzulo, G., Rigoli, F., & Chersi, F. (2013). The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. *Frontiers in psychology*, 4.
- Polonio, L., & Coricelli, G. (2015). Testing the level of consistency between choices and beliefs in games using eye-tracking. *International Journal of Game Theory*, 1-40.
- Polonio, L., Di Guida, S., & Coricelli, G. (2015). Strategic sophistication and attention in games: an eye-tracking study. *Games and Economic Behavior*, 94, 80-96.
- Polonioli, A. (2016). Reconsidering the normative argument from bounded rationality. *Theory & Psychology*, 26(3), 287-303.
- Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature neuroscience*, 16(9), 1170-1178.
- Preciado, D., Munneke, J., & Theeuwes, J. (2017). Mixed signals: The effect of conflicting reward-and goal-driven biases on selective attention. *Attention, Perception, & Psychophysics*, 1-14.
- Przyrembel, M., Smallwood, J., Pauen, M., & Singer, T. (2012). Illuminating the dark matter of social neuroscience: considering the problem of social interaction from philosophical, psychological, and neuroscientific perspectives. *Frontiers in Human Neuroscience*, 6.
- Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, 9(7), 545-556.
- Reber, J., Feinstein, J. S., O'Doherty, J. P., Liljeholm, M., Adolphs, R., & Tranel, D. (2017). Selective impairment of goal-directed decision-making following lesions to the human ventromedial prefrontal cortex. *Brain*, 140(6), 1743-1756.

- Redish, A. D. (2016). Vicarious trial and error. *Nature reviews. Neuroscience*, 17(3), 147.
- Rescorla, R. A. (1972). A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement. *Classical conditioning II: Current research and theory*.
- Rilling, J. K., & Sanfey, A. G. (2011). The neuroscience of social decision-making. *Annual review of psychology*, 62, 23-48.
- Roth, A. E., & Erev, I. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and economic behavior*, 8(1), 164-212.
- Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, 15(8), 549-562.
- Rushworth, M. F., Mars, R. B., & Sallet, J. (2013). Are there specialized circuits for social cognition and are they unique to humans?. *Current opinion in neurobiology*, 23(3), 436-442.
- Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, 5(17), 61-71.
- Santiesteban, I., Banissy, M. J., Catmur, C., & Bird, G. (2012). Enhancing social ability by stimulating right temporoparietal junction. *Current Biology*, 22(23), 2274-2277.
- Savage, L. 1954. *The Foundations of Statistics*. New York: John Wiley.
- Saxe, R., & Houlihan, S. D. (2017). Formalizing emotion concepts within a Bayesian model of theory of mind. *Current Opinion in Psychology*, 17, 15-21.
- Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in cognitive sciences*, 19(2), 65-72.
- Schilbach, L. (2014). On the relationship of online and offline social cognition. *Frontiers in human neuroscience*, 8.
- Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). A second-person neuroscience in interaction. *Behavioral and brain sciences*, 36(4), 441-462.
- Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human orbitofrontal cortex represents a cognitive map of state space. *Neuron*, 91(6), 1402-1412.
- Schuck, N. W., Gaschler, R., Wenke, D., Heinzle, J., Frensch, P. A., Haynes, J. D., & Reverberi, C. (2015). Medial prefrontal cortex predicts internally driven strategy shifts. *Neuron*, 86(1), 331-340.
- Schultz, W. (2015). Neuronal reward and decision signals: from theories to data. *Physiological Reviews*, 95(3), 853-951.
- Schultz, W. (2016). Dopamine reward prediction-error signalling: a two-component response. *Nature reviews. Neuroscience*, 17(3), 183.

- Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, 42, 9-34.
- Schurr, F., Tam, B. P., & Maloney, L. T. (2013, January). Learning Patterns in Noise: Environmental Statistics explain the Sequential Effect. In *Proceedings of the Cognitive Science Society* (Vol. 35, No. 35).
- Schwarb, H., & Schumacher, E. H. (2012). Generalized lessons about sequence learning from the study of the serial reaction time task. *Advances in cognitive psychology*, 8(2), 165.
- Seid-Fatemi, A., & Tobler, P. N. (2014). Efficient learning mechanisms hold in the social domain and are implemented in the medial prefrontal cortex. *Social cognitive and affective neuroscience*, 10(5), 735-743.
- Seip, K. L., & Grøn, Ø. (2016). Leading the game, losing the competition: identifying leaders and followers in a repeated game. *PloS one*, 11(3), e0150398.
- Seo, H., Cai, X., Donahue, C. H., & Lee, D. (2014). Neural correlates of strategic reasoning during competitive games. *Science*, 346(6207), 340-343.
- Seo, H., Kim, S., Cai, X., Abe, H., Donahue, C. H., & Lee, D. (2017). Neural Correlates of Strategic Decision-Making in the Primate Prefrontal Cortex. In *The Prefrontal Cortex as an Executive, Emotional, and Social Brain* (pp. 3-15). Springer Japan.
- Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, (0).
- Shteingart, H., & Loewenstein, Y. (2014). Reinforcement learning and human behavior. *Current Opinion in Neurobiology*, 25, 93-98.
- Schuck, N. W., Gaschler, R., Wenke, D., Heinzle, J., Frensch, P. A., Haynes, J. D., & Reverberi, C. (2015). Medial prefrontal cortex predicts internally driven strategy shifts. *Neuron*, 86(1), 331-340.
- Smith, D. V., Clithero, J. A., Boltuck, S. E., & Huettel, S. A. (2014). Functional connectivity with ventromedial prefrontal cortex reflects subjective value for social rewards. *Social cognitive and affective neuroscience*, 9(12), 2017-2025.
- Sonsino, D. (1997). Learning to learn, pattern recognition, and Nash equilibrium. *Games and Economic Behavior*, 18(2), 286-331.
- Sonsino, D., & Sirota, J. (2003). Strategic pattern recognition—experimental evidence. *Games and Economic Behavior*, 44(2), 390-411.
- Spiliopoulos, L. (2012). Pattern recognition and subjective belief learning in a repeated constant-sum game. *Games and economic behavior*, 75(2), 921-935.
- Spiliopoulos, L. (2013). Beyond fictitious play beliefs: Incorporating pattern recognition and similarity matching. *Games and Economic Behavior*, 81, 69-85.

- Spiliopoulos, L. (2013). Strategic adaptation of humans playing computer algorithms in a repeated constant-sum game. *Autonomous agents and multi-agent systems*, 1-30.
- Spiliopoulos, L. (2016). The determinants of response time in a repeated constant-sum game: A robust Bayesian hierarchical model. *Browser Download This Paper*.
- Spunt, R. P., & Adolphs, R. (2014). Validating the why/how contrast for functional MRI studies of theory of mind. *Neuroimage*, 99, 301-311.
- Spunt, R. P., & Adolphs, R. (2017). A new look at domain specificity: insights from social neuroscience. *Nature Reviews Neuroscience*, 18(9), 559-567.
- Stahl, D. O., & Wilson, P. W. (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1), 218-254.
- Stalnaker, T. A., Cooch, N. K., & Schoenbaum, G. (2015). What the orbitofrontal cortex does not do. *Nature neuroscience*, 18(5), 620-627.
- Stanley, D. A. (2015). Getting to know you: general and specific neural computations for learning about people. *Social cognitive and affective neuroscience*, 11(4), 525-536.
- Stanley, D. A., & Adolphs, R. (2013). Toward a neural basis for social behavior. *Neuron*, 80(3), 816-826.
- Starkweather, C. K., Babayan, B. M., Uchida, N., & Gershman, S. J. (2017). Dopamine reward prediction errors reflect hidden state inference across time. *Nature neuroscience*, 20(4), 581.
- Stefanics, G., Kimura, M., & Czigler, I. (2011). Visual mismatch negativity reveals automatic detection of sequential regularity violation. *Frontiers in human neuroscience*, 5.
- Stöttinger, E., Filipowicz, A., Danckert, J., & Anderson, B. (2014). The effects of prior learned strategies on updating an opponent's strategy in the rock, paper, scissors game. *Cognitive science*, 38(7), 1482-1492.
- Strombach, T., Weber, B., Hangebrauk, Z., Kenning, P., Karipidis, I. I., Tobler, P. N., & Kalenscher, T. (2015). Social discounting involves modulation of neural value signals by temporoparietal junction. *Proceedings of the National Academy of Sciences*, 112(5), 1619-1624.
- Sul, S., Tobler, P. N., Hein, G., Leiberg, S., Jung, D., Fehr, E., & Kim, H. (2015). Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality. *Proceedings of the National Academy of Sciences*, 112(25), 7851-7856.
- Sun, Y., O'Reilly, R. C., Bhattacharyya, R., Smith, J. W., Liu, X., & Wang, H. (2015). Latent structure in random sequences drives neural learning toward a rational bias. *Proceedings of the National Academy of Sciences*, 112(12), 3788-3792.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1, No. 1). Cambridge: MIT press.

- Suzuki, S., Adachi, R., Dunne, S., Bossaerts, P., & O'Doherty, J. P. (2015). Neural mechanisms underlying human consensus decision-making. *Neuron*, *86*(2), 591-602.
- Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., Ichinohe, N., Haruno, M., ... & Nakahara, H. (2012). Learning to simulate others' decisions. *Neuron*, *74*(6), 1125-1137
- Suzuki, S., Jensen, E. L., Bossaerts, P., & O'Doherty, J. P. (2016). Behavioral contagion during learning about another agent's risk-preferences acts on the neural representation of decision-risk. *Proceedings of the National Academy of Sciences*, *113*(14), 3755-3760.
- Tamir, D. I., Thornton, M. A., Contreras, J. M., & Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences*, *113*(1), 194-199.
- Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. *Psychological Review*, *124*(4), 410.
- Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279-1285.
- Ting, C. C., Yu, C. C., Maloney, L. T., & Wu, S. W. (2015). Neural mechanisms for integrating prior knowledge and likelihood in value-based probabilistic inference. *Journal of Neuroscience*, *35*(4), 1792-1805.
- Tobler, P. N., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2006). Human neural learning depends on reward prediction errors in the blocking paradigm. *Journal of Neurophysiology*, *95*(1), 301-310.
- Tusche, A., Böckler, A., Kanske, P., Trautwein, F. M., & Singer, T. (2016). Decoding the charitable brain: empathy, perspective taking, and attention shifts differentially predict altruistic giving. *Journal of Neuroscience*, *36*(17), 4719-4732.
- Unsworth, N., & Engle, R. W. (2005). Individual differences in working memory capacity and learning: Evidence from the serial reaction time task. *Memory & cognition*, *33*(2), 213-220.
- Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *Neuroimage*, *48*(3), 564-584.
- Velez-Ginorio, J., Siegel, M., Tenenbaum, J. B., & Jara-Ettinger, J. Interpreting actions by attributing compositional desires.
- Vickery, T. J., Kleinman, M. R., Chun, M. M., & Lee, D. (2015). Opponent identity influences value learning in simple games. *Journal of Neuroscience*, *35*(31), 11133-11143.
- Viejo, G., Khamassi, M., Brovelli, A., & Girard, B. (2015). Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning. *Frontiers in behavioral neuroscience*, *9*.

- Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior*, 2nd rev.
- Vostroknutov, A., Polonio, L., & Coricelli, G. (2017). *Observational Learning and Intelligence*. Working paper
- Wake, S. J., & Izuma, K. (2017). A common neural code for social and monetary rewards in the human striatum. *Social Cognitive and Affective Neuroscience*, *12*(10), 1558-1564.
- Wan, X., Cheng, K., & Tanaka, K. (2015). Neural encoding of opposing strategy values in anterior and posterior cingulate cortex. *Nature neuroscience*, *18*(5), 752-759.
- Wang, R., Shen, Y., Tino, P., Welchman, A. E., & Kourtzi, Z. (2017). Learning predictive statistics: strategies and brain mechanisms. *Journal of Neuroscience*, *37*(35), 8412-8427.
- Wang, Z., Xu, B., & Zhou, H. J. (2014). Social cycling and conditional responses in the Rock-Paper-Scissors game. *Scientific reports*, *4*, 5830.
- Wallin, A., Swait, J., & Marley, A. A. J. (2017). Not Just Noise: A Goal Pursuit Interpretation of Stochastic Choice.
- Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*(3-4), 279-292.
- Wikenheiser, A. M., & Schoenbaum, G. (2016). Over the river, through the woods: cognitive maps in the hippocampus and orbitofrontal cortex. *Nature reviews. Neuroscience*, *17*(8), 513.
- Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., & Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron*, *81*(2), 267-279.
- Wittmann, Marco K., et al. "Self-other mergence in the frontal cortex during cooperation and competition." *Neuron* 91.2 (2016): 482-493.
- Wyart, V., & Koechlin, E. (2016). Choice variability and suboptimality in uncertain environments. *Current Opinion in Behavioral Sciences*, *11*, 109-115.
- Xue, G., He, Q., Lu, Z. L., Levin, I. P., Dong, Q., & Bechara, A. (2013). Agency modulates the lateral and medial prefrontal cortex responses in belief-based decision making. *PloS one*, *8*(6), e65274.
- Yu, A. J., & Cohen, J. D. (2009). Sequential effects: superstition or rational behavior?. In *Advances in neural information processing systems* (pp. 1873-1880).
- Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, *22*(2), 513-523.
- Yoshida, W., Seymour, B., Friston, K. J., & Dolan, R. J. (2010). Neural mechanisms of belief inference during cooperative games. *Journal of Neuroscience*, *30*(32), 10744-10751.

Zaki, J., & Ochsner, K. (2011). Reintegrating the study of accuracy into social cognition research. *Psychological Inquiry*, 22(3), 159-182.

Zaki, J., Schirmer, J., & Mitchell, J. P. (2011). Social influence modulates the neural computation of value. *Psychological science*, 22(7), 894-900.

Zednik, C., & Jäkel, F. (2016). Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese*, 193(12), 3951-3985.

Zhu, L., Mathewson, K. E., & Hsu, M. (2012). Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning. *Proceedings of the National Academy of Sciences*, 109(5), 1419-1424.

Zsuga, J., Biro, K., Papp, C., Tajti, G., & Gesztelyi, R. (2016). The “proactive” model of learning: Integrative framework for model-free and model-based reinforcement learning utilizing the associative learning-based proactive brain concept. *Behavioral neuroscience*, 130(1), 6.



## Résumé

Les interactions sociales humaines reposent sur notre capacité à apprendre et à ajuster nos comportements en réponse à ceux d'autrui. Les jeux stratégiques offrent un cadre pertinent pour étudier les processus cognitifs sous-tendant la représentation des intentions d'autrui, ainsi que les actions adaptées par lesquelles ces intentions se traduisent. Ces dernières décennies, le champ de l'économie comportementale a montré que les comportements humains dévient quasi systématiquement des prescriptions d'optimalité (équilibre) formulées par la théorie des jeux. Sur la base de récentes avancées en sciences cognitives, nous avons proposé que l'étude des sources de variation comportementale entre les individus pourrait fournir des informations cruciales à notre compréhension des limites de l'apprentissage social humain, et nous permettre de mieux comprendre cette non-convergence vers des interactions mutuellement bénéfiques. Dans ce travail de thèse, nous avons combiné des outils computationnels issus des neurosciences cognitives au cadre formel de l'économie comportementale dans le but d'étudier la façon dont les humains diffèrent dans leur compréhension du comportement d'autrui au cours d'interactions stratégiques compétitives. Dans un premier temps, nous avons abordé la question de l'interaction entre l'environnement de jeu et l'hétérogénéité de l'apprentissage stratégique. Nos résultats ont montré que, lors d'une interaction compétitive répétée, la structure (règle) du jeu peut influencer le niveau d'engagement dans un mode d'apprentissage stratégique sophistiqué, et expliquer les déviations par rapport à l'équilibre. Nos données suggèrent que les participants occupant une position désavantagée dans l'interaction stratégique sont contraints par la sophistication de leur apprentissage. Leurs opposants, bien qu'avantagés, doivent tout de même s'engager dans un apprentissage stratégique sophistiqué pour adapter leur comportement et maximiser leurs gains. Cette étude a ainsi révélé pour la première fois l'impact des différences interindividuelles dans l'apprentissage stratégique sur les déviations des décisions par rapport à l'optimalité, et éclaire les processus responsables de l'émergence de dynamiques de choix leader-follower. De plus, nos résultats suggèrent qu'une analyse coût-bénéfice pourrait sous-tendre l'engagement des joueurs stratégiques dans des processus d'apprentissage plus sophistiqués. Dans un second temps, nous avons testé l'hypothèse selon laquelle la profondeur (le niveau de sophistication) de l'apprentissage stratégique n'est pas le seul facteur permettant la compréhension des intentions d'autrui au cours d'une interaction stratégique, mais que cette compréhension repose également sur la capacité à détecter et exploiter des patterns dans son comportement. Nous avons observé que les participants étaient capables de détecter des régularités statistiques dans le comportement de l'opposant, mais également que cette aptitude n'était pas corrélée à un engagement plus faible dans un apprentissage stratégique sophistiqué, suggérant ainsi que les humains peuvent combiner des informations provenant de deux types d'apprentissage pour améliorer la précision de leurs croyances vis-à-vis d'autrui au cours de prises de décision sociales.

**Mots Clés** Apprentissage humain, Interactions sociales, Jeux stratégiques, Économie comportementale, Neurosciences computationnelles, Neuroéconomie

## Abstract

Social interactions rely on our ability to learn and adjust on the spot to the other's behavior. Strategic games provide a useful framework to study the cognitive processes involved in the representation of the other's intentions and their translation into the most adapted actions. In the last decades, the growing field of behavioral economics provided evidence of a systematic departure of human's behavior from the optimal prescription formulated by game theory. Based on recent advances in cognitive sciences, we hypothesized that characterizing the source of heterogeneity in behavior might provide key insights to understand the boundaries over human social learning, and therefore deviation from mutually beneficial interactions. We first address the question of the interplay between the game environment and the heterogeneity in formation of high-order beliefs over the opponent's behavior through strategic learning. We show that in a competitive repeated interaction, the payoff structure of the underlying game can influence the engagement in strategically sophisticated learning and explain deviation from game optimality (equilibrium). Our data suggest that participants in a disadvantaged role are constraints in their learning sophistication, and thus in the overcoming of their position, by their own cognitive capacities. Their opponents, albeit advantaged, still need to engage in strategically sophisticated learning but to follow and adjust their behavior in order to maximize their earnings. This study provides the first evidence of the key implication of strategic learning heterogeneity in equilibrium departure and provide insight to explain the emergence of a leader-follower dynamics of choice. In addition our results suggest that a cost-benefit analysis might drive the engagement of strategic players in a more sophisticated learning process. In a second step, we investigated the hypothesis that the depth of strategic learning is not the only factor in play to grasp the other's mind during competitive interaction, but that the capacity to detect and exploit patterns in her behavior is also important. We found that not only subjects were able to detect patterns in the opponent's behavior, but that the capacity to do so was not correlated to a lower engagement in sophisticated strategic learning, therefore suggesting that humans can combine information from both types of learning to improve belief accuracy during social decision making.

## Keywords

Human learning, Social interactions, Strategic games, Behavioral economics, Computational neuroscience, Neuroeconomics