UNIVERSITÉ DE GENÈVE

FACULTÉ DE PSYCHOLOGIE
ET DES SCIENCES DE L'ÉDUCATION

DOCTORAT EN NEUROSCIENCES
des Universités de Genève
et de Lausanne

UNIVERSITÉ DE GENÈVE

UNIL | Université de Lausanne

UNIVERSITÉ DE GENÈVE

FACULTÉ DE PSYCHOLOGIE
ET DES SCIENCES DE L'ÉDUCATION

Professeur Theodor Landis, directeur de thèse
Docteure Sara Gonzalez Andino, co-directrice de thèse

DECISION-MAKING IN ECONOMIC GAMES: NEURAL UNDERPINNINGS OF RATIONALITY DEVIATIONS AND INTER-INDIVIDUAL DIFFERENCES

THÈSE

Présentée à la
Faculté de Psychologie et des Sciences de l'Éducation

de l'Université de Genève

pour obtenir le grade de
Docteure en Neurosciences

par

Hélène TZIEROPOULOS ÖSTERLÖF

de Grèce, Suède et Lausanne (VD)

Thèse N° 50

2010

UNIVERSITÉ
DE GENÈVE

FACULTÉ DE PSYCHOLOGIE
ET DES SCIENCES DE L'ÉDUCATION

DOCTORAT EN NEUROSCIENCES
des Universités de Genève et de Lausanne

Thèse de HELENE TZIEROPOULOS OSTERLOF

Intitulée : Decision-making in economic games : Neural underpinnings of rationality deviations and inter-individual differences

*La Faculté de psychologie et des sciences de l'éducation, sur préavis du jury de thèse formé par

Prof. Théodore Landis, Directeur de thèse, Service de neurologie de l'hôpital universitaire de Genève
Dr Sara Gonzalez Andino, co-directrice de thèse, faculté de médecine, Université de Genève
Prof. Peter Bossaerts, California Institute of Technology, Pasadena, Californie
Prof. Claude-Alain Hauert, FPSE, Université de Genève
Prof. Alessandro Villa, faculté des Hautes Etudes Commerciales, Université de Lausanne

autorise l'impression de la présente thèse, sans prétendre par là émettre d'opinion sur les propositions qui y sont énoncées.

Genève, le 22 janvier 2010

Le Doyen :

Bernard Schneuwly

Thèse No 50
No d'immatriculation 01-330.083

N.B. La thèse imprimée doit porter la déclaration précédente * et remplir les conditions énumérées dans les « Informations aux étudiants relatives aux thèses de doctorat à l'Université de Genève ».

Remerciements

Oh, I get by with a little help from my friends
Yes, I'm gonna try with a little help from my friends[*]

Mes remerciements s'adressent en premier lieu à Sara Gonzalez Andino qui m'a dès le premier jour soutenue, peu importe les circonstances et les choix que j'ai dû faire. J'ai eu beaucoup de plaisir à travailler avec quelqu'un d'à la fois aussi brillant qu'humain, et je la remercie de m'avoir accueillie au sein de son laboratoire pour ces années remplies de discussions portant autant sur la solution inverse que sur des aspects beaucoup plus banals de la vie: la recherche n'est définitivement pas une usine à chorizo. Rolando Grave de Peralta a toujours été présent pour me guider dans des choix plus «statistiques», mais aussi pour m'expliquer avec une patience infinie des concepts ardus pour une psychologue – son usage de la métaphore ayant toujours facilité le message.

Cette thèse n'aurait peut-être pas non plus vu le jour sans la rencontre avec le Professeur Peter Bossaerts à Champéry: son intérêt et ses conseils ont été autant d'éléments qui ont rendu possible la suite de ce travail. Je tiens également à remercier les membres de ma commission de thèse, le Dr Claire Bindschaedler, le Professeur Patrik Vuilleumier et tout particulièrement le Professeur Claude-Alain Hauert qui m'a suivie depuis longtemps et toujours avec intérêt. Je remercie également le Professeur Alessandro Villa d'avoir accepté de faire partie du jury, et finalement, je tiens à remercier sincèrement mon directeur de thèse le Professeur Theodor Landis, qui a montré son soutien à plusieurs reprises durant ces dernières années. Cette thèse a été intégralement financée par le fonds européen BACS FP6-IST-027140.

Mes premières années à l'hôpital ont été teintées d'éclats de rires, d'overdoses de Mikado et parfois de larmes à peines retenues dans un petit bureau partagé avec Mélanie Genetti. Sans elle, les bonnes décisions n'auraient peut-être pas été prises et pour cela je la remercie sincèrement. S'en sont suivies deux autres années dans un nouveau bureau, dont les journées ont été quotidiennement illuminées par des téléphones impromptus, des e-mails spontanés et des pauses de midi avec Roland Vocat et Arnaud Saj qui m'ont intégrée dans un laboratoire d'amitié et de vie quotidienne, et sans qui mes années à l'hôpital n'auraient pas eu la même saveur. Je remercie finalement Mélanie Michel pour son soutien et sa présence qui a bercé ces années de belles discussions et de beaucoup de rires et aussi Chloé de Balthasar pour sa douceur, son écoute et sa maladresse légendaire.

---

[*] The Beatles, *With a little help from my friends,* 1967

En vrac, je tiens encore à remercier Quentin Noirhomme pour avoir passé des heures à m'expliquer Matlab, Alexandra Darque pour avoir fait pareil avec la chimioréception, Danièle Bühler pour son soutien logistique, Karim N'Diaye pour s'être toujours donné la peine de critiquer avec douceur, et aussi Dominique Müller et Jean-Marie Annoni pour leur soutien dans des moments critiques.

Pour plein de fous rires et discussions qui ont rendu ces trois dernières années chaleureuses (repas de midi, apéros, Diablerets/Champéry, book club, 7 ½, et autres) et aussi pour avoir participé à certaines de mes expériences, je tiens à remercier Camille, Virginie, Karsten, Amal, Agustina, Chiara, Marian, Cécile, Martin, Yann, Tonia, Vincent, Markus, Aurélie, Verena, Nadia, Tatiana et Anne-Sophie et plein de sujets anonymes; côté FAPSE, Virginie, Raoul, Diego et Estelle.

Je remercie Couleur 3 pour ses émissions qui m'ont quotidiennement fait rire aux éclats et pour avoir permis à mon amour pour la musique de faire rempart à la folie latente due à l'isolement prolongé dans mon bureau. J'en profite pour m'excuser auprès des voisins pour toutes les fois ou j'ai chanté trop fort en nettoyant des données EEG.

Hors de l'enceinte de l'hôpital, je remercie bien sûr tous mes amis qui m'ont soutenue durant ces années et surtout qui m'ont apporté l'équilibre dont j'avais besoin pour mener une vie accomplie dans tous les domaines…Je remercie avec toute mon amitié et du fond de mon coeur Damaris, Joëlle F., Marie-France, Catherine, Tania, Élise, Jessica, Joëlle G., Tomasz, Hugues, Yann, Ken, Andrew, Ida, Cécilia, Henrik, Jonathan, Jean, Lorenzo, et Kenneth particulièrement pour avoir relu mon travail, ainsi que tous ceux qui étaient présents le 25 avril 2009 à Morges. Je tiens aussi à remercier Per et Gunilla pour m'avoir chaleureusement accompagnée ces sept dernières années.

Ma famille nucléaire partage avec Claes cette confiance inébranlable qu'ils m'ont accordée depuis toujours. Dans tous mes moments de doute, eux n'ont jamais douté. Mes parents m'ont soutenue inconditionnellement dans tous mes choix, tout en me guidant en douceur vers les meilleurs. Ma sœur Katrin et mon frère Thomas ont toujours été là pour rire de moi lorsque je donnais trop d'importance à des futilités…je ne remercierai jamais assez ces quatre personnes qui ont fait tout ce que je suis.

Claes a vécu ces sept dernières années à mes côtés, et avec une infinie patience et une gentillesse sans limite, m'a écoutée et soutenue dans ce long projet comme dans le reste de ma vie. Je le remercie pour son soutien et sa présence, parmi tant d'autres choses qu'il serait difficile d'énumérer ici. Därför att jag skulle aldrig klarat det utan dig.

Résumé

L'étude de la prise de décision a intéressé plusieurs champs de recherche. Tant les psychologues que les économistes, entre autres, ont toujours cherché à décortiquer les mécanismes sous-jacents à la prise de décision. Les neurosciences ont offert, de manière inattendue, un point de rencontre entre ces deux disciplines: l'étude du comportement économique et ses corrélats neuronaux, la neuroéconomie. Cette nouvelle discipline s'est intéressée en premier lieu aux fondements des décisions «irrationnelles» prises par une grande majorité de sujets lorsqu'ils sont confrontés à une série de paradigmes issus de la théorie des jeux. En effet, cette dernière, se reposant sur l'idée que l'homme n'est concerné que par ses intérêts propres, prédit un comportement essentiellement calculateur et rationnel. Cependant, les observations empiriques divergent fortement de ces prédictions. Plusieurs facteurs ont été proposés pour expliquer ces divergences: l'impact des émotions sur la prise de décision ou, plus généralement, l'idée que la plupart des décisions sont prises de manière rapide et automatique. Ce mode de fonctionnement est normalement bénéfique, mais peut aussi parfois mener à des erreurs d'appréciation, raison pour laquelle un mode de contrôle devrait prendre le dessus et rediriger le comportement vers des stratégies plus adaptées.

L'impact direct des émotions sur la prise de décision a été largement étudié. Cependant, la simple appréhension de l'émotion qui pourrait découler d'une mauvaise décision (le regret ou la déception) peut aussi fortement biaiser celle-ci. Cet épiphénomène des émotions n'a été que rarement étudié comportementalement, et jamais électrophysiologiquement. Notre première étude s'est donc intéressée à l'influence de la déception passée sur la prise de décision future: elle a démontré que le simple fait de vivre une déception modifiait fortement les attentes quant à un futur échange et ce dans le but d'éviter de revivre une seconde déception. En effet, dans une version du *Trust Game* adaptée pour l'électro-encéphalographie (EEG), nous avons confirmé l'hypothèse que diminuer ses attentes quant à l'issue d'un échange futur est un moyen de prévention efficace contre la déception. Cependant, utilisée à outrance par une partie des sujets, cette technique les a éloignés du but principal du jeu, la maximisation de leurs gains. Nous avons comparé l'activité électrique des sujets très sensibles à la déception à celle des sujets plus résistants et avons mis en évidence la présence d'une carte topographique de l'activité électrique spécifique aux sujets résistants et à la déception. Nous avons interprété ces résultats dans le cadre de la théorie du système duel: alors qu'un groupe de sujets «sensibles» n'utilise qu'un mode automatique de prise de décision, des mécanismes de contrôle (ainsi que leurs corrélats neuronaux reflétés par la carte spécifique) interviennent chez l'autre groupe de sujets lorsqu'ils sont confrontés à une déception, peut-être pour les empêcher de diminuer automatiquement leurs attentes quant aux échanges futurs.

Pour valider cette interprétation, nous avons demandé aux participants de revenir jouer une seconde fois au même jeu. Nous leur avons expliqué que nos résultats nous semblaient surprenants, et leur avons demandé de bien garder à l'esprit qu'il n'y avait aucun lien entre les résultats d'un échange et les échanges suivants. Lors de cette seconde expérience, les sujets qualifiés de sensibles dans la première étude ont montré un comportement davantage rationnel, mais aussi la présence de la carte trouvée chez le groupe «rationnel» de la première étude. Ainsi, cette seconde expérience nous a permis de valider le lien entre la présence de cette carte de contrôle et un comportement plus avantageux.

Ces résultats suggèrent également qu'un simple changement de consigne peut provoquer l'apparition d'un système de contrôle sur le comportement automatique. Pour investiguer cette question, nous avons crée une version adaptée à l'EEG d'un autre jeu, le *Ultimatum Game*. Les participants ont joué deux blocs: le premier selon les consignes classiques, tout en sachant qu'ils allaient probablement répliquer le comportement irrationnel et largement documenté qui consiste à rejeter certaines offres. Avant de jouer le second bloc, nous avons souligné auprès des sujets l'irrationalité de leur comportement, leur laissant le choix de prendre en compte ce point de vue ou non. Les résultats ont à nouveau mis en exergue de grandes différences interindividuelles. Alors que certains sujets ont accepté toutes les offres dans le second bloc, et sans signes de conflit manifestes, d'autres sujets n'ont pas pu/voulu s'empêcher de rejeter les offres inéquitables. La comparaison des cartes a révélé à nouveau des différences électrophysiologiques dans la même fenêtre temporelle que la première étude et la localisation des générateurs sous-jacents à montré que les sujets ayant rejeté des offres dans le second bloc ont désactivé des zones frontales, typiquement liées au contrôle du comportement. Grâce à cette étude nous avons également découvert un réseau de structures dont l'activation corrèle fortement avec le degré d'équité de l'offre.

La quatrième étude nous a permis de révéler un mécanisme de codage par les ondes thêta dans les régions hippocampique/parahippocampique de trois patients. Nous avons d'abord trouvé des grandes réponses sélectives à l'issue des échanges du *Trust Game* dans les contacts proches de l'hippocampe. Puis nous avons corrélé l'amplitude de l'activité thêta avec le changement des attentes entre deux essais consécutifs et avons trouvé une corrélation significative entre cette dernière (dans l'hippocampe) et les changements dans les attentes quant à l'issue d'un échange à venir. Ainsi, le comportement observé dans la première étude serait sous-tendu par un mécanisme d'apprentissage déclenché lorsque l'issue réelle d'une décision diverge trop des attentes du sujet.

Nos recherches démontrent ainsi l'importance de prendre en compte les différences interindividuelles lors de futures études ou modélisations de la prise de décision.

What piece of work is a man!

How noble in reason,

How infinite in faculties,

In form and moving

How express and admirable,

In action how like an angel

In apprehension how like a god:

The beauty of the world

The paragon of animals![*]

[*] Willliam Shakespeare, *Hamlet,* Act II scene II. Hair, *What a piece of work is man,* 1968.

Table of contents

INTRODUCTION

Yesterday evening my husband Claes and I were feeling lazy about the idea of cooking, so we phoned a pizzeria and asked them to prepare a *Pavarotti* and a *Mimi* for 9 pm. Arriving there a few minutes earlier than expected, the pizzas weren't ready so the waiter asked us if we wanted to drink something. Claes ordered a beer and asked for the bill but the waiter with a simple gesture made us understand that we did not have to pay for the drink. Though, when pizzas were ready, Claes paid for them and added a tip almost as high as the prize of the beer, exchanging open smiles with the waiter.

A win-win situation.

The waiter gets a "huge" tip (given the service) and Claes pays less than expected for his beer. Did the waiter expect Claes to act this way? What was at stake for him? His boss would probably not fire him for giving away one beer, but still, if he acted this way with all the customers he might lose his job. On which elements did he base his judgment to trust Claes? Had Claes not reciprocated, would the waiter still trust other customers, or rather be submerged by disappointment and not give a chance to the following ones? Why did Claes reciprocate? In a strictly self-interested perspective, wouldn't it be more reasonable to leave a lower tip, just enough to keep up appearances?

Would all waiters take the same risk, and all customers reciprocate?

Decision-making involving monetary gains and losses is not only about traders, markets, stock exchanges, mortgages, bankruptcies and other Madoff(s): it is a part of our everyday life.

As such, it has been extensively studied in psychology, economics, and more recently neuroscience. The interest of understanding, modelling and in the best cases being able to predict the outcome of decision-making is obvious. As decisions are at the core of society's evolution and humans' interactions, understanding the components of the underlying process is crucial in some cases (e.g. financial crisis, wars, Milgram's experiments) as well as less honourable in others (how to lead a consumer to the decision of buying my brand-new toothpaste?). Still, almost every aspect of the world today is the reflection of human decisions.

The rise and fall of Homo Economicus

Blame it all upon
A rush of blood to the head[*]

Common wisdom exhorts us to calm down before reacting or making a decision following an emotional arousal. Thus, emotions have always been considered as disruptive elements in the process of decision-making. That is probably why - during the first attempts to build economical models of decision-making - emotions have been simply removed from the general picture. For instance, in the models based upon game theory which were developed in the second half of the 20th century, humans were replaced by rational agents, being purely self-interested and described later as "cold gain maximizers" (Thaler, 2000). The Homo Economicus was born, a hybrid sharing some characteristics with humans but lacking an essential one to resemble them: emotions. This cold calculator dramatically simplified the work of modelling decision-making. Indeed, without emotions, what is left can be easily predictable; decisions must be the result of a calculation of costs and benefits that any human or machine is capable of performing rather quickly. For instance, the concept of expected value is the idea that when an agent must choose between two options, he will compute the utility (their desirability) of both actions' outcome, weight them by their probability of occurrence and finally select the one which offers the highest gain. The expected value was then integrated into expected utility theory, which is more sophisticated as it takes into account more parameters than the expected value, but still emotion does not have its place amongst them. This theory, first presented by Bernoulli in 1738, remained the predominant reference in choices modelling for 250 years (Kahneman, 2003).

Unfortunately, once the models were developed and the testing phase started, humans failed to reproduce the Chimera's behaviour. Showing sensitivity to injustice, low resistance to offended pride, frame-dependant reasoning and stronger reaction to losses compared to equivalent gains, laboratories' subjects were far from being economically rational.

For instance, in the domain of decision-making under uncertainty, when confronted to choosing between receiving 50$ or a 50% chance of receiving 110$ the majority of people will choose to receive 50$ although the expected value of this first option is smaller than that of the gamble (Sanfey & Chang, 2008). Examples are also found in the domain of reasoning. Consider the following problem: A bat and a ball cost $ 1.10 in total. The bat costs $ 1.00 more than the ball. How much does the ball cost?

[*] Coldplay, *A Rush of Blood to the Head,* 2002

The spontaneous tendency (refrained by only a small percentage of students) is to answer ten cents (Frederick, 2005).

A very famous task in logical reasoning is the Wason Selection Task (Wason, 1968). Four cards are displayed on a table. The subject knows that each card has a letter on one side and a number on the other. The rule is the following: "if there is a D on one side of any card, then there is a 3 on its other side". The four cards show a B, a D a 3 and a 7. Wason asked his students which card should be turned over in order to verify if the rule were valid. Although the correct answer is D and 7, most of the students answered only D, or D and 3 (although "D implying 3" does not mean that 3 implies D). Tested on two different versions of the task, only 16.7% of the students mentioned that the card 7 had to be turned over to verify the rule.

One of the most important works that deeply shattered the perspective of the "rational agent" is a decades-long program led by Kahneman and Tversky on "bounded rationality" (Kahneman, 2003). Between approximately 1973 and 1992, their main topics of research concerned heuristics and bias in decision-making under uncertainty, choice under risk (and the development of the famous prospect theory), and framing effects. Their starting point - and null hypothesis - was always a model of how a rational agent would behave in those various contexts.

The prospect theory for instance demonstrated that contrary to the expected utility theory, the value function (which assigns a value to an outcome) is not symmetrical. Indeed, the loss function is steeper than the gain function, meaning that a loss is twice as painful as an equivalent gain is pleasurable; this is the loss aversion. Moreover, whereas in expected utility model, utility is weighted by the raw probability of occurrence of the outcome, Kahneman and Tversky (1979) demonstrated that people tend to overweight small probabilities and underweight large ones. Some critics emerged, arguing that a lack of motivation (or a lack of involvement) might be behind the violation of the normative principles. Although pertinent, neither this claim (Camerer & Hogarth, 1999), nor the use of techniques such as asking the participants to justify their choices or assessing only experts (Shafir & LeBoeuf, 2002) did change the occurrence of those classical biases.

At the same time, the testing of paradigms directly issued from game theory yielded as well – with regards to Homo Economicus' expected behaviour – aberrant results. Game theory is basically the mathematical modelling of strategic behaviour when agents have to make choices that will affect or will be affected by choices of others. The principal application is to find equilibrium, i.e. the strategy that

each player will adopt and maintain throughout the game. Those games are very interesting because they involve a reasoning process as well as a social component (at least two players are required).

For instance, the Trust Game and the Ultimatum Game are good examples of the gap between theory's predictions and laboratories results in the domain of strategic decision-making.

Described in 1995 in the same setting as we will use it (Berg, Dickhaut, & McCabe, 1995) the Investment Game (based on the Trust Game, TG) was created in order to answer questions about the factors influencing the likelihood of trust in economic transactions. In this two-player game, each round is played with a new partner and anonymously. At the beginning of each round, Player 1 (the Investor) has a certain amount at his disposal and can invest any part of it in the game. The part invested is tripled by the experimenter and given to Player 2 (the Trustee). The Trustee can in turn give back the amount of his choice to the Investor. Let us see an example in which the Investor has 10 CHF at his disposal. In the best case scenario, he invests everything; the Trustee receives 30 CHF and reciprocates the trust by giving back the half. Both players end up with 15 CHF, it is a win-win situation. However, a rational Trustee focused on maximizing his own gains would never reciprocate – there is no reason to do so, as there will be no consequences (single-shot interaction) and the anonymity is guaranteed. Knowing this, the Investor should never invest – thus the unique Nash equilibrium in this game is to invest O. However, in this first experiment, 30 out of 32 participants invested money in the game (on average an amount close to the half of the total). Since then, this spontaneous tendency to invest in the game has been reported in an large number of studies (e.g. Delgado, Frank, & Phelps, 2005; King-Casas et al., 2005; McCabe, Houser, Ryan, Smith, & Trouard, 2001; van't Wout & Sanfey, 2008).

In the same vein, the Ultimatum Game (UG) is again an anonymous, single-shot two-player game, in which Player 1 has a certain amount at his disposal and must propose a share to Player 2 (Güth, Schmittberger, & Schwarze, 1982). If Player 2 accepts the proposal, the share is done accordingly. However, if he refuses, both players end up with nothing. A selfish income-maximizer should accept all kind of offer, even very low, as it is always a positive gain compared to its actual state (0). Knowing this, Player 1 should always propose the smallest possible amount. Again, classical results are quite different from this prediction (Camerer & Thaler, 1995). Actually, Players 1 tend to propose rather fair offers (30 to 40 percent of the total amount), which coincides with Players 2 behaviour as they tend to massively reject offers judged unfair (less than 20 percent of the total amount), although ruining their chances to maximize their gains by doing so.

Confronted by those unexpected (but so human!) behaviours, researchers have built up a series of variations around the original paradigms in order to understand how subjects can deviate from the predictions to such an extent. Some fundamental questions on human nature were also at the centre of the debate: are we deeply altruistic, to the point of spontaneously proposing fair offers and punishing even at our own costs what we perceive as unfairness? Or is it a learned behaviour, useful for societies to function correctly (punish the free-riders)? Do humans spontaneously trust each other? What are the necessary conditions to trust someone (Alesina & La Ferrara, 2002)? Do factors such as attractiveness change the behaviour (Solnick & Schweitzer, 1999)? If the emotion is expressed verbally, does it diminish the tendency to punish (Xiao & Houser, 2005)? Amounts at stake have been raised (until sometimes the equivalent of three months of salary, Cameron, 1999) but the behaviour seemed robust (even in some small-scale societies living for instance in tropical forests or deserts, Henrich et al., 2005). Interestingly, some authors (Kagel, Kim, & Moser, 1996) reported that between being really fair or just looking fair but acting selfishly, subjects chose the second option. This result suggests that it is more for the sake of manners than for a real concern for equity that people propose fair offers. The debate has gone as far as testing those games on chimpanzees, the results showing that they are indeed closer to Homo Economicus than humans are (K. Jensen, Call, & Tomasello, 2007), although capuchin monkeys seem somehow sensitive to unequal rewards distributions (Brosnan & De Waal, 2003).

Thus emotions are disruptive to a rational decision-making process, such as demonstrated with the TG and the UG. They lead the decision-maker to behave in an "irrational" way, refusing offers or trusting strangers for no economical reason. On the other side, one of the most influential theories on the role of emotions in decision-making came to the opposite conclusion: emotions represent a good signal that should be listened to when making decisions; they are not only present but necessary and beneficial.

Indeed, parallel to these laboratory's discoveries, clinical observations and experimentation led to the Somatic Markers Hypothesis (Bechara & Damasio, 2005; Bechara, Damasio, Damasio, & Anderson, 1994). The starting point is the observation of patients with damage to the ventromedial prefrontal cortex (VMPFC), showing a fascinating dissociation between normal IQ and severe impairment in decision-making and emotion regulation. Those patients tend to make "bad" decisions, i.e. leading to losses (in the financial domain as well as in their personal lives, such as family, friends and social status). Their choices are no longer advantageous as if they were unable to learn from previous mistakes. The Somatic Marker Hypothesis postulates that this inability in decision-making reflects "*a defect in an*

emotion mechanism that rapidly signals the prospective consequences of an action, and accordingly assists in the selection of an advantageous response option".

Using a gambling task in which the subjects have to pick up one card out of four different decks, they demonstrated the role of emotion in preventing the subjects from making a wrong decision (through hunches and even subconscious pre-hunches). Patients with damage to the VMPFC or to the amygdala failed to trigger those somatic signals (as shown by recordings of their skin conductance responses, Bechara, Damasio, Damasio, & Lee, 1999) which should normally appear just before choosing a bad deck. They therefore persisted in choosing bad decks - offering larger rewards compared to the other decks, but more frequent and higher penalties - being counterproductive in a long-term perspective. Interestingly, the authors showed in a subsequent study that even when patients became aware of the rule of the game, they persisted in choosing the bad decks, as if the knowledge of acting wrong, decoupled from the emotion, was not sufficient alone to redirect the behaviour in an advantageous manner.

To reconcile those contradictory roles of emotions, Damasio and his colleagues proposed to distinguish emotions that are part of the decision-making process from those that are externally induced (Bechara & Damasio, 2005). They illustrate their proposal as follows: when driving in a hurry, one might decelerate after imagining the shame of being caught by the police or the fear of having an accident. In this case, the emotions (somatic states) are integral to the decision-making process and bias the behaviour in an advantageous way. This interference of emotions is completely different from the case where while driving, someone receives a phone call announcing the death of a friend. In this latter situation, the emotion generated by an external factor and having nothing to do with decision-making might be disruptive and have bad consequences.

Be they beneficial or disruptive, one certainty about emotions has been acquired through this large body of literature: they cannot be ignored. Those discoveries, taken together, sounded the death knell for the Homo Economicus.

Observing the behaviour of brain-damaged patients is one possibility to infer which parts of the brain (areas and/or connecting fibres) are necessary to make decisions as predicted by a theory, or as expected according to the behaviour of a control population. On the other side, being able to look at the brain activity of control subjects during the decision phase is obviously a powerful way to learn, understand and model human decision-making. Thus, the development of neuroimaging methods has provided researchers with a highly exploited tool to study economics and gave rise to a new discipline: neuroeconomics.

A deeper look into the brain: use of imagery and the birth of neuroeconomics

And so, and now, I'm sorry I missed you
I had a secret meeting in the basement of my brain[*]

Indeed, thanks to neuroimagery, researchers were now provided with a tool to investigate the neural correlates underlying the unpredictable behaviour observed - among others - in the TG and in the UG (Kenning & Plassmann, 2005).

One of the first studies using fMRI on the Trust Game (McCabe et al., 2001) demonstrated that in the subjects willing to invest, prefrontal regions were more active when playing with a human compared to a computer counterpart. This was again a piece of evidence against the view that humans were purely rational decision-makers: if so, then the human factor should not be relevant. Subsequent studies thus focused on human interaction, helping to define the structures underlying reciprocity and trust in the Trust Game. In a multi-round version of the TG where the Investor played 10 rounds with the same Trustee, the head of the caudate nucleus (dorsal striatum) has been reported as computing information about the fairness of a social partner's decision, and the intention to repay his decision with trust (King-Casas et al., 2005). Interestingly another study the same year demonstrated as well the role of the caudate nucleus in processing feedback information (here, the outcome of the TG) especially when the feedback is relevant in order to learn and better adapt choices (Delgado et al., 2005).

It is obvious that the neuroimagery results of Trust Game (due to its specific setting which induces expectations and feedbacks such as rewards and punishments) must be related to the literature on the prediction error signals, i.e. a class of electrophysiological signals recorded when a reward is expected but does not materialise and vice-versa. For instance, single cell recordings in non-human primates show that the firing rate of midbrain dopamine neurons increases when unexpected rewards occur, but decreases when expected rewards are absent (Schultz, 1998). Interestingly these neurons project to both ventral and dorsal striatum, including the caudate nucleus. Thus, the studies mentioned before, by showing that the activation of the caudate nucleus predicts the intention to reciprocate and the assessment of the fairness of the offer, indicate that elements of the classical reward circuitry cited above are involved in the TG and might even guide subjects in their decision-making process.

[*] The National, *Secret Meeting,* 2005

A series of studies focused on trust and how it is granted to other individuals with the use of oxytocin, a neuropeptide secreted by many mammals to strengthen pair bonding, positive social interactions, maternal care, social attachment and so on (Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005). The authors showed that the administration of oxytocin to human subjects while playing the TG significantly increased their investments. The mechanism proposed by the authors is that increased trust is a consequence of diminished fear of being betrayed. The same experiment combined with fMRI showed that control subjects decreased their trust levels after being betrayed, on the contrary to subjects in the oxytocin condition (Baumgartner, Heinrichs, Vonlanthen, Fischbacher, & Fehr, 2008). This difference in trust modulation as a function of past betrayal was associated with lower activations in the amygdala, in midbrain regions (brainstem effectors sites) and in the dorsal striatum of the subjects under oxytocin. Thus regions linked to reward prediction but also to emotions might modify the effect of oxytocin on trust.

Interestingly, the reverse effect had been previously reported (Zak, Kurzban, & Matzner, 2005): subjects receiving a signal of trust, compared to subjects receiving similar amounts of money but without trust signals, showed increased levels of oxytocin. The same group tested the hypothesis that another hormone, the dihydrotestosterone (DHT) is linked to negative social interaction, i.e. distrust (Zak, Borja, Matzner, & Kurzban, 2005). DHT is described as a highly reactive hormone whose level rises when falling in defeat or before an athletic match. Using a TG, the authors showed that men, but not women, increased their levels of DHT after distrust. This behaviour was interpreted as an aggressive reaction to distrust (based on the link between testosterone and aggression), apparently absent in women - although they also reported that they disliked being distrusted.

Other kinds of research investigated the bases of trust behaviour. For instance, one study on the genetic bases of trust with 682 pairs of monozygotic and dizygotic twins suggested that cooperative behaviour - i.e. individual's interpersonal trust and willingness to reciprocate trust - is heritable (Cesarini et al., 2008). Indeed, their analyses revealed that genetic differences were a more important source of variation in the behaviour than differences in the environment.

Although human behaviour in the TG is probably much more complex than hormone secretion and neuronal firing, each one of these components is a part of the observable whole, i.e. trust granting, reaction to betrayal, etc. As such, no information should be ignored, all the more so as linking observations at all levels might be the key to finally understanding the basic mechanisms involved in decision-making in the TG among other situations.

The Ultimatum Game did not escape neuroeconomists either. In a pioneering experiment (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003) a network comprising the insula, the dorsolateral prefrontal cortex (DLPFC) and the anterior cingulate cortex (ACC) has been defined as responsible for decision-making in the UG. The authors demonstrated that the insula was sensitive to the fairness of an offer, which is an important result as the insula has been traditionally linked to negative emotions like disgust (Wicker et al., 2003). When anterior insula activation is higher than that of the DLPFC, the unfair offer is rejected, whereas when it is lower, the unfair offer is accepted. Their interpretation is that the DLPFC, responsible for the control of behaviour, incites the subject to accept the offer (although unfair), in order to fulfil the rational goal of maximizing the subject's gains. At the same time, the insula reflects the emotional negative reaction to unfairness, and if stronger than the DLPFC, the offer is rejected. Thus the decision of the subject results from the balance between those two levels of activity, whereas the ACC is activated because of the conflicting aspect of the situation. Interestingly, in a Prisoner's Dilemma game – another game issued from game theory, which can be compared to an UG where players must decide simultaneously - anterior insula activation has been reported when cooperation was not reciprocated (Rilling, Goldsmith et al., 2008).

In the light of this first study, the interpretation seems obvious: subjects' natural tendency is to reject unfair offers under the impulse of negative emotions. Fortunately the DLPFC works as a regulator and sometimes prevents this impulsion, for the sake of the rational goal maintenance. The idea that emotions must be controlled in order to achieve the cognitive goal of maximizing the gains found support in another study on patients with lesions to the ventromedial prefrontal cortex (Koenigs & Tranel, 2007). As postulated in the Somatic Markers Hypothesis, the VMPFC is crucial for integration of emotional signals and emotion regulation. Patients with VMPFC lesions often demonstrate at the same time reduced social emotions (such as shame and guilt, which leads them to be sometimes more "rational" than control subjects, Koenigs et al., 2007) and poor regulation of negative emotions resulting in "hyper-reactivity" to little frustrations. In this study, the patients showed a higher rejection rate of unfair offers compared to other patients and to control subjects: they diverged from those groups precisely when emotions should be regulated. Finally, in a behavioural study on the UG (van't Wout, Kahn, Sanfey, & Aleman, 2006) the authors reported that skin conductance responses were higher for unfair than for fair offers (but only when playing against human partners), suggesting again that a strong negative affect is linked to the rejection of unfairness in the UG.

However, two other studies led to the opposite conclusion (Knoch, Pascual-Leone, Meyer, Treyer, & Fehr, 2006; van't Wout, Kahn, Sanfey, & Aleman, 2005). The authors showed that applying Transcranial Magnetic Stimulation (TMS) on the right dorsolateral prefrontal cortex (rDLPFC) results in a decrease of the rejection rate of unfair offers. As TMS disrupts the good functioning of the targeted area ("simulating" the effects of a lesion), those results suggest, contrary to previous studies, that the spontaneous tendency is to accept any kind of offer, but the control normally exerted by the rDLPFC leads to rejecting unfair offers, as if its role was to implement fairness-related behaviour and prevent the economic temptation to accept all offers. A recent study also suggested that emotions were not the critical factor explaining the rejection of unfair offers: in this paradigm, an increase in skin conductance was associated with rejection of unfair offers only when the offer concerned the subject himself (Civai, Corradi-Dell'acqua, Gamer, & Rumiati, 2009). Indeed, in a *third-party* condition where the subject had to decide for another person, although the unfair offers were still rejected, no increase in skin conductance response was recorded. The authors suggest that other factors than the emotional arousal might be the ground for rejection (which is a questionable interpretation: since the emotion has been felt once by the subject, it does not seem necessary to feel it for the *third-party* too in order to reject unfairness). Finally, as for the TG, a study on monozygotic and dizygotic twins has shown that more than 40% of the variation in rejection behaviour can be explained by genetic effects (Wallace, Cesarini, Lichtenstein, & Johannesson, 2007).

Thus, summarizing all the data on brain areas involved only in those two games is a quite difficult task. Moreover, one must consider that when studies address "decision-making", it is not always clear to which step of the process they refer. Indeed, decision-making involves many stages such as the representation of the decision problem, the assignment of value to each possible action, the comparison of those different values to select the action, the desirability of the outcome and finally the processing of the feedback in order to improve quality of future decisions (Rangel, Camerer, & Montague, 2008).

Some consistencies across studies nonetheless do appear (Rilling, King-Casas, & Sanfey, 2008; Sanfey, 2007):

- Areas linked to the reward circuitry (Figure 1), i.e. mesencephalic dopamine projections to ventral and dorsal striatum - including the caudate nucleus (Haruno et al., 2004) and the nucleus accumbens (Rilling et al., 2002) - are clearly involved in feedback processing, probably by providing a signal that helps to adjust behaviour (this will be discussed more in details in a following chapter). This system is involved in much more than "basic" rewards. As mentioned earlier, the caudate nucleus is now considered as a key structure in social prediction error (Delgado et al., 2005; King-Casas et al., 2005).

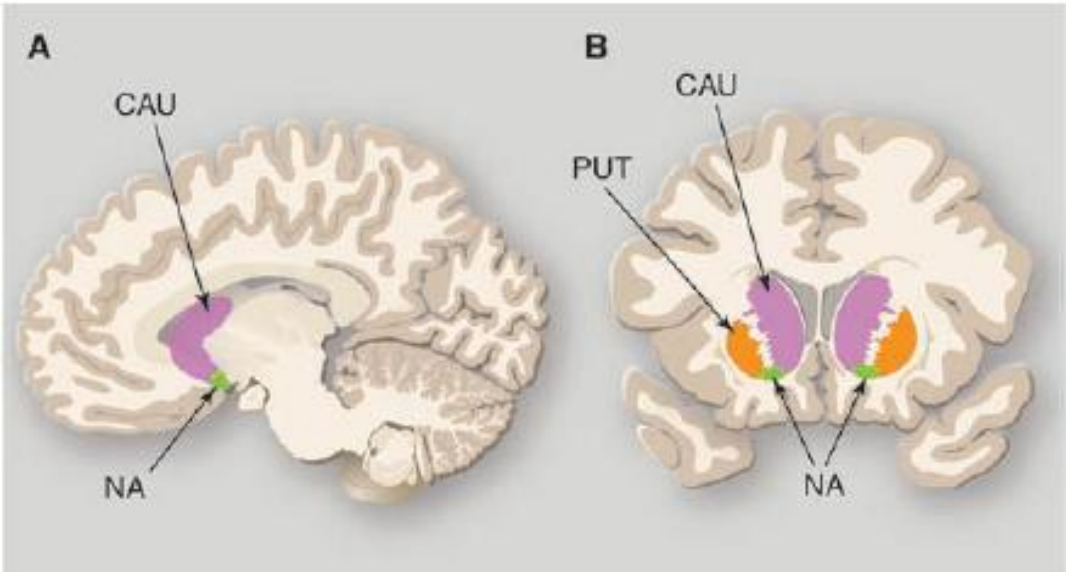


Fig1. *Sagittal (A) and coronal (B) sections of the human brain showing subcomponents of the striatum involved in reward processing: caudate nucleus (CAU), nucleus accumbens (NA) and putamen (PUT). Adapted from Sanfey (2007).*

- Areas linked to emotion processing (Figure 2): we have seen that the amygdala might impact decision-making as described by the Somatic Marker Hypothesis or by the studies on oxytocin. More recently, attention has been paid to the anterior insula which often appears to be activated when a negative outcome is discovered. As mentioned earlier, insula's activation has been traditionally linked to aversive stimuli, and has been proposed as scaling the magnitude of unfairness. Recently a broader role has been proposed: insula reflects degrees of unpleasantness in subjective evaluation (Grabenhorst & Rolls, 2009). For a review on the role of the insula see Singer, Critchley and Preuschoff (2009).

- Areas of the prefrontal cortex linked to behavioural control (Figure 2): the dorsolateral prefrontal cortex (DLPFC) is known for its exertion of cognitive control, for instance over emotions (see for example Hare, Camerer, & Rangel, 2009). Lesions to the ventromedial prefrontal cortex (VMPFC) as well as TMS studies demonstrate that its disruption leads to significant changes in decision-making. Finally, the orbitofrontal cortex (OFC) might encourage respect of the norms because of its sensitivity to punishment.

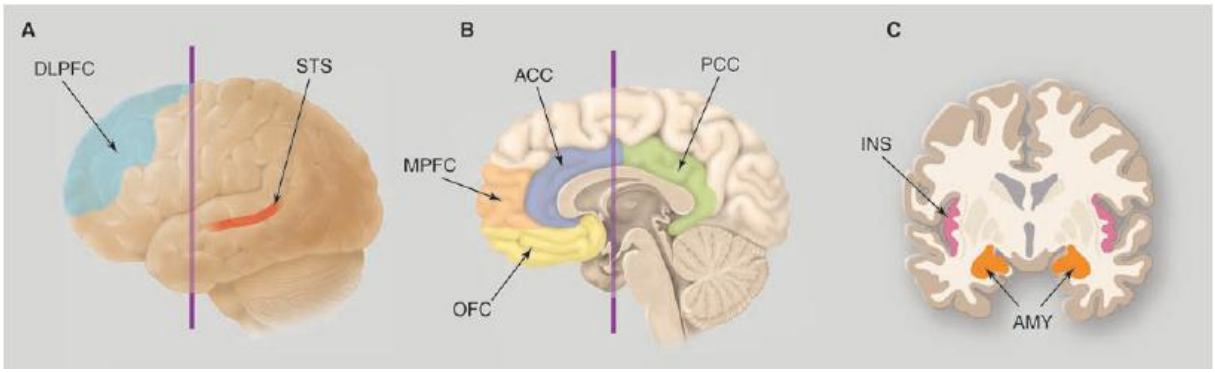


Fig2. *Lateral (A), sagittal (B) and coronal (C) sections of human brain showing areas commonly activated in social decision-making and Theory of Mind tasks: (A) Dorsolateral prefrontal cortex (DLPFC), superior temporal sulcus (STS). (B) Orbitofrontal cortex (OFC), medial prefrontal cortex (MPFC), anterior cingulate cortex (ACC) and posterior cingulate cortex (PCC). (C) Insula (INS) and amygdala (AMY). Adapted from Sanfey (2007).*

Moreover, neurons in the OFC encode economic values in an absolute fashion (independently of other choices' values, Padoa-Schioppa & Assad, 2006; Padoa-Schioppa & Assad, 2008). This implication of the OFC has also been shown in an fMRI study (Fujiwara, Tobler, Taira, Iijima, & Tsutsui, 2008). Indeed, in the lateral OFC, absolute losses correlated positively with a scale of introversion, whereas relative losses correlated positively with a scale of neuroticism. Interestingly those results suggest the possibility that personality traits impact loss processing in different ways. Another study demonstrated its role in the willingness to pay, supporting the hypothesis that the OFC is also involved in encoding the values of the goals (Plassmann, O'Doherty, & Rangel, 2007) .

The Anterior Cingulate Cortex (ACC) has drawn considerable interest recently for its role in guiding behaviour, detecting conflicts, regulating emotions, etc. For instance its involvement has been shown in both cognitive up- and down-regulation of emotions (Ochsner et al., 2004). In the broader frame of cognitive control, the ACC has been held responsible for the conflict detection. In the Conflict Monitoring Hypothesis (Botvinick, Braver, Barch, Carter, & Cohen, 2001), "*The conflict monitoring system first evaluates current levels of conflict, then passes this information on to centres responsible for control, triggering them to adjust the strength of their influence on processing*.". To support their theory, the authors relied on a vast literature review describing the intervention of the ACC in three types of situations: overriding predominant but task-irrelevant responses; choosing between equally acceptable responses and error commission. Therefore, the ACC seems to show strong activation right before a shift in behaviour is observed across a large panel of different tasks. The authors conclude that the ACC has a role in conflict detection but that it intervenes as well in the control itself, although existing support for this hypothesis is less obvious. For the authors, its connectivity to the prefrontal cortex might corroborate this idea.

Indeed two recent papers (Ridderinkhof, van den Wildenberg, Segalowitz, & Carter, 2004; Rushworth, Behrens, Rudebeck, & Walton, 2007) - by interrogating the different roles of the OFC and the ACC in guiding behaviour - demonstrate that this hypothesis has gained ground. Although both regions are important in reward-based decision-making, it seems that their respective connections to different areas explain that the OFC is more important when guidance is based upon associations of reinforcements with stimuli whereas the ACC intervenes in associations of reinforcement with actions (Rushworth et al., 2007). The second study suggests that in addition, the ACC may be more important for context-dependant representations of choice-outcomes values than the OFC, which functions more independently of the context (Buckley et al., 2009). It seems anyway that both regions cannot be ignored as their role in decision-making is undeniable. For a review on the contributions of various prefrontal areas on cognitive control see Ridderinkhof et al. (2004).

Finally, an overlap exists between the structures involved in decision-making in social context and those of social cognition (Frith & Singer, 2008), particularly in theory of mind (ToM) studies (ToM being the ability to take another person's perspective into account, and to understand his beliefs and intentions). This overlap had been studied with scanned participants playing a UG and a Prisoner's Dilemma Game with both human and computer counterparts (Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004). The authors found in both games activations in anterior paracingulate cortex and posterior superior temporal sulcus (STS) that are traditionally linked to ToM studies (see Figure 2. They did not find any activation in the temporo-parietal junction which is usually also observed in ToM studies).

All economists do not share the point of view that understanding decision-making's neural substrates will really improve the understanding of economic behaviour (Maskin, 2008). Indeed, if the aim is to predict which option a subject will choose, it does not help to know which parts of the brain can predict it (unless economists have a scanner at their disposal…which in any case cannot be integrated in a mathematical model). However, discoveries at the brain level can be at the basis of the development of new hypotheses, and can contribute to deciding between a good explanation and a bad one. Most importantly, by considering the agent as a Homo Neurobiologicus (whose social as well as economical behaviour is the result of neurobiology) rather than a Homo Economicus, the economic models might become much more realistic (Kenning & Plassmann, 2005).

For instance, some suggest that "*neuroeconomics could model the details of what goes on inside the consumer mind just as organizational economics models activity inside firms*" (Camerer, Loewenstein, & Prelec, 2004).

Those same authors listed some reasons for which neuroscience might be very useful to economics. For instance, the measurement of brain activity is more reliable than self-reports or surveys in giving indices about which variables are relevant to economical decision-making. As well, it may be discovered that some quasi-similar behaviours rely on completely different structures and vice-versa, that some structures take part in many different processes. Understanding how other internal factors such as hormones or mood affect decision-making can also improve the building of realistic models (Camerer et al., 2004). Maybe the influence of these discoveries on economic models will not be direct, but by influencing first psychology, neuroscience might slowly modulate some of the conceptions that are still often the basis of economical models and that have had hard times recently, such as the immovable rationality of the decision-making agent.

This is indeed one of the greatest controversial subjects: we will see now how a different conception of the decision-making process might reconcile many conflicting views.

Dual system account in decision-making models

*Wisely and slow: they stumble that run fast**

The claim that humans were completely irrational has raised a great amount of voices against it. Indeed, the "rationality debate" has shaken the world of many different scientific communities for decades, as rationality is often considered as a key feature that distinguishes humans from animals (see Shafir & LeBoeuf, 2002, for a summary of main objections).

First, economically and individualistically irrational does not mean irrational at all levels. For instance, altruistic punishment (the subject penalizes himself to punish someone else, like refusing an unfair offer in the UG) might promote collaboration in a larger scale perspective (Bowles & Gintis, 2002, 2004; Boyd, Gintis, Bowles, & Richerson, 2003; Fehr & Camerer, 2007) such as groups or society. By punishing free-riders (the non-cooperators), whatever the costs are, cooperation is promoted. It has been even argued that strong reciprocity is necessary for the survival of the species in case of threatening events (Gintis, 2000). In that sense, sometimes, the benefits of the group must take precedence over the benefits of the individual. For instance, in a study of a public goods game where 84.3% of the subjects showed altruistic punishment (Fehr & Gachter, 2002), the authors reported that the subjects who punished the most often were also the ones who contributed the most to the public good. More interestingly, the punishment of non-cooperators increased the amount invested by all the participants in subsequent games. Another study on 15 different populations across the world showed that all populations (but to a different degree) show costly punishment behaviour, and that costly punishment correlates with altruistic behaviour (Henrich et al., 2006). Note that controversy exists: those results have been nuanced by another study showing that in a repeated setting, although punishment increased cooperation, it did not increase the average payoff of the group (Dreber, Rand, Fudenberg, & Nowak, 2008). A recent report suggests that rewarding the cooperators rather than punishing the free-riders is a better way to promote public cooperation (Rand, Dreber, Ellingsen, Fudenberg, & Nowak, 2009) and moreover this strategy results in higher payoffs when interactions are repeated. This is all the more interesting since punishment promotes cooperation only if antisocial punishment (sanctioning people who act pro-socially) is not massively applied (Herrmann, Thoni, & Gachter, 2008).

* William Shakespeare, *Romeo and Juliet,* Act II, scene II.

In any case, this does not mean that altruistic punishment is a rational calculation. Indeed, the participants in Fehr and Gachter's (2002) experiment reported high levels of anger against the free-riders. Thus, this behaviour, considered as irrational in a strictly individualistic perspective, might be viewed as useful in a larger perspective, although is it accompanied with strong negative feelings. Indeed, it has been shown that watching a free-rider receiving electrical shocks increases activity in reward systems and decreases activity in empathic ones (Singer et al., 2006). Even the act of inflicting punishment on a non-reciprocator is correlated with higher activity in the striatum (de Quervain et al., 2004). For recent reviews focusing on punishment see Seymour, Singer and Dolan (2007) and on the nature of human altruism see Fehr and Fischbacher (2003).

Second, in many situations, humans are not that bad. A small detail such as how the problem is presented can completely change the deal. For instance in the Wason Selection Task, if the abstract aspect of the problem is removed by proposing it in more realistic terms, the performance changes drastically. This has been reported in a study where the participants had to imagine that they were policemen entering a bar with the aim of checking the following rule: If a person is drinking a beer, then that person must be over 19 years of age. The four cards presented to the subjects were annotated "Drinking Beer", "Drinking Coke", "22 years of age", "16 years of age". The subjects spontaneously gave the correct answer, i.e. they must check the age of the customer drinking beer, and the beverage of the customer who is 16 years old (Griggs & Cox, 1982). Indeed it has been shown that depending on the context in which the rule is given, different areas are recruited into the reasoning task (Canessa et al., 2005). Examples like this one have been reported many times, and have been the core material for some defenders of human rationality.

We have seen that changing the frame can change the prominent response. Other studies demonstrate that switching from probabilities to frequencies might also strongly affect the supposedly bad reasoning of humans. For instance, Cosmides and Tooby (1996) have presented a great amount of experiments suggesting that alternative conclusions should be drawn about human rationality than those classically proposed by Kahneman and Tversky. The authors conclude that:

"*It may be time to return to a more Laplacian view, and grant human intuition a little more respect than it has recently been receiving. The evolved mechanisms that undergird our intuitions have been subjected to millions of years of field testing against a very rich and complexly structured environment. With only a few hundred years of normative theorizing under our belts, there may still be aspects of real-world statistical problems that have escaped us. Of course, no system will be completely error-free, even*

under natural conditions. But when intuition and probability theory appear to clash, it would seem both logical and prudent to at least consider the possibility that there may be a sophisticated logic to the intuition. We may discover that humans are good intuitive statisticians after all."

One way to partially reconcile these views is to consider that the processes underlying decision-making are not unitary. As Kahneman (2003) proposes: "*The central characteristics of agents is not that they reason poorly but that they often act intuitively. And the behaviour of these agents is not guided by what they are able to compute, but by what they happen to see at a given moment*."

The dual system approach proposes that the systems involved in decision-making are differentiated by the way they process information and guide behaviour. The first one, System 1, can be considered as the "default" mode; it is responsible for fast and automatic processing, and guides behaviour with the help of heuristics, mental shortcuts, reflexes, intuition, and so on. It is said to be fast (Todorov, Mandisodza, Goren, & Hall, 2005; Willis & Todorov, 2006), associative (Sloman, 1996), unconscious (Dijksterhuis, Bos, Nordgren, & van Baaren, 2006), affective (Sanfey & Chang, 2008), highly dependent upon the context and evolutionarily old (even shared with animals). On the other hand, System 2, of a higher cognitive level, intervenes when something in the situation demands more than the usual routines, such as adaptation, control over spontaneous behaviour, mental effort. System 2 is supposedly less dependent upon the context, more conscious, evolutionarily new (unique to human), slow, rule-based. According to Stanovich (2004), System 2 processing "*allows us to sustain the powerful context-free mechanisms of logical thought, inference, abstraction, planning, decision-making and cognitive control*". One of the most important properties of System 2 is its capacity to override System 1 processes when an inappropriate activation can lead to negative results. For an overview of the features that are most often attributed to System 1 and 2 organized in clusters, see Table 1.

Note that some of these distinctions are subject to controversy. For instance, it is often said that System 1 is evolutionarily old and shared with other animals whereas System 2 is more recent and purely human. Some forms of implicit processing such as the human belief system are nonetheless probably recent and not shared with other animals (Goel & Dolan, 2003).

System 1	System 2
Cluster 1 (Consciousness)	
Unconscious (preconscious)	Conscious
Implicit	Explicit
Automatic	Controlled
Low effort	High effort
Rapid	Slow
High capacity	Low capacity
Default process	Inhibitory
Holistic, perceptual	Analytic, reflective
Cluster 2 (Evolution)	
Evolutionarily old	Evolutionarily recent
Evolutionary rationality	Individual rationality
Shared with animals	Uniquely human
Nonverbal	Linked to language
Modular cognition	Fluid intelligence
Cluster 3 (Functional characteristics)	
Associative	Rule based
Domain specific	Domain general
Contextualized	Abstract
Pragmatic	Logical
Parallel	Sequential
Stereotypical	Egalitarian
Cluster 4 (Individual differences)	
Universal	Heritable
Independent of general intelligence	Linked to general intelligence
Independent of working memory	Limited by working memory capacity

Table 1. *"Clusters of attributes associated with dual systems of thinking". Adapted from Evans (2008).*

The controversy over their distinctive characteristics leads to an important point in the definition of those "two" systems. As argued by Stanovich (2004) "System 1" is not a good terminology as it might lead to think that it is a single cognitive system. It is probably rather an "*autonomous set of systems*" that all meet some of the criteria for being considered as a System 1 processes (fast, automatic etc.). The frontiers are not so clear, all the more so since some automatic processes (but not all) of System 1 were first conscious and controlled and became automated (e.g. driving). However, it is also known that we can acquire implicit knowledge without knowing any explicit rule of it (Evans, 2008).

The idea that two different kinds of processes are at the basis of human behaviour is rather old and its origin almost impossibly retraceable. Many diverse domains appealed to such dichotomy to model behaviour, knowledge, memory, attitudes, etc. For instance, in the domain of information processing, one of the most influential theory describes a division of human memory into "*a labile control processes and a learned or inherent structural components*". The authors further postulate that "*Automatic processing is activation of a learned sequence of elements in long-term memory that is initiated by appropriate inputs and then proceeds automatically—without subject control, without stressing the capacity limitations of the system, and without necessarily demanding attention. Controlled processing is a temporary activation of a sequence of elements that can be set up quickly and easily but requires attention is capacity-limited (usually serial in nature), and is controlled by the subject*" (Schneider & Shiffrin, 1977).

In social psychology, the use of direct questionnaire to invest beliefs towards minorities or socially discriminated groups became difficult with the slow development of social norms and the trauma left by the Second World War. Some researchers started to think that although people were not (or less) overtly racist anymore, hidden stereotypes might have survived somewhere in their belief system. They had to develop tests able to reveal the real attitudes of the subjects. Although developed forty years later, the most popular test on attitudes is the Implicit Association Test (Greenwald, McGhee, & Schwartz, 1998). This test measures implicit attitudes towards different "objects" by assessing the automatic associations between this object and an evaluative attribute. The authors reported for instance that people are faster at associating the word "black" with a negative attribute than with a positive one, or with the word "white", although those same subjects were not showing any prejudice on explicit scales of beliefs towards minorities.

Another example is found in the domain of persuasion; the very influential Heuristic-Systematic Model describes two basic modes through which a perceiver will define his attitudes about a new piece of information (Chaiken, 1980). The systematic mode refers to a rather analytic mode in which the subject puts in much cognitive effort to understand the message; he might thus be influenced by the real content of the information. On the other side, the heuristic mode will yield to the subject some heuristics on which he can rely without really processing the information (such as "experts can always be trusted" or "consensus means the right opinion"). The judgment issued from this processing mode will be based upon those heuristics rather than personal opinions or real processing of the information (Chaiken & Trope, 1999).

This dual system approach entered the field of economy mainly through the work of Kahneman and Tversky, who used it to explain many of their astonishing results. Although in psychology one of the most relevant aspects of the theory is the implicit/explicit dimension, the interest changes depending on the field. For instance, in reasoning, the most important distinction might be the use of heuristics (mental shortcuts) versus deeper processing (systematic), whereas in decision-making it might be the automatic versus control dimension.

Typically, Kahneman and Tversky (1973) demonstrated that some classical errors were due to the use of inappropriate heuristics, such as the "representativeness heuristic" which consists of overweighting representativeness and underweighting basic rate when assessing probabilities for an element to belong to a certain group (Kahneman & Tversky, 1973). Another example is the violation of the conjunction rules (the probability to observe simultaneously two events cannot be superior to the probability to observe one of these events alone) because again of the same heuristic (Tversky & Kahneman, 1983). Many other heuristics explaining biases in probabilities estimation have been described, such as the "availability heuristic" and the "adjustment and anchorage heuristic" (Tversky & Kahneman, 1974).

It is important to keep in mind that the situations in which System 1 leads to wrong or inadequate decisions are exceptions. Most of the time, relying on this system helps to behave in accordance with the context, without putting too much cognitive effort into it and without transforming humans into irrational beings. The Somatic Marker Hypothesis is one of the best demonstrations that not relying on intuitions can lead to bad decision-making: the hunches must indeed be listened to. Even if it seems close to the debate on the role of emotions in decision-making, the debate is broader here. Nonetheless, emotions or affective processing are often associated to System 1. Other examples without the emotional factor also show that System 1 is normally a good leader. For instance, a study on groups such as paramedics and fire officers showed that in emergency situations, the expert retrieves corresponding schemas of actions that help to take fast decisions and find solutions (Evans, 2008). There is sometimes explicit reasoning in the application, but the key to success is this process of automatically finding the schemas corresponding to the situation.

Finally, some factors have been linked to the facility of use of each system. For instance, time pressure (Finucane, Alhakami, Slovic, & Johnson, 2000), involvement in two cognitive tasks simultaneously or being in a good mood favour the use of System 1 (Kahneman, 2003). On the other hand, the facility of System 2 has been more often correlated with inter-individual differences such as exposure to statistical

thinking (Nisbett, Krantz, Jepson, & Kunda, 1983), intelligence (Frederick, 2005) or "need for cognition" (the extent to which people enjoy effortful reasoning, Shafir & LeBoeuf, 2002).

Dual system in Neuroeconomics

Obviously, as neuroeconomics tends to understand all the neural processes underlying decision-making, the idea of defining those two systems in the brain has recently received a great interest in neuroscience (Evans, 2003, 2008; Kahneman, 2003; Lieberman, 2007; Loewenstein, Rick, & Cohen, 2008; Rustichini, 2008; Sanfey & Chang, 2008; Todorov, Harris, & Fiske, 2006). Still, dual system evidences at the level of the brain are scarce and the results are often controversial and/or not fully convincing.

For instance, Sanfey et al's study (2003) has been referred to as a demonstration of dual system evidence in the brain, as both the DLPFC and the ACC have been linked to System 2 and the anterior insula to System 1 (Lieberman, 2007). On the other side, as those structures are involved in both accept/reject decisions (it is their relative activation that determines the decision) they can also be considered as parts of one system. Indeed, demonstrating the existence of two different systems implies that at least one structure must be present in one system and not in the other, if an overlap between the systems is admitted.

The study from de Martino and his colleagues (De Martino, Kumaran, Seymour, & Dolan, 2006) is probably the closest to being clear evidence for dual system process (Kahneman & Frederick, 2007). The authors studied the "framing effect", i.e. the sensitivity to the context in which a problem is presented to the subject. As mentioned earlier, System 1 is supposedly very sensitive to the context contrary to System 2. They showed that activation of the amygdala was greater when subjects choose according to general behavioural tendency (risk-averse in gain condition and risk-seeking in loss condition) whereas ACC activity was enhanced when subjects played against this tendency. They next classified the subjects according to a "rationality index" (based on their behavioural sensitivity to the framing effect), and found a positive correlation between this index and the activity in the right orbital and medial frontal cortex (OMPFC) and in the ventromedial prefrontal cortices (VMPFC). This is an elegant demonstration of the involvement of the amygdala in the use of heuristics (when subjects show risk-averse and risk-seeking behaviour according to the context). At the same time, ACC activity appears when subjects run counter this heuristic reasoning, suggesting that a conflict is detected between the tendency to either use heuristics or higher level reasoning. The OMPFC might integrate emotional and cognitive inputs helping a more "rational" behaviour to emerge, as rationality index was

correlated with sensitivity to the framing effect. Note that the results of this study have been contradicted by another study in which losses and gains were coded by almost the same areas (striatum, medial OFC, ventromedial PFC, ventral ACC) and where no structure traditionally linked to negative emotions (insula, amygdala) was active. The authors insist on differentiating anticipated, experienced and decision utilities to account for those distinctions. The mechanisms which come into play when anticipating or experiencing a gain or a loss might be different from the mechanisms triggered at the precise moment when the decision needs to be taken (Tom, Fox, Trepel, & Poldrack, 2007).

In the domain of intertemporal choice other evidence has been reported. For instance, one experiment focused on inequity in time discounting (McClure, Laibson, Loewenstein, & Cohen, 2004): deciding between receiving 1 $ today or 2 $ tomorrow should be equivalent to deciding between receiving 1 $ in one year or 2 $ in one year and a day. However, people traditionally chose to receive 1 $ immediately in the first case, but rather 2 $ in one year and a day in the second case. The authors demonstrate that some limbic structures rich in dopamine innervations are preferentially activated in the case of immediate rewards than in trials with no reward. They also found that when areas in the prefrontal cortex were more active than limbic structures subjects tend to choose the long-term reward. To answer to serious doubts cast by their use of gift certificates as rewards (not immediate consumption) the authors re-ran the experiment and replicated the previous results with the use of juice and thirsty participants (McClure, Ericson, Laibson, Loewenstein, & Cohen, 2007). However, we are confronted to the same problem of interpretation here: the structures involved in both cases (immediate and long-term rewards) are identical, so they can be seen as one system in which the respective activation of each element determines behaviour. Other studies in this domain have found dual system processes (Knutson, Rick, Wimmer, Prelec, & Loewenstein, 2007), but all those results are nonetheless matter of debate and controversy (see for example Glimcher, Kable, & Kenway, 2007).

Finally, in the domain of moral dilemmas, a conflict between affective and deliberative processes has been suggested when making a decision. For instance, people are ready to kill one person by pushing on a button if it saves the lives of five other people. However, in the same situation, people refuse to push somebody off a bridge (instead of pressing a button) to save the same five lives (Thomson, 1985). This paradoxical behaviour might find its source in a conflict between deliberative judgment (which is easier when the action is to press a button rather than pushing someone off a bridge) and affective processing (enhanced by the fact that pushing someone is emotionally extremely disturbing compared to pressing on a button). Indeed, the authors tested if personal moral dilemmas (involving personally harming someone) activated different areas than “impersonal” moral dilemmas (Greene, Sommerville,

Nystrom, Darley, & Cohen, 2001). They found that personal dilemmas were linked to a greater activation of classical emotional networks than impersonal dilemmas and thus suggested that the emotional response elicited by personal moral dilemmas must be cognitively overcome if the subject wants to make a decision that is contrary to his emotions. This dual system approach accounts also for results on patients with ventromedial prefrontal cortex who make more utilitarian judgments than control population (Greene, 2007; Koenigs et al., 2007).

With this large body of data, Liebermann (2007) proposed a classification of the areas depending on their belonging to X-systems (X for reflexive) and C-systems (C for reflective). According to him, X-system is composed of the amygdala, the basal ganglia, the ventromedial prefrontal cortex, the lateral temporal cortex and the dorsal anterior cingulate cortex (dACC). C-system comprises the rostral anterior cingulate cortex (rACC), the lateral and medial prefrontal cortex, the lateral parietal cortex and the medial temporal lobe, including the hippocampus but not the amygdala (Figure 3). Although reviewing social cognitive science paradigms, Lieberman uses the terminology (reflexive and reflective) that usually relates to reasoning. This demonstrates the overlap between the descriptions of dual system models in different discipline (Evans, 2008).

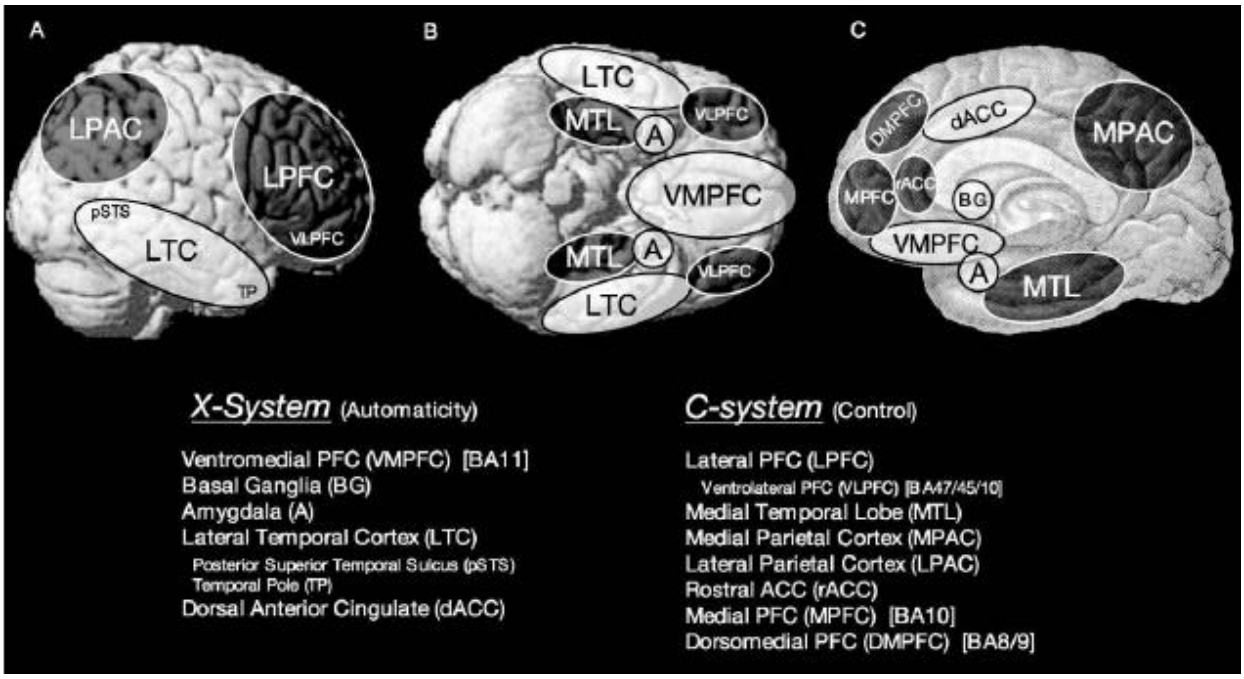


Fig3. *Hypothesised neural correlates of the C-system supporting reflective social cognition (analogous to controlled processing) and the X-system supporting reflexive social cognition (analogous to automatic processing) displayed on a canonical brain rendering from (A) lateral, (B) ventral, and (C) medial views. Adapted from Lieberman (2007).*

Finally, it is still unknown how those systems interact to influence behaviour (Loewenstein et al., 2008). Evans (2008) refers to two different hypotheses as follows: the "parallel-competitive" hypothesis supposes that both systems compete to influence behaviour each one in a different way. The "default-interventionist" hypothesis defends that one system is the default mode and is overridden by the other system only if necessary. Those hypotheses have not been tested empirically until now.

Results of neuroeconomics studies did in turn inspire economic models which integrate the idea that judgment and behaviour are the result of the interaction of multiple processes (see for instance Benhabib & Bisin, 2005; Bernheim & Rangel, 2004; Brocas & Carrillo, 2008; Fudenberg & Levine, 2006).

We have seen that decision-making can either result from a rather fast reaction to a situation or from a deeper analysis of it. Emotions can occur at any point during this process; they can be thwarted or not, and can as well be facilitating or disruptive elements. Indeed, the mild and maybe even unconscious hunches described in the Somatic Markers Hypothesis are at many levels clearly distinguishable from a rush of blood to the head under which decision is completely biased and probably not in a favourable way.

Interestingly, far from the debate on dual system, some researchers attempted to explain the deviations from expected utility by an aspect of emotions that must be tackled before concluding this introduction.

The underrated power of apprehension

Δεν ελπίζω τίποτα. Δε φοβούμαι τίποτα. Είμαι λεύτερος. [*]

Indeed, an often underestimated (or voluntarily excluded) epiphenomenon of emotions is the power they exert only by being anticipated. Negative emotions such as regret or disappointment might strongly influence decision-making by generating avoidance behaviour. In a formal attempt to integrate this aspect in decision-making models, two groups developed regret and disappointment theories which took into account the power of apprehension. First, Bell (1985) formalized disappointment as "*a psychological reaction to an outcome that does not match up to expectations; the greater the disparity, the greater the disappointment*" (its positive counterpart being called elation). He added that "*People who are particularly averse to disappointment may learn to adopt a pessimistic view about the future*". Disappointment is different from regret, the latter occurring when one can compare the outcome of the selected option to that of the rejected one. A second model was proposed less than one year later (Loomes & Sugden, 1986). Both models have in common the idea that subjects anticipate "post-decisional emotions" such as regret and disappointment, and thus take them into account at the moment when the decision is made. They diverge principally in considering or not that this is "irrational": "*In our theory, people seek consistently to maximise expected satisfaction, where that expectation includes the anticipation of possible disappointment and elation. We cannot see any reason for regarding such a maximand as irrational; nor do we think that any simple experience of satisfaction, whatever its source, can be designated either rational or irrational*." (Loomes & Sugden, 1986).

More than one decade later, a team of researchers in the Netherlands made a first attempt to integrate some notions of regret and disappointment theories with the notions of counterfactual thinking issued from norm theory (Kahneman & Miller, 1986). Counterfactual thinking is the "*psychological process of comparing the obtained outcome with other possible outcomes*" which implies "*mentally mutating one or more aspects of a past event*". Depending on the element which is mutated in this process of rerunning a past event, the experienced emotion can be different (Zeelenberg et al., 1998).

For example, whereas shame is linked to counterfactual thoughts in which the mutated elements are related to the self, guilt is associated with the mutation of elements concerning the behaviour (Niedenthal, Tangney, & Gavanski, 1994). The results of Zeelenberg's study indicate that, in the same

[*] *I hope for nothing. I fear nothing. I am free.* Epitaph on the grave of Nikos Kazantzakis, Heraclion, Crete

vein, regret is most often related to behaviour-based counterfactual thoughts (the actions of the subject are mutated) whereas disappointment is related to counterfactual thoughts in which aspects of the situation (beyond the control of the subject) are changed (Zeelenberg et al., 1998). This confirms Bell's (1985) definition: disappointment being the difference between the expected outcome and the real outcome, and the only way to change the real outcome not depending on the subject, those definitions coincide. Regret, on the other hand, arises when the subject compares the outcome of the chosen option to the outcome of the rejected one. In this case, in order to change the result, the element to mutate in the past event is the subject's own action (chose the other option), which also implies more responsibility than in the case of disappointment (Zeelenberg, van Dijk, Manstead, & van der Pligt, 2000). The authors conclude that although their own research was focused on counterfactual thoughts about past events, the process might be similar for future events, as disappointment and regret theories also mention the impact of the anticipation of these emotions on decision-making. Their prediction is that the anticipation of disappointment is closely related to risk aversion, and that "*people who pre-compute situation-focused counterfactuals will anticipate this disappointment, and will consequently reduce the amount of risk they are willing to take*".

A series of studies had already been published during those years, mainly on the effect of anticipated regret (for an excellent review see Zeelenberg, 1999) but also on the effect of the anticipated pleasure originating from a positive outcome (Mellers, Schwartz, & Ritov, 1999). Still, no empirical research had been conducted on disappointment, thus, Zeelenberg's team hypothesis remained to be tested until a pioneering paper (van Dijk, Zeelenberg, & van der Pligt, 2003).

The participants (psychology students) had first to answer to a questionnaire including two questions which asked them to rank the relevance of a career in psychology and the relevance of a career in law. Then, supposedly unrelated to the first task, the participants filled out an "intelligence test" composed of some items of a classical IQ test. Participants in the *self-relevant* condition had this test presented as very relevant for being a good psychologist, whereas in the *self-irrelevant* condition, they read that the results of this test were very relevant for lawyers. The variable relevance was crossed with another variable, the feedback (the test score), which was either given in 30 minutes, or received at home two weeks later. After this test, the participants had to report how good they thought they were at the intelligence test. Thanks to a clever manipulation, the experimenters obtained twice this answer from each participant with an interval of 30 minutes between the two acquisitions. The results show that the participants in the *self-relevant* x *imminent feedback* conditions significantly decreased their estimates about their own performance between the two acquisitions, and were the only ones to do so. In other

words, the authors infer that lowering one's expectations is a way of avoiding disappointment, but this strategy is used only when the context is relevant for the subject, and only when the comparison between the expectations and the reality is imminent. Interestingly, the authors mention other strategies such as investing extra-effort to obtain the desired outcome, or changing *a posteriori* the probabilities of occurrence of an event so that disappointment looks unavoidable (Tykocinski, 2001). However, when the threat of being disappointed is imminent, and as disappointment is the gap between the expected and the real outcome, diminishing this gap by lowering expectations seems an efficient and easy way of avoiding this negative emotion. It is nonetheless dangerous, as at its extreme extent it can lead to feelings of helplessness (if no aspects of the situation can be mutated, see Seligman & Maier, 1967).

Regret and disappointment as anticipated emotions have almost never been studied in neuroscience. One of the only studies comprising economic games, neuroscience and regret demonstrated the involvement of the orbitofrontal cortex (OFC) in the experience of regret (Camille et al., 2004). In this study, the authors used a task in which control subjects and brain-damaged patients had to choose between two gambles. In a *complete feedback* condition, the results of both the chosen and the rejected gambles were displayed. In a *partial feedback* conditions, subjects knew only about the chosen option's outcome. Although able to emotionally respond to absolute gains and losses, OFC patients were not sensitive to the value of the unchosen outcome, they seemingly did not feel regret at all. A model was then tested to determine the influence of anticipated regret (and disappointment experienced in the *partial feedback* condition). Anticipated regret came out as a determining factor in normal control subjects whereas OFC patients' decisions were based only on the expected values of the gambles. Disappointment anticipation was not significant for either group. Disappointment and regret being differentially affected in control subjects and in patients, the authors concluded that these two emotions must be generated by distinct neural networks.

The involvement of the OFC in the experience and in the anticipation of regret has been confirmed with the use of neuroimagery (Coricelli et al., 2005; Coricelli, Dolan, & Sirigu, 2007). The paradigm was very similar to the one described above. Regret was measured in the *complete feedback* condition as the discrepancy between the actual outcome and the outcome of the unchosen gamble. Disappointment was measured in the *partial feedback* condition as the difference between the two possible outcomes of the same gamble, if unfavourable to the participant. The magnitude of disappointment was correlated to enhanced activity in the middle temporal gyrus and in the dorsal brainstem (linked to the processing of aversive stimuli such as pain). Regret was correlated to the activity in the medial OFC as well as dorsal anterior cingulate cortex and anterior hippocampus activity. The authors also tested a model of choice in

which the anticipation of disappointment was integrated. When subjects did not choose to maximize expected value but rather to avoid disappointment, a greater activation of the substantia nigra was found, an element of the dopamine network previously mentioned. This activity was also observed after the outcome, when the subjects assessed gains and losses in terms of prediction error (more active for gains and relatively deactivated for losses). When subjects decided to minimize future regret instead of maximizing gains, a greater activity was found again in the substantia nigra but also in the dorsal anterior cingulate. Finally, the cumulative effect of regret was observed in a modulation of OFC, amygdala, inferior parietal lobule and right somatomotor cortex activity. Thus, the authors deducted that anticipation and experience of regret share common neural substrates. Anticipation might involve reactivating the processes underlying experienced regret. The results of those studies combined with other findings on the OFC suggest that the OFC is not only involved in the pre-decisional phase but rather at all levels of the decision-making process. Indeed, the OFC might modulate the basic responses to losses and gains, integrating all cognitive and emotional elements in the process of decision-making.

Recently, the subjective experiences of regret and disappointment have been compared while subjects were scanned during a gambling task (Chua, Gonzalez, Taylor, Welsh, & Liberzon, 2009). The authors used the same *partial/complete feedback* paradigm to induce either disappointment or regret (and their positive correlates, i.e. elation and rejoice). They report that both regret and disappointment activated the anterior insula and the dorsomedial prefrontal cortex, but that regret generated higher activations. This corresponded to the behaviour of the subjects as they reported greater dislike of their choices (and the outcome of their choices) in the cases of regret compared to disappointment. Note that this seems natural as in the disappointment condition they were not informed about the outcome of the other option. On the contrary to Coricelli and al's study (2005), they did not find that regret activates the medial OFC (but lateral OFC) which might be explained by differences in the experimental setting. This study did unfortunately not investigate these emotions in their anticipating role. It nonetheless suggests once again that when studied together, disappointment seems to have less effect than regret.

Brain signals linked to feedback processing

Where is the what if the what is in why?[*]

As mentioned earlier, although there are almost no studies on disappointment, this emotion as formally defined by Bell cannot be completely distinguishable from the phenomenon of reward prediction error linked to the dopamine system. In the Trust Game for instance, the mismatch between the expected and the real outcome might have similar signature in the brain as a feedback error.

On this topic the literature is vast and complex as it comprises research from very different fields. On one side, neurophysiological studies on dopaminergic single neuron describe mechanisms such as change in firing rate according to the expected presence or not of a reward (Fiorillo, Newsome, & Schultz, 2008; M. Matsumoto & Hikosaka, 2009; Schultz, 1998; Zaghloul et al., 2009). On the other side, electroencephalography studies have long ago described an event-related potential, the Feedback Related Negativity (FRN) that appears on fronto-central electrodes following feedback error and which supposedly reflects activations of the ventral striatum, the anterior cingulate cortex and medial prefrontal cortex (see Falkenstein, Hoormann, Christ, & Hohnsbein, 2000 for a tutorial). Finally, neuroimagery (fMRI) studies report blood-oxygen-level dependant signals in the striatum which also supposedly reflect reward/error prediction processing (see for instance Hester, Barre, Murphy, Silk, & Mattingley, 2008; Samanez-Larkin et al., 2007; Seymour, Daw, Dayan, Singer, & Dolan, 2007). Many debates are related to this theme.

First, there are various interpretations of those "error" signals, and theories such as error detection, conflict detection or reinforcement learning still fight against each other for the best explanation. Briefly, in the error detection theory, the ERN (a component similar to the FRN but more specific to motor action errors) reflects the mismatch between required and executed responses. Indeed, it was first considered as belonging to a group of mismatch-related signals such as the Mismatch Negativity (Falkenstein et al., 2000; Näätänen, Gaillard, & Mäntysalo, 1978). In the conflict detection view, the ACC uses this signal to monitor action and exert cognitive control over the behaviour (Botvinick et al., 2001). The reinforcement learning theory holds that FRN reflects the impact of negative prediction error signal on the anterior cingulate cortex. This signal is conveyed by the midbrain dopamine system and generated as a reward

[*] Moloko, *Where is the What if the What is in Why?,* 1995

prediction error when outcomes are worse than expected (Holroyd & Coles, 2002). Again, ACC will modify performance as a function of this signal.

Second, questions about the way and the precise location where these signals are processed do also shake the community: is there a network whose activation reflects positive rewards and de-activation negative signals (reward prediction error)? Or are those signals processed in two different networks? Are we sure that those areas are not rather responding to the expectancy dimension (for a glimpse on the debate, see Hajcak, Moser, Holroyd, & Simons, 2006; Holroyd & Krigolson, 2007; Holroyd, Nieuwenhuis, Yeung, & Cohen, 2003; Nieuwenhuis, Holroyd, Mol, & Coles, 2004; Yeung & Sanfey, 2004)?

To illustrate this, here is a sample of conclusions drawn in one year only: gains and losses are represented in a similar manner but in slightly different regions of the striatum (Seymour, Daw et al., 2007); the signal in posterior medial prefrontal cortex is greater for errors subsequently corrected than for errors that were repeated (Hester et al., 2008); neural mechanisms of feedback processing differ between gains and losses (Cohen, Elger, & Ranganath, 2007); activity in the ACC reflects the salience of a new piece of information, as well as variation in its signal reflects individuals' learning rates (Behrens, Woolrich, Walton, & Rushworth, 2007); the amplitude of FRN after losing against a computer predicts whether the subject will change decision behaviour on the following trial (Cohen & Ranganath, 2007); the FRN reflects the gain/loss dimension whereas another component, the P200, reflects the expected/unexpected dimension (Polezzi, Lotto, Daum, Sartori, & Rumiati, 2008). A study on monkeys with electrodes implanted in the ACC showed that its activity signals violated expectations but also discriminates between losses and gains (Sallet et al., 2007).

There is to our knowledge only one EEG study on the Ultimatum Game (Polezzi, Daum et al., 2008). The authors observed an FRN and concluded that it distinguishes between fair and other kinds of offers (mid-values and unfair), but their definition of FRN might not be compatible with the one used in other studies.

Indeed, to add complexity to this puzzling domain, EEG studies are hardly comparable as the definition of the components - as well as the parameters used to process the electrical data - is extremely variable. In addition, the source localization algorithms are often used in such a way that they yield inconsistent results. Moreover, it has been shown that cognitive strategies can regulate both physiological (skin conductance) and neural (striatum) signals (Delgado, Gillis, & Phelps, 2008), which

complicates again a clear comparison of the results yielded by different paradigms. Recently it has been suggested to denote Outcome-Related Potentials (Negativity/Positivity) all the signals that are linked to outcomes (Kamarajan et al., 2009). This terminology presents the advantage to encompass both the quality (gain/loss) and the quantity (small/large) dimensions. This is a first step, but maybe there should first be a consensus on the definition itself of those components.

The study of oscillatory activity represents another technique to investigate the neural correlates of decision-making and feedback processing, more precisely how different areas communicate in different frequency bands. This adds interesting information to the debate, as it might help to disentangle the nature of those signals. For instance, using wavelet-based time-frequency analysis, a study revealed the distinction between two mediofrontal oscillatory components. The first one in the beta range (12-30 Hz) was associated with gains whereas the second one, in the theta range (4-8 Hz) was associated with losses. The presence of high frequency oscillations (beta band) is interpreted by the authors as indicating that positive feedback is mediated by those oscillations that couple fronto-striatal areas involved in reward processing (Marco-Pallares et al., 2008). Moreover, another study reported that disadvantageous decision-making in the Iowa Gambling Task was associated with an increase in the theta/beta ratio (Schutter & Van Honk, 2005). High ratios might thus indicate weak inhibitory control over motivational drives, leading to an increased reward-dependency and a reduced sensitivity to punishment. A previous study with the Iowa Gambling task had reported that an alpha-band component reflects the mismatch between expected and actual outcomes (Oya et al., 2005). Recently, another study questioned more precisely the role of theta band in feedback-related processes: is it associated to conflict between responses, reinforcement learning processes or error detection (Tzur & Berger, 2009) ? Their results indicate that theta rhythms reflect general evaluation mechanisms, not only evaluation of motor action as in the ERN. The authors propose that the same mechanism is involved when comparing an expected to a presented stimulus than when comparing an expected to a performed action. The greater the mismatch, the greater the neural power and phase synchrony in the theta band will be. Some months later, another study focused on the generator of the theta band activity linked to feedbacks (Christie & Tata, 2009). The authors reported that theta signals following gains and losses originate from right medial prefrontal cortex, and possibly anterior cingulate. They moreover claimed that those generators of feedback-induced theta are anatomically different from the generators of the FRN.

It is maybe needless to say that the debate is still open, and that a combination of different techniques might be the ideal way to unravel where, how and why the processing of “physiologic disappointment” occurs.

A final word

We have seen thus far that humans are not rational agents as expected by the models issued from game theory. For instance, emotions play a complex role in decision-making; they can be disruptive or of a great help, depending on many factors. This has been demonstrated by a great body of research in neuroeconomics. We have also seen that the biases in decision-making are not the sole result of the impact of emotions but in a much broader perspective, they can be explained by the fact that people generally rely on fast and automatic ways of processing information. This is normally very helpful and efficient, but the system it not infallible, which is why a second system is supposed to take control of the situation (of the behaviour) when needed. Finally, a last aspect has been described: the mere anticipation of negative emotions as a strong biasing factor. Many studies have been conducted on the power of anticipated regret, and sometimes on anticipated disappointment. Unfortunately, disappointment has recently seemed to be the poor cousin of regret, as the effects of its anticipation are always lesser. This is surprising as behavioural studies as well as mathematical models predict that anticipated disappointment might have an impact on decision making.

It is precisely in this literature that we found the breach in which we inserted our first study. The following section presents the questions under study for each experiment of this thesis.

Questions under study

Study 1: The Impact of Disappointment in Decision-Making: Inter-Individual Differences and Electrical Neuroimaging

Although literature on disappointment and regret exists, both emotions are often studied concomitantly. This is a strong limitation as regret implies personal "mistakes" whereas this notion of responsibility is much less present in the case of disappointment. Thus, when experiencing both emotions in the same study, disappointment always seems less painful. That is maybe why when modelling these emotions no effect is found for disappointment alone as an anticipated emotion. To address this question, we designed an experiment in which disappointment alone - and its positive correlate, elation - were under study. We hypothesised that disappointment has a biasing effect as well as an anticipated emotion but that this effect is usually hidden by the powerful feeling of regret. Second, regarding the vast literature on social games, we also hypothesised that this effect will be observed only when playing against a human counterpart and not against a computer.

Study 2: Validation of the association between the presence of the fronto-central map and a lower sensitivity to previous disappointment (a higher DTT)

The first study led to unexpected results and changed the course of this thesis. Indeed, we found that some subjects were highly sensitive to disappointment and showed a strong bias in their behaviour following the disappointing trials. They lowered their expectations about the following trial as a function of previous disappointment. Another group of subjects were much more resistant to previous disappointment, and were more successful throughout the game. Those subjects relied on two systems (that we interpreted as one automatic and one controlled) whereas the "impulsive" first group seemed to rely only on one system (interpreted as automatic). A necessary step (and elegant demonstration) to validate our interpretation of the previous results would consist in being able to demonstrate this link ("rational" behaviour – controlled system) in the "sensitive" participants. For that, we invited the subjects of Study 1 to participate to a second study, and compared the behaviour and electrophysiological activity of the "sensitive" subjects across the two studies. We hypothesised that if their "irrational" behaviour was pointed out it might be voluntarily modified. Moreover, if our conclusions were correct, this would be reflected by the use of System 2.

Study 3: Knowing what to do but doing the opposite: Rejection of unfairness in the Ultimatum Game

In this study the aim was to assess to what extent it was possible to have a homogeneous group of subjects relying either on System 1 or on System 2 depending on the frame. More precisely, we decided to use the debate around human "real" nature in the Ultimatum Game to test the existence of both systems in a different task. We hypothesised that if the real tendency of the subjects was to accept all offers, but manners or social norms were at the basis of the rejection of unfair offers, letting the subjects know the game theory's predictions would allow them to take advantage of this situation and accept all offers as they would instinctively do. On the other hand, if people are really upset by unfairness, but the fact of knowing game theory's predictions nonetheless lead them to accept unfair offers, this behaviour should be accompanied by signs of conflict detection and/or control over behaviour. By accepting unfair offers for the sake of "rationality", a mechanism of control over emotions should appear, and this might be the reflection of the intervention of System 2.

Study 4: Coding mechanisms in Local Field Potentials (LFPs) behind disappointment/elation asymmetrical effect

Although inter-individual differences in reaction to disappointment appeared in Study 1, some common structures might nonetheless code disappointment before a subgroup of subjects exert control over their behaviour. This (these) common structure(s) might code differently negative from positive and neutral outcomes as the behaviour is subsequently modified depending on the outcome category. The frequency bands and the timing of neural oscillations constitute one of the richest coding mechanisms used by living systems. In this last study, we were interested in better understanding the asymmetry between the effects of disappointment and elation on behaviour, with the hypothesis that an identical asymmetry should be observed in neural signals. To provide a precise picture of these mechanisms we relied on the wide band time-frequency analysis of intra-cranial recordings made on three patients evaluated in the Presurgical Unit of the Geneva University Hospitals between June 2008 and August 2009.

STUDY 1

THE IMPACT OF DISAPPOINTMENT IN DECISION-MAKING: INTER-INDIVIDUAL DIFFERENCES AND ELECTRICAL NEUROIMAGING

In this first study, we wanted to test the hypothesis that disappointment and its anticipation can generate the effects previously described (van Dijk et al., 2003) when it is studied independently of regret. We also wanted to study the neural correlates of this phenomenon.

More precisely, we predicted that in a multi-round Trust Game, a strongly disappointing outcome might lead subjects to diminish their expectations about the following trial in order to avoid experiencing disappointment again, while always playing with a new Trustee in every trial. In addition, we hypothesised that this effect should not be observed if participants play against a computer. In this latter case, they should rather opt for strategies such as investing extra cognitive effort in the task, which can be reflected by longer reaction times or a more exploratory behaviour.

Methods

Participants and Experimental Design

Thirty-two healthy young volunteers (mean age 25 $\pm$ 3.8, 17 females) were recruited by advertisements posted at different faculties of the University of Geneva. Four of the participants were left-handed. They had no history of neurological problems. The whole experiment was approved by the local ethics committee (Geneva University Hospitals). A written informed consent was signed by all participants before starting the experiment. After signing the informed consent, the participants read the instructions.

Inducing and quantifying disappointment through the Trust Game.

In our version of the TG (Figure 1), participants (Investors) had to decide how much money out of an initial endowment (10 CHF/trial) they wanted to share with an unknown opponent (Trustee) on the sole basis of a picture of his/her face presented for half a second. The faces presented to the subjects were all displaying a neutral expression and were selected from the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development

Program Office (P.J. Phillips, 1998). The amount invested was tripled and given to the Trustee, according to the classical Trust Game rules.

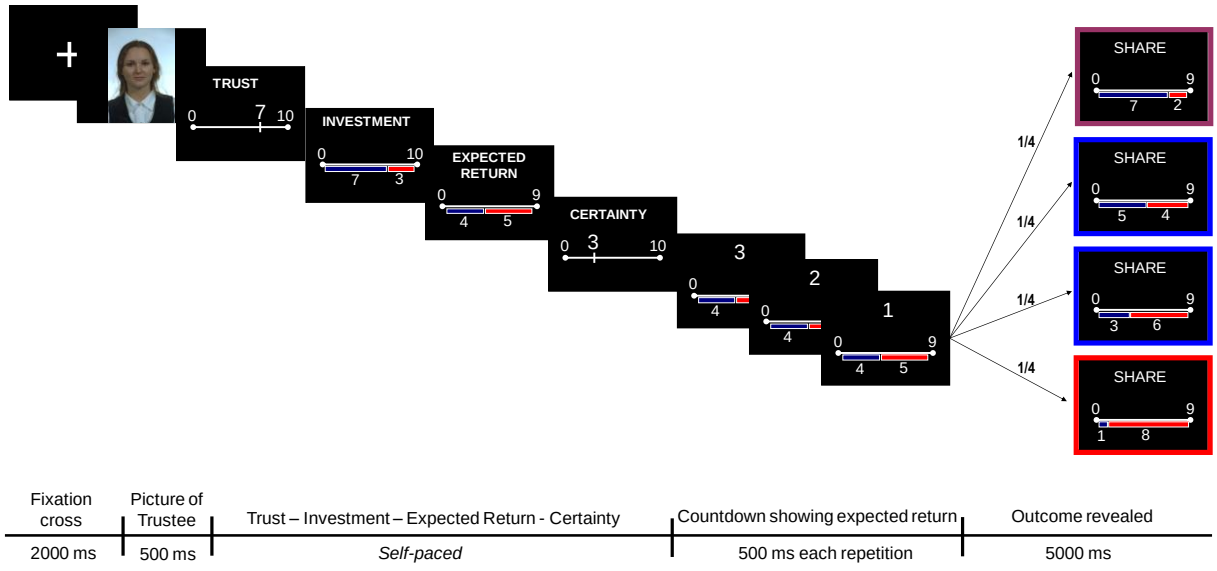


Fig1. Time course of one trial: *Each trial begins with a picture of the Trustee (500 ms.). The Investor has to rate on a 10 points-scale how trustworthy he assesses the face to be. Then, the Investor decides how much he is willing to invest in this trial (10 points- scale). The amount invested is tripled and given to the Trustee. After this exchange, the Investor indicates how much he thinks that the Trustee will give him back, on a scale going from 0 to 3 times the investment. Finally, the Investor has to report his self-confidence about all the choices he has been making during this trial (10 points-scale). While waiting for the outcome (the final share = the Trustee's decision), a countdown screen flashes three times showing the expected return, so that the Investor keeps his expectations in mind. The outcome is then revealed, framed in different colours depending on the outcome. The Investor plays 240 times, each time against a different Trustee, to whom the re-payment is randomly assigned.*

In this version of the Trust Game, we asked the Investor to provide numerical estimates of the following variables:

1) *Trustworthiness (TR):* how trustworthy is the Trustee; 2) *Investment (INV):* how much of his/her endowment is to be shared with the Trustee; 3) *Expectation (ER):* how much does s/he hope to get back from the Trustee (the share) and 4) *Confidence (CF):* to what extent is s/he certain about these decisions. The scalp electroencephalogram (EEG, 64 channels) and reaction times (RTs) were recorded during all steps of the game.

According to the expectations of the Investor (value entered as Expected Return), we manipulated the outcomes (the money given back by the Trustee to the Investor) to elicit different emotions in the Investors, namely: 1) elation (outcomes much larger than expected, 60 trials); 2) disappointment (outcome much lower than expected, 60 trials); 3) neutral-positive feeling (outcome similar to

expectations, slightly more than expected, 60 trials); 4) neutral-negative (slightly less than expected, 60 trials). We used the disparity, i.e., the difference between the actual outcome and the expected one, as our quantitative measure of disappointment/elation for each trial. Cut-off levels to determine these four categories were set prior to the experiment as following: neutral negative (letting INV denote investment, if ER>INV, interval = [ER-(0,4*INV);ER], or interval = [ER;(ER*0.4)]); neutral positive (if ER>INV, interval = [ER;ER+(0.4*INV)] or [ER;ER+(0.4*ER)]); negative/ disappointing (outcome much lower than expected; interval = [0; smallest value (even INV or ER)/2]); and positive/elation (outcome much higher than expected; if ER>INV: interval = [ER + (possible maximum – ER)/2;maximum possible] or interval = [2*INV;2.5*INV]). Consequently, the Investor played 240 trials. For each trial, a new picture (face of the Trustee) was randomly assigned to the outcome, unknown to the Investor.

From the 32 participants, 19 played a variant of the game named Experimental Condition (EC). They were told that the Trustees with whom they were going to play had already participated in a similar study so that for each possible value of investment we (the experimenters) already knew how much the Trustees were willing to give back in return. Thus, the participants thought that they were playing with human opponents although not in real-time. The remaining subjects (13) played a version of the game named Control Condition (CC) in which they were explicitly informed that we programmed the game according to predefined rules. The development of the game and the images presented were identical in both conditions to minimize potential differences in EEG patterns. Participants were told that they would be paid according to their performance, thus the final goal of the experiment was to maximize their gains. For the EC participants we asserted that this goal could only be reached by an accurate assessment of the trustworthiness of the Trustees. For the CC participants it is by understanding the encoded strategy that they would be able to fulfil the long-term purpose of maximizing their gains.

To test if participants in the EC lowered their expectations as a function of previous disappointment we defined a measure of the change in strategy (CS) during the experiment. We used the difference between the value of the previous trial and that of the current trial for the 3 variables under study: trustworthiness, investment and expectation. A negative value of the CS thus indicates that the Investor has lowered his trustworthiness rating/ investment/ expectation in the actual trial compared to the previous one, whereas a positive value means that the he increased his trustworthiness rating/ investment/ expectation in the current trial compared to the previous one.

At the end of the session we asked the participants to fill in a questionnaire in which they had to report: a) the emotion that best described what they felt in the cases where they invested substantially but

received little in return; b) if they thought that there was a link between their assessment of the faces and the outcome, or if they felt that it was random; c) if they thought that they were influenced by previous outcomes experienced throughout the game.

This experiment was run using Cogent 2000 developed by the Cogent 2000 team at the FIL and the ICN and Cogent Graphics developed by John Romaya at the LON at the Wellcome Department of Imaging Neuroscience.

EEG recordings

The EEG was recorded at 1024 Hz (5th order sinc filter with a -3 dB point at 1/5th of the sampling frequency) using 64 BioSemi sintered Ag-AgCl electrodes. The electrodes were mounted on the manufacturer-provided cap according to an extended 10-20 system. The Biosemi system uses a common mode sense (CMS) active electrode as the reference transformed in our case to the average reference during offline analysis where artefact-contaminated trials were rejected. Epochs of 1100 ms (100 ms baseline) were extracted after notch filtering at 50 Hz and superior harmonics. No baseline correction was applied. Averages of the epochs were computed for every subject and outcome. Bad electrodes interpolation was based on spherical splines.

Seven subjects were excluded from the analysis because of excessive (more than 1/3 of the trials) contamination of their EEGs by artefacts (as determined by visual inspection of raw EEG. Electrophysiological results are then based on a sample of 25 subjects (14 of the EC and 11 of the CC).

Data Analysis

All analysis for EEG and behavioural data were carried out using custom-built Matlab routines. To test the similarity or dissimilarity of the topography of instantaneous scalp maps we used the absolute value of the correlation coefficient between maps. The correlation coefficient can be then used as the test statistics to build a non-parametric multivariate randomization test to evaluate the statistical significance of differences between instantaneous scalp maps within two or more groups. This test, known as a randomized one-way multivariate analysis of variance (Mielke, 2001), can be applied to each measured topography, i.e., to each time frame. In this test, we assume that no differences exist between the groups (hypothesis Ho) and compute the distribution of the test statistic by permuting members from one group to the other 1500 times.

Note that the ERP topography (as a landscape) is reference-independent. A reference is a constant value added or subtracted to the voltage instantaneously recorded at each electrode. As such, a reference can modify the absolute value for each map (the individual waveshape and components recorded at each electrode) but never the topography of the map. The relative ranking of the values - i.e. the landscape - remains unaffected by the reference and therefore the topographies can be compared to the ones reported in other studies. Finally, the statistical analyses based on topographical comparisons inherit this property of independence: changing the reference cannot modify the obtained results. Since electric fields at the scalp bear a linear relationship with the underlying generators, non-trivial differences in scalp topography necessarily reflect differences in underlying generators. Trivial differences should be understood in this context as simple polarity inversions or scaling (multiplicative) factors.

Results

Behavioural Results

When asked to evaluate their feelings after adverse unexpected outcomes, the vast majority (30/ 32) of the subjects spontaneously reported negative emotions, with disappointment being the prevalent one. In a forced-choice version of this question, they restated their answer by choosing disappointment (15/32 subjects) over anger (7/32) and betrayal (6/32). A complete overview of the questionnaire is given in Supplementary Table S1.

Highly significant linear correlations ($P<10^{-8}$) were found between trustworthiness ratings and investments (mean r = 0.65 in EC and 0.61 in CC), and between trustworthiness ratings and expected returns (mean r = 0.72 in EC, 0.69 in CC). Correlations were particularly strong between investments and expected returns (mean r = 0.79 in EC and 0.76 in CC). Surprisingly, the effect of the condition seemed minimal as participants in CC generated correlation values similar to those in the EC. Likewise, most participants (9/13) in the CC reported a causal link between the face and the outcomes rather than the random assignment actually implemented. The total gains, the amount of trustworthiness ratings, investment or expected return showed no behavioural differences ($P>0.05$, Kruskal-Wallis) between participants of the EC and CC. The comparison of RTs between subjects of EC and CC revealed no significant differences either (Figure 2).

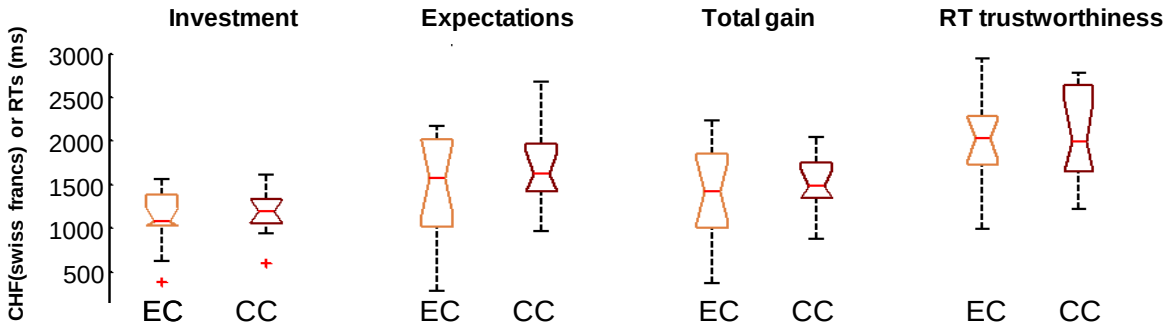


***Fig2. Comparison of the different variables between EC and CC**: Boxplots showing the distributions of values for total investment, total expectations, total gains and mean reaction times (RTs for trustworthiness ratings when subjects are grouped according to the experimental (EC) or control (CC) conditions. No differences are found between subjects playing EC or CC. RTs in this and following figures are expressed in milliseconds. Gains, investments or trustworthiness ratings are expressed in Swiss Francs and in the absolute values cumulated over the whole game or specific stages of it (when indicated).*

We then looked at the influence of previous disappointment on current decisions (Figure 3). This analysis was not restricted to EC participants as no significant differences between EC and CC participants' behaviour were found.

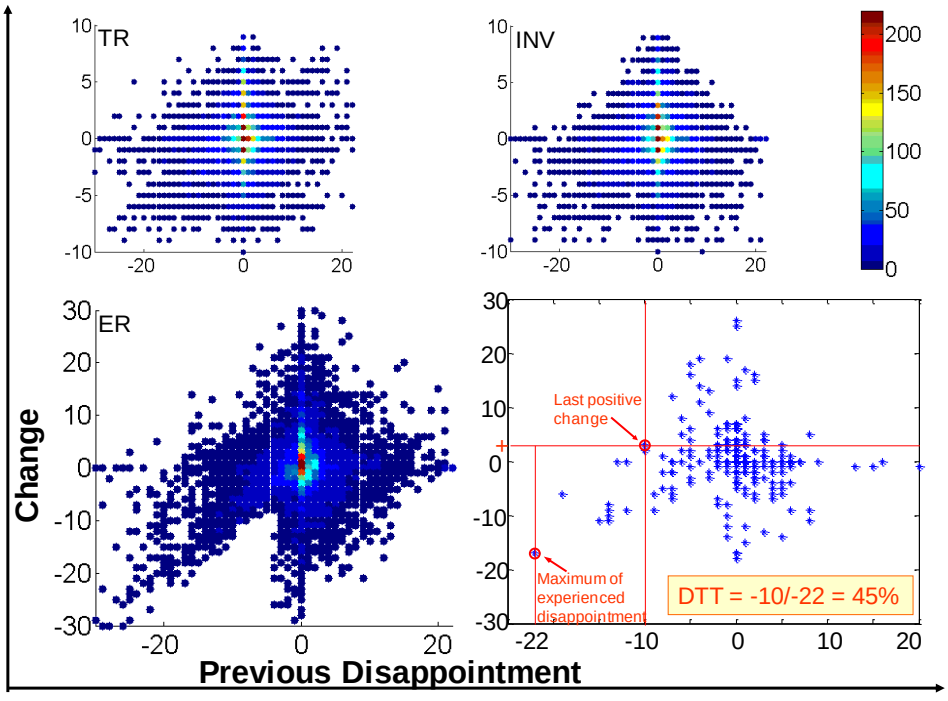


***Fig3. The influence of disappointment on upcoming decisions:** Difference of trial n-1 and trial n (y axis) as a function of previous disappointment (x axis) for one individual subject (bottom, right), and for the 3 variables Trustworthiness Ratings (TR), Investment (INV) and Expected Return (ER) for all subjects. Colours of the scale are proportional to the density of observations for the x,y point. Disappointment Tolerance Threshold (DTT) is defined for each participant as the x-value at which the last positive change in ER is observed (y axis) divided by the maximum disappointment experienced. For example here, the DTT of the individual subject is: -10/-22 = 0.45 or 45%.*

The plot of CS for the variable expected return (expectations) as a function of the disappointment experienced in the previous trial reveals a strong bias in the behaviour of the subjects following a disappointing outcome (Figure 3). Indeed, beyond an individual-specific level of disappointment, only negative CS values of expected return are observed for all participants, irrespective of the condition (EC/CC, see Supplementary Figures S1 and S2). This tendency is also present for investments and to a lesser extent for trustworthiness ratings. Noteworthy, this bias is not observed at intermediate levels of disappointment/elation (neutral outcomes) where a seemingly random distribution of choices is observed for all variables (TR, INV and ER). To rule out the possibility that this behaviour reflects an increase in cautiousness due to having invested a large amount in the previous trial (and would be consequently unrelated to the outcome of this trial) we also analysed the behaviour following elation (positive surprise, when the subject receives much more than expected). Indeed, to obtain an outcome much larger than expected subjects must also invest large amounts, so this effect of cautiousness might be observed too. However, Figure 3 indicates that this relation is not observed for elation (the upper right panel of the ER plot (expectations) is not empty). Moreover, this threshold effect is absent for 17 out of the 32 participants (no positive limit after which the expectations systematically decrease). Finally, the converse effect is more often observed: some subjects tend to increase their expectations after a strong positive outcome (as seen for instance in Supplementary Figure S3). The modification in behaviour revealed by Figure 3 is therefore a consequence of disappointment.

We found no behavioural indication of the use of another strategy by the participants of the CC. No particular exploratory behaviour was observed, nor any differences in RTs, total gains and total investments or in the answers to the questionnaire. The individual plots of CS as a function of prior disappointment failed to reveal any particular differences: there is always a limit to the tolerance of disappointment beyond which previous history modifies subsequent behaviour. However, while some Investors showed this behaviour after a small amount of disappointment, others seemed less influenced by it. Moreover, the Investors who reacted easily to disappointment often increased their expectations and investments after experiencing elation (Supplementary Figures S3 and S4). Thus, whereas the experimental manipulation failed to evoke a more cognitive strategy in CC participants, inter-individual differences emerged as a new unanticipated factor in this task. Indeed, the largest behavioural differences amongst Investors were observed across the individuals rather than across the experimental conditions.

We therefore classified the subjects according to their behaviour following disappointment. For that, we defined a measure aimed to reflect the individual impact of previous experience on decision-making that

we termed the Disappointment Tolerance Threshold (DTT). With this measure we tried to characterize the extent to which each subject deviates from the uniform circular distribution expected when trials are independent. We defined the DTT for each participant as the x-value at which the last positive change in ER is observed (y axis) divided by the maximum of disappointment experienced (see Figure 3). Dividing by maximum of disappointment allowed us to create a measure between 0 and 1 that facilitated the comparison of the subjects. Note that the DTT is defined for the expected return for the sake of compatibility with the formal definition of disappointment (Bell, 1985).

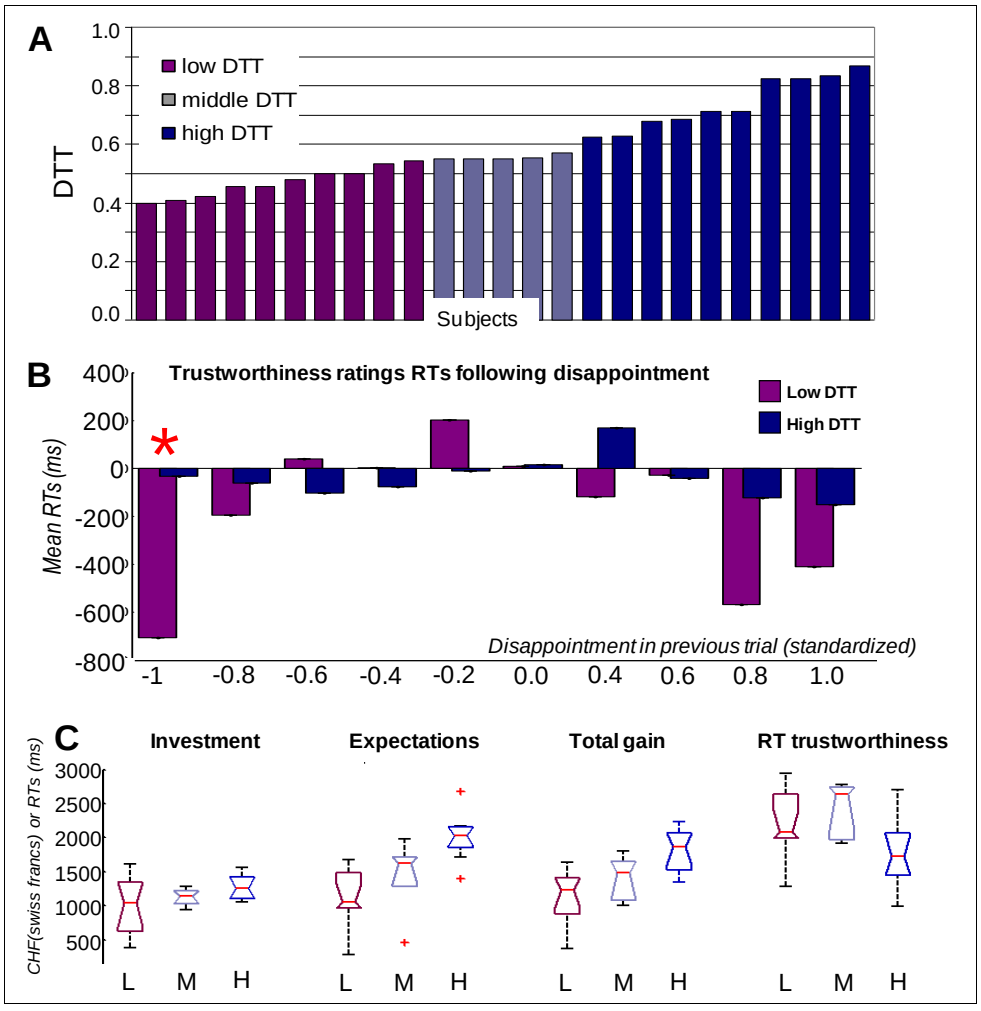


Fig4. Impact of disappointment sensitivity on behaviour: (A) *Segregation of the subjects on the basis of behaviour. Disappointment Tolerance Thresholds are shown on y-axis.* ***(B)*** *Mean (over subjects) Reaction Times after standardization by individual mean reaction times over the whole task. The disappointment in previous trial was individually normalized and divided into ten bins of equal length.* ***(C)*** *Boxplots showing the distributions of values for total investment, total expectations, total gains and mean reaction times (RTs) for trustworthiness ratings when subjects are grouped by DTT values.*

Three groups of subjects were created. The first category (low-DTT) includes participants with the 10 lowest thresholds who require little disappointment to modify their behaviour. The High-DTT category consists of participants with the 10 highest DTTs. Finally, the remaining 5 participants constitute the middle category. The plot of the DTT over subjects (Figure 4A) revealed no clear-cut division: this classification might seem somewhat arbitrary since the lowest DTT of the middle category is the same as the higher DTT of the lower category. However, we needed large extreme groups as to allow for robust statistical analysis of EEG. To validate this classification, we used a more complex measure based upon estimating the probability of observing only negative changes in ER for disappointment values in the lower quartile. This measure confirmed the ranking based on DTTs.

A last issue had to be disentangled concerning this behavioural modification following disappointment. Indeed, this behaviour can be either a consequence of cognitive updating of the estimates (i.e. the subject thinks that he granted trust too easily in the previous trial, so he diminishes it in the following) or simply an impulsive reaction (a rather automatic reaction leading to a negative outcome). Consequently, we analyzed two factors considered to be key elements of impulsive decisions (Evenden, 1999): 1) RTs, reflecting non-thoughtful decisions elicited after emotional outcomes and 2) counter-productive behaviour, i.e. failing to maximize one's gains. Figure 4B shows that Low-DTT subjects are significantly faster to evaluate trustworthiness after experiencing disappointment than after neutral or favourable outcomes. Moreover, they are also significantly faster than the High-DTT group ($P<0.01$, indicated by an asterisk on top of the bar). They take approximately half of the time to evaluate next Trustee after emotional outcomes (average: 2014 ms after neutral and 1296 ms after disappointing outcomes) although they are generally slower than High-DTT to assess the trustworthiness of the faces when all trials are considered together (Figure 4C, last boxplots). It is the reaction to disappointment – and elation to a lesser extent - that leads to shorter RTs in the Low-DTT group, rather than a general trend to take faster decisions. Importantly, the RTs of High-DTT group in the trials where they lowered their expectations after disappointment are not shorter than their average RTs.

To assess the second aspect, i.e. counter-productiveness, we compared the total investment, total expectation and total gains for the Low, Middle and High-DTT groups (Figure 4C). The total gains are significantly different between the three groups ($P = 0.0003$, Kruskal-Wallis), with the largest gains observed for the group with the High-DTT, intermediate gains for the Middle-DTT group and the lowest gains for the Low-DTT group. The comparison of the total investment between the three groups showed no significant differences ($P=0.19$, Kruskal-Wallis) indicating that the smaller gains of the Low-DTT

cannot be explained by an overall more cautious strategy of investment but again, by a specific behaviour after disappointment.

Electrophysiological results

After assessing that no topographical differences existed between participants in the EC and in the CC we explored potential differences in underlying neural systems between Low and High-DTT subjects. No significant differences were observed between the Low and High-DTT groups at other stages than the outcome evaluation phase. At this stage (outcome presented at time 0), differences in single ERPs amplitude and in scalp topographies appeared, both within and across groups, as detailed in Figure 5.

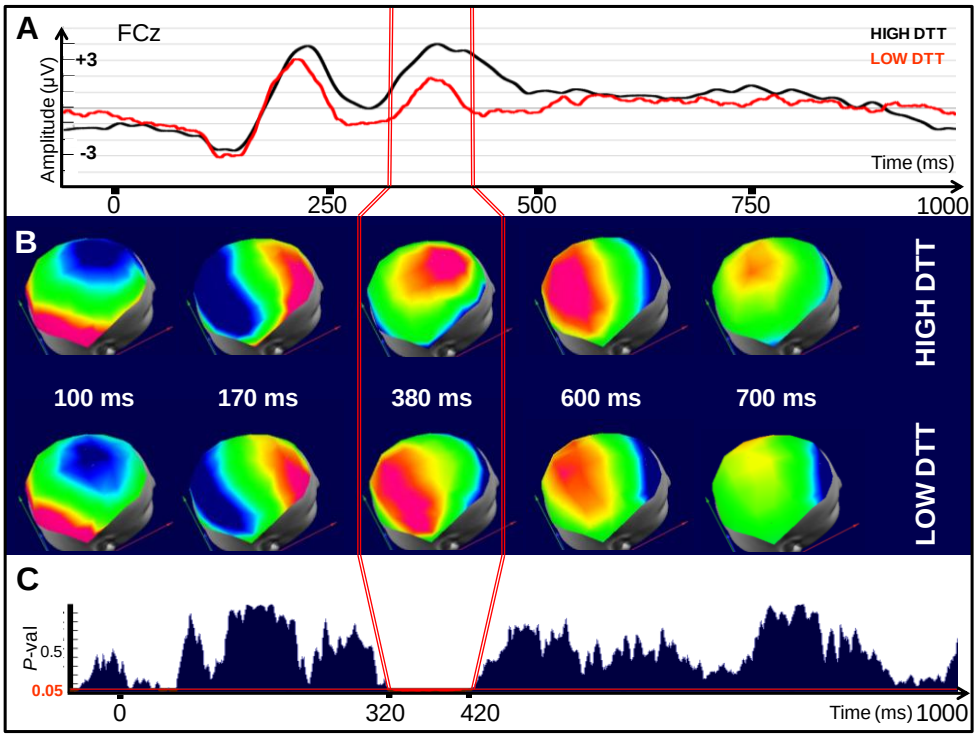


Fig5. Electrophysiological differences between individuals with High and Low Tolerance to Disappointment: ***(A)*** *Signals recorded at a frontocentral electrode (FCz) after presentation of the disappointing outcomes at time 0.* ***(B)*** *Evolution through time of the corresponding Event-Related Potential maps* ***(C)*** *P-value for a randomized MANOVA performed on the maps. Threshold for significance (P<0.05) is indicated by the red line.*

After disappointing outcomes, differences were found between groups (High vs Low DTTs) both in the amplitude of the mean ERPs on fronto-central electrodes (around 350ms, see for instance FCz in Figure 5A) and in the scalp topographies (320-420 ms, $P<0.01$, randomized MANOVA, Figure 5B). More precisely, significant differences between 300 and 400 ms ($P<0.01$, Friedman's nonparametric two-way analysis of variance) were found on the following electrodes: Fp1, F1, F3, FC1, Fpz, AF8, AF4, AFz, Fz,

F2, F4, FC4, FC2, FCz, Cz, where High DTTs' amplitude was greater than low DTTs' amplitude. Conversely, at the same timing, significant differences (P<0.01, Friedman's nonparametric two-way analysis of variance) with greater amplitude for the Low DTTs compared to High DTTs were found on the following parieto-occipital electrodes: P1, P3, P5, P7, PO7, Iz, Oz, POz, Pz, TP8, CP6, P6, P8, PO8 and O2. Neutral outcomes did not yield any differences in topographies (Supplementary Figures S5 and S6). A very brief topographic difference (380-400 ms) was observed between groups for the case of elation (Supplementary Figure 7). No significant differences in ERPs were observed around this time interval.

The within-group analysis of topographies following the different outcomes (elation, disappointment, neutral positive, and neutral negative) revealed significant differences for the same interval within the High DTT group but not within the Low DTT group. The post-hoc analysis (based on paired comparisons using the same test) identified the disappointment condition as the source of the differences (Figure 6).

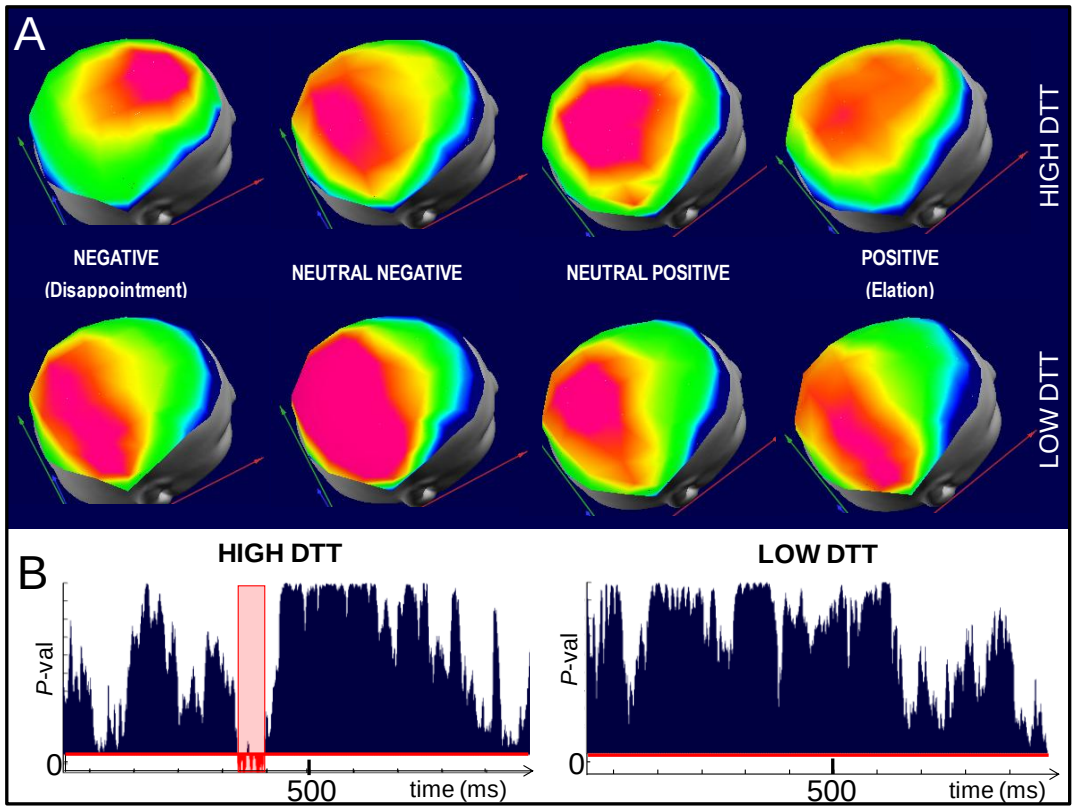


Fig6. Two different neural systems are used by High-DTT players: (A) *The map with fronto-central maximum appears only when High-DTT participants are confronted by disappointment. A different map is observed after neutral and satisfactory outcomes for the High-DTT which is identical to the map seen for all outcomes in the Low-DTT group.* ***(B)*** *Randomized Manova reveals that the fronto-central map is significantly different (P<0.05, in red) from the three other maps of the High-DTT group (left), whereas no significant differences are observed between conditions in the Low-DTT group (right).*

Discussion

Disappointment as a biasing factor in social decision-making

We manipulated the emotional and the social factors to investigate the strategies adopted by the subjects to cope with disappointment. Contrary to our prediction (a smaller influence of previous disappointment in CC players) we did not observe differences in the measured behavioural parameters or EEG topographies between the EC and CC. A likely explanation is the display of a Trustee's face in both variants of the game. Indeed, previous studies have shown that the mere presence of a picture depicting a face is sufficient to modify human behaviour (Bateson, Nettle, & Roberts, 2006) which might explain the difference between our results and other studies reporting differences between playing against humans and machines (McCabe et al., 2001; Sally, 1995). Moreover, the results of the questionnaire showed that CC players imagined links between the type of face and the outcome, and unexpectedly invested according to their trustworthiness ratings. Taken together these elements suggest that CC participants played as EC participants because the CC still involved a very strong social component.

Although not in terms of reduced expectations to avoid disappointment, a few studies already reported a decrease of trust following a non-cooperative interaction. In one study for instance (Baumgartner et al., 2008), control subjects compared to subjects under oxytocin significantly diminished their trust ratings after learning that half of their positive moves were not reciprocated. However, this behaviour seems rather rational as the feedback was given after a first block of 6 trials: the subjects naturally inferred that they were *overall* too trustworthy and consequently diminished their positive moves, probably not impulsively. Besides, it is the other group (OT condition) that showed shorter RTs in the second block compared to the first one (the feedback was given between blocks). In the same vein, another study (Delgado et al., 2005) reported shortened RTs when subjects had positive *a priori* towards their game partners. Unfortunately the authors do not mention if the converse effect was found, i.e. shortened RTs when non-cooperating with bad partners. In any case, these studies show that when an *a-priori* (or a chemical substance) that facilitates the trust decision is available, RTs are systematically shortened.

One study described almost the same bias as the one reported here (inexplicable decrease of trust towards an unknown partner because of the previous trial's outcome) but this influence was minor compared to the influence of the trustworthiness of the face, as rated by independent participants (van't Wout & Sanfey, 2008). Indeed the interaction between trustworthiness of the face and previous outcome

was not significant, suggesting that the trustworthiness rating was clearly more decisive for the investment decision than previous trial's outcome. Unfortunately the authors did not analyse RTs. However, this result is important as it shows that the effect described in our study has been replicated. Interestingly, it has been shown that suffering a major negative experience in the past year diminishes the trust granted to people in general, and that the variable which correlates the most with low trust is financial misfortune (Alesina & La Ferrara, 2002).

Inter-individual differences in the tolerance to disappointment

Beyond this general decrease in expectations following breaches in trust, we found large inter-individual differences in the tolerance to disappointment. Our results consequently support Bell's (1985) assumptions: individuals more sensitive to disappointment adopt a more pessimistic view about the future. Indeed, within our sample, the group of Investors with lower DTTs invested less and expected less after being highly disappointed, which led them to earn significantly less money than the High-DTT group. In other words, their pessimistic views on future unknown Trustees induced by prior experience made them less effective in achieving their goals of maximizing gains. These elements combined with the shorter RTs (often considered as a core aspect of impulsivity, Whiteside & Lynam, 2001) favour an explanation in terms of impulsivity rather than being the reflection of the cognitive reassessment of the subject's trust estimates (as for instance could be the case in Delgado et al's study (2005)). This behaviour was probably not fully conscious as only a few participants reported a possible influence of outcomes in prior trials while most of them acknowledged that they might have adapted their choices as a function of personal feelings about categories of trustees ("Old people never betray").

A dual system account for biased decisions

The dual system theory (Evans, 2008) provides a perfectly fitted explanation to the maladaptive behaviour of the Low-DTT group against the (mostly) well-adapted choices of the High-DTT group. According to this model, two neural systems co-exist and even compete to drive decisions. While different attributes and names have been given to these two systems most authors agree that one of these systems is responsible for fast, automatic and reflexive behaviour while the other is responsible for slow, reflective and conscious processing. Our electrophysiological results strongly corroborate both this theory and the behavioural results. First, the event-related potential appearing around 350 ms after the presentation of the outcome can be related to the family of feedback-related potentials (Falkenstein et al., 2000). It is present in both groups of subjects, reflecting the fact that they both discover a negative feedback. There is however a difference in the amplitude, the one of Low-DTT group being smaller. Interestingly, it has been shown recently that one of those feedback-related signals, the

Outcome-Related-Positivity, is smaller and appears earlier in impulsive subjects (Kamarajan et al., 2009). This is crucial as these characteristics have been found for the ERP of the Low-DTT group (compared to High-DTT group). This electrophysiological cue adds value to the interpretation of an impulsive behaviour in the Low-DTT.

Moreover, the topographical analysis revealed that at least two different neural systems (networks) were active in the High-DTT group during the evaluation of the outcome. The existence of two different scalp topographies *necessarily* implies that there are differences in the brain networks used in each case. It does not however imply that both networks might not share common structures or more importantly that both networks could not co-exist in time. The degree of involvement of each network in decisions is determined by internal emotional states and the control network involved in detecting errors and controlling impulses might explain the gradual transition from impulsiveness to a more controlled behaviour. Indeed, the plot of DTT values over the participants suggests a gradual transition rather than clear-cut groups. This transition might be due to inter-individual differences in the degrees of involvement of both networks.

For neutral or satisfactory outcomes, where the inhibition of the automatic behaviour is unnecessary, the fast and reflexive system seemed sufficient for all subjects (statistically identical maps across and within groups for neutral outcomes). The automatic system apparently encouraged the participants to keep going on as long as the real outcome was close to the expectations (neutral cases) and provoked approach/avoidance reactions when the outcomes were better/worse than expected. This interpretation is also supported by the fact that Low-DTT individuals tended to increase their expectations after experiencing elation in the previous trial. In contrast, the High-DTT group used the reflective system when confronted to unexpected outcomes, which helped them to adapt their behaviour and avoid prejudiced social decisions. The short (20 ms) difference exclusive to High-DTT participants during elation further supports this interpretation. Thus, the particular role of the reflective system in this context seems to be to withhold the reflexive behaviour by inhibiting the impulsive reaction to systematically decrease expectations about an unknown Trustee.

Importantly, if the results of this study were to support an emotional vs. rational explanation we should have found differences when comparing emotional to neutral outcomes in the group showing the bias (the Low-DTTs). However Low-DTT map topographies showed no differences as a function of previous trial's outcome. The only plausible explanation of the fact that their high sensitivity to disappointment (strongly impacting their behaviour) is not reflected in their maps is that the same neural network

(system) is responsible for all the aspects of their behaviour. In this context the theory of an automatic system as described earlier fits perfectly the data whereas there is little support for an interpretation in terms of emotional vs rational dual-system.

The co-existence of both systems within identical individuals (the High-DTT group) rules out *per se* that anatomical differences are at the basis of electrophysiological differences. Similarly, the equal distribution of EC and CC participants into High and Low-DTT groups indicates that map differences are not due to the use of empathy/mirroring networks (Frith & Singer, 2008). Finally the electrophysiological analysis was designed to exclude differences due to trivial changes in generators such as inversion of the dipolar moment (leading to polarity inversions in the maps). Consequently the significant differences detected between the scalp topographies of Low and High-DTT participants or within the High-DTT group necessarily reflect differences in the neural networks implicated in the processing of the outcomes.

Scalp potential of High-DTT subjects in the case of extremely disappointing outcomes resembles the map linked to error or conflict detection elucidated in other studies (Cohen & Ranganath, 2007; Holroyd, Hajcak, & Larsen, 2006; Holroyd et al., 2003). This map has been consistently localized to medial structures in the frontal wall, in particular to the anterior cingulate cortex (ACC). Interestingly, the ACC seems to have a general role in representing and updating action values (Ito, Stuphorn, Brown, & Schall, 2003; K. Matsumoto, Suzuki, & Tanaka, 2003; Walton, Devlin, & Rushworth, 2004) and lesions to this structure in macaques lead to deficits in integration of the most recent outcome to guide choices (Kennerley, Walton, Behrens, Buckley, & Rushworth, 2006). This fronto-central map is present only in High-DTT subjects although both groups showed a feedback-potential at this time point. A plausible explanation might be that the amplitude of the ERP waveform reflects the combination of the activation of two (or more) different generators in High-DTT but not in the Low-DTT. Indeed, both groups might receive one error signal but in High-DTT group, another structure comes into play to exert control over behaviour, and its activity amplifies the potential recorded at the surface of the scalp. This could explain 1) that the ERP peaking at 350 ms after outcome presentation is observed in both groups (it partly reflects an error signal) 2) the difference in amplitudes between both groups (combined with the impulsivity of Low-DTT which decreases the signal) and 3) the fact that although both groups show an error-signal, the fronto-central map is observed only in High-DTT group.

Interestingly, another study (De Martino et al., 2006) showed a correlation between the activation of the prefrontal cortex (medial and orbital parts) and resistance to the frame effect (be risk-averse in a sure

option but risk-seeking in a gamble option). The authors hypothesised that the most rational individuals are the ones who better identify their emotions and the moment when they should be controlled in order to avoid bad consequences. Our results clearly support this idea. Moreover, the first neuroimaging study on the Trust Game (McCabe et al., 2001) reported that only the subjects who invested in the game (7 out of 12) showed a greater activation of the prefrontal areas when playing against a human rather than a computer. The five remaining subjects did not show significant differences between these conditions. This can be interpreted almost in the same way as our results: some subjects, i.e. those who did not cooperate, relied on an automatic system (not investing at all) irrespective of the condition, whereas the other group used another system when playing with humans.

While additional neuroimaging studies are required to identify the key components of the reflective/reflexive systems, our results suggest that differences in their activation are short-lived (100 ms) and therefore hard to detect with low temporal resolution imaging techniques.

This study is the first one to provide such a complete combination of behavioural and electrophysiological elements (on one hand the link between impulsivity (RTs), lower feedback potentials, counter-productivity, and the strategy of lowering expectations to avoid disappointment, and on the other hand the presence of a map potentially reflecting conflict detection and control exertion linked to a more rational behaviour) in favour of the dual system model in decision-making, (Sanfey & Chang, 2008), specifically by showing the involvement and interplay of two different neural systems within the same individuals (the High DTTs). Our results help to clarify circumstances under which each system emerges and the consequences of relying upon one or another system. Clearly our data better matches a “default-interventionist” functioning mode rather than a “parallel-competitive” mode (see Evans, 2008, hypotheses). Indeed, it seems that in most situations, System 2 does not intervene (for instance in neutral feedbacks) and no signs of conflict were detected.

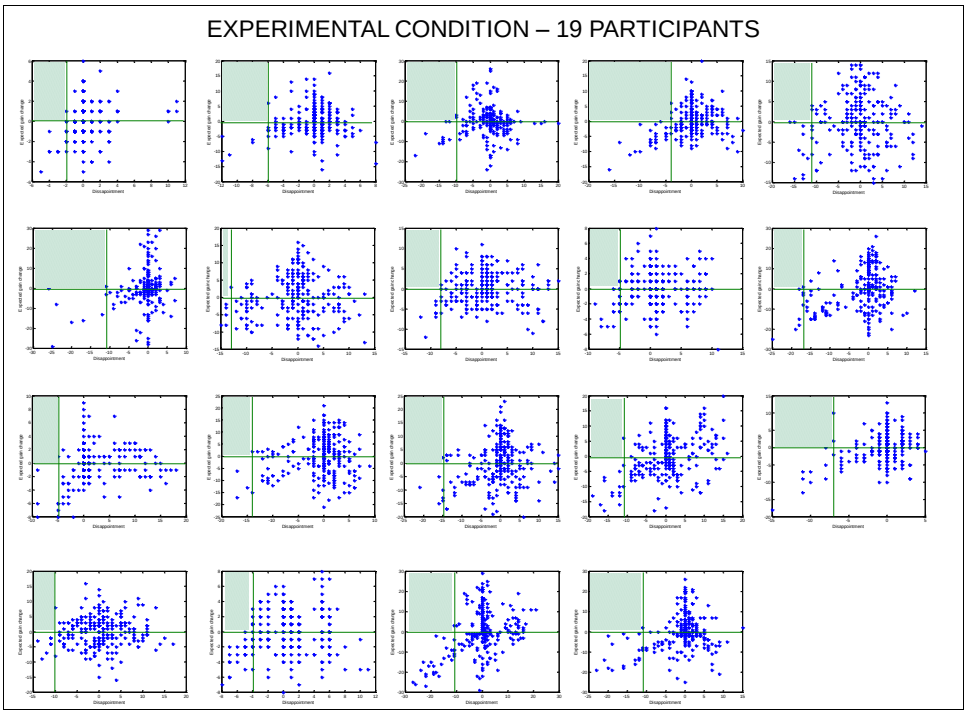
Whereas some previous studies investigated close emotions such as regret (Camille et al., 2004; Coricelli et al., 2007), or related topics such as unfair revenge (Herrmann et al., 2008), no study has, to our knowledge, explored disappointment (defined in a formal way and induced by the context) and its neural correlates in a similar fashion.

Supplementary Material

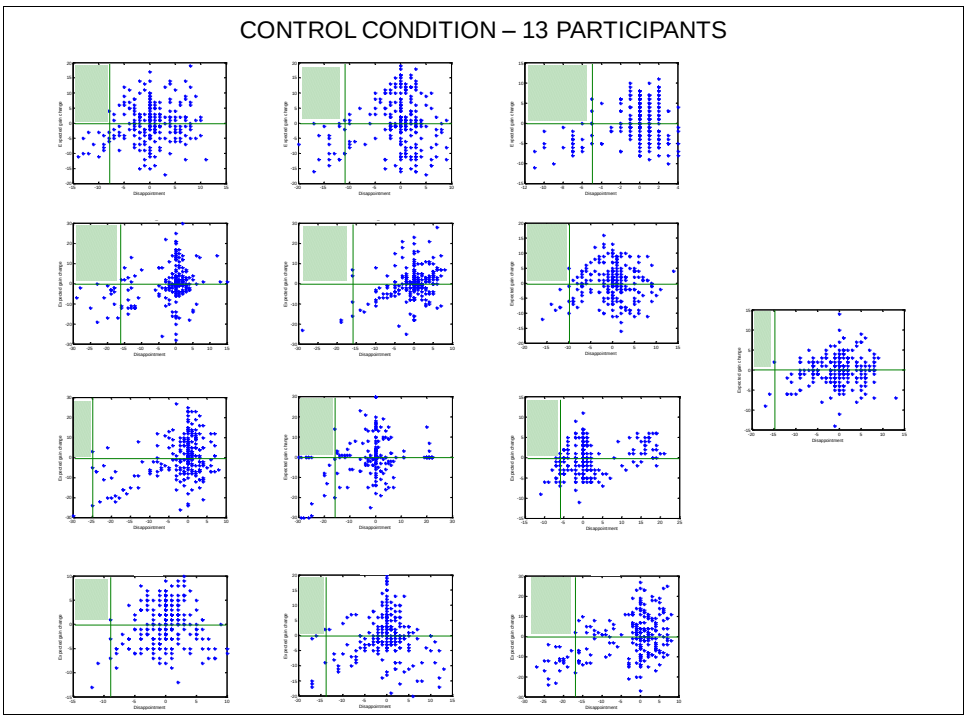
Supplementary Table S1: Synthesis of the answers to the questionnaire. CA indicates that the participant mentioned a category of people in this answer (e.g.: "Old people never betray") to justify either a change in his strategy or a link between his judgment of the face and the outcome. PO indicates that the participant mentioned the impact of previous outcomes to justify a change in strategy.

subject	age	gender	laterality	condition	link judgment-face	change in strategy	emotion	forced-choice	DTT
1	26	M	left	CONTR	yes (CA)	yes, volontarily	disappointment	disappointment	
2	27	M	right	CONTR	yes (CA)	yes, volontarily	disappointment	disappointment	LOW
3	26	M	right	CONTR	yes	yes, volontarily	disappointment	disappointment	
4	20	F	right	CONTR	no	yes, volontarily	disappointment	disappointment	HIGH
5	24	F	right	CONTR	yes	yes, volontarily	frustration	disappointment	HIGH
6	24	M	left	CONTR	no	yes, volontarily	frustration	disappointment	
7	21	F	right	CONTR	yes	yes, volontarily	frustration	regret	
8	20	F	right	CONTR	no	yes, volontarily	frustration	anger	
9	22	F	right	CONTR	yes	yes, volontarily	frustration	betrayal	HIGH
10	19	M	left	CONTR	yes (CA)	yes, volontarily	irritation	anger	LOW
11	20	F	right	CONTR	yes (CA)	yes, volontarily	irritation	anger	
12	19	M	right	CONTR	maybe	yes, volontarily	amusement	disappointment	HIGH
13	26	M	right	CONTR	no	yes, volontarily	laughter	betrayal	LOW
14	26	F	right	EXP	more or less	yes (CA+PO)	disappointment	disappointment	HIGH
15	30	M	right	EXP	yes	no	disappointment	disappointment	HIGH
16	33	M	right	EXP	yes (CA)	yes (CA+PO)	disappointment	disappointment	HIGH
17	24	F	right	EXP	yes	yes, depending on the results	disappointment	disappointment	HIGH
18	30	M	right	EXP	yes	yes	disappointment	disappointment	LOW
19	25	F	right	EXP	yes	yes (CA)	disappointment	disappointment	LOW
20	27	F	right	EXP	yes		disappointment	disappointment	LOW
21	26	M	right	EXP	yes	yes, unvolontarily	disappointment	disgust	LOW
22	31	M	right	EXP	yes	yes (CA)	despair	regret	HIGH
23	27	M	right	EXP	yes	no	frustration	anger	HIGH
24	25	F	left	EXP	yes	Yes, beyond control, PO	frustration	betrayal	
25	24	F	right	EXP	yes	yes, volontarily even if irrational	frustration	disappointment	
26	32	F	right	EXP	no		anger	anger	LOW
27	30	F	right	EXP	no		anger	betrayal	
28	27	F	right	EXP	yes	yes (PO)	irritation	anger	
29	23	F	right	EXP	yes	not really, sometimes PO	betrayal	betrayal	LOW
30	27	M	right	EXP	yes	yes, volontarily	negative	anger	LOW
31	25	F	right	EXP	yes (CA)	maybe	astonishment	regret	
32	20	F	right	EXP	maybe	yes (PO)	disillusion	betrayal	

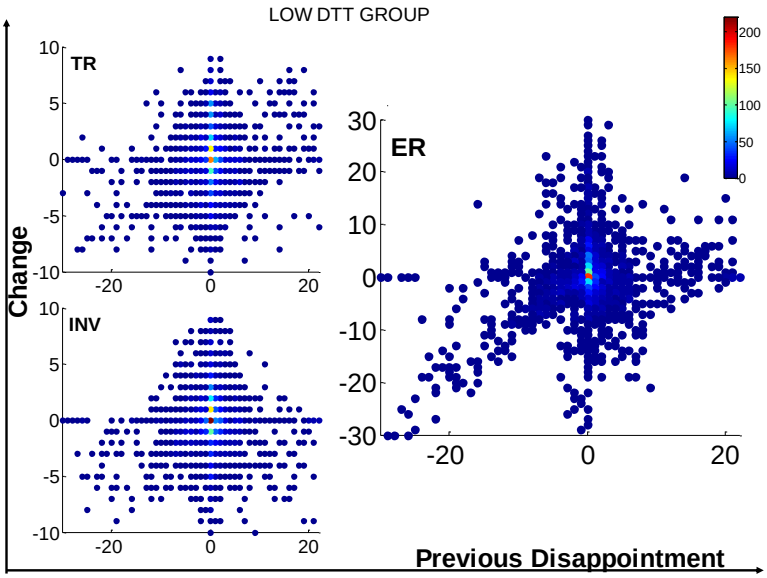
Supplementary Figure S1: DTT for single subjects in the Experimental Condition. The dashed green zone represents the absence of positive change in expectations in actual trial compared to previous trial, and is present for all 19 subjects of the experimental condition (EC).



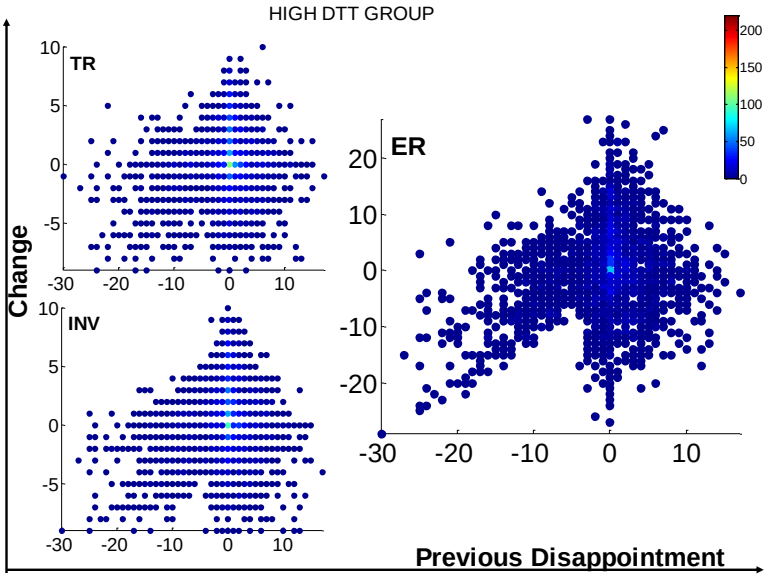
Supplementary Figure S2: DTT for single subjects in the Control Condition. The dashed green zone represents the absence of positive change in expectations in actual trial compared to previous trial, and is present for all 11subjects of the control condition (CC).



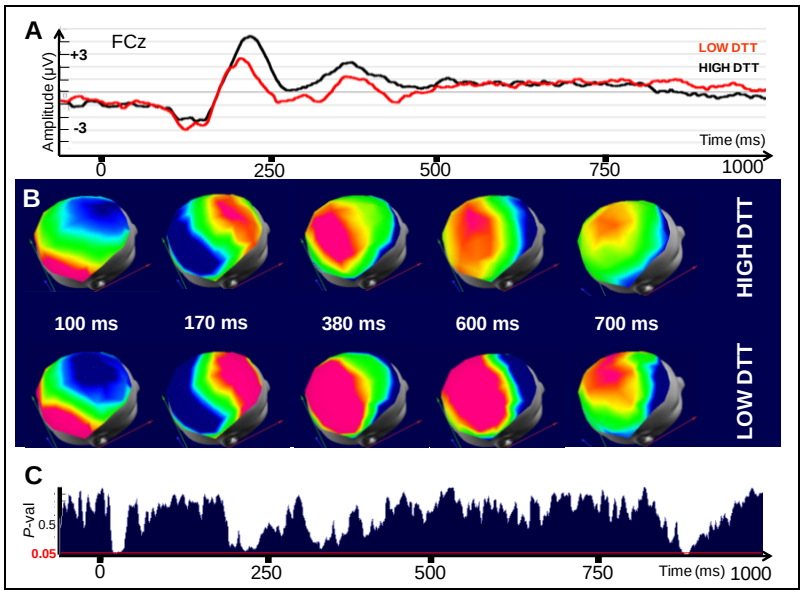
Supplementary Figure S3: Cs plots for Low-DTT group. Change in the 3 variables under study as a function of the disappointment experienced in previous trial. TR stands for Trustworthiness Ratings, INV for the investment and ER for the expected return (expectations). A clear tendency to raise the TR and ER after elation (great positive value on X axis) is observed.



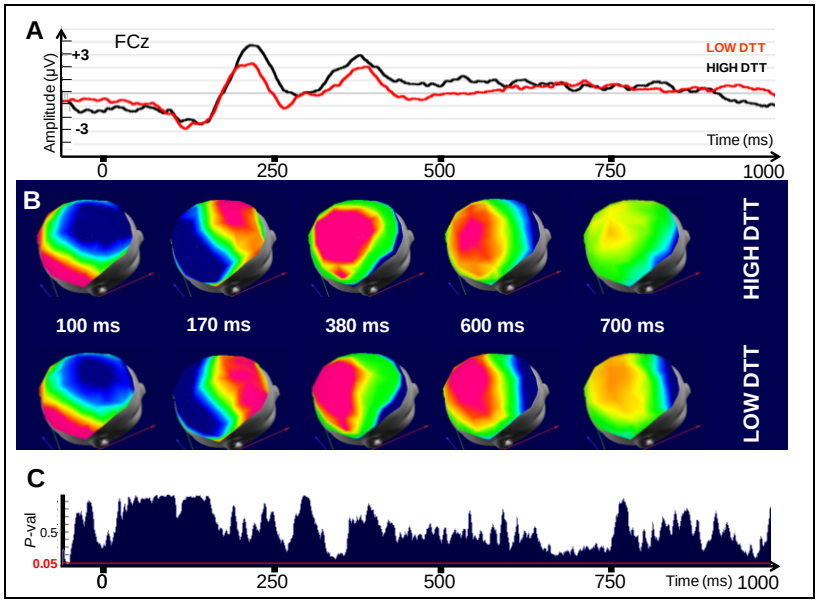
Supplementary Figure S4: Cs plots for High-DTT group. Change in the 3 variables under study as a function of the disappointment experienced in previous trial. TR stands for Trustworthiness Ratings, INV for the investment and ER for the expected return (expectations). Here there is no tendency to increase any of the variables after elation. This demonstrates a lighter influence of previous history for this group.



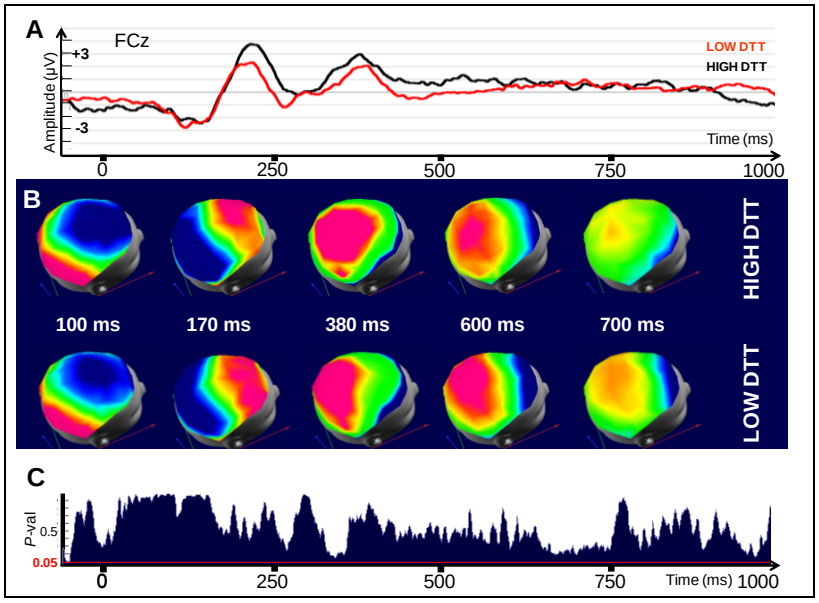
Supplementary Figure S5: Electrophysiological comparison between High and Low-DTT groups: (A) Signals recorded at a frontocentral electrode (FCz) after presentation of the Neutral Negative outcomes at time 0. (B) Evolution through time of the corresponding Event-Related Potential maps (C) P-value for a randomized MANOVA performed on the maps. Threshold for significant differences (P<0.05) is indicated by the red line. No differences are observed.



Supplementary Figure S6: Electrophysiological comparison between High and Low-DTT groups: (A) Signals recorded at a frontocentral electrode (FCz) after presentation of the Neutral positive outcomes at time 0. (B) Evolution through time of the corresponding Event-Related Potential maps (C) P-value for a randomized MANOVA performed on the maps. Threshold for significant differences (P<0.05) is indicated by the red line. No differences are observed.



Supplementary Figure S7: Electrophysiological comparison between High and Low-DTT groups: (A) Signals recorded at a frontocentral electrode (FCz) after presentation of the Positive (Elation) outcomes at time 0. (B) Evolution through time of the corresponding Event-Related Potential maps (C) P-value for a randomized MANOVA performed on the maps. Threshold for significant differences (P<0.05) is indicated by the red line. A significant difference is observed during 20 ms (380-400 ms).



STUDY 2

VALIDATION OF THE ASSOCIATION BETWEEN THE PRESENCE OF THE FRONTO-CENTRAL MAP AND A LOWER SENSITIVITY TO PREVIOUS DISAPPOINTMENT (A HIGHER DTT)

In the second study, our main goal was to confirm the interpretation in terms of dual system of Study 1. More specifically, the aim was to confirm the link between the High-DTT behaviour and the map that we hypothesised to reflect the intervention of System 2.

Methods

Participants and Experimental Design

For that, we contacted the participants of Study 1 and asked them to participate to a second experiment, Trust Game 2. We first debriefed the subjects on the results of Study 1, without mentioning the distinction between the two groups. We just said that we were surprised by the results because the participants seemed to be influenced by the outcome of the preceding trials, whereas we expected them to play each trial independently as the same Trustee never appears twice. They were asked to keep this in mind and to play the game once again.

Five out of the ten members of Low-DTT group accepted to play for a second time and we kept their recordings for a comparison with Study 1. The participants read the instructions (identical as in Study 1) and then played the Trust Game which was exactly the same as in Study 1, except that we deleted the slide on self-confidence and accelerated the countdown before presentation of the outcome in order to shorten the total time of the experiment. These changes cannot affect the results as they concern slides that are out of the analysed time frame (the 1000 ms epoch following the outcome presentation).

DTTs and map comparisons could not be statistically tested due to small sample size.

Results

Behavioural Results

Correlations between the 3 variables (trustworthiness, TR; investment, INV; and expected return, ER) were again extremely high (mean r: TR-INV = 0.83, TR-ER = 0.75, INV-ER = 0.87, $P < .0001$). Only one subject showed smaller correlations (respectively 0.6, 0.4 and 0.59, still significant at $P < 0.001$). The comparison of total gains for each subject between TG1 and TG2 (Figure 1) showed a systematic increase. Significance tested with Kruskal-Wallis non-parametric one-way ANOVA is confirmed for three subjects ($P < .05$ for S1 and $P < .001$ for S4 and S5). The comparison approached significance for S2 (P=0.06).

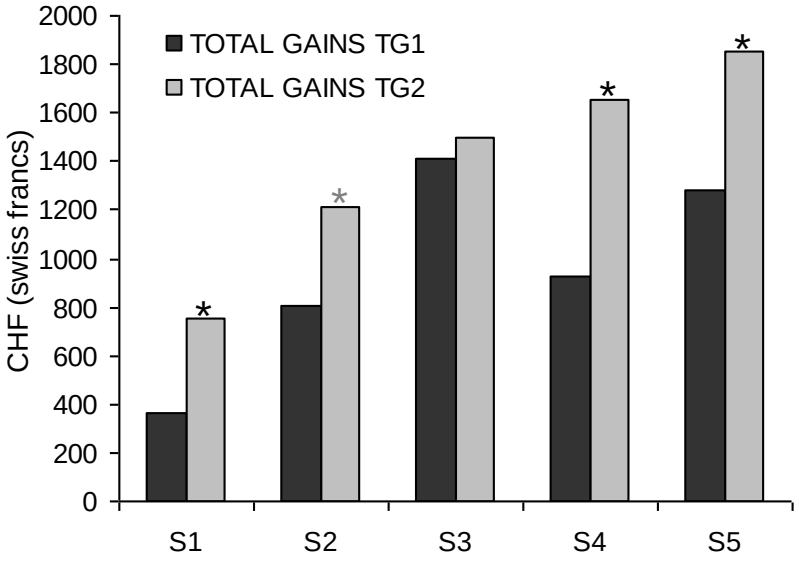


***Fig1**. Comparison of the total gains between TG1 and TG2 in Swiss francs shows that subjects better fulfilled the goal of maximizing their gains in TG2*

Mean Disappointment Tolerance Threshold (DTT) of these five subjects increased in Study 2 compared to Study 1 (respectively, 0.45 in Study 1 and 0.47 in Study 2, see Figure 2.) However, one subject (the same who did not show strong correlations) decreased his DTT. Without this subject, the mean DTT over the 3 other subjects is 0.58, the last subject having completely abolished the DTT effect.

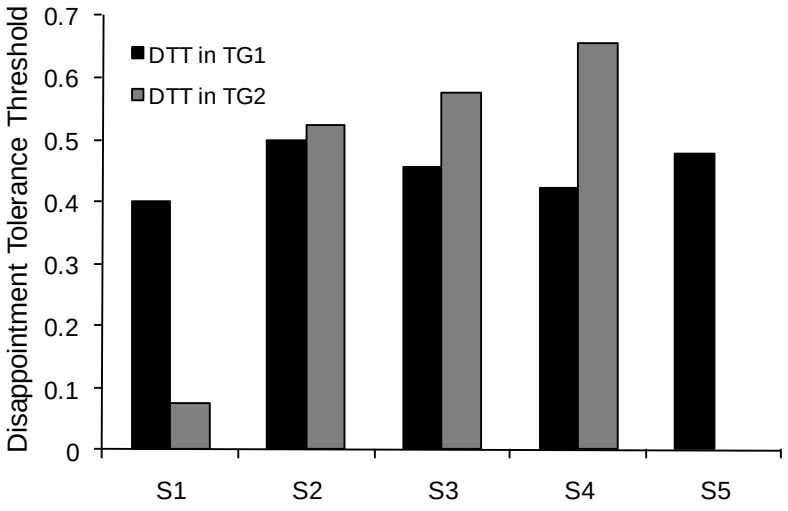


Fig2. *The increase in total gains in TG2 compared to TG1 is accompanied with higher DTTs in TG2*

Electrophysiological results

Figure 3 shows the comparison for each subject (S1 to S5) of the electrical topographic map recorded in Study 1 to the one recorded in Study 2.

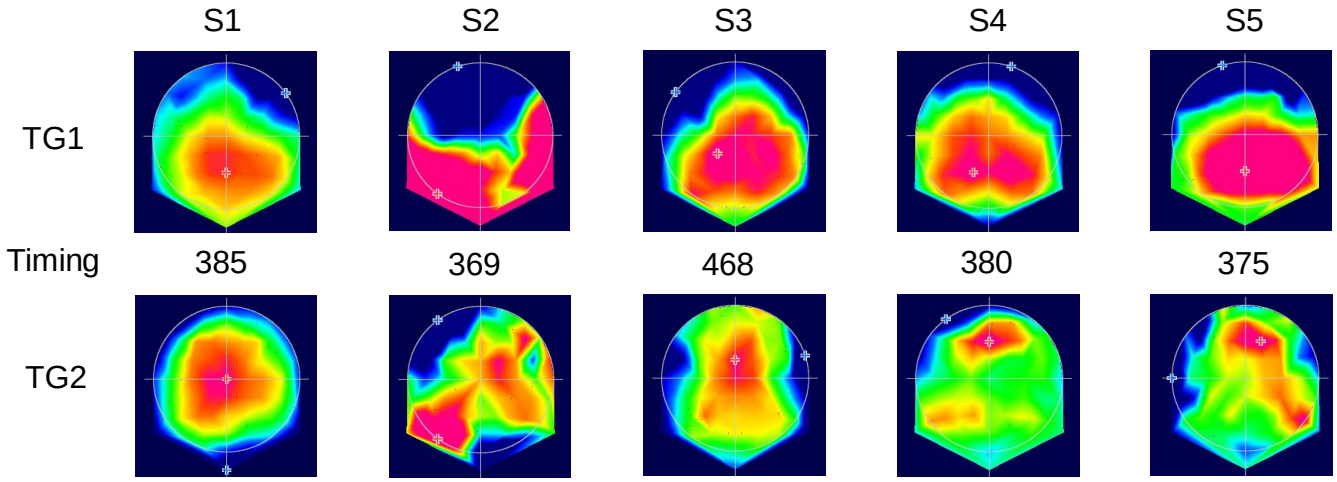


Fig3*. Topographical electric maps for the 5 subjects of Low-DTT group. Comparison of the maps between the first experiment (TG1) and the second experiment (TG2) for each subject (S1-S5). The timing is different as we chose the moment where the "System 2" map appeared.*

Although the difference between the maps cannot be statistically tested, a shift of the maximal activity of the map towards anterior parts of the brain and a fronto-central location is observed for the five subjects.

Discussion

This second experiment confirms the link between higher gains, the increase of DTT (the reduced sensitivity to disappointment) and the presence of a map showing maxima around fronto-central electrodes. One subject showed a decrease in DTT which might seem paradoxical with presence of the fronto-central map. However, this subject also showed reduced correlations between the three variables under study. Looking closely at his results, it appears that this participant, in order to "break" the link between previous and actual trial, assessed the face according to his opinion but invested alternatively very high and very low amounts. This explains the lower correlations, as well as the presence of the map, as this strategy obviously implied some control exertion over spontaneous behaviour. The fact that correlations between trustworthiness ratings and investments are still extremely high for all subjects confirms that they participated actively in the task and still relied on their impressions of the faces to decide the amount to invest.

This study indicates that, irrespective of the original source of inter-individual differences found in Study 1, the behaviour in the Trust Game (and its neural underpinnings) can be modulated by the context. One question that still remains unclear is why in the first study these two groups acted differently. One step was to define that one group behaved more impulsively in Study 1, but it is hard to know whether this behaviour reflected personality traits or the state of the subject at the moment of the experiment. Indeed, it has been shown that being in a good mood, for instance, favours the use of System 1, as well as being under time pressure or having to perform two tasks simultaneously (Finucane et al., 2000; Kahneman, 2003). Although we cannot assert that subjects were all in a bad or a good mood in Study 1, no differences in the experimental setting (such as time pressure of involvement in dual tasks) can account for the inter-individual differences. On the other hand, the facility of use of System 2 has been linked to personality traits such as IQ or "need for cognition" (Frederick, 2005; Shafir & LeBoeuf, 2002). One explanation of our results is that subjects of the Low-DTT group are less prone to use System 2 because of such inter-individual differences. However, this question needs further investigation, for instance with the use of questionnaires on impulsivity and general intelligence.

These results are crucial because they show that an experimental manipulation can induce the use of System 2 which is manifested both behaviourally and electrophysiologically. More than confirming the interpretation of Study 1, they open a new research avenue on behavioural and electrophysiological inter-individual differences in the use of dual system.

STUDY 3

KNOWING WHAT TO DO BUT DOING THE OPPOSITE: REJECTION OF UNFAIRNESS IN THE ULTIMATUM GAME

The results of Study 2 suggest that a simple change of instructions can induce the emergence of System 2. We wanted to further investigate this possibility with the use of another paradigm. For that, in the present study, we focused on how the classical results of the Ultimatum Game might be affected if the subjects were aware of the game theory's prediction. More precisely, we wanted to disentangle the two following hypotheses:

(a) **EMOTION** If a strong negative emotion is provoked by unfairness but the subjects feel compelled to accept unfair offers to fulfil the goal of maximizing their gains, then a control over their impulse to reject the offer must be exerted for the sake of "rationality". This might be reflected by behavioural indicators such as lengthened reaction times and neurophysiological indicators of control over emotions (Sanfey et al., 2003).

(b) **SOCIAL NORMS** On the other hand, if the natural tendency were to accept all offers and the only reason beyond the rejection are manners (learned social norms of fairness), then the subjects might live very comfortably the experience of accepting all offers, having the good excuse of following the instructions. Reactions times should be shorter, and no signs of conflict or behavioural control should be observed (Knoch et al., 2006).

Another way to investigate the mechanisms beyond unfairness rejection is to identify structures reacting to unfairness. We hypothesised that if the "emotion" hypothesis is correct, then structures linked to emotions should correlate with unfairness, whereas if the "social norm" hypothesis is true, no indication of an emotional processing should be present.

Methods

To address these issues we conducted an EEG experiment in which we manipulated the information given to the subjects between two blocks of the same Ultimatum Game.

Participants

Sixteen young healthy volunteers (11 female, 4 left-handed, mean age 24.06) with no history of neurological disorders participated in this experiment. They were recruited through an advertisement posted at the University of Geneva and all signed an informed consent before starting the experiment. This experiment was approved by the local ethics committee.

The Ultimatum Game

The UG is a two-player game in which Player 1 has at his disposal a certain amount of money and must propose a share of his choice to Player 2. If Player 2 accepts the offer the share is done accordingly, but if he refuses both players earn nothing.

Procedure

In our version of the UG, participants always played the part of Player 2. There were 6 different types of offers: the "fair" ones (Player 2 receives 50%, 45% or 40% of the amount) and the "unfair" ones (Player 2 receives 15%, 10% or 5% of the amount). The total amount could be 10, 15, 20, 25 or 30 Swiss Francs (CHF) which resulted in 30 different offers in total (6 levels of fairness * 5 possible total amounts). We added two kinds of offers that we will denote as superfair: those two offers were advantageous to the subject (60% or 55% of the total was given to the subject). To be realistic, the superfair offers were presented in reduced frequency compared to other offers (only 8 times in each block). This manipulation allowed us to have 8 degrees of fairness to test, and to investigate the reaction of the subjects when confronted to offers that are more advantageous for them than for the other players (Players 1).

After reading the instructions, the subjects played the first block (120 offers, 60 fair). The principles of the game were briefly explained to them and they were told that the people against whom they were going to play had already given us their sharing proposal. Thus we suggested that the participants were going to play against "real" players although not in real-time.

Each round started with a fixation cross, followed by the picture of Player 1 (different in each round) displayed for 500ms (Figure 1, (Phillips, Wechsler, Huang, & Rauss, 1998)). The offer then appeared and the subject answered by pressing 2 different buttons (accept/reject) on the keyboard (controlled variable: for 8 subjects, the key "v" corresponded to "accept" and the key "n" corresponded to "reject" and for the other half it was the opposite). This experiment was implemented using Cogent 2000 developed by the Cogent 2000 team at the FIL and the ICN and Cogent Graphics developed by John Romaya at the LON at the Wellcome Department of Imaging Neuroscience.

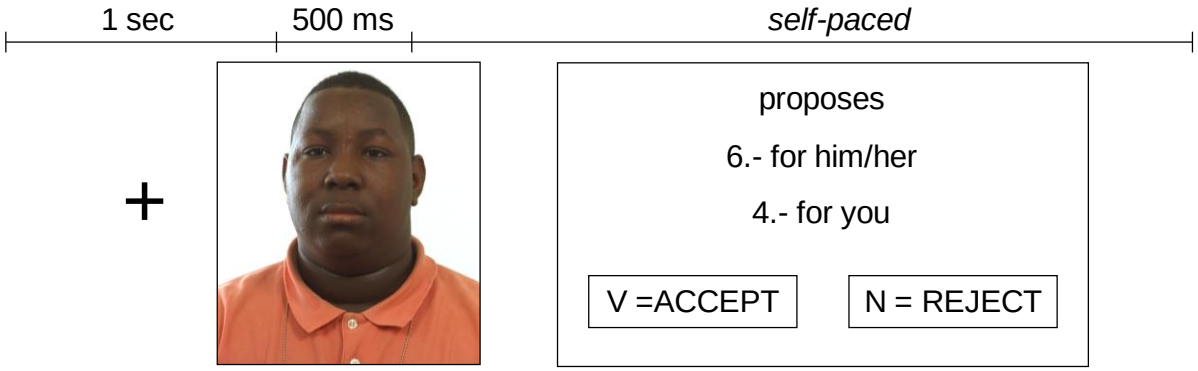


Fig1. *Time line of a trial in the Ultimatum Game.*

After the first block, participants were asked if they had rejected some offers, and if so, to explain why and which feelings were associated to this situation. Finally, they were asked if, in their opinions, their behaviour was justifiable or not.

After answering this questionnaire, the participants read a little text explaining that the classical results of the UG have always been a matter of considerable astonishment amongst the scientific community. Indeed, although it has always been predicted that the Players 2 should accept all kinds of offers, experimental results always disconfirmed this theory as people tended to reject low offers. This is surprising as the Players 2 have nothing to lose by accepting even small offers, and this behaviour has always been considered as "irrational" by economists. The participants were then asked to keep this in mind while playing the second block. The second block was similar to the first one (120 offers, 60 fair).

EEG data acquisition

The EEG was recorded at 1024 Hz (5th order sinc filter with a -3 dB point at 1/5th of the sampling frequency) using 64 BioSemi sintered Ag-AgCl electrodes. The electrodes were mounted on the manufacturer-provided cap according to an extended 10-20 system. The Biosemi system uses a common mode sense (CMS) active electrode as the reference. This reference was transformed in our case to the average reference during offline analysis, where visual inspection was used to reject trials containing eye blinks or artefacts. Epochs of 1100 ms (100 ms baseline) were extracted after notch filtering at 50 Hz and superior harmonics for each trial, time 0 representing the moment when the offer is revealed. No baseline correction was applied since this can distort the maps for source localization (Wendel et al., 2009). Averages of these epochs were calculated first for each subject and each type of outcome, and then for all the subjects together (Grand Mean). All analysis of EEG and behavioral data were carried out using custom built Matlab routines. Bad electrodes interpolation was based on spherical splines.

Data analysis

Reaction Times (RTs) and responses (accept/reject) were recorded and the following comparisons were performed: rejected/accepted * fair/unfair * block1/2. Normality was tested and rejected for all RTs distributions (Lilliefors test $P<0.001$). Thus we used a non-parametric one-way ANOVA known as Kruskal-Wallis test to perform the comparisons on RTs. To compare the acceptance rates of different offers/conditions we used the Fischer Exact One-Tailed Test.

The following comparisons were performed between subject's average ERPs in both blocks together: fair vs unfair offers, fair vs unfair vs superfair offers, the fairest (50-50) vs the most unfair (5-95) offer, accepted vs rejected offers. The ERPs comparisons were performed with a Friedman non-parametric ANOVA with a $P< 0.05$ threshold of significance on the average over subjects of each electrode at each time point. For each of these analyses, topographical map comparisons were performed using a randomized MANOVA procedure. Indeed, a good measure of the similarity or dissimilarity of the topography of instantaneous scalp maps, insensitive to simple polarity inversions or scaling factors, is the absolute value of the correlation coefficient between maps. The correlation coefficient can be then used as the test statistics to build a non-parametric multivariate randomization test to evaluate the statistical significance of differences between instantaneous scalp maps within two (or more) conditions. This test, known as a randomized one-way multivariate analysis of variance (Mielke, 2001) can be applied to each measured topography, i.e., to each time frame. In the test, we assume that no differences exist between the conditions (hypothesis Ho) and compute the distribution of the test

statistic by permuting members from one condition to the other 1500 times. An important feature of the method is that if a difference is found between map topographies, it necessarily implies a difference in at least one underlying generator responsible of the activity recorded at the surface of the scalp. This allows assessing if different structures or networks are involved depending on the condition. We thus compared the maps for a time frame of 1000 ms following the presentation of the offer (including pre-stimulus period of 100 ms), depending on the response of the subject (accepted vs rejected) and on the block.

Finally, to identify a network of structures responding to the fairness of the offer, we performed an estimation of the intracranial Local Field Potentials within the whole brain volume (Grave de Peralta Menendez, Gonzalez Andino, Morand, Michel, & Landis, 2000; Grave de Peralta Menendez, Murray, Michel, Martuzzi, & Gonzalez Andino, 2004). Once the sources were determined, we calculated the correlation (Kendal *tau* Rank Correlation) between the levels of fairness of the offer (6 levels: the two superfair offers are too rare to be included in this analysis) and the activity of the generators involved in the decision phase (time 0 corresponds to the moment when the offer is revealed to the subject). We kept only those correlations whose significance were under the threshold of 0.01 and who lasted more than 20 ms. Here, only the early correlations (<500 ms) are reported to avoid possible confounds with the motor responses and only the very early ones (<200 ms) will be discussed. Whenever a significant correlation was detected in the pre-stimulus period it was considered as an artefact and removed from the whole analysis.

Results

Behavioural Results

Classical UG results are reproduced in Block 1

In the first block, the results were similar to classical findings in the UG. Acceptance rate for fair offers was 87% and rejection rate for unfair offers was 83%. Average RTs were significantly shorter for the acceptance of fair offers than for the rejection of unfair offers (respectively 1188.2<1320.14 ms, P=0.002).

Superfair offers yield mixed behaviour in Block 1

The first superfair offer (60-40) was slightly less often accepted than the fair offers (acceptance rate of 84% compared to 86% for all other fair offers, not significant), and the RTs were not different. However,

the second superfair offer (55-45) showed significantly longer RTs than fair offers to be accepted (*P*= 0.007) and a significantly higher acceptance rate (97% of those offers were accepted compared to 86% for all fair offers, *P* = 0.005, and 84% for the fairest, *P* = 0.015).

The fairest offer is processed faster than the other offers in Block 1

The RTs were then analyzed as function of the degree of fairness (1-8). The fairest offer (50-50) yielded shorter reaction times than all the other offers (except for superfair (60-40)). No other differences between the categories were observed.

Fair offers are accepted faster and rejected slower than unfair offers in B1

To better define the differences in RTs we looked separately at rejected and accepted offers.

For the accepted offers, the fairest offer (50-50) took significantly less time to be accepted than one of the superfair (55-45) and than all unfair offers. The second fairest offer (45-55) was shorter than the second most unfair offer (10-90), and the third fairest offer (60-40) was shorter to accept than the two most extreme unfair offers. For the rejected offers, 2 differences were significant: the "40-60" offer took significantly more time to be rejected than the "90-10" and the "95-5" offers.

Awareness of the consequences of rejection reduced but did not eliminate adverse reactions to unfairness in Block2

In the second block, acceptance rates for fair offers significantly rose up to 98% (compared to 87% in the first block, *P* < 0.001) whereas rejection rate for unfair offers significantly dropped to 36% (compared to 83% in the first block, *P* < 0.001). Figure 2 summarizes these results. Fair offers were still accepted faster than unfair offers were rejected (respectively 742.19 ms and 863.87 ms, *P* < 0.001). Interestingly, accepting unfair offers was also done faster than rejecting them (811.74 ms vs 863.87 ms., *P* < 0.001). No statistics could be done on "fair-rejected" offers as they were almost inexistent.

For the two superfair offers, acceptance rates were lower than for normal fair offers (94% for both superfair offers compared to 98% for all fair offers, *P*=0.07), because one of the subjects systematically rejected them (the 15 other subjects accepted all of them). When accepted, RTs were significantly shorter for the first superfair offer (60-40) than for the other superfair offer (55-45, *P*<0.001) and for all other fair offers together (*P*=0.005).

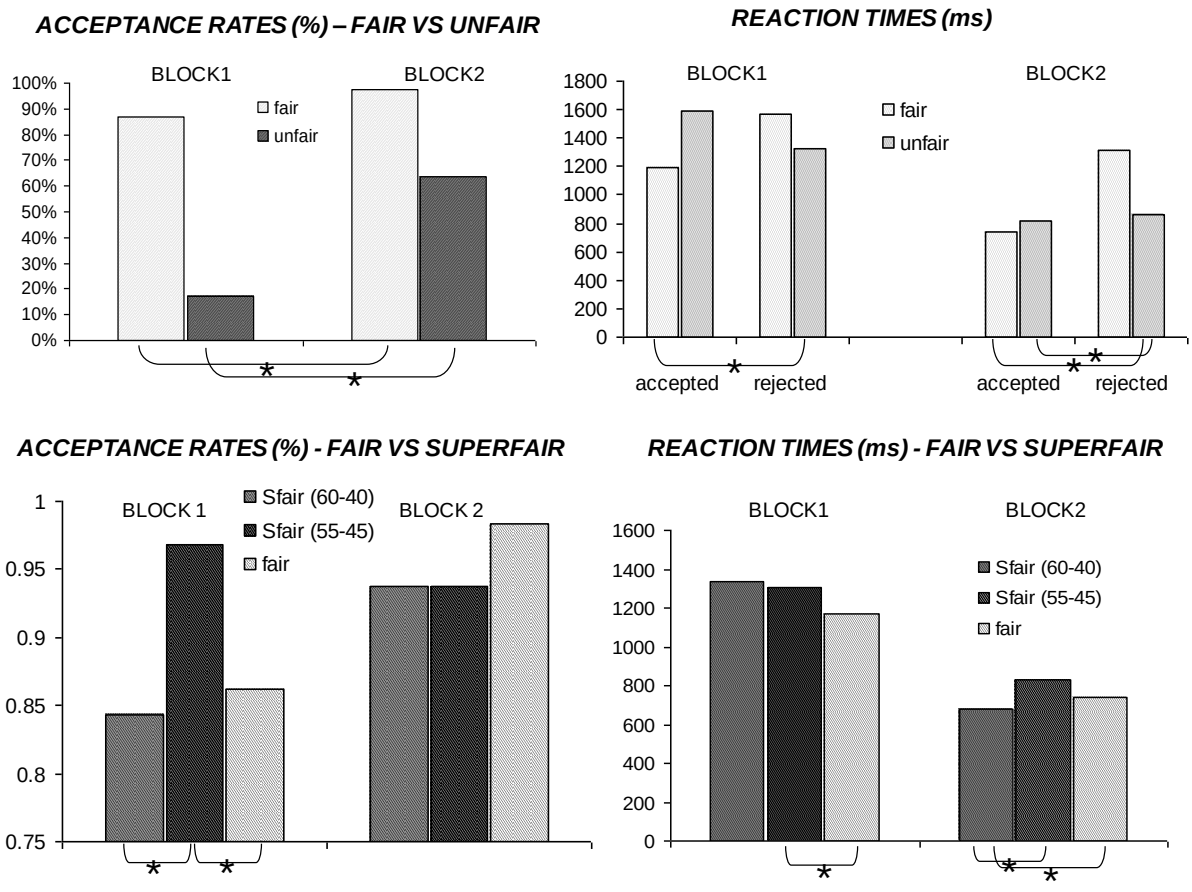


Fig2. *Acceptance rates and RTs for the two blocks. Significant differences are indicated with asterisks.*

A detailed analysis was also conducted depending on the subjects' responses (accepted/rejected), but no differences were found in the RTs of rejected offers, no matter the fairness of the offers. For the accepted offers, one difference was found between the most superfair (60-40) offer (shorter to accept) than the second most fair offer (45-55), longer to accept, consistent with previous results.

Behavioural inter-individual differences

The manipulation (becoming aware of the game theory's prediction and the "irrationality" of rejecting offers) drastically changed the behaviour of 6 subjects out of 16; these 6 subjects accepted all offers in the 2nd block (Group A). The 10 other subjects (Group B) still refused some offers in the 2nd block, although some of them decreased their rejection rates (from roughly 88% to 31%). We therefore also analysed their RTs separately.

In the second block, the 6 subjects who accepted all offers (Group A) were significantly faster at accepting unfair offers compared to Group B (758.64 ms.< 887.72 ms., P<0.001), whereas there were no significant differences in their RTs for the acceptance of fair offers. As Group A did not reject any offers in the second block, no comparisons on rejected offers were possible to perform. Interestingly,

going back to the first block to compare the RTs of these two groups, the opposite tendency is found: the subjects who accepted all offers in the first block are significantly *slower* than the other group on almost all comparisons ($P < 0.001$ for all the comparisons except for "fair-rejected" offers which are very rare in the first block (13%), see Figure 3). Importantly, as the buttons used to answer were balanced, those differences cannot be due to the dominance of one hand over the other.

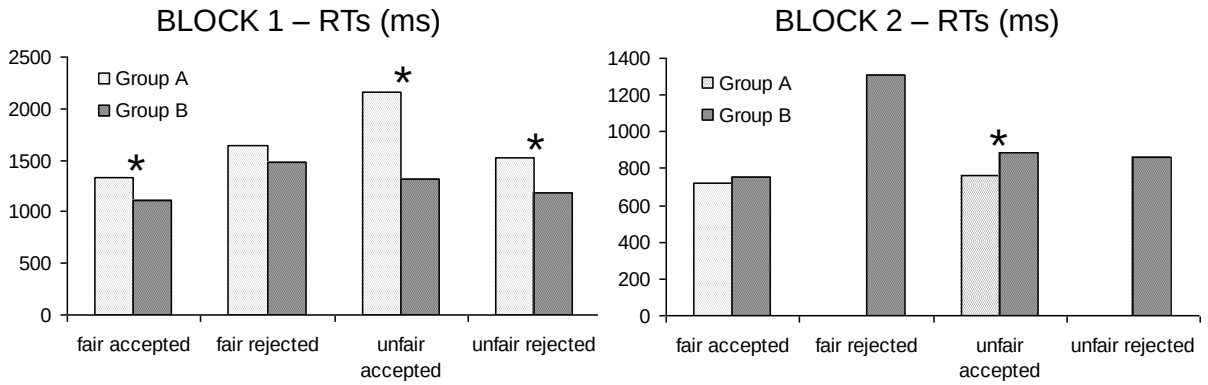


***Fig3**. Inter-individual differences in RTs: comparison of behaviour across blocks once the subjects are divided into two groups according to their behaviour in block 2. Significant differences (P < 0.001) are indicated with an asterisk. In the first block, participants of Group A (who accepted all the offers in block 2) are slower than Group B. This trend is inversed in the second block for the unfair offers.*

Electrophysiological results

We initially performed all the comparisons based on our classification of fairness (8 levels) but also based on the subjects responses (accepted/rejected). Although both will be described in the following section, we believe that comparisons based on subjects responses better reflect their perception and evaluation of fairness. Here, only differences found at least on 4 electrodes and lasting more than 20 ms are reported.

ERPs and map comparisons of fair vs unfair offers

The comparison of the average of fair offers and the average of unfair offers for both blocks and all subjects together revealed some significant differences between 500 and 550 ms on frontal and fronto-central electrodes (Az, Fz, F4, F6, FT8, FC6, FCz, Fz). Map comparisons did not yield any significant results. The subtraction of fair to unfair offers in B1 or in B2 did not reveal a precise map related to conflict detection

ERP and map comparisons of fair vs unfair vs superfair offers

ERPs comparison between these 3 conditions revealed consistent significant differences between superfair offers and fair/unfair offers. Chronologically, early differences appeared between 70 and 100

ms and were focused on 4 parieto-occipital electrodes (O1, POz, P10, and O2). These differences propagated in time and space, showing significant differences between approximately 120 and 170 ms on 8 parieto-occipital electrodes (P9, PO7, PO3, O1, Iz, POz, PO8, O2), and finally between 220 and 270 ms on 25 electrodes, including frontal, fronto-central, parietal and temporal electrodes.

ERPs and map comparisons of the fairest offer (50-50) to the most unfair offer (95-5)

This comparison yielded significant differences on parietal electrodes between 350 and 400 ms. The map comparison of these two offers did yield an interval of significant difference between 740 and 790 ms.

We then decided to compare electrophysiological responses according to the decision of the subjects (which better reflect their appreciation of the fairness of the offers) rather than our initial fair/unfair categorisation.

Comparison of accepted vs rejected offers in B1 and B2, all subjects together

In the first block, ERPs comparison of accepted vs rejected offers (all subjects together) revealed differences on 4 fronto-central (Fpz, F2, Fz, FCz) electrodes between 465 and 495 ms. Map comparison of accepted vs rejected offers did not yield any significant difference, neither block 1 vs block 2 accepted offers (Figure 4, upper panel). The comparison of block 1 vs block 2 rejected offers, and accepted vs rejected offers in block 2 were not performed because the technique requires equivalent groups. The subtraction of the rejected offers to the accepted offers revealed a map with frontal maximum corresponding to the differences found in the ERPs (between 465 and 495ms). In block 2, those comparisons could be done only for the group of 10 subjects who still rejected offers. The ERPs comparisons again revealed significant differences on frontal and fronto-central electrodes (Fz, F2, F4, FC2, FCz, Cz) between 480 and 500 ms. The subtraction of the rejected offers to the accepted ones at this precise time window showed a map with frontal maxima as in the first block.

Electrophysiological inter-individual differences

As we found significant differences in the behaviour of two subgroups of subjects, we performed the electrophysiological analyses accordingly to this division (Figure 4, middle panel). For Group A, whose members accepted all offers in B2, the MANOVA results on scalp topographies show no significant differences between responses or between blocks. For Group B, the topographical comparison between "Accepted vs. Rejected" offers did not yield any significant differences neither in B1 nor in B2. However a difference appeared in the comparison between blocks (Figure 4, middle panel): when comparing accepted offers in B1 to accepted offers in B2, a significant difference appeared between 335 and 365 ms

($P < 0.05$). At the same timing a trend is observed for the comparison of rejected offers in B1 and in B2 ($P = 0.057$). This trend becomes significant between 335 and 360 ($P < 0.05$) when taking into account only those offers which were rejected in both blocks (meaning that some rejected offers in the first block were then accepted in block 2, which is why the mere comparison "rejected in B1 vs. rejected in B2" yields only a trend).

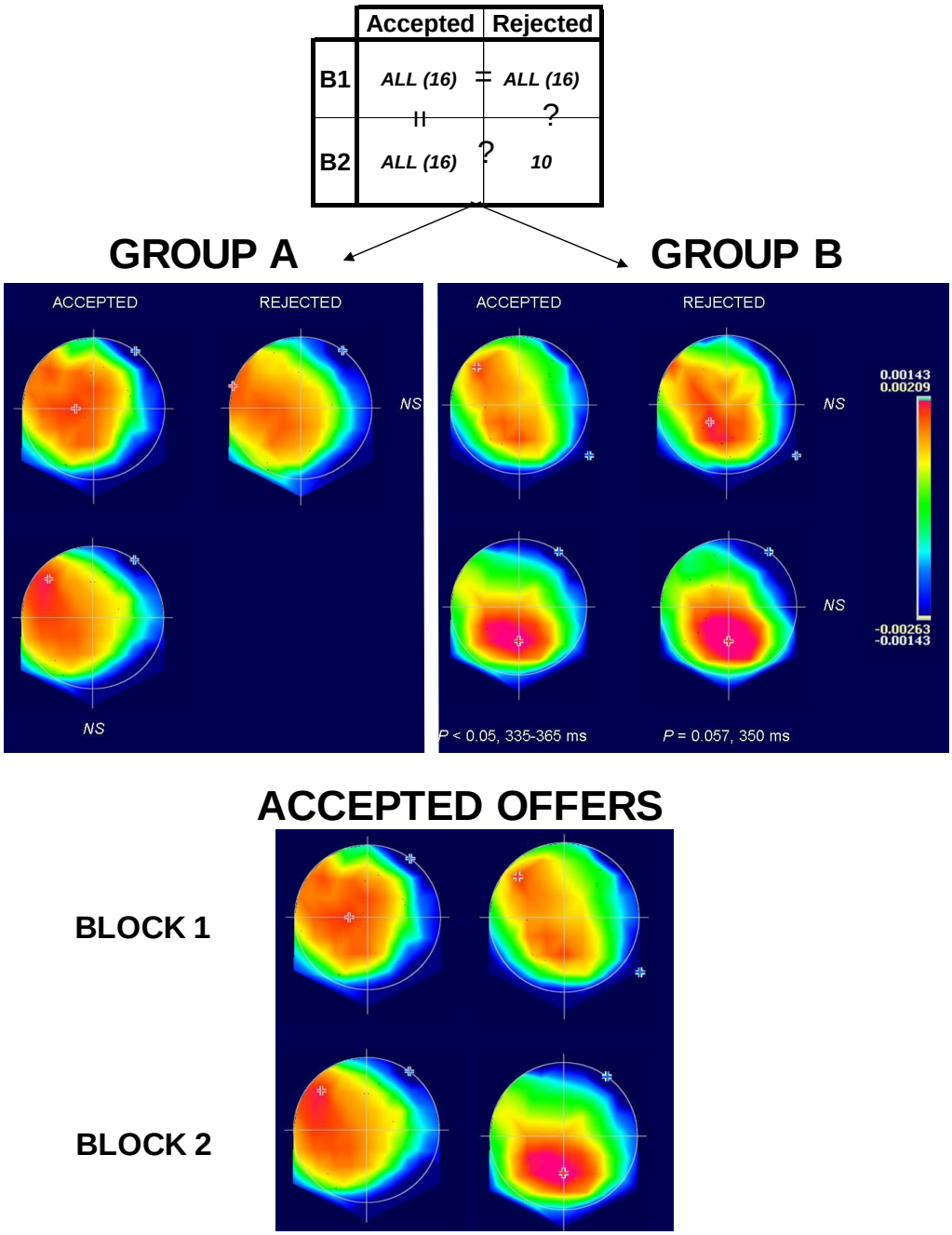


Fig4. *Upper panel (table) indicates the comparisons where no significant differences were found and those which could not be tested because of the inequality of the group's size. The second panel displays the MANOVA results once the groups are divided according to their behaviour in block 2. There are no significant differences for Group A. For Group B, significant differences are found between 335 and 365 ms for the comparison of accepted offers in block 1 and block 2. For the comparison of rejected offers, there is a trend around 350 ms. The last panel displays the maps related to accepted offers, in block 1 and 2 and for Group A (left) and Group B (right).*

Note that no differences were found in the pre-stimulus period. Consequently, differences observed in the topographies cannot reflect differences in the general state of this group of subjects between block 1 and block 2. The last panel of Figure 4 displays the maps of the accepted offers, in both blocks and for both groups to facilitate their visual comparison. Indeed, the conflict or need for control described in the "emotion" hypothesis arises when unfair offers are accepted. It is consequently in the maps of accepted offers that it should be observed.

To define which generators were responsible for the differences found in the topographical maps of the group who still rejected offers in the second block, we performed a source localization analysis on this group. For statistical power and as differences were found both for accepted and rejected offers, all offers were considered together. The results show that the main difference between B1 and B2 is a decrease in left prefrontal areas in B2 compared to B1 (superior and mid- prefrontal cortex (PFC), mid-orbitofrontal cortex (OFC)). Areas more active in the second block compared to the first one include bilateral lingual, L calcarine, L pre-and postcentral, bilateral fusiform, L superior occipital, L precuneus, R cuneus, L supplementary motor area, L mid-temporal, L inferior and superior parietal (Figure 5).

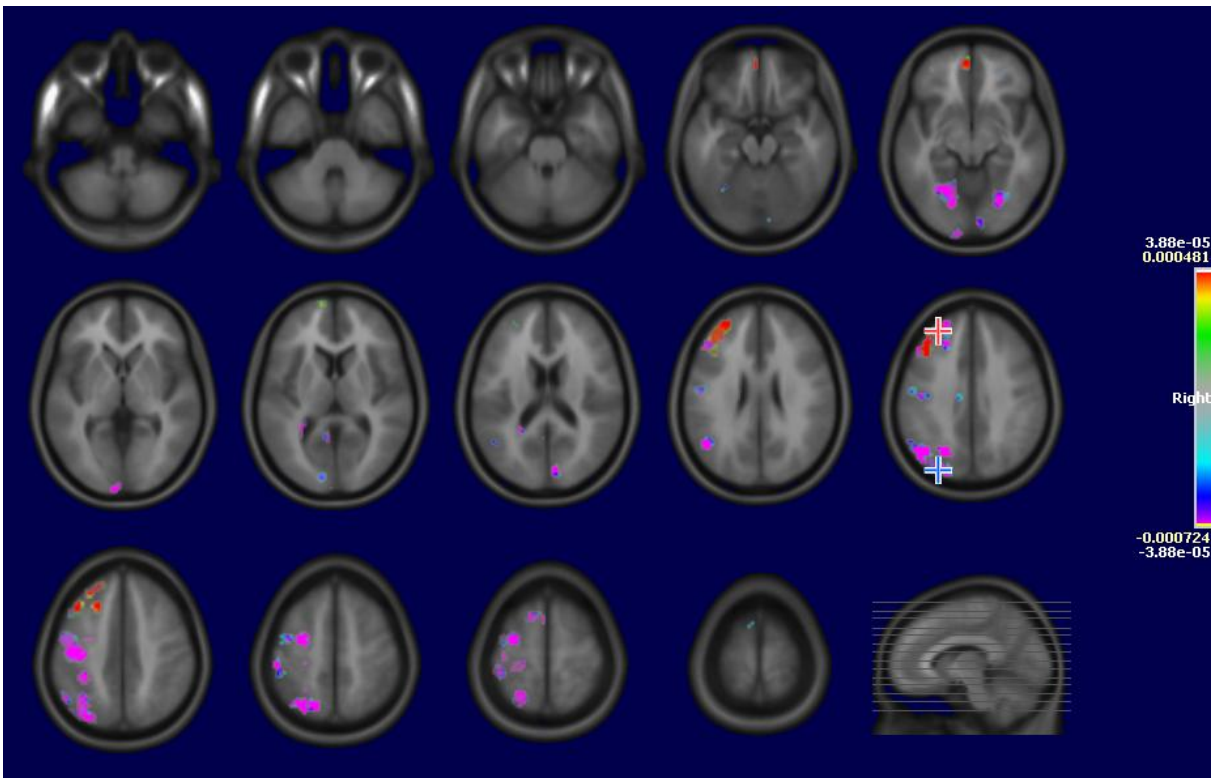


Fig5. *Brain voxels showing significant differences (P<0.05) between blocks for participants who still refused offers in the second block (Group B). Estimates of intracranial EEG for both blocks are obtained using ELECTRA inverse solution. Voxels where the amplitude of the estimated sources is significantly larger in the first block are coloured in the red-orange spectrum and voxels showing the opposite effect (B2>B1) in the blue-violet spectrum.*

Identifying a network correlating with levels of fairness

Knowing that the instructions given between the blocks strongly influenced the behaviour of the subjects and moreover, in a different way depending on the group, we will report here only the correlations found in block 1. Note that we performed the correlations between the degree of fairness of the offers (not the behavioural accept/reject distinction) and the activations.

The first significant correlation is positive and takes place between 150 and 180 ms. The network comprises (right) middle and (left) inferior orbito-frontal cortices, left anterior insula, bilateral anterior and left middle cingulate cortex, left superior temporal pole and left caudate nucleus (Figure 6).

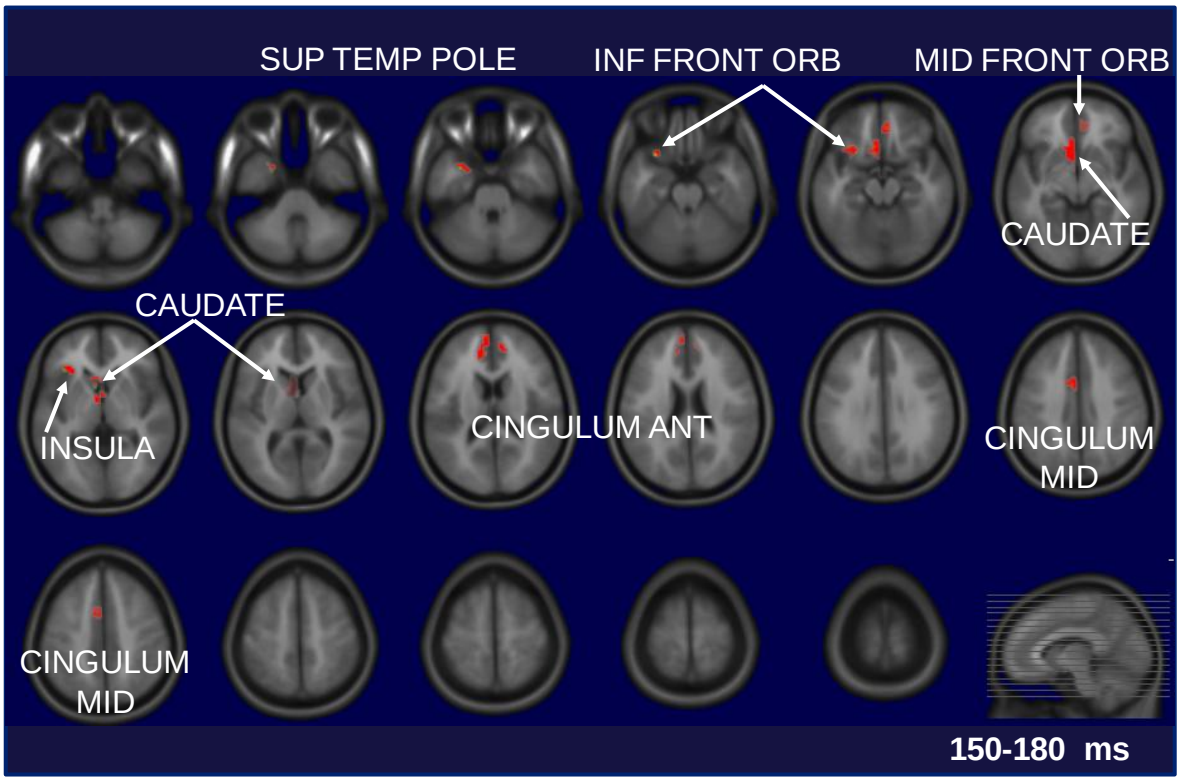


***Fig6**. Brain voxels showing significant correlation between the amplitude of the estimated intracranial EEG and the level of fairness during the interval comprised between 150-180 ms after offer presentation.*

Five other correlations have been found implying diverse structures. Some are nonetheless consistent through different time windows, such as bilateral postcentral gyri, anterior and middle cingulate cortices, right frontal and temporal structures (for a complete overview, see Table 1 in the Supplementary material). They will not be further discussed as their temporal proximity to the timing of responses does not allow to exclude possible confounds linked to motor response preparation or execution.

Discussion

Basic reaction to fairness, unfairness and "superfairness"

Although classical reactions to fair and unfair offers have been reported in a large number of studies, spontaneous unfair offers towards Player 1 (superfair offers for Player 2) have scarcely been tested. Rejections of "hyperfair" offers have been reported in some experiments (in Papua New Guinea, Russia and China, see Heinrich et al, 2006). However, in Europe and United States this behaviour was until now only detected by using more sensitive bargaining instruments (Andreoni, Castillo, & Petrie, 2003; Huck, 1999). The participants of our study thus seem to be a particular group. Indeed, in the first block, when the offer was very close to a perfect share (55-45) but advantageous for the participants, the RTs significantly increased whereas the mean acceptance rate rose up to 97%. Thus, when very close to a perfect share, the subjects hesitated but accepted getting more than Player 1. This cannot be explained only by the surprise, as when the offer was even more advantageous for the subjects (60-40), RTs and acceptance rates were no longer different from those of fair offers. This rejection of offers clearly advantageous to the participant goes against the view that subjects exhibit purely selfish behaviour: the game is anonymous and the participant never plays twice against the same counterpart, so no reputation or fear of reprisal can motivate this behaviour.

One subject spontaneously tackled this topic when answering to question 2 (*have you rejected any offers? If yes, can you justify the reason behind your rejection and the feelings linked to this situation?*): "Yes, when the share was not equitable and there was more money for me, it embarrassed me, I had the feeling of taking advantage of the other person". This subject is indeed the only one who still rejected superfair offers in the second block (and rejected them all). Thus the participants might be "embarrassed" by the fact of taking advantage of someone else. However, although anonymous to the other player, we cannot exclude that the subjects' embarrassment was not only altruistic but also linked to the fear of what the experimenter would think about their behaviour.

This hypothesis is all the more interesting since in the second block, a different pattern was observed: the RTs were shorter for the most superfair offer (even if the acceptance rate was lower, we know that this is due to one subject only). This indicates that in the second block, the fact that the subjects felt pressure to accept all offers combined with the fact that an offer is clearly advantageous to them facilitated the decision. The morally reprehensible behaviour of accepting unfairness when the subjects benefits from it (but rejecting it in the other direction) is compensated for by the excuse that it would be

"irrational" to reject any offer. There is no more reason to be embarrassed as subjects are only "following the instructions". The ERPs computed on the two superfair offers showed many significant differences with the other offers (fair/unfair): first focused on parieto-occipital electrodes, they progressively spread in time and space to finally encompass an area of 25 electrodes. A difference in amplitude might reflect the recruitment of a larger number of neurons, which could be due to the surprise as those offers are less frequent than the other kinds of offers.

More generally, in the second block, rejecting unfair offers took more time than to accept both unfair and fair offers, which signals the possible conflict induced by this situation. However, as subjects reacted differently, it is more relevant to interpret these results separately.

Dual System & Inter-Individual Differences

In Group A, the subjects seemed to adhere to the idea that rejecting unfairness is irrational, considering the goal of maximizing one's gains. They changed their way of playing the game by accepting all offers in the second block. This change in behaviour was surprisingly not reflected by electrophysiological changes when comparing map topographies. In Group B, however, being aware of the rationality expectations did not have the same effect: the networks involved in decision-making were significantly different from those used in the first block (a difference in scalp topographies implying necessarily the intervention of at least one different generator).

In order to better understand the differences between the blocks we performed an estimation of the Local Field Potentials on both blocks for the two groups. Our results first showed that - consistent with the map comparisons - there were no robust differences in Group A between block 1 and block 2. For the other group, the main difference between blocks consisted of a de-activation of prefrontal areas in block 2. This de-activation of frontal areas was concomitant with longer RTs than the other group, and an inconsistent (sometimes rejecting even more unfair offers in the second block than in the first one) and counter-productive behaviour (by rejecting some offers, the subjects of Group B necessarily earned less than those of Group A).

Thus, an intriguing issue is that the subjects who changed their behaviour (Group A) did not show electrophysiological differences (that would reflect an increase or decrease in the exerted control) whereas the group who behaved similarly in both blocks (still rejecting unfairness in block 2, Group B) showed electrophysiological differences not consistent with signs of a greater conflict.

To account for these results, two crucial points must be considered: first, in our analyses, all accepted offers were averaged together, no matter if they were fair or unfair. Second, the "emotion" hypothesis suggests that the need for control over negative emotions arises when subjects accept unfair offers for the sake of fulfilling their rational goal of gain maximisation.

In the first block, around 20% of the unfair offers were accepted, and in the second block, Group A accepted all offers, including the unfair ones. However, in Group B, 4 out of the 10 subjects *decreased* their acceptance rates of unfair offers in block 2 compared to block 1, while another subject of this group rejected all the unfair offers, as in block 1. Consequently, according to the aforementioned hypothesis, half of the Group B members exerted *less* control in block 2 compared to block 1. This is consistent with the results of inverse solution revealing a disengagement of prefrontal areas in the second block.

For Group A, who accepted all offers in the second block, the absence of signs of greater conflict or control in block 2 compared to block 1 might be explained by the following reasons: a) again, both kinds of offers are considered together (unfair as well as fair offers) and b) the decision to accept everything was probably taken even before the beginning of the block, which also explains that their RTs were dramatically lower than those of the first block and those of the other group (there is, strictly speaking, no more decision-making in this second block. It is even possible that the subjects did not read the offer at all).

Regarding dual system theories, these results thus suggest that in the first block (for all subjects) and in the second block (for the subjects who accepted all offers) automatic approach/avoidance behaviour (accept fairness – reject unfairness) co-existed with controlled behaviour (accept unfairness) to guide behaviour. Shorter RTs in the second block reflect the absence of a trial-by-trial decision, as this latter was taken even before the beginning of the second block. For the other group, the lesser activation of control areas (control system) might reflect the fact that half of its members relied only (or at least much more frequently than in the first block) on the automatic system and rejected unfairness more often than in the first block. As grouped with other subjects who showed behaviour halfway between the two extremes (accept all unfair offers – reject all unfair offers) the RTs were slower than that of Group A, but still faster than in the first block.

It is important to note that although a difference in map topographies necessarily involves a difference in the underlying generators, the converse is not true. Consequently, although the maps of rejected offers are not statistically different from those of the accepted offers, this does not exclude the observable map

from being the result of different networks - for instance, the conflict elicited by the rejection of fairness which occurred many times when offers were rejected, or the pleasure arising from punishing the opponent by rejecting his offer (de Quervain et al., 2004; Singer et al., 2006). Besides, ERPs comparison showed significant differences between accepted and rejected offers in both blocks. As the size of the sample cannot guarantee robust statistics, this question should be further investigated in future research.

This first set of analyses better corroborates the "emotion" hypothesis than the "social norms" hypothesis, as 1) some subjects still rejected unfairness although it was clearly stated that they jeopardized their chances to fulfil their goal by doing so 2) apparently, the other group of subjects still had to exert a form of control over their behaviour in order to accept unfairness.

Another way to assess the basis of rejection is to define which parts of the brain react to unfairness (emotions, or "cold" value assessor?). The results of the correlation analysis offered even more support to the "emotion" hypothesis.

Networks involved in assessing the fairness of the offer

Although the association between the fairness of an offer and the activations of a set of different regions has been reported (Padoa-Schioppa & Assad, 2008; Polezzi, Lotto et al., 2008; Rilling et al., 2004; Sanfey et al., 2003), a precise timing of their respective intervention was still missing to have a clearer depiction of the phenomenon. One study in particular reported a correlation between insula activation and the degree of fairness of the offers, but this link was established only on two values (the activation of insula was greater for a 9:1 offer than for an 8:2 offer). The activation of the insula was then interpreted as reflecting the negative emotion triggered by the unfairness of the situation, whereas the ACC activation would reflect the conflict between the rational goal and the emotional impulse and the DLPFC the control exerted over emotions when unfair offers are accepted (Sanfey et al., 2003).

Our data shows a first positive correlation taking place between 150 and 180 ms after offer presentation. Importantly, as the offers have been classified from most fair (rank 3) to most unfair (rank 8), a positive correlation means that the higher the unfairness, the higher the amplitude. The network comprises (bilateral) middle and (left) inferior orbitofrontal cortices, left anterior insula, bilateral anterior and left middle cingulate cortex, left superior temporal pole and left caudate nucleus.

The correlation between insula activation and unfairness might indeed reflect the negative emotion induced by the situation (Grabenhorst & Rolls, 2009; Wicker et al., 2003). The correlation between ACC (classically linked to conflict detection, Botvinick et al., 2001) and the unfairness is consistent with the idea that the greater the unfairness, the greater the conflict.

Left superior temporal pole has been linked to TOM networks (H. L. Gallagher & Frith, 2003) as being responsible for storing personal semantic and episodic memories to which one should refer in order to understand another person's intentions. In this experiment, we cannot exclude such mechanisms when the subject discovers the other player's offer. Caudate nucleus activation has been linked to reward-based behavioural learning (Haruno et al., 2004; Packard & Knowlton, 2002) as midbrain dopamine cells project to both ventral and dorsal striatum. Most importantly, its role in processing feedback has been demonstrated in a social context. Indeed, in a multi-round version of the Trust Game in which participants were scanned with fMRI, the head of the caudate nucleus has been shown as receiving or computing information about the fairness of the opponent's decision and as a predicator of the intention to pay back the decision with trust (King-Casas et al., 2005). Another study confirmed the involvement of caudate nucleus in differentiating between positive and negative outcomes in a social game but only if participants had no a priori on their partners - supplementary information thus diminishing the role of the structures involved in processing feedbacks (Delgado et al., 2005). The caudate nucleus has thus been defined as a key structure registering social prediction errors in order to better guide future decisions (Rilling, King-Casas et al., 2008). Globally, the fact that reward and memory structures correlate with the unfairness of the offer might also be explained by the fact that even if fair and unfair offers are equally represented in our game, people generally expect the others to be fair and consequently, a defection that violates social norms is always more surprising than a fair behaviour (Chang & Sanfey, 2009).

Finally the involvement of the orbitofrontal cortex is not surprising as recordings in non-human primates show that neurons of the orbitofrontal cortex encode economic value (Padoa-Schioppa & Assad, 2006) irrespective of visuospatial factors and motor responses but also independently of other values (i.e. they encode the absolute value of the offer, Padoa-Schioppa & Assad, 2008). Although a precise timing is not reported, the peak of firing rate of the OFC neurons is situated around 150-200 ms.

We have thus identified two different time windows where important processes occur.

First, between 150 and 180 ms the fairness of the offer is assessed by a network which shows greater activation to the most unfair offers and includes parts of the orbitofrontal and the cingulate cortices, left anterior insula, left superior temporal pole and left caudate nucleus. This network is compatible with previous findings (Sanfey et al., 2003) but in addition we suggest that other structures are sensitive to the fairness of the offer and we moreover precisely define the timing of this network's activation.

Second, we identified a time window (300-400 ms) during which inter-individual differences appear and impact the decision-making process. These differences are seen both in the behaviour and in the underlying neural networks, with one group seemingly relying on a combination of automatic and controlled mechanisms whereas the other one deactivated the areas of control, letting an automatic system alone guide the decision-making. Consequently we suggest that inter-individual differences should be systematically taken into account when attempting to model human decision-making. Indeed, our data - by offering a demonstration of the existence of a dual system of decision-making - also points out that its respective use can considerably vary across subjects.

Taken together, our results support the "emotion" hypothesis over the "social norm" hypothesis. The correlation between insula and the degree of fairness of the offers, as well as the written reports of the subjects and the greater activation of prefrontal areas when unfairness is accepted strongly suggest that first, there is an emotional reaction to unfairness and second, the control exerted over this emotion will determine the adequacy of the subsequent behaviour.

Supplementary material

1. Table of correlations between fairness and activity

Timing of the correlation (ms)	Positive/Negative	Region of activation
150-180	positive	R Mid Orbitofrontal, L Inferior Orbitofrontal, L Insula, B Anterior Cingulate Cortex, L Mid Cingulate Cortex, L Caudate, L Superior Temporal Pole
241-256	negative	Poscentral Gyrus L
317-334	negative	R Inferior Orbitofrontal , R Mid Cingulate Cortex
361-378	positive	R Mid Temporal, R Postcentral Gyrus
428-450	positive	R Postcentral Gyrus, R Inferior Parietal
451-486	positive	B Inferior, Mid & Superior Temporal, B Parahippocampal Gyri, B Fusiform Gyri, R Mid Orbitofrontal, B Mid Frontal, B Inf Frontal Tri, R Occipital, L Inferior, Mid & Sup Occipital, B Lingual, B Insula, R Calcarine, R Thalamus, R Caudate, R Rolandic Operculum, B Cuneus, B Angular Gyri, L Supramarginal Gyrus, R Precentral Gyrus, L Superior Parietal, R Mid Cingulum
494-525	positive	L Superior Temporal Pole, L Inferior, Mid & Superior Temporal, L Superior & Inferior Orbitofrontal, L Superior & Mid Frontal, L Superior Medial Frontal, L Inferior Frontal Tri, L Putamen, L Caudate, L Insula, L Postcentral Gyrus, L Mid Occipital
	negative	R Rolandic Operculum, R Mid Temporal

2. Answers of the subjects to the questionnaire (non translated)

Question: **Avez-vous rejeté certaines offres? Si oui, pouvez-vous expliquer les raisons de vos refus et les sentiments liés à ces situations? Pensez-vous que ce comportement est justifiable?**

S1: Oui lorsque l'offre me paraissait injuste. Le sentiment d'**injustice** me faisait refuser l'offre.

S2: Oui, j'ai rejeté certaines offres. Car l'autre partenaire me proposait peu d'argent, par rapport à ce qu'il se gardait. J'ai accepté les propositions dans lesquelles on «gagnait» pareil ou à peu près autant. Ce comportement peut être justifiable car si deux personnes s'associent, ce n'est pas pour qu'il y **en ait une qui profite de l'autre**.

S3: Oui, j'ai rejeté des offres. J'ai rejeté celles où je n'avais presque rien à gagner, où cela n'avait **pas d'intérêt** pour moi. Je pense que c'est justifiable dans la mesure où je rejette les offres où l'autre gagne beaucoup plus que moi, que ce n'est pas du tout **équitable.**

S4: Oui j'ai en rejeté car mon partenaire ne me proposait pas un partage assez **équitable**. Pour moi c'est ou 50-50, ou moi plus, ou alors je ne dois pas avoir moins que les deux tiers de la somme. Mais sur les 6.- 4.-, la somme (4.-) ne représente pas assez. Non, c'est injustifié car la personne me propose un partage sans retour *(NDA: c'est-à-dire comme c'est une seule interaction, il devrait tout accepter, m'a-t-il expliqué).*

S5 : Oui, j'ai refusé plusieurs offres, car je trouvais que le partage n'était pas **équitable**. J'étais embêtée d'avoir à refuser une offre, mais je ne suis pas tellement d'accord de **me laisser avoir**. Avant de commencer le jeu j'ai pensé que je n'avais qu'à accepter toutes les offres, et constituer ainsi un capital maximum; mais j'ai également constaté que si je refusais certaines offres je ne perdais finalement pas grand-chose, puisque celles-ci n'étaient pas très élevées, alors que celui ou celle qui a fait la proposition perdait beaucoup plus. J'ai éprouvé alors une certaine satisfaction à lui faire payer le prix de son avarice.

S6: Oui, à cause des disproportions exagérées dans le partage. J'ai la sensation que ce n'est **pas juste** du fait d'une trop grande différence, donc qu'on **profite** de moi. A ce stade je ne sais pas si ce comportement est justifiable du fait que je ne sais pas s'il y a une contrepartie à l'offre qu'on me fait. Si non, ce serait très absurde de refuser de l'argent même si le partage est disproportionné.

S7: Oui. Quand le partage n'étais pas équitable 1) plus d'argent pour moi: cela me gêne, j'ai l'impression de «profiter» de la personne ; 2) trop de différence entre la somme que la personne garde et moi-même: j'ai l'impression que **ça ne vaut pas le coup**. Ou: Quand la personne m'a semblée triste, démunie (habits et expressions faciales), handicapée (j'aurais honte de prendre de l'argent).
Oui je pense que c'est justifiable en fonction des valeurs humaines que j'ai.

S8: Oui. J'ai refusé lorsque le partage ne me semblait pas équitable (p.ex. 1.- pour moi et 9.- pour lui/elle). Sinon j'acceptais, même si j'avais une proposition un peu plus faible. Ce comportement est justifiable dans ces conditions de jeu! Dans le sens où **je n'ai pas envie de me faire avoir**. Mais dans d'autres conditions, par exemple si le jeu était réel, j'aurais plus souvent accepté l'offre car même 2 petits francs m'auraient permis de m'acheter un café.

S9: Oui, plusieurs. Principalement celles qui étaient très en dessous de l'offre que devait avoir mon partenaire. Pourquoi devrais-je avoir une offre inférieure à la sienne ? Il n'y a aucune raison. C'est une sorte d'**injustice**.

S10: Rejet des offres qui étaient **trop discrépantes** (grande différence entre ce qui était pour eux et pour moi). En principe, si la différence était inférieure à 4 j'acceptais *(NDA : je pense qu'elle voulait écrire «refusais» ici)* surtout pour les offres de valeurs peu importante. Quand la valeur était plus élevée, j'acceptais une différence un peu plus grande.

S11: Oui. Les parts n'étaient pas égales, les personnes faisant les offres n'inspiraient pas **confiance**. Comportement plus ou moins justifiable. J'ai rejeté les offres avant tout par rapport à mon intuition.

S12: Oui, quand la somme proposée est peu importante. Je trouve des fois que **c'est inégalitaire**. Je pense que ce comportement est justifiable. Des fois je refuse une offre (mais rarement) en fonction du visage de la personne.

S13: Oui. Pour jouer le jeu, question de quand-même regarder les photos et les offres. Mais je crois que ça ne vous a pas rendu service en fait…Je ne refusais que si l'offre était, en plus de paraître **injuste** après une photo **peu sympathique**, **ridiculement basse**, pour ne pas renoncer à une somme relativement intéressante.

S14: J'ai refusé certaines offres car le gain que j'obtenais n'était pas égal ou supérieur à 50%. J'ai essayé de décider sir une base rationnelle. J'ai donc essayé **d'exclure mes sentiments** dans la prise de décision.

S15: J'ai refusé certaines offres car je n'avais **pas confiance** en les personnes qui s'affichaient à l'écran. Oui c'est justifiable car il est important dans la vie de tous les jours d'avoir de l'intuition (la personne nous semble-t-elle rassurante ou non?)

S16: Oui, j'ai refusé les offres dérisoires. Exemple: si l'on me propose 75 centimes je vais refuser car je me sentirais **offensé.**

STUDY 4

CODING MECHANISMS IN LOCAL FIELD POTENTIALS (LFPS) BEHIND DISAPPOINTMENT/ELATION ASYMMETRICAL EFFECT

In this study, we were interested in better understanding the neural basis of the strong impact of disappointment in the Trust Game. Indeed, although we found large inter-individual differences in reaction to disappointment in Study 1, some structures (common to all subjects) might nonetheless differently code negative, positive and neutral outcomes as the behaviour is subsequently modified depending on the outcome category. We hypothesised that this effect should be observed in the neural signals recorded with electrodes implanted in deep structures of the human brain.

Methods

Trust Game

The Trust Game was similar to that played by the participants of Study 1. The number of trials varied depending on the availability of the patient but the proportion of outcome types remained stable (¼ disappointment, ¼ neutral negative, ¼ neutral positive, ¼ elation).

Patients

All the patients signed an informed consent approved by the local ethics committee before starting the experiment. They were all hospitalized in the Presurgical Unit of the Geneva University Hospitals for pharmaco-resistant epilepsies. The goal of their stay in the unit was to precisely localize the epileptic focus in order to remove it surgically. The localization process is done by implanting strips of 8-10 electrodes in diverse structures of the brain and recording the activity over a significant period of time.

Patient 1 was a 54-year-old woman who played 2 blocks of 60 trials each. Electrodes were implanted in the anterior and the posterior bilateral hippocampus, and in the bilateral amygdala. The epileptogenic trigger zone was located in the mesial temporal lobe (a sclerosis of the left hippocampus was identified). Left hippocampus and left amygdala were surgically removed and the crises disappeared.

Patient 2 was an 18-year-old man who played 2 blocks of 40 trials each. Electrodes were implanted in the left (anterior and posterior) and in the right hippocampus, in the bilateral amygdala, in the left fronto-

orbital and fronto-lateral cortex. The epileptogenic trigger zone was located in the left anterior temporal lobe whose external part was surgically removed (sparing the left hippocampus).

Patient 3 was a 20-year-old man who played 2 blocks of 40 trials each. Electrodes were implanted in the bilateral hippocampus, in the bilateral amygdala, in the right fronto-orbital and the right temporal pole. The epileptogenic trigger zone was located in the left amygdala. The patient did not undergo a surgical operation.

Data Analysis

Reaction Times between patients and experimental group (N=19) were compared with Wilcoxon signed rank test.

General EEG pre-processing

Mean event-related potentials (ERPs) were obtained by averaging segments of the EEGs aligned on the onset of the outcome using custom Matlab scripts. EEG traces were 1500 ms with 500 ms as pre-stimulus period (time 0 = outcome presentation). All epochs were visually inspected before averaging to eliminate any potential interictal epileptic events (e.g. spikes). To eliminate as much as possible effects due to volume conduction and highlight local events we transformed the EEG traces into bipolar recordings by referencing each electrode to its closest electrode from the recording strip. A notch filter at 50 Hz and superior harmonics was used to eliminate power source contamination. To ensure that responses were specific to the evaluation of the outcome rather than visually evoked we subtracted the 500 ms pre-stimulus mean from the EEG signal at all points (baseline correction). As the outcome presentation was preceded by the countdown in which the expected return is flashed three times to remind to the subjects, visual responses cannot account for potential responses found in the post-stimulus analyzed period.

Detecting brain areas with outcome-related responses

To investigate which contacts (brain regions) showed outcome-related responses (ERPs) we transformed the mean ERPs into z-scores by using the mean and variance computed over the 500 ms baseline that preceded outcome presentation. A contact was considered as showing outcome-related responses at some time frame if the mean ERP deviated in $\pm$ 2.85 standard deviations from the baseline mean. The deviation had to remain between previous limits for at least 20 ms to consider the site as outcome responsive.

Detecting brain areas with selective outcome-related responses

To detect contacts where the ERP responses showed selectivity for the type of outcome, we compared, on a time-frame by time-frame basis, the median of the distribution of voltages recorded over the three following categories: 1) Disappointment, where subjects received much less than expected (reward expected and not received); 2) Neutral, receiving approximately what they expected; 3) Elation, where they received much more than expected (unexpected reward). Since voltage distributions across trials were not following a normal distribution with unspecified mean and variance (Lilliefors test, $P > 0.35$, mean over all timeframes) we relied on a non-parametric version of the repeated-measure ANOVA, the Kruskall-Wallis test, to compare the ERPs across outcome categories. The Bonferroni correction was used to adjust for the multiple tests over each time frame. Statistical differences across categories were accepted when the adjusted p-values were smaller than 0.01.

Time-frequency analysis based on Stockwell Transform and Phase Locking Factor (PLF)

ERP analysis alone ignores activity that is not tightly locked to the event, and contains no information concerning modulations of the phase and power of oscillatory activity. We therefore expressed the EEG traces as time-frequency representations (TFRs) using the Stockwell-transform, which provides information about the phase and power of each frequency component of the EEG over time.

The Stockwell transform (S-transform), developed in 1996 for analyzing geophysics data (Stockwell, Mansinha, & Lowe, 1996) is a modification of the Continuous Wavelet Transform. In contrast with the Wavelet Transform (which describes signals in terms of scales and dilations) the S-transform deals directly with time and Fourier frequencies. In the S-transform, the mother wavelet is created from two components: an oscillatory exponential kernel $exp\{-i\ 2\ pi\ f\ t\ \},$ which determines frequency, and a Gaussian envelope, which is translated across the signal in order to localize power and phase. The complex-valued output is directly invertible into the Fourier transform spectrum, but is characterized by the frequency-dependent time resolution of the wavelet transform. Also, the S-transform produces absolutely referenced local phase information (Stockwell, 2007), since the oscillatory kernel is not translated across the signal as in wavelets. Previous reason makes the S-transform particularly suitable for the analysis of phase locking.

Evoked potentials have been proposed to result from phase locking or phase reset of electroencephalographic (EEG) activities within specific frequency bands (Sayers, Beagley, & Henshall, 1974). Since then, considerable evidence for such stimulus-induced phase locking effects has been accumulated by several research groups, using varied signal analysis methods (Haenschel, Baldeweg,

Croft, Whittington, & Gruzelier, 2000; Jansen, Agarwal, Hegde, & Boutros, 2003; Makeig et al., 2002; Winterer et al., 2000). Additionally, the phase rather than the amplitude of induced oscillations seems to correlate with the generation of action potentials.

For each frequency, the S-transform produces a complex time series $w(t,k)$, where t represents the time point within trial k. The instantaneous phase is defined as the angle, i.e.

$$\varphi(t,k) = \arctan\frac{\mathrm{Im}(w(t,k))}{\mathrm{Re}(w(t,k))}$$

The phase locking factor (also known as inter-trial coherence) was defined as:

$$PLF(t,f) = \frac{1}{n}\left|\sum_{k=1}^{n}\frac{\varphi_k(t,f)}{|\varphi_k(t,f)|}\right|$$

The Phase Locking Factor measures the consistency across trials of the EEG spectral phase at each frequency and time point. The PLF takes values from 0 to 1 with one representing perfect phase coherence across trials. The presence of phase locking at a particular frequency can be evaluated by applying the Rayleigh test for non-uniformity of the phase. Significance was set to 0.01 adjusted by the number of frequencies considered (from 1 to Nyquist frequency).

Frequencies and contacts influencing behavioural changes after unexpected outcomes

According to our experimental design, two different variables might influence the expectations of the Investor about the amount given in return by the Trustee: 1) the outcome of the previous trial and 2) the trust granted to the current Trustee according to his/her face. To investigate how previous outcome influences expectations independently of current player's trust we used partial correlation analysis. Partial correlation analysis is aimed at finding linear partial correlation coefficients between pairs of variables after removing the effects of other variables. The basic goal is to remove spurious correlations (i.e. correlations explained by the effect of other variables) so as to reveal hidden correlations. To compute partial correlations we relied on the Matlab function "partialcorr". Using this function we computed the sample Spearman (rank) correlation coefficients between the power of the oscillations of each fixed time-frequency pair in trial i-1 and the update in expectations (temporal derivative in expectations) from trial i-1 to trial i. The nuisance (controlled) variable was the trust assessment in trial i. The correlation was considered significant when $P<0.01$. The *p*-values for rank partial correlations were computed using a Student's *t* distribution for a transformation of the correlation. This is exact if the power distribution across trials follows a normal distribution and so does the trust (Lilliefors, $P < 0.05$).

Results

Behavioural results - Questionnaire

Question 1: In the trials where you invested a great amount and received a small one in return, which emotion corresponds best to your feelings at that moment?

Patient 1: Disappointment
Patient 2: Anger
Patient 3: Disappointment

Question 2: During the experiment, did you have the feeling that there was a link between your trustworthiness ratings and the outcomes, or do you think that the outcomes were randomly assigned to the faces?

Patient 1: Yes, we must base our judgment on the first impression of the face
Patient 2: Yes, there was a small link
Patient 3: No, I didn't think about this. The questions had no logic for me, totally random

Question 3: If you had to choose between the following emotions, which one corresponds best to your feelings when you invested a lot and received a little amount in return (disappointment-disgust-regret-betrayal-anger)?

Patient 1: Disappointment
Patient 2: Betrayal
Patient 3: Anger

Question 4: During the experiment, do you think that you modified your way of playing as a function of the preceding trials? If it is the case, was it voluntary or beyond your control?

Patient 1: No
Patient 2: I tried voluntary to extract information from the outcomes, on the basis of faces' expressions
Patient 3: Yes, I changed my way of playing as a function of my opponent's faces. I play more aggressively when they look like shady persons. This change was totally involuntary.

Correlations

High significant correlations were found between the three variables: trust (TR), investment (INV), expected return (ER), suggesting that the patients based their investment behaviour on their trustworthiness ratings of the different Trustees (Table 1). Some significant correlations were also found with the variable certainty (CT).

		PA1	*PA2*	*PA3*
Block1	TR-INV	**0.8**	**0.85**	**0.83**
	TR-ER	**0.77**	**0.9**	**0.65**
	INV-ER	**0.96**	**0.96**	**0.86**
	TR-CT		0.23	
	INV-CT		0.2	
	ER-CT		0.25	-0.23
Block2	TR-INV	**0.93**	**0.9**	
	TR-ER	**0.91**	**0.9**	
	INV-ER	**0.97**	**0.96**	**0.95**
	TR-CT		**0.49**	
	INV-CT	**-0.33**	**0.57**	
	ER-CT	**-0.28**	**0.54**	

Table 1. *Significant linear correlations between the different behavioural parameters in both experimental blocks. Correlations under the significance threshold of P < 0.001 are shown in bold, otherwise P < 0.05.*

Reaction Times

Patients were generally slower than the subjects of the experimental group (EC groups) of our first study as shown in Figure 1.

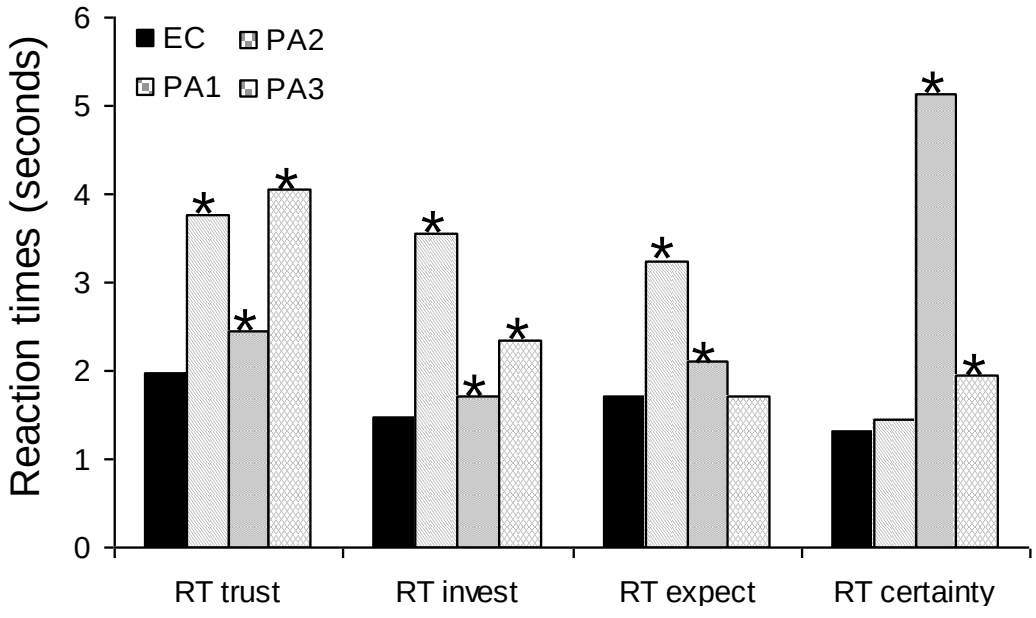


Fig1. *Reaction Times (RT) measured in seconds for the three patients (grey shadings) and a control group formed by 19 healthy volunteers (black). Significant differences (P < 0.05, Kruskal-Wallis) between the EC group and the patients are indicated by an asterisk.*

Totals

Patient's total Trust, Investment, Expectations, Certainty and Gains were compared to the experimental group of our first study (Figure 2).

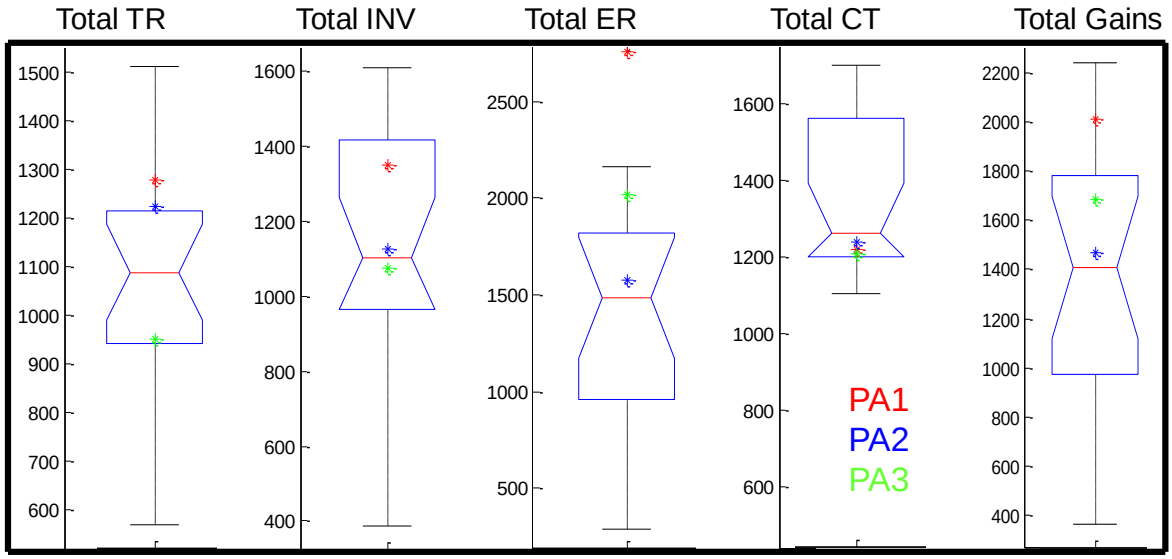


***Fig2**. Patients' total trust, investment, expectations, confidence and gains compared to the EC group. All scales represent the amount in Swiss francs except for the variable certainty (added values of the 10 points scales).*

Impact of previous disappointment on current expectation: the Disappointment Tolerance Threshold

The behavioural results of the three patients demonstrated that they were sensitive to previous disappointment in the same way that the participants of Study 1. Although variable, a DTT was observed for all patients (see Figure 3 below); the change observed for the variable Expectation was significantly greater when the outcome of the previous trial was disappointing compared to neutral, and also compared to elation for patients 1 and 3.

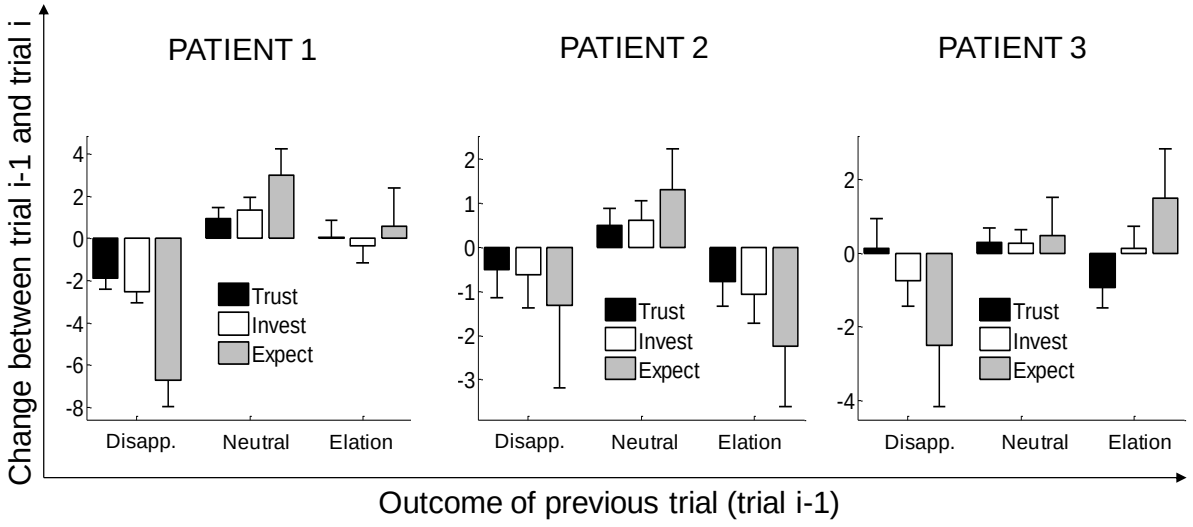


***Fig3**. Changes in the three variables (Trust, Investment and Expectations) between previous trial and current trial, according to the outcome of the previous trial. Note that the expectations' decisions following a positive outcome (elation) strongly varies between the patients.*

Electrophysiological Results

After visual inspection of raw EEG data, 14 trials were rejected for the first patient, 7 for the second patient and 2 for the third patient.

Medial temporal lobe (MTL) contacts show Outcome-related potentials (ERPs) and Outcome Selectivity: The strongest and most consistent ERP responses across patients were observed in the medial temporal lobe (Figure 4, page 89). Medial temporal lobe (MTL) ERPs were characterized by an initial component peaking around 200 ms. This component was followed by a larger and more sustained component with maxima around 600 ms which started to significantly deviate from baseline at around 450 ms. Polarity inversions were often seen between neighbour contacts and the peak-to-peak amplitude on the responses was very large (~200μV) for bipolar contacts. Both the size of the response and the polarity inversion suggest that the response is local to the neighbourhood of the recording tip instead of an effect due to volume conduction.

The initial component of the response at 200 ms was weakly (P<0.05) selective for the outcome category (Figure 5, page 90). In two patients, the initial response differed between elation and the other outcomes. In the third patient, we observed a significant difference between expected (neutral) and unexpected (elation and disappointment) outcomes. The second component showed significant selectivity for the unexpected outcomes.

Outcome-related responses were occasionally observed in the orbitofrontal and frontal cortex (Figure 5). On bipolar recordings, frontal responses were overall of small amplitudes and had little reproducibility across patients. This lack of reproducibility is likely due to the positioning of the contacts that necessarily varied from one patient to the other.

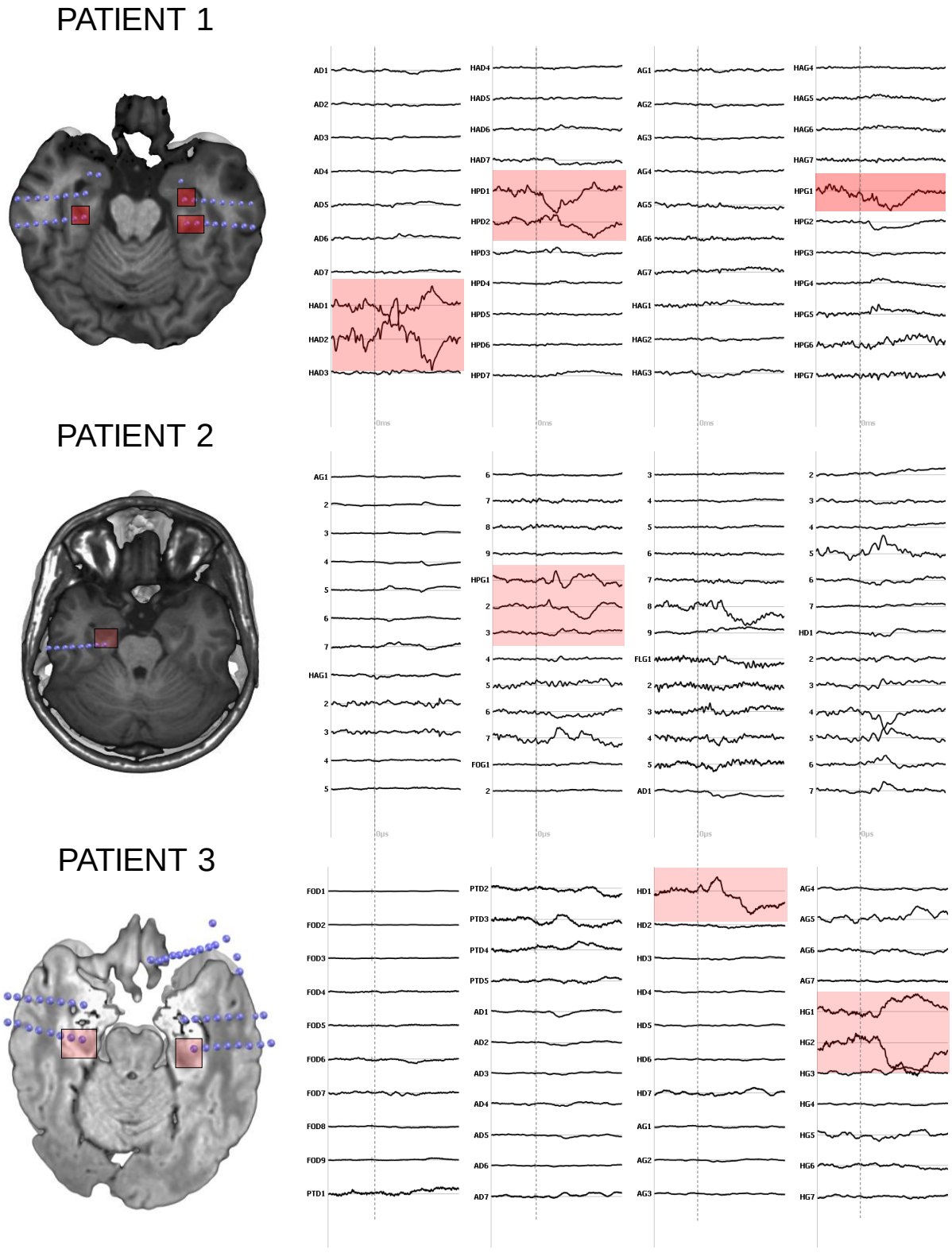


Fig4. *Outcome-related potentials observed in the three patients after transforming the original data into bipolar recordings to emphasize locally-generated events. Outcome is given at time zero and is indicated by a light grey vertical line. The 500 ms pre-stimulus where visual responses should be present if existing is shown. High amplitude responses are systematically observed in contacts at the hippocampal/parahippocampal area (shaded in light red in both the ERPs and the anatomical image of the patient). Note that polarity inversions across contiguous hippocampal contacts are observed for the three patients suggesting that Outcome-related potentials are locally generated by the hippocampus.*

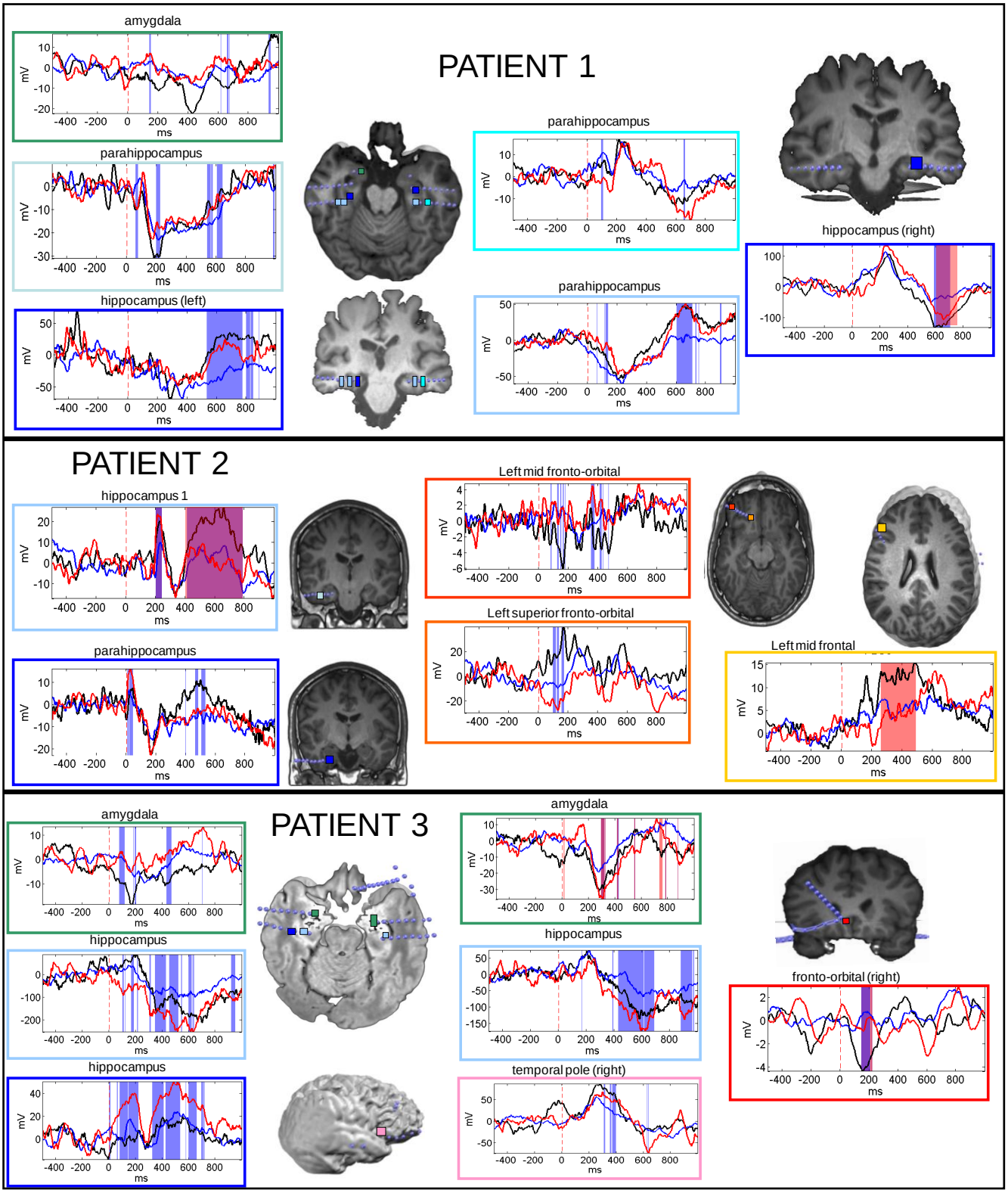


Fig5. *Outcome-related potentials showing selective responses. Traces are unfiltered (except for notching at 50 Hz and its harmonics). Mean responses to disappointment are drawn in black, to elation in red. Blue traces mark the mean responses when there is no mismatch between patients' expected and received reward. Small coloured squares overlaid on the patient anatomical image indicate the location of the contact where the traces were recorded (bounding box colour coincides with colour of the square in the anatomical image). Periods where the three categories significantly differ (Kruskal-Wallis, P<0.01) are highlighted in blue/violet.*

Outcome-related inter-trial phase alignment of Theta and Alpha Oscillations

Theta and alpha band oscillations showed significant phase alignment across trials for the three patients. No consistent effects were seen for the other frequency bands. The localization of the contacts and the timing of the theta and alpha phase locking were however different.

Theta phase locking was maximal around 4 Hz and started right after outcome presentation (Figure 6a). It was significant on MTL contacts (Figure 6b) with maxima over hippocampal contacts in the left hemisphere, and remained significant during the first 500 ms after outcome presentation.

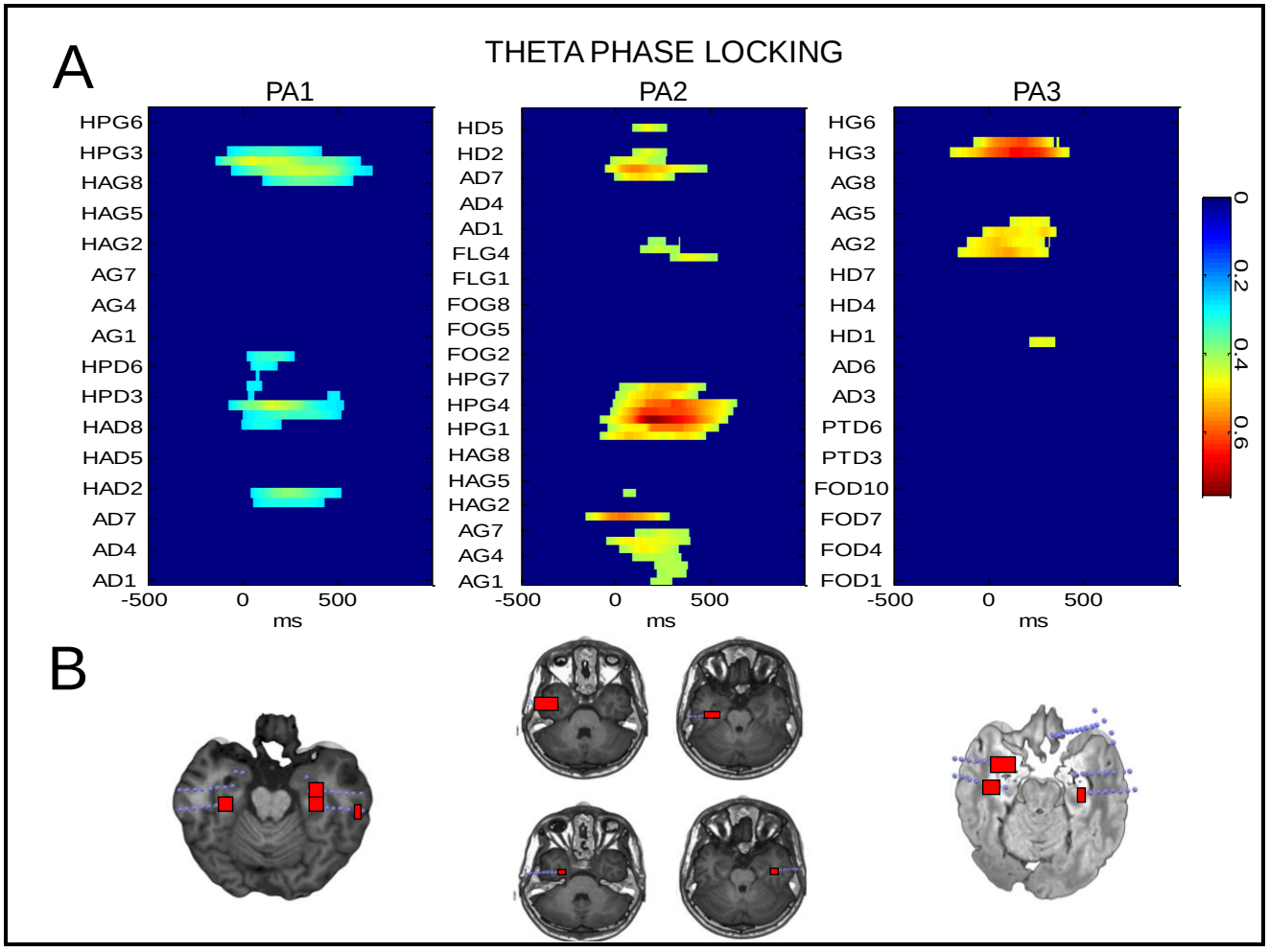


Fig6. *The phase of theta oscillations becomes significantly aligned across trials after presentation of the Outcome (time zero). Each column depicts the results for a single patient. (**A**) Plots of the Phase Locking Index across trials observed within the theta band (6 Hz) as a function of time (horizontal axis) and space (contacts, vertical axis). Only significant values (P<0.01, adjusted, Rayleigh test for non-uniformity of the phase) are shown. (**B**) Axial MRI cuts showing the position of the contacts were the Phase Locking was highly significant (P<0.01).*

Alpha band phase alignment (maximal at 8Hz) had a slightly later onset (around 50 ms after outcome presentation) than the theta phase locking (Figure 7a). Alpha phase locking was highly significant on frontal contacts and on the deepest, more anterior MTL contacts placed at/near the amygdala (Figure 7b). Its duration was shorter than theta phase locking, disappearing at around 400ms.

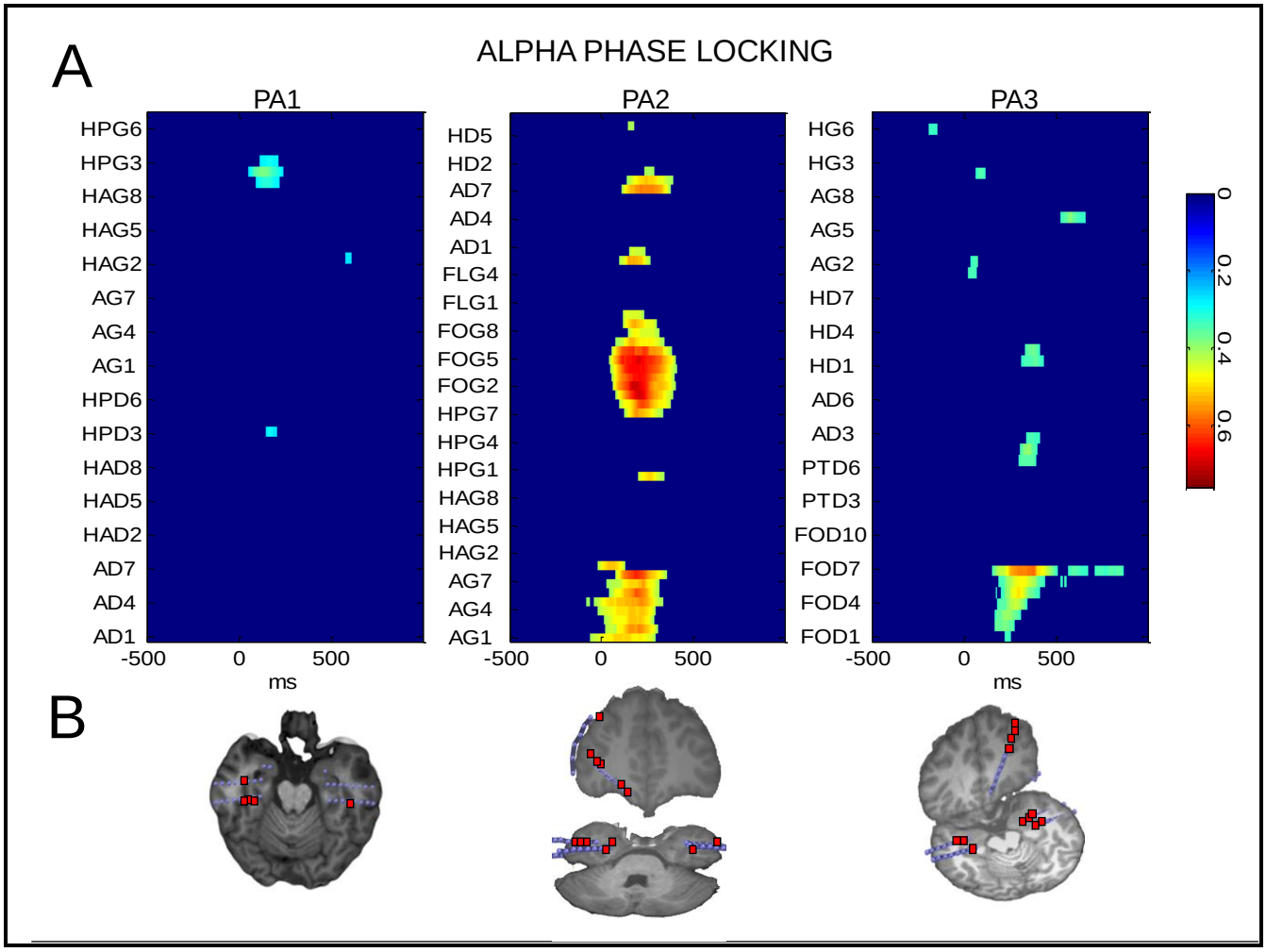


Fig7. *The phase of alpha oscillations becomes significantly aligned across trials after presentation of the Outcome (time zero). The contacts showing alpha phase alignment are different from those showing theta phase alignment. Alpha Phase Locking is maximal over frontal electrodes and anterior contacts (amygdala) in the MTL. (**A**) Significant values of the Phase Locking Index (8Hz) as a function of time and space. (**B**) Axial MRI cuts showing the precise location of the contacts where significant alpha Phase Locking was detected (P<0.01).*

Theta/Gamma oscillations on MTL contacts influence behavioural changes after unexpected outcomes

Contacts placed on the rhinal cortex showed a significant correlation between on one side, the theta and the high gamma oscillations after outcome presentation and on the other side, changes in the expected return in the following trials (Figure 8). The correlation pattern in the time-frequency domain revealed an interesting structure since theta correlation was typically accompanied or followed by correlation bursts

within the high gamma band that often extended well into the epsilon oscillations range (above 100 Hz). Correlations were significant in the period from 300 to 800 ms after outcome presentation and were quite reproducible across patients except for shifts in latency. No significant correlations were observed in cortical areas or frontal contacts.

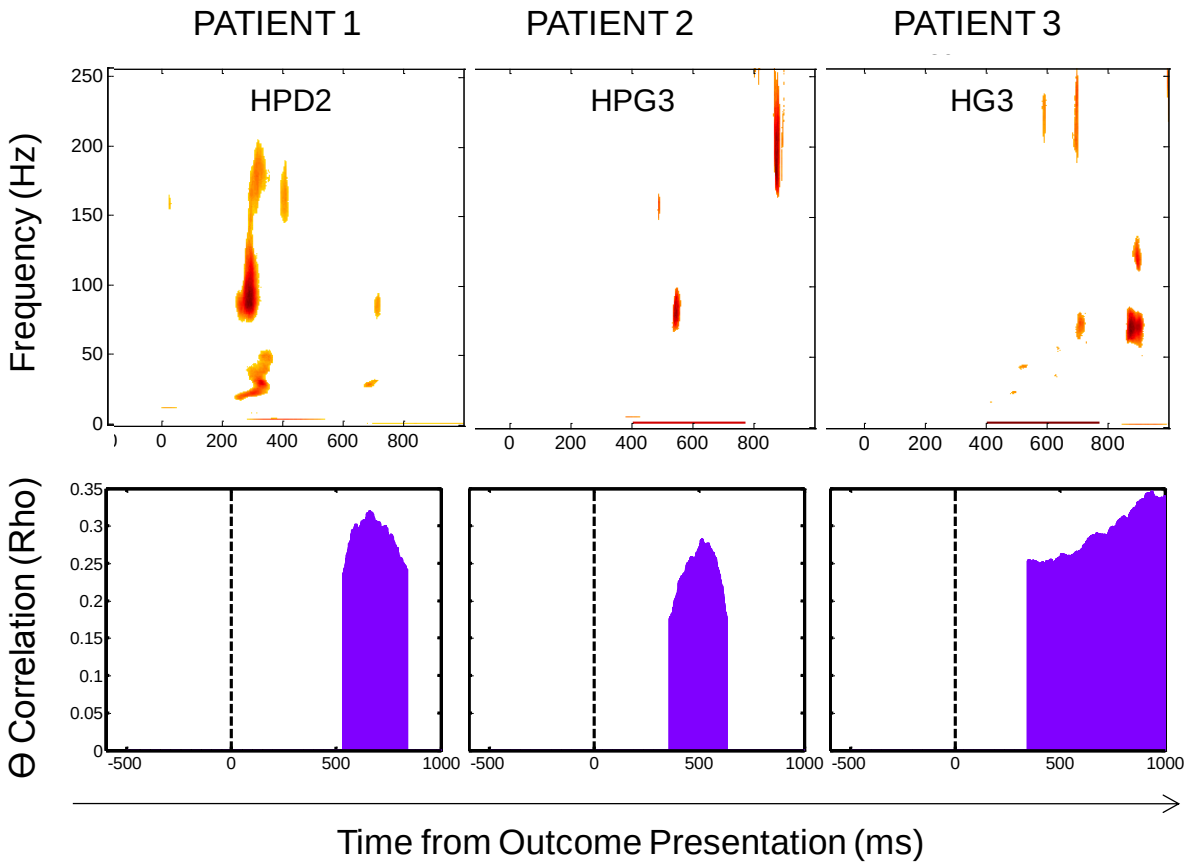


***Fig8**. Theta/Gamma oscillations during outcome evaluation lead to behavioural adjustments in expected returns in upcoming trials. The upper panel displays the time intervals and frequencies at which the power of the oscillations during the outcomes presentation significantly correlates (P<0.01) with changes in expected return (effects of trust are removed by partial correlation analysis). The lower panels display the precise timing of significant correlations for the theta oscillations (4-8 Hz) (after adjustment by the number of tested frequencies using Bonferroni criteria).*

Discussion

Disappointment is, according to the formal definition (Bell, 1985), the difference between the expectation about the outcome of a decision and the actual outcome of this decision. Therefore, disappointment is the emotion that best reflects a prediction error signal. Indeed, according to the reinforcement learning theory, the discrepancy between an expected reward and the actual reward generates an error signal which helps to adjust behaviour in future decisions or actions (Schultz, 1998). This learning process (memory formation and consolidation) involves many different areas of the brain, including the rhinal cortex (perirhinal and entorhinal) that surrounds the hippocampus. For instance, ablation of peri- and entorhinal cortices prevents the formation of associations between visual cues and their subsequent rewards (Liu et al., 2004). Indeed, the rhinal cortex is rich in dopamine innervations coming from the ventral tegmental area (Insausti, Amaral, & Cowan, 1987). Moreover, single neuron recordings in macaques have shown that entorhinal neurons respond not only to reward-cues (Mogami & Tanaka, 2006) but also directly to rewards (Sugase-Miyamoto & Richmond, 2007). Recently, these reward-related responses in the hippocampus have been linked to associative learning, showing that a subpopulation of hippocampal neurons increases its firing rate in error trials whereas another subpopulation increases its activity following correct trials (Wirth et al., 2009). This coding mechanism was observed only in a learning context and only for the “correct trials” neurons, suggesting that these neurons do not convey general information about the successful trial completion but are specifically reactive to learning (unlike the “error” cells which also responded in a simple reward task without learning). They also showed that the “correct trials” neurons changed their stimulus-selective properties in the trials where behavioural learning occurred. Finally, a last population of “changing cells” was identified (whose selectivity changes over time), which signals a continuum in the learning-related hippocampal neurons.

The strong hippocampal responses found in our study were not present during the visual cues announcing the upcoming reward (the countdown showing expectations), but were clearly dependent upon the reward presentation itself. We thus show that the rhinal cortex, more than the amygdala and the orbitofrontal cortex (OFC, see Gottfried, O'Doherty, & Dolan, 2003), conveys information about the mismatch between expectations and reality, at least in the context of the Trust Game. Indeed, responses recorded in the rhinal cortex were systematically stronger than those recorded in the amygdala and in the prefrontal areas. Moreover, the latencies of these outcome-related responses

recorded in the rhinal cortex are not in agreement with the idea that it receives information from the amygdala and the OFC.

In a first time-frame, around 200 ms, these responses were present independent of the type of outcomes, and in a second time frame (between 400 and 600 ms) they were emotion-specific (specific to unexpected outcomes). These late responses were dependent on the discrepancy between the expectations and the actual outcome rather than on the valence (positive/negative) of the outcome.

To better understand the coding mechanisms and neural role of rhinal signals in the Trust Game we analyzed the phase alignment of oscillations across trials and the partial correlation between frequency amplitude in trial i-1 and change in expectations between trial i-1 and trial i. We found significant phase locking across trials in the theta range (4-8 Hz) in the hippocampal regions, starting very early (around 200 ms) after outcome presentation. This coincides with the early responses found in the ERPs of the similar electrodes that were not specific to a particular type of outcome. The link between theta oscillations in the rhinal cortex, the encoding of error prediction signals (disappointment and elation) and the behavioural changes in trials following large error prediction signals further supports a role of theta oscillations in the rhinal cortex in memory encoding. The rhinal cortex seems to be important to retain outcome-related information during large time intervals and to update behaviour accordingly (between the presentation of the outcome in trial i-1 and the expectations in trial i). The observation of significant outcome-related theta phase locking in the rhinal area is novel but not unexpected. Indeed, theta oscillations have been recurrently related to working memory or to memory encoding and retrieval (Axmacher, Mormann, Fernandez, Elger, & Fell, 2006; Hasselmo, Bodelon, & Wyble, 2002; O. Jensen & Lisman, 1998). Within the rhinal area, theta rhythm seems to modulate gamma activity (O'Keefe & Recce, 1993; Siapas, Lubenov, & Wilson, 2005; Skaggs, McNaughton, Wilson, & Barnes, 1996) and intervene in learning (Buzsáki, 2006). Moreover, a temporal coding mechanism - through the combination of the theta phase and the timing of place cells' spikes in the hippocampus - has been shown to provide to rats their own location in space (Huxter, Burgess, & O'Keefe, 2003; Siapas et al., 2005; Skaggs et al., 1996).

Another phase locking was observed mainly on fronto-orbital electrodes (but also on some amygdala/ hippocampus contacts), in the alpha band range around 200 ms after outcome presentation. Importantly, one previous recording obtained with electrodes implanted in the human brain reported a strong correlation between an alpha band ERP's amplitude and the error prediction signal (Oya et al., 2005). The authors used a classical Iowa Gambling Task and showed that the strongest correlation

between the amplitude of the ERP (on isolated alpha band) and the prediction error signal was found when the patient chose a bad deck of cards but did not receive the expected "punishment", i.e. a large loss. Our result (alpha oscillations responding to outcomes in the prefrontal cortex) are consistent with their observation, although sample size did not allow us to test the specificity of this phase alignment to a particular category of outcome (expected or not, positive or negative).

To conclude, our results show that:

1) There is an intervention of the rhinal cortex not only in the association of a cue to an upcoming reward, but directly in the processing of the outcome. Although this effect has been shown in non-human primates (Sugase-Miyamoto & Richmond, 2007) it is to our knowledge, its first direct demonstration in the human brain. Importantly, the patients received no monetary compensation for their participation in the task. This implies that the hippocampal responses were not linked to financial reward but actually constituted a learning signal derived from the mismatch between expectations and outcomes.

2) Indeed, theta rhythm seemed to play a role in the evaluation of the outcome. This role, previously suggested from EEG recordings in humans (Tzur & Berger, 2009), is here extended to the retention of the outcome information during the delay between the feedback about the subject's decisions and the formation of new beliefs in the following trial. We have thus linked the processing of the outcome to a learning process, as it modifies the behaviour in the following trial. This implication of hippocampal activity in learning place-objects associations has been previously shown (Wirth et al., 2009) but we extend it here to a context where there are no spatial elements.

3) Finally, it is important to clarify that this encoded information is more than the association between the faces and outcome for the following reasons: 1) the effect of the trust variable was excluded from the correlation analysis and 2) most of the time, the face presented in the following trial did not belong to the same category (gender, ethnicity, age, etc) as the face in the current trial. We cannot exclude that a more general learning process occurred to encode those kinds of associations throughout the whole game, but they have in all cases a lesser effect upon behaviour than the outcome of the previous trial.

GENERAL DISCUSSION, CONCLUSIONS AND PERSPECTIVES

Experience is not what happens to a man:
it is what a man does with what happens to him[*]

From our first study we learned that disappointment strongly impacts the decision-making process, leading subjects to lower their expectations about upcoming social interactions. We could have found further support for this interpretation from the analysis of the correlations. Indeed, if consequently to a disappointing outcome the correlations between trustworthiness and investment were lower than usual, it would have been a supplementary evidence for the impact of previous disappointment in decision-making. However, all the behavioural data gathered in this study seem to confirm the following adage: "*Blessed is he who expects nothing, for he shall never be disappointed".* Most interestingly, we found large inter-individual differences in the tolerance to disappointment, with some subjects letting an automatic approach/avoidance system guide their decisions and others being able to put cognitive brakes on their impulses and better adapt to the situation. Actually, we interpreted our results within the frame of dual system theories (Evans, 2008), postulating that two modes co-exist to guide decision-making. One mode allows making fast decisions through automatic or routine mechanisms (System 1) and the other exerts control over these mechanisms when they might lead to maladaptive decisions (System 2). One important step was to confirm the link between the presence of a fronto-central map that we supposed to reflect the intervention of System 2 and the more rational behaviour of a subgroup of subjects. Indeed, although this different map necessarily reflected a change in the underlying generators, we had no guarantee that this change was related to the intervention of a control system rather than the activation of new brain areas still related to automatic decisions (System 1).

We found confirmation of our interpretation in the second study, where the link between the electrophysiological indicator of the intervention of System 2 (the fronto-central map) and a more rational behaviour (higher Disappointment Tolerance Thresholds, and significantly higher gains for almost all subjects) was again established. Our data also suggested that a manipulation such as changing the frame (here, the instructions given to the subjects) can change the neural underpinnings of decision-making *in the same subjects*. This is crucial since the study showing the most similar results to ours (De Martino et al., 2006) demonstrated that first, changing the frame can change the responses of the subjects, and second, that the subjects who are less sensitive to the frame effect show greater

[*] Aldous Huxley, *Texts and Pretexts,* 1932

activation of the orbitofrontal cortex (OFC, supposedly reflecting a greater intervention of System 2). Here, we take a step further by showing that a change in task instructions can elicit the use of System 2 in subjects who *a priori* tended to rely on System 1.

In Study 3 we wanted to further investigate if the emergence of System 2 can be favoured by manipulating the instructions (and therefore the information) given to the subjects. Although not expected, we again found large inter-individual differences in the reaction to the frame effect. The first block replicated the classical results of the Ultimatum Game (UG), whereas in the second block, being aware of the irrationality of rejecting unfairness led a subgroup of subjects to accept all the offers, irrespective of their fairness degree. These subjects, showing "rational" behaviour according to game theory, did not show greater signs of control (greater activation of prefrontal areas and/or lengthened reaction times) and were much faster than in the first block. We suggested that the origin of this promptness and of the lack of control indicators had to be found in the fact that they took the decision to accept everything right after reading the instructions, and therefore before the beginning of the second block. The decision-making process being done prior to the block, the observed electrophysiological map in the second block stemmed from a combination of automatic and controlled processes, as in the first block. However, there would have been another way to test the hypotheses of dual system in the Ultimatum Game:

Indeed, in this study, we did not compare "basic" inter-individual differences in the UG, because first it was not the purpose of this experiment and second, we were expecting, according to previous literature (Sanfey et al., 2003), to see greater signs of control *for all the subjects together* in the second block compared to the first one (expecting that all the subjects would accept unfair offers in the second block). But as we found inter-individual differences in the second block, we also investigated this aspect. Consequently, the division of subjects in the first block was done *a posteriori,* according to their behaviour in the second block. By doing so, it is possible that we grouped subjects together (in the first block) who in fact played the game very differently. To investigate basic inter-individual differences in the use of dual system in the UG, we could have classified the subjects in the first block (before the manipulation) according to their rate of acceptance of unfair offers. It is maybe at this point that we might have found evidence of a more controlled behaviour in some subjects (i.e. those who accepted the greatest number of unfair offers), and its electrophysiological correlates. If this analysis were to reveal inter-individual differences it might establish a strong link between this study, our studies with the Trust Game, but also the study on the frame effect previously mentioned (De Martino et al., 2006). Indeed, de Martino and his colleagues concluded that "*Our results raise an intriguing possibility that*

more "rational" individuals have a better and more refined representation of their own emotional biases that enables them to modify their behaviour in appropriate circumstances, as for example when such biases might lead to suboptimal decisions", and this hypothesis is strongly corroborated by our first and second studies.

Yet, we found some striking similarities between Study 1 and Study 3. In both experiments, inter-individual differences - revealed both by behaviour and electrophysiology - unexpectedly appeared again as a main factor influencing decision-making. Probably more remarkable is the similarity in the timing at which those differences appear, between 300 and 400 ms after the presentation of the offer in the Ultimatum Game or the outcome in the Trust Game. The identification of such a precise time frame is possible only with neuroimaging tools of high temporal resolution. This result might guide future research towards finding inter-individual differences in other aspects of personality occurring at the same time interval. Indeed, we conclude that there is a specific time frame during which differences in decision-making appear, but we can only speculate on their origin. A likely reason is that members of each group perceived differently unfairness/betrayal, and consequently reacted differently. It might also happen that all individuals equally assessed unfairness but the cognitive control encouraged by the use of System 2 intervened more often in some individuals than in others (indeed, Study 3 showed that the network of structures correlating with fairness was active before the time frame in which we found inter-individual differences). As our studies were not designed to study inter-individual differences, our understanding of the data is limited by the absence of psychological tests on personality traits (such as IQ, impulsivity, need for cognition, etc), which have traditionally been linked to the facility of use of System 2.

Finally, the last study offered a great insight into the neural mechanisms involved in the Trust Game, and particularly why expectations were universally decreased after disappointment (the DTT effect). Our results showed that even if involuntary, subjects encode and learn from negative reward prediction errors. After disappointing outcomes, theta oscillations in the hippocampal/parahippocampal regions seemed to act as a coding mechanism that retained information across trials and influenced upcoming behaviour. Importantly, our version of the TG is very annoying for the subjects because they cannot reduce the uncertainty about the behaviour of the Trustees. Indeed, as the Trustee changes in every trial, and as (unknown to the participants) there is no systematic link between any feature of the face and the outcome, we have voluntarily put the subjects in a position where all their efforts to reduce uncertainty were in vain. Psychologically, this discomfort was reduced by the creation of imaginary rules reported by many subjects, such as "Old people never betray", "Asians are treacherous". One recent

study reported that following an Ultimatum Game, subjects showed enhanced memory for partners who violated their expectations during the game, but no effect linked to the fairness of the offer was found (Chang & Sanfey, 2009). Our results are consistent with the idea that the element encoded when discovering the outcome is the mismatch between the prediction of the subject and the reality. This memory trace is reactivated in the subsequent trial in order to (wrongly) redirect behaviour towards trusting, investing and in particular, expecting less from the new Trustee. However, there are some evidences to think that theta rhythms encoded the outcome itself rather than the association between a category of face and the outcome. First, the disappointment experienced in the previous trial had a much greater effect on the following trial's expectations than any other variable (trust was excluded by the partial correlation procedure). Second, we did not find a systematic bias against a particular category of Trustees.

An interesting perspective would be to study the evolution of striatal signals in our version of the Trust Game, for instance on the basis of inverse solution as the striatum is rarely the target of intracranial recordings in patients. Previous literature reported changes in the striatal signals that reflect first a "social" prediction error and then shift in time to anticipate the other player's actions (Delgado et al., 2005, King-Casas et al., 2005). As nothing can help to improve trust estimation in our study (outcomes are randomly assigned to the different faces), these signals should remain constant throughout the game and not diminish over time.

Importantly, the negative mismatch (disappointment) had the same effect on all subjects as they all reduced their trust, their investment and most of all their expectations in the following trial. We hypothesise that trust and investment were automatically reduced after disappointing outcomes in order to be coherent with the upcoming decrease in expectations. Indeed, the effect is found in expectations more than in the other variables because depending on the face, subjects cannot completely reduce their trust ratings (for instance when the following face is an old lady, it is harder to be suspicious about her trustworthiness). However we need to emphasize that elation, the "positive" mismatch, did not induce such a systematic behavioural adaptation in our sample. It seems then that there is a real distinction between positive and negative prediction errors or at least between their respective effects upon behaviour. Indeed, the authors who reported the superiority of the effect of expectation violation in face recognition also reported that distinct neural networks were involved in the processing of positive and negative expectation violation (Chang & Sanfey, 2009). Our intracranial results do not completely support this idea but instead show that responses within the same area might differ in ERP polarity and thus in the underlying phase of slow oscillations.

Interestingly we did not find consistent results in the amygdala, although this structure is often considered as essential both in emotions and in reward processing. Moreover, according to previous results (Baumgartner et al., 2008), we could have expected a lower activation of amygdala in the subjects who did not modify their behaviour as a function of previous disappointment. Indeed this study demonstrated that normal subjects decreased trust after betrayal, and that this difference in trust modulation as a function of previous betrayal was associated with lower activations in the amygdala. The absence of consistent effects in the amygdala might be due to many reasons, including the following: 1) the contacts were not systematically correctly located in the amygdala; moreover, the amygdala was probably part of the epileptogenic region for two of the three patients 2) the amygdala is a complex structure, encompassing many roles, e.g. in the attentional domain (M. Gallagher & Holland, 1994), in reinforcement learning (Paton, Belova, Morrison, & Salzman, 2006), in impressions formation (Schiller, Freeman, Mitchell, Uleman, & Phelps, 2009) and cannot be reduced to its "emotional" reactivity (Phelps, 2006). For instance, although the amygdala has been often associated with fear detection, one recent single case study reported that a patient with complete amygdala removal was not affected in his fear detection abilities (Tsuchiya, Moradi, Felsen, Yamazaki, & Adolphs, 2009), suggesting that the amygdala is more important for its modularity role than for processing emotions. However, in our study, some significant effects specific to emotions (or to disappointment alone) compared to neutral outcomes were nonetheless found in the amygdala (as in the OFC), but they were not systematic and difficult to compare from one patient to the other.

Finally, we would like to emphasize the question of the incentives and of the supposedly lack of resemblance between laboratories experiments and real life situations (Falk & Heckman, 2009). In our fourth study, patients knew from the beginning that there was no financial retribution for their participation, whereas subjects in Study 1 expected to be paid as a function of their performance. Despite this difference, both patients and control subjects reacted identically to losses, be they financial or symbolic. This result is consistent with some studies reporting classical results in the UG although no real money is at stake (Murnighan & Saxon, 1998), but also with the studies showing that increasing the amounts at stake doesn't change the rejection behaviour either (Camerer & Thaler, 1995). Taken together, our data suggest that what is overall relevant to human behaviour is the negative emotion that accompanies the mismatch between expectations and reality rather than the financial incentives by themselves. Another determinant factor seems to be the availability of the information about the consequences of decisions. Indeed, in Studies 2 and 3, subjects completely changed their way of playing the games after listening to the experimenter's instructions, suggesting that the awareness of the counter-productiveness of certain behaviours or the fear of seeming irrational in front of the

experimenter modulate the behaviour strongly than the presence or the absence of financial incentives. Concerning the issue of the lack of realism in laboratory settings, our results indicate that subjects were actually playing the games quite dutifully, reporting strong negative emotions and showing irrational behaviour: *"After all, experiments do not need to physically resemble real life situations, but only to create psychologically meaningful situations"* (Todorov et al., 2006). In the Ultimatum Game for instance, it is always hard to know if Player 1 (the Proposer) is acting like a sophisticated profit maximizer who only fears rejection of his offer by Player 2, or if he proposes fair offers for altruistic reasons. The Dictator Game (identical to UG but where Player 2 has no choice but to accept the offer) indicates that both arguments are true: proposals are smaller than in the UG, but still positive (Camerer & Thaler, 1995). However the most intriguing behaviour in the UG has always been the rejection of unfair offers by Player 2. One subject of our third study summarized perfectly the dilemma occurring in the decision-maker's mind when an unfair offer is presented:

"Yes, I refused some offers because I thought that the proposal was not fair. I was bothered to have to refuse an offer, but I don't really agree to be taken advantage of. Before starting the game, I thought that I should accept all offers so as to build a maximum capital; but then I realized that I did not lose that much by refusing some offers, as they were not high, whereas the one who proposed the share would lose a lot more. ***I thus experienced some satisfaction to make him pay the price of his cupidity****."*

This quotation is consistent with anterior research reporting activations of reward areas and de-activation of empathic networks when punishing free-riders (de Quervain et al., 2004; Singer, 2006). It also perfectly clarifies why subjects classically reject unfair offers coming from humans but not from computers (Sally, 1995). "*People are punishing unfairness, not rejecting inequality"* (Camerer & Thaler, 1995). Note that the debate on human "real nature" beyond the rejection of unfairness is somehow irrelevant. The aforementioned confrontation of diverse hypotheses seems to imply that emotions can be contrasted to a learned behaviour (through implementation of respect for social norms). However, the fact of feeling a negative emotion, for instance when discovering an unfair proposal, cannot exclude *per se* the fact that this emotion was originally absent but built up through education. We can learn to be *truly* upset by a situation. One interesting way to address this question is to look at the data from toddlers and non-human primates' studies. We have seen that studies on non-human primates yield diverging results (Brosnan & De Waal, 2003; K. Jensen et al., 2007). One study with children ranging from age 5 to 12 reported that kindergartners accepted smaller offers than older ones (Murnighan & Saxon, 1998). This result was found when playing with real money but also with M&M's, and was replicated with teenagers (6th graders stated that they were willing to accept lower offer than 9th graders

and college students) playing with money. Thus, rejection of unfairness – even if accompanied with negative emotions – might be rather a learned behaviour than an innate impulse. However more research is needed, for instance longitudinally, to validate this hypothesis.

To conclude, throughout this research work, inter-individual differences appeared as a main factor in decision-making, with differences across subjects found at the behavioural as well as at the neural level. Although they are widely studied in some areas of research (indeed, differential psychology is a branch of psychology specialized in studying inter-individual differences), they have barely been studied in the neuroscience of decision-making. Closely related topics such as inter-individual differences in reward-processing have been more often investigated, but generally on clinical populations, to show for instance abnormal reward processing in mania (Abler, Greenhouse, Ongur, Walter, & Heckers, 2008), schizophrenia (Juckel et al., 2006) or substance dependence (Bjork, Smith, & Hommer, 2008). However personality traits in the healthy population might as well impact these processes. A few studies have addressed inter-individual differences in neural reward processing and showed influences of impulsivity (Martin & Potts, 2004), extraversion (Cohen, Young, Baek, Kessler, & Ranganath, 2005), risk aversion (Tobler, O'Doherty, Dolan, & Schultz, 2007) and academic motivation (Mizuno et al., 2008).

One recent paper addressed the question of how inter-individual differences in approach/avoidance personality traits (in particular reward sensitivity) can modulate neural reward processing (Simon et al., 2009). The effect of inter-individual differences in reward sensitivity upon decision-making had previously been shown in the Iowa Gambling Task, at the behavioural level (Franken & Muris, 2005). In this fMRI study (Simon et al., 2009), the authors reported that the higher the subjects were on a reward-seeking behavioural measure, the more the ventral striatum and the orbitofrontal cortex were active during the receipt of a reward. The subjects scoring high on this scale also showed lesser sensitivity to negative outcomes, suggesting that inter-individual differences in approach/avoidance behaviour can be measured at the neural level of appetitive functioning (the response to incentives).

Importantly, the main conclusion of this area of research is that inter-individual differences should be taken into account when studying the neural processing of reward, which is indeed a main conclusion in our work addressing decision-making. While we did not aim to study inter-individual differences, they appeared as inescapable in our understanding of the decision-making process. A remaining question is the sources of the observed differences. We attributed them to the use of cognitive control over spontaneous behaviour, but several issues remain unclear and should be investigated in future research. For instance, what *basically* triggered the use of System 2 in High-DTT subjects of Study1? Is it a lesser sensitivity to the outcome of one's decision (inter-individual differences in reward processing)

that allows some individuals to act more "rationally" than others? Or did both groups process the outcomes in the same way, but differences emerged in the "lesson" they learned from their mistakes? Besides, what is the source of the inter-individual differences in the behavioural reactions to elation? Is it the opposition between two general tendencies, one being to "not try your luck twice" and the other to tempt more and more as long as you win? Is the reaction to elation the reflection of a risk-averse versus risk-taking personality? Obviously the asymmetry between negative (all the subjects reacting equally to disappointment) and positive error prediction needs to be further investigated. In Study 3, why did some subjects follow the instructions and some subjects do the opposite? How is it that a subgroup of subjects rejected more unfair offers in the second block than in the first one? Were they "resisting" to the instructions? Were they upset by the suggestion of their irrationality? Reward sensitivity as a source of inter-individual neural and behavioural differences is a promising and a crucial research perspective as real life applications can be extended, for instance, to addictive behaviour and, in the context of neuroeconomics, to pathological gambling. As Dostoevsky* brilliantly described in the following scene, something is still beyond our understanding of decision-making when it comes to human gambling behaviour:

"[…] *a very young man who was plunging heavily* […] *His eyes kept flashing and his hands shaking; yet all the while he staked without any sort of calculations – just what came to his hand, as he kept winning and winning, and raking and raking his gains.* […] *For a few minutes the Grandmother watched him. "Go and tell him" suddenly she exclaimed with a nudge at my elbow, " – go and tell him to stop, and to take his money with him, and go home".* […] *"On the left, among the players at the other half of the table, a young lady was playing* […]. *Taking some gold and a few thousand-franc notes out of her pocket – would begin quietly, coldly, and after much calculation, to stake, and mark down the figures in pencil on a paper, as though striving to work out a system according to which, at given moments, the odds might group themselves. Always she staked large coins, and either lost or won one, two, or three thousand francs a day, but no more; after which she would depart. The Grandmother took a long look at her. "That woman is not losing," she said.* "

And the Grandmother started to gamble frenetically.

* Fyodor Dostoevski, *«The Gambler»,* 1867

REFERENCES

Abler, B., Greenhouse, I., Ongur, D., Walter, H., & Heckers, S. (2008). Abnormal reward system activation in mania. *Neuropsychopharmacology, 33*(9), 2217-2227.

Alesina, A., & La Ferrara, E. (2002). Who trusts others? *Journal of Public Economics, 85*(2), 207-234.

Andreoni, J., Castillo, M., & Petrie, R. (2003). What Do Bargainer's Preferences Look Like? Experiments with a Convex Ultimatum Game. *American Economic Review, 93*(3), 672-685.

Axmacher, N., Mormann, F., Fernandez, G., Elger, C. E., & Fell, J. (2006). Memory formation by neuronal synchronization. *Brain Research Reviews, 52*(1), 170-182.

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biol Lett, 2*(3), 412-414.

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., & Fehr, E. (2008). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron, 58*(4), 639-650.

Bechara, A., & Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior, 52*(2), 336-372.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition, 50*(1-3), 7-15.

Bechara, A., Damasio, H., Damasio, A. R., & Lee, G. P. (1999). Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. *J Neurosci, 19*(13), 5473-5481.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nat Neurosci, 10*(9), 1214-1221.

Bell, D. E. (1985). Disappointment in Decision Making under Uncertainty. *Operations Research, 33*(1), 1-27.

Benhabib, J., & Bisin, A. (2005). Modeling internal commitment mechanisms and self-control: A neuroeconomics approach to consumption-saving decisions. *Games and Economic Behavior, 52*(2), 460-492.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior, 10*(1), 122-142.

Bernheim, B., & Rangel, A. (2004). Addiction and Cue-Triggered Decision Processes. *The American Economic Review, 94*(5), 1558-1590.

Bjork, J. M., Smith, A. R., & Hommer, D. W. (2008). Striatal sensitivity to reward deliveries and omissions in substance dependent patients. *Neuroimage, 42*(4), 1609-1621.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict Monitoring and Cognitive Control. *Psychological Review, 108*(3), 624-652.

Bowles, S., & Gintis, H. (2002). Homo reciprocans. *Nature, 415*(6868), 125-128.

Bowles, S., & Gintis, H. (2004). The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology, 65*(1), 17-28.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proc Natl Acad Sci U S A, 100*(6), 3531-3535.

Brocas, I., & Carrillo, J. D. (2008). The Brain as a Hierarchical Organization. *The American Economic Review, 98*(4), 1312-1346.

Brosnan, S. F., & De Waal, F. B. (2003). Monkeys reject unequal pay. *Nature, 425*(6955), 297-299.

Buckley, M. J., Mansouri, F. A., Hoda, H., Mahboubi, M., Browning, P. G., Kwok, S. C., et al. (2009). Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. *Science, 325*(5936), 52-58.

Buzsáki, G. (2006). *Rhythms of the Brain* (1 ed.): Oxford University Press, USA.

Camerer, C., & Hogarth, R. M. (1999). The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. *Journal of Risk and Uncertainty, 19*, 7-42.

Camerer, C., Loewenstein, G., & Prelec, D. (2004). Neuroeconomics: Why economics needs brain. *The Scandinavian Journal of Economics, 106*(3), 555-579.

Camerer, C., & Thaler, R. H. (1995). Anomalies: Ultimatums, Dictators and Manners. *The Journal of Economic Perspectives, 9*(2), 209-219.

Cameron, L. A. (1999). Raising the Stakes in the Ultimatum Game: Experimental Evidence From Indonesia. *Economic Inquiry, 37*(1), 47-59.

Camille, N., Coricelli, G., Sallet, J., Pradat-Diehl, P., Duhamel, J. R., & Sirigu, A. (2004). The involvement of the orbitofrontal cortex in the experience of regret. *Science, 304*(5674), 1167-1170.

Canessa, N., Gorini, A., Cappa, S. F., Piattelli-Palmarini, M., Danna, M., Fazio, F., et al. (2005). The effect of social content on deductive reasoning: an fMRI study. *Hum Brain Mapp, 26*(1), 30-43.

Cesarini, D., Dawes, C. T., Fowler, J. H., Johannesson, M., Lichtenstein, P., & Wallace, B. (2008). Heritability of cooperative behavior in the trust game. *Proc Natl Acad Sci U S A, 105*(10), 3721-3726.

Chaiken, S. (1980). Heuristic Versus Systematic Information Processing and the Use of Source Versus Message Cues in Persuasion. *Journal of Personality and Social Psychology, 39*(5), 752-766.

Chaiken, S., & Trope, Y. (1999). *Dual-Process Theories in Social Psychology* (1 avril 1999 ed.). New York: Guilford Publications.

Chang, L. J., & Sanfey, A. G. (2009). Unforgettable ultimatums? Expectation violations promote enhanced social memory following economic bargaining. *Frontiers in Behavioral Neuroscience, 3*(36).

Christie, G. J., & Tata, M. S. (2009). Right frontal cortex generates reward-related theta-band oscillatory activity. *Neuroimage, 48*(2), 415-422.

Chua, H. F., Gonzalez, R., Taylor, S. F., Welsh, R. C., & Liberzon, I. (2009). Decision-related loss: regret and disappointment. *Neuroimage, 47*(4), 2031-2040.

Civai, C., Corradi-Dell'acqua, C., Gamer, M., & Rumiati, R. I. (2009). Are irrational reactions to unfairness truly emotionally-driven? Dissociated behavioural and emotional responses in the Ultimatum Game task. *Cognition*.

Cohen, M. X., Elger, C. E., & Ranganath, C. (2007). Reward expectation modulates feedback-related negativity and EEG spectra. *Neuroimage, 35*(2), 968-978.

Cohen, M. X., & Ranganath, C. (2007). Reinforcement learning signals predict future decisions. *J Neurosci, 27*(2), 371-378.

Cohen, M. X., Young, J., Baek, J. M., Kessler, C., & Ranganath, C. (2005). Individual differences in extraversion and dopamine genetics predict neural reward responses. *Brain Res Cogn Brain Res, 25*(3), 851-861.

Coricelli, G., Critchley, H. D., Joffily, M., O'Doherty, J. P., Sirigu, A., & Dolan, R. J. (2005). Regret and its avoidance: a neuroimaging study of choice behavior. *Nat Neurosci, 8*(9), 1255-1262.

Coricelli, G., Dolan, R. J., & Sirigu, A. (2007). Brain, emotion and decision making: the paradigmatic example of regret. *Trends Cogn Sci, 11*(6), 258-265.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58*(1), 1-73.

De Martino, B., Kumaran, D., Seymour, B., & Dolan, R. J. (2006). Frames, biases, and rational decision-making in the human brain. *Science, 313*(5787), 684-687.

de Quervain, D. J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., et al. (2004). The neural basis of altruistic punishment. *Science, 305*(5688), 1254-1258.

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nat Neurosci, 8*(11), 1611-1618.

Delgado, M. R., Gillis, M. M., & Phelps, E. A. (2008). Regulating the expectation of reward via cognitive strategies. *Nat Neurosci, 11*(8), 880-881.

Dijksterhuis, A., Bos, M. W., Nordgren, L. F., & van Baaren, R. B. (2006). On making the right choice: the deliberation-without-attention effect. *Science, 311*(5763), 1005-1007.

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature, 452*(7185), 348-351.

Evans, J. S. (2003). In two minds: dual-process accounts of reasoning. *Trends Cogn Sci, 7*(10), 454-459.

Evans, J. S. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu Rev Psychol, 59*, 255-278.

Evenden, J. L. (1999). Varieties of impulsivity. *Psychopharmacology (Berl), 146*(4), 348-361.

Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science, 326*(5952), 535-538.

Falkenstein, M., Hoormann, J., Christ, S., & Hohnsbein, J. (2000). ERP components on reaction errors and their functional significance: a tutorial. *Biol Psychol, 51*(2-3), 87-107.

Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends Cogn Sci, 11*(10), 419-427.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature, 425*(6960), 785-791.

Fehr, E., & Gachter, S. (2002). Altruistic punishment in humans. *Nature, 415*(6868), 137-140.

Finucane, M. L., Alhakami, A., Slovic, P., & Johnson, S. M. (2000). The Affect Heuristic in Judgments of Risks and Benefits. *Journal of Behavioral Decision Making, 13*(1), 1-17.

Fiorillo, C. D., Newsome, W. T., & Schultz, W. (2008). The temporal precision of reward prediction in dopamine neurons. *Nat Neurosci*.

Franken, I. H. A., & Muris, P. (2005). Individual differences in decision-making. *Personality and Individual Differences, 39*(5), 991-998.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives, 19*(4), 25-42.

Frith, C. D., & Singer, T. (2008). The role of social cognition in decision making. *Philos Trans R Soc Lond B Biol Sci, 363*(1511), 3875-3886.

Fudenberg, D., & Levine, D. K. (2006). A Dual-Self Model of Impulse Control. *The American Economic Review, 96*(5), 1449-1476.

Fujiwara, J., Tobler, P. N., Taira, M., Iijima, T., & Tsutsui, K. (2008). Personality-dependent dissociation of absolute and relative loss processing in orbitofrontal cortex. *Eur J Neurosci, 27*(6), 1547-1552.

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends Cogn Sci, 7*(2), 77-83.

Gallagher, M., & Holland, P. C. (1994). The amygdala complex: multiple roles in associative learning and attention. *Proc Natl Acad Sci U S A, 91*(25), 11771-11776.

Gintis, H. (2000). Strong Reciprocity and Human Sociality. *Journal of Theoretical Biology, 206*(2), 169-179.

Glimcher, P. W., Kable, J., & Kenway, L. (2007). Decision theory: New methods, new insights. Neuroeconomic studies of impulsivity: Now or just as soon as possible? *The American economic review, 97*(2), 142-147.

Goel, V., & Dolan, R. J. (2003). Explaining modulation of reasoning by belief. *Cognition, 87*(1), B11-22.

Gottfried, J. A., O'Doherty, J., & Dolan, R. J. (2003). Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science, 301*(5636), 1104-1107.

Grabenhorst, F., & Rolls, E. T. (2009). Different representations of relative and absolute subjective value in the human brain. *Neuroimage, 48*(1), 258-268.

Grave de Peralta Menendez, R., Gonzalez Andino, S. L., Morand, S., Michel, C. M., & Landis, T. (2000). Imaging the electrical activity of the brain: ELECTRA. *Hum Brain Mapp, 9*(1), 1-12.

Grave de Peralta Menendez, R., Murray, M. M., Michel, C. M., Martuzzi, R., & Gonzalez Andino, S. L. (2004). Electrical neuroimaging based on biophysical constraints. *Neuroimage, 21*(2), 527-539.

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends Cogn Sci, 11*(8), 322-323; author reply 323-324.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105-2108.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *J Pers Soc Psychol, 74*(6), 1464-1480.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology, 73*(3), 407-420.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization, 3*, 367-388.

Haenschel, C., Baldeweg, T., Croft, R. J., Whittington, M., & Gruzelier, J. (2000). Gamma and beta frequency oscillations in response to novel auditory stimuli: A comparison of human electroencephalogram (EEG) data with in vitro models. *Proc Natl Acad Sci U S A, 97*(13), 7645-7650.

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biol Psychol, 71*(2), 148-154.

Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science, 324*(5927), 646-648.

Haruno, M., Kuroda, T., Doya, K., Toyama, K., Kimura, M., Samejima, K., et al. (2004). A neural correlate of reward-based behavioral learning in caudate nucleus: a functional magnetic resonance imaging study of a stochastic decision task. *J Neurosci, 24*(7), 1660-1665.

Hasselmo, M. E., Bodelon, C., & Wyble, B. P. (2002). A proposed function for hippocampal theta rhythm: separate phases of encoding and retrieval enhance reversal of prior learning. *Neural Comput, 14*(4), 793-817.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., et al. (2005). "Economic man" in cross-cultural perspective: behavioral experiments in 15 small-scale societies. *Behav Brain Sci, 28*(6), 795-815; discussion 815-755.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., et al. (2006). Costly punishment across human societies. *Science, 312*(5781), 1767-1770.

Herrmann, B., Thoni, C., & Gachter, S. (2008). Antisocial punishment across societies. *Science, 319*(5868), 1362-1367.

Hester, R., Barre, N., Murphy, K., Silk, T. J., & Mattingley, J. B. (2008). Human medial frontal cortex activity predicts learning from errors. *Cereb Cortex, 18*(8), 1933-1940.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109*(4), 679-709.

Holroyd, C. B., Hajcak, G., & Larsen, J. T. (2006). The good, the bad and the neutral: electrophysiological responses to feedback stimuli. *Brain Res, 1105*(1), 93-101.

Holroyd, C. B., & Krigolson, O. E. (2007). Reward prediction error signals associated with a modified time estimation task. *Psychophysiology, 44*(6), 913-917.

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., & Cohen, J. D. (2003). Errors in reward prediction are reflected in the event-related brain potential. *Neuroreport, 14*(18), 2481-2484.

Huck, S. (1999). Responder behavior in ultimatum offer games with incomplete information. *Journal of Economic Psychology, 20*(2), 183-206.

Huxter, J., Burgess, N., & O'Keefe, J. (2003). Independent rate and temporal coding in hippocampal pyramidal cells. *Nature, 425*(6960), 828-832.

Insausti, R., Amaral, D. G., & Cowan, W. M. (1987). The entorhinal cortex of the monkey: III. Subcortical afferents. *J Comp Neurol, 264*(3), 396-408.

Ito, S., Stuphorn, V., Brown, J. W., & Schall, J. D. (2003). Performance Monitoring by the Anterior Cingulate Cortex During Saccade Countermanding. *Science, 302*(5642), 120-122.

Jansen, B. H., Agarwal, G., Hegde, A., & Boutros, N. N. (2003). Phase synchronization of the ongoing EEG and auditory EP generation. *Clin Neurophysiol, 114*(1), 79-85.

Jensen, K., Call, J., & Tomasello, M. (2007). Chimpanzees are rational maximizers in an ultimatum game. *Science, 318*(5847), 107-109.

Jensen, O., & Lisman, J. E. (1998). An oscillatory short-term memory buffer model can account for data on the Sternberg task. *J Neurosci, 18*(24), 10688-10699.

Juckel, G., Schlagenhauf, F., Koslowski, M., Wustenberg, T., Villringer, A., Knutson, B., et al. (2006). Dysfunction of ventral striatal reward prediction in schizophrenia. *Neuroimage, 29*(2), 409-416.

Kagel, J. H., Kim, C., & Moser, D. (1996). Fairness in Ultimatum Games with Asymmetric Information and Asymmetric Payoffs. *Games and Economic Behavior, 13*(1), 100-110.

Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics. *The American Economic Review, 93*(5), 1449-1475.

Kahneman, D., & Frederick, S. (2007). Frames and brains: elicitation and control of response tendencies. *Trends Cogn Sci, 11*(2), 45-46.

Kahneman, D., & Miller, D. T. (1986). Norm Theory: Comparing reality to its alternatives. *Psychological Review, 93*(2), 136-153.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4).

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica, 47*(2), 263-291.

Kamarajan, C., Porjesz, B., Rangaswamy, M., Tang, Y., Chorlian, D. B., Padmanabhapillai, A., et al. (2009). Brain signatures of monetary loss and gain: outcome-related potentials in a single outcome gambling task. *Behav Brain Res, 197*(1), 62-76.

Kennerley, S. W., Walton, M. E., Behrens, T. E. J., Buckley, M. J., & Rushworth, M. F. S. (2006). Optimal decision making and the anterior cingulate cortex. *9*(7), 940-947.

Kenning, P., & Plassmann, H. (2005). Neuroeconomics: an overview from an economic perspective. *Brain Res Bull, 67*(5), 343-354.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. *Science, 308*(5718), 78-83.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science, 314*(5800), 829-832.

Knutson, B., Rick, S., Wimmer, G. E., Prelec, D., & Loewenstein, G. (2007). Neural predictors of purchases. *Neuron, 53*(1), 147-156.

Koenigs, M., & Tranel, D. (2007). Irrational economic decision-making after ventromedial prefrontal damage: evidence from the Ultimatum Game. *J Neurosci, 27*(4), 951-956.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., et al. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature, 446*(7138), 908-911.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature, 435*(7042), 673-676.

Lieberman, M. D. (2007). Social cognitive neuroscience: a review of core processes. *Annu Rev Psychol, 58*, 259-289.

Liu, Z., Richmond, B. J., Murray, E. A., Saunders, R. C., Steenrod, S., Stubblefield, B. K., et al. (2004). DNA targeting of rhinal cortex D2 receptor protein reversibly blocks learning of cues that predict reward. *Proc Natl Acad Sci U S A, 101*(33), 12336-12341.

Loewenstein, G., Rick, S., & Cohen, J. D. (2008). Neuroeconomics. *Annu Rev Psychol, 59*, 647-672.

Loomes, G., & Sugden, R. (1986). Disappointment and Dynamic Consistency in Choice under Uncertainty. *The Review of Economic Studies, 53*(2), 271-282.

Makeig, S., Westerfield, M., Jung, T. P., Enghoff, S., Townsend, J., Courchesne, E., et al. (2002). Dynamic brain sources of visual evoked responses. *Science, 295*(5555), 690-694.

Marco-Pallares, J., Cucurell, D., Cunillera, T., Garcia, R., Andres-Pueyo, A., Munte, T. F., et al. (2008). Human oscillatory activity associated to reward processing in a gambling task. *Neuropsychologia, 46*(1), 241-248.

Martin, L. E., & Potts, G. F. (2004). Reward sensitivity in impulsivity. *Neuroreport, 15*(9), 1519-1522.

Maskin, E. (2008). Economics. Can neural data improve economics? *Science, 321*(5897), 1788-1789.

Matsumoto, K., Suzuki, W., & Tanaka, K. (2003). Neuronal Correlates of Goal-Based Motor Selection in the Prefrontal Cortex

10.1126/science.1084204. *Science, 301*(5630), 229-232.

Matsumoto, M., & Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature, 459*(7248), 837-841.

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proc Natl Acad Sci U S A, 98*(20), 11832-11835.

McClure, S. M., Ericson, K. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2007). Time discounting for primary rewards. *J Neurosci, 27*(21), 5796-5804.

Mellers, B., Schwartz, A., & Ritov, I. (1999). Emotion-Based Choice. *Journal of Experimental Psychology: General, 128*(3), 332-345.

Mielke, P. W., Berry, K.J. (2001). *Permutation Methods: A Distance Function Approach*. New-York: Springer-Verlag.

Mizuno, K., Tanaka, M., Ishii, A., Tanabe, H. C., Onoe, H., Sadato, N., et al. (2008). The neural basis of academic achievement motivation. *Neuroimage, 42*(1), 369-378.

Mogami, T., & Tanaka, K. (2006). Reward association affects neuronal responses to visual stimuli in macaque te and perirhinal cortices. *J Neurosci, 26*(25), 6761-6770.

Murnighan, J. K., & Saxon, M. S. (1998). Ultimatum bargaining by children and adults. *Journal of Economic Psychology, 19*(4), 415-445.

Näätänen, R., Gaillard, A. W. K., & Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica, 42*(4), 313-329.

Niedenthal, P. M., Tangney, J. P., & Gavanski, I. (1994). "If only I weren't" versus "if only I hadn't": distinguishing shame and guilt in counterfactual thinking. *J Pers Soc Psychol, 67*(4), 585-595.

Nieuwenhuis, S., Holroyd, C. B., Mol, N., & Coles, M. G. (2004). Reinforcement-related brain potentials from medial frontal cortex: origins and functional significance. *Neurosci Biobehav Rev, 28*(4), 441-448.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The Use of Statistical Heuristics in Everyday Inductive Reasoning. *Psychological Review, 90*(4), 339-363.

Ochsner, K. N., Ray, R. D., Cooper, J. C., Robertson, E. R., Chopra, S., Gabrieli, J. D., et al. (2004). For better or for worse: neural systems supporting the cognitive down- and up-regulation of negative emotion. *Neuroimage, 23*(2), 483-499.

O'Keefe, J., & Recce, M. L. (1993). Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus, 3*(3), 317-330.

Oya, H., Adolphs, R., Kawasaki, H., Bechara, A., Damasio, A., & Howard, M. A., 3rd. (2005). Electrophysiological correlates of reward prediction error recorded in the human prefrontal cortex. *Proc Natl Acad Sci U S A, 102*(23), 8351-8356.

P.J. Phillips, H. W., J. Huang, P. Rauss. (1998). The FERET database and evaluation procedure for face recognition algorithms. *Image and Vision Computing J, 16*(5), 295-306.

Packard, M. G., & Knowlton, B. J. (2002). Learning and memory functions of the Basal Ganglia. *Annu Rev Neurosci, 25*, 563-593.

Padoa-Schioppa, C., & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature, 441*(7090), 223-226.

Padoa-Schioppa, C., & Assad, J. A. (2008). The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. *Nat Neurosci, 11*(1), 95-102.

Paton, J. J., Belova, M. A., Morrison, S. E., & Salzman, C. D. (2006). The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature, 439*(7078), 865-870.

Phelps, E. A. (2006). Emotion and cognition: insights from studies of the human amygdala. *Annu Rev Psychol, 57*, 27-53.

Phillips, P. J., Wechsler, H., Huang, J., & Rauss, P. (1998). The FERET database and evaluation procedure for face recognition algorithms. *Image and Vision Computing J, 16*(5), 295-306.

Plassmann, H., O'Doherty, J., & Rangel, A. (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *J Neurosci, 27*(37), 9984-9988.

Polezzi, D., Daum, I., Rubaltelli, E., Lotto, L., Civai, C., Sartori, G., et al. (2008). Mentalizing in economic decision-making. *Behav Brain Res, 190*(2), 218-223.

Polezzi, D., Lotto, L., Daum, I., Sartori, G., & Rumiati, R. (2008). Predicting outcomes of decisions in the brain. *Behav Brain Res, 187*(1), 116-122.

Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D., & Nowak, M. A. (2009). Positive interactions promote public cooperation. *Science, 325*(5945), 1272-1275.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nat Rev Neurosci, 9*(7), 545-556.

Ridderinkhof, K. R., van den Wildenberg, W. P., Segalowitz, S. J., & Carter, C. S. (2004). Neurocognitive mechanisms of cognitive control: the role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. *Brain Cogn, 56*(2), 129-140.

Rilling, J. K., Goldsmith, D. R., Glenn, A. L., Jairam, M. R., Elfenbein, H. A., Dagenais, J. E., et al. (2008). The neural correlates of the affective response to unreciprocated cooperation. *Neuropsychologia, 46*(5), 1256-1266.

Rilling, J. K., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., & Kilts, C. (2002). A neural basis for social cooperation. *Neuron, 35*(2), 395-405.

Rilling, J. K., King-Casas, B., & Sanfey, A. G. (2008). The neurobiology of social decision-making. *Curr Opin Neurobiol, 18*(2), 159-165.

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. *Neuroimage, 22*(4), 1694-1703.

Rushworth, M. F., Behrens, T. E., Rudebeck, P. H., & Walton, M. E. (2007). Contrasting roles for cingulate and orbitofrontal cortex in decisions and social behaviour. *Trends Cogn Sci, 11*(4), 168-176.

Rustichini, A. (2008). Dual or unitary system? Two alternative models of decision making. *Cogn Affect Behav Neurosci, 8*(4), 355-362.

Sallet, J., Quilodran, R., Rothe, M., Vezoli, J., Joseph, J. P., & Procyk, E. (2007). Expectations, gains, and losses in the anterior cingulate cortex. *Cogn Affect Behav Neurosci, 7*(4), 327-336.

Sally, B. (1995). When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences. *Organizational Behavior and Human Decision Processes, 63*(2), 131-144.

Samanez-Larkin, G. R., Gibbs, S. E., Khanna, K., Nielsen, L., Carstensen, L. L., & Knutson, B. (2007). Anticipation of monetary gain but not loss in healthy older adults. *Nat Neurosci, 10*(6), 787-791.

Sanfey, A. G. (2007). Social decision-making: insights from game theory and neuroscience. *Science, 318*(5850), 598-602.

Sanfey, A. G., & Chang, L. J. (2008). Multiple systems in decision making. *Ann N Y Acad Sci, 1128*, 53-62.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science, 300*(5626), 1755-1758.

Sayers, B. M., Beagley, H. A., & Henshall, W. R. (1974). The mechansim of auditory evoked EEG responses. *Nature, 247*(441), 481-483.

Schiller, D., Freeman, J. B., Mitchell, J. P., Uleman, J. S., & Phelps, E. A. (2009). A neural mechanism of first impressions. *Nat Neurosci, 12*(4), 508-514.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and Automatic Human Information Processing: I. Detection, Search and Attention. *Psychological Review, 84*(1), 1-66.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *J Neurophysiol, 80*(1), 1-27.

Schutter, D. J., & Van Honk, J. (2005). Electrophysiological ratio markers for the balance between reward and punishment. *Cognitive Brain Research, 24*(3), 685-690.

Seligman, M. E., & Maier, S. F. (1967). Failure to escape traumatic shock. *J Exp Psychol, 74*(1), 1-9.

Seymour, B., Daw, N., Dayan, P., Singer, T., & Dolan, R. (2007). Differential encoding of losses and gains in the human striatum. *J Neurosci, 27*(18), 4826-4831.

Seymour, B., Singer, T., & Dolan, R. (2007). The neurobiology of punishment. *Nat Rev Neurosci, 8*(4), 300-311.

Shafir, E., & LeBoeuf, R. A. (2002). Rationality. *Annu Rev Psychol, 53*, 491-517.

Siapas, A. G., Lubenov, E. V., & Wilson, M. A. (2005). Prefrontal phase locking to hippocampal theta oscillations. *Neuron, 46*(1), 141-151.

Simon, J. J., Walther, S., Fiebach, C. J., Friederich, H. C., Stippich, C., Weisbrod, M., et al. (2009). Neural reward processing is modulated by approach- and avoidance-related personality traits. *Neuroimage, 49*(2), 1868-1874.

Singer, T., Critchley, H. D., & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends Cogn Sci, 13*(8), 334-340.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature, 439*(7075), 466-469.

Skaggs, W. E., McNaughton, B. L., Wilson, M. A., & Barnes, C. A. (1996). Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus, 6*(2), 149-172.

Sloman, S. A. (1996). The Empirical Case for two Systems of Reasoning. *Psychological Bulletin, 119*(1), 3-22.

Solnick, S. J., & Schweitzer, M. E. (1999). The Influence of Physical Attractiveness and Gender on Ultimatum Game Decisions. *Organ Behav Hum Decis Process, 79*(3), 199-215.

Stanovich, K. (2004). *The Robot's Rebellion: Finding Meaning in the Age of Darwin* (1 edition, May 15, 2004 ed.). Chicago: University of Chicago Press.

Stockwell, R. G. (2007). A basis for efficient representation of the S-transform. *Digital Signal Processing, 17*(1), 371-393.

Stockwell, R. G., Mansinha, L., & Lowe, R. P. (1996). Localization of the complex spectrum: the S transform. *IEEE Transactions on Signal Processing, 44*(4), 998-1001.

Sugase-Miyamoto, Y., & Richmond, B. J. (2007). Cue and reward signals carried by monkey entorhinal cortex neurons during reward schedules. *Exp Brain Res, 181*(2), 267-276.

Thaler, R. H. (2000). From Homo Economicus to Homo Sapiens. *Journal of Economic Perspectives, 14*(1), 133-141.

Thomson, J. J. (1985). The Trolley Problem. *The Yale Law Journal, 94*(6), 1395-1415.

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2007). Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. *J Neurophysiol, 97*(2), 1621-1632.

Todorov, A., Harris, L. T., & Fiske, S. T. (2006). Toward socially inspired social neuroscience. *Brain Res, 1079*(1), 76-85.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*(5728), 1623-1626.

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science, 315*(5811), 515-518.

Tsuchiya, N., Moradi, F., Felsen, C., Yamazaki, M., & Adolphs, R. (2009). Intact rapid detection of fearful faces in the absence of the amygdala. *Nat Neurosci, 12*(10), 1224-1225.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*(4157), 1124-1131.

Tversky, A., & Kahneman, D. (1983). Extentional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review, 90*(4), 293-315.

Tykocinski, O. E. (2001). I never had a chance: Using hindsight tactis to mitigate disappointments. *Personality and Social Psychology Bulletin, 27*(3), 376-382.

Tzur, G., & Berger, A. (2009). Fast and slow brain rhythms in rule/expectation violation tasks: focusing on evaluation processes by excluding motor action. *Behav Brain Res, 198*(2), 420-428.

van Dijk, W. W., Zeelenberg, M., & van der Pligt, J. (2003). Blessed are those who expect nothing: Lowering expectations as a way of avoiding disappointment. *Journal of Economic Psychology, 24*(4), 505-516.

van't Wout, M., Kahn, R. S., Sanfey, A. G., & Aleman, A. (2005). Repetitive transcranial magnetic stimulation over the right dorsolateral prefrontal cortex affects strategic decision-making. *Neuroreport, 16*(16), 1849-1852.

van't Wout, M., Kahn, R. S., Sanfey, A. G., & Aleman, A. (2006). Affective state and decision-making in the Ultimatum Game. *Exp Brain Res, 169*(4), 564-568.

van't Wout, M., & Sanfey, A. G. (2008). Friend or foe: the effect of implicit trustworthiness judgments in social decision-making. *Cognition, 108*(3), 796-803.

Wallace, B., Cesarini, D., Lichtenstein, P., & Johannesson, M. (2007). Heritability of ultimatum game responder behavior. *Proc Natl Acad Sci U S A, 104*(40), 15631-15634.

Walton, M. E., Devlin, J. T., & Rushworth, M. F. (2004). Interactions between decision making and performance monitoring within prefrontal cortex. *Nat Neurosci, 7*(11), 1259-1265.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology, 20*(3), 273-281.

Wendel, K., Vaisanen, O., Malmivuo, J., Gencer, N. G., Vanrumste, B., Durka, P., et al. (2009). EEG/MEG Source Imaging: Methods, Challenges, and Open Issues. *Comput Intell Neurosci*, 656092.

Whiteside, S., & Lynam, D. (2001). The five factor model and impulsivity: using a structural model of personality to understand impulsivity. *Personality and Individual Differences, 30*(4), 669-689.

Wicker, B., Keysers, C., Plailly, J., Royet, J. P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust. *Neuron, 40*(3), 655-664.

Willis, J., & Todorov, A. (2006). First impressions: making up your mind after a 100-ms exposure to a face. *Psychol Sci, 17*(7), 592-598.

Winterer, G., Ziller, M., Dorn, H., Frick, K., Mulert, C., Wuebben, Y., et al. (2000). Schizophrenia: reduced signal-to-noise ratio and impaired phase-locking during information processing. *Clin Neurophysiol, 111*(5), 837-849.

Wirth, S., Avsar, E., Chiu, C. C., Sharma, V., Smith, A. C., Brown, E., et al. (2009). Trial Outcome and Associative Learning Signals in the Monkey Hippocampus. *Neuron, 61*(6), 930-940.

Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proc Natl Acad Sci U S A, 102*(20), 7398-7401.

Yeung, N., & Sanfey, A. G. (2004). Independent coding of reward magnitude and valence in the human brain. *J Neurosci, 24*(28), 6258-6264.

Zaghloul, K. A., Blanco, J. A., Weidemann, C. T., McGill, K., Jaggi, J. L., Baltuch, G. H., et al. (2009). Human substantia nigra neurons encode unexpected financial rewards. *Science, 323*(5920), 1496-1499.

Zak, P. J., Borja, K., Matzner, W. T., & Kurzban, R. (2005). The Neuroeconomics of Distrust: Sex Differences in Behavior and Physiology. *The American Economic Review: Papers and Proceedings, 95*(2), 360-363.

Zak, P. J., Kurzban, R., & Matzner, W. T. (2005). Oxytocin is associated with human trustworthiness. *Horm Behav, 48*(5), 522-527.

Zeelenberg, M. (1999). Anticipated Regret, Expected Feedback and Behavioral Decision Making. *Journal of Behavioral Decision Making, 12*(2), 93-106.

Zeelenberg, M., van Dijk, W. W., Manstead, A. S. R., & van der Pligt, J. (2000). On bad decisions and disconfirmed expectancies: The psychology of regret and disappointment. *Cognition and Emotion, 14*(4), 521-541.

Zeelenberg, M., van Dijk, W. W., van der Pligt, J., Manstead, A. S. R., van Empelen, P., & Reinderman, D. (1998). Emotional Reactions to the Outcomes of Decisions: The Role of Counterfactual Thought in the Experience of Regret and Disappointment. *Organ Behav Hum Decis Process, 75*(2), 117-141.