Scanning the Humean Brain:
The Neural Basis of Social Conventions and Norms [*]

Francesco Guala
Department of Economics, University of Milan

Tim Hodgson
School of Psychology, University of Exeter

[Preliminary draft – April 2010]

Abstract

Hume famously described social institutions as conventional equilibria in repeated coordination games. He pointed out that conventions are threatened by occasional changes in incentives that induce players to deviate from the established path. According to Hume's followers these temptations are partly offset by expectations of conformity that turn conventions into social norms. In this paper we present neuroscientific evidence that partly confirms this Humean insight. Breaches of convention prompt reactions similar to violations of social norms. However, conformity to conventions is also sustained by a natural human tendency to follow past regularities and to avoid uncertain prospects.

David Hume is the philosopher of *induction*, *passion*, and *custom*. These three pillars support every archway in Hume's philosophical edifice, from epistemology to ethics, social philosophy, and political science. By highlighting the role of habits, induction and sentiments, Hume left for his followers the major task of articulating in detail the ways in which they jointly support and constrain human cognition, sociality, and morality. For Hume's legacy includes also philosophical naturalism, and naturalists expect philosophical inquiry to make headway slowly – but steadily – with the progress of science.

In this paper we focus on a specific aspect of Hume's legacy, which is also one of the central puzzles of modern social theory. Over the last three decades the emergence and resilience of social conventions and norms has been widely studied by economists, philosophers, psychologists, evolutionary biologists and anthropologists. Progress has been fostered especially by the development of new tools in the area of evolutionary game theory, which have allowed the simulation of hypothetical scenarios using rigorous mathematical and computational techniques.[1]

In this paper we tackle these questions from a different angle: relying on the methods of neuroscience, we try to identify the proximate causal mechanisms that are responsible for the emergence and consolidation of social conventions and norms through the repeated interactions of small groups of individuals. Contemporary research in social neuroscience emphasizes the role played by circuits governing emotional reactions in the limbic brain

[*] Funded by ESRC/MRC grant RES-000-22-2392. Tim Miller's and Hannah Enke's assistance in running the experiments is gratefully acknowledged. We take full responsibility for all the remaining mistakes.
[1] E.g. Binmore (1998, 2006), Skyrms (1996, 2004), Boyd and Richerson (2005), Gintis (2009).

(Adolphs 2003), and higher cognitive functions in the frontal cortex that enable strategizing in complex social settings (Frith 2007). Some of this research has relied on experimental designs that have been widely used by economists and psychologists interested in the foundations of human sociality. Social dilemma games such as the Prisoners' Dilemma and the Ultimatum Game have featured prominently in this tradition, and inevitably social neuroscience has inherited their virtues and limitations.

It is by no means obvious, for example, that social dilemma games are good models for the repeated interactions that constitute the bulk of our social life. Games like the Prisoner's Dilemma and the Ultimatum Game are particularly unfavourable settings for pro-social behaviour (or "cooperation", for short). Each player's self-interest is opposed to the interests of the other players, and the individually rational strategy is never to cooperate in these games. This is quite unrealistic, for in many real-life situations cooperation is both collectively and individually beneficial, and each player has strong reasons to build a reputation of reliable cooperator.

One swift solution is simply to abandon social dilemma games and model life as a repeated coordination game where the interests of all players are always aligned (Skyrms 2004, Binmore 2006). This sounds too easy, however: to stipulate that cooperation is *always* individually optimal seems as unrealistic as to assume that it never is – we all face dilemmas of cooperation that try our social conscience every now and then. A realistic middle ground is to recognize that *both* coordination and social dilemmas feature in our daily interactions, and human sociality has evolved in the context of complex games of this kind.

In the next two sections we introduce a complex game described by Hume in his social and political writings. Hume's game consists in a series of coordination stage-games occasionally interrupted by "Temptation rounds" that offer the opportunity to free ride at the expense of the other players. Studying Hume's game in the neuroscience lab reveals that pro-social behaviour is partly sustained by emotions, as social neuroscientists have repeatedly emphasized. More precisely, social conformity derives from a natural human aversion to ambiguity and uncertainty, modulated by a neural "braking system" located in the human amygdala. The amygdala sustains a propensity to follow past regularities and social customs that explains the resilience of conventions beyond the narrow boundaries of the coordination games in which they have evolved. Neural evidence thus confirms two important insights of Hume and his followers, namely, that conventions tend to turn into norms, and that our affective propensities play a key role in the emergence and resilience of social institutions.

1. Social coordination

Hume is considered one of the great precursors of the game-theoretic analysis of social institutions.[2] In a famous paragraph of the *Treatise of Human Nature*, he compares social coordination with the action of two rowers in a boat:

> Two men, who pull the oars of a boat, do it by an agreement or convention, tho' they have never given promises to each other. Nor is the rule concerning the stability of possession the less deriv'd from human conventions, that it arises gradually, and acquires force by a slow progression, and by our repeated experience of the inconveniences of transgressing it. On the contrary, this experience assures us still

[2] See Gauthier (1979), Sugden (1986), Binmore (1998).

> more, that the sense of interest has become common to all our fellows, and gives us a confidence of the future regularity of their conduct: And 'tis only on the expectation of this, that our moderation and abstinence are founded. (Hume 1740: Part II, Section II)

Two important ideas are expressed in this passage: first, social institutions like property rights are *conventions* that arise from repeated coordination (the rowers' analogy). Secondly, *inductive learning* ("experience") is a crucial mechanism sustaining human conventions.

Hume's insights were revived in the twentieth century, when rational choice theorists were finally able to model and analyze rigorously social coordination. A simple symmetric coordination game is represented in Table 1. Each player has two possible strategies, Left or Right, and the numbers in each cell represent their payoffs (the first number for the row player, the second one for column).

	Left	Right
Left	1, 1	0, 0
Right	0, 0	1, 1

Table 1: A simple coordination game

Standard game theory, to be sure, does not give particularly helpful advice in situations of this kind: a perfectly rational calculator cannot do better than flipping a coin and choosing a strategy at random. In a seminal study of coordination games Thomas Schelling (1960) however noticed that in many real-life interactions we are much more successful than purely rational calculators. He argued that seemingly irrelevant features of the environment, such as the position of the objects of choice or the way they are labelled, function as cues that help us converge on a common solution. A strategy that is made salient by such features is called a "focal point", and in the course of repeated interaction is likely to become a point of attraction for the individuals in a given population (see also Sugden 1986; Bacharach and Bernasconi 1997).

Schelling's hypothesis was mainly based on anecdotal examples and rudimentary experiments, but recent more systematic studies have confirmed its validity (e.g. Mehta et al. 1994). Among the "irrelevant" details that can make a strategy salient, Schelling argued, *history* plays an important role: "Precedence seems to exercise an influence that greatly exceeds its logical importance or legal force", and "there is […] a strong attraction to the *status quo ante*" (1960: 67-8). This idea was further articulated by David Lewis (1969) in his classic book on *Convention*. Lewis, who was mainly interested in the conventional nature of language, argued that a strategy can become salient simply in virtue of the fact that it was played by a sufficiently large number of people in previous rounds of the game. When this happens, we shall say that a "convention" has emerged in a given population.

2. Hume's game

In normal circumstances conformity with a convention is supported by rational self-interest: we drive on the left because we want to avoid accidents; we say "cat" rather than "tac" because we want to be understood by our interlocutors; we wear black at funerals because we want to communicate our grief (cf. Hume 1740, Part II, Section II; Lewis 1969). But rational

self-interest, on its own, is probably too slender a basis for sociality. There are always "trembling hand" accidents, to begin with, when individuals deviate by mistake from the established equilibrium. These deviations may have only minor effects in payoff terms, but can nevertheless generate confusion in the other players, insinuating doubts regarding the kind of game that is actually being played. And sometimes the payoff structure *does* change, because of some change in players' preferences or in their material incentives. Even though coordination and cooperation are advantageous to all in the long run, a single individual may forget it when faced with an attractive opportunity to deviate in the short run. This may trigger further breaches of conventions, and undermine the mutual expectations of conformity that constitute the basis of long-term social cooperation.

Hume was aware of the problem. His vision of social life can be best represented as a sequence of coordination games occasionally interrupted by "Temptation rounds" that offer an opportunity to gain at the expense of other players:

> All men are sensible of the necessity of justice to maintain peace and order, and all men are sensible of the necessity of peace and order for the maintenance of society. Yet […] such is the frailty or perverseness of our nature! It is impossible to keep men faithfully and unerringly in the paths of justice. Some extraordinary circumstances may happen, in which a man finds his interests to be more promoted by fraud or rapine, than hurt by the breach which his injustice makes in the social union. (Hume 1777: Part I, No. V)

The game in Figure 2 tries to capture the structure of the predicament highlighted by Hume (we shall call it "Hume's game"). Over an initial sequence of uninterrupted coordination rounds, two players have the chance to create a convention of the Schelling-Lewis type. Over time, however, the sequence is interrupted by "Temptation rounds" offering the possibility of deviation to one player (Row) at the expense of the other (Column). Let us suppose for simplicity that only the row player (the "Potential Deviant") is aware of the imminent change of payoffs before a Temptation round. The other player (Column) is taken by surprise, and offers Row a free-riding opportunity.

	Left	Right
Left	50p, 50p	0, 0
Right	0, 0	50p, 50p

Coordination round (C)

	Left	Right
Left	50p, 50p	£2, 0
Right	£2, 0	50p, 50p

Temptation round (T)

1 2 3 4 5 6 7 8 **9** 10 11 12 **13** 14 15 16 **17** 18 19 20 **21**
C→C→C→C→C→C→C→C→**T**→C→C→C→**T**→C→C→C→**T**→C→C→C→**T**

Figure 2: Sequence of Coordination and Temptation rounds in Hume's game

Although the Potential Deviant faces an incentive to deviate at each Temptation round, she is also potentially vulnerable to sanctions. Column may withdraw cooperation following a breach of convention: a "trigger strategy", in game-theoretic jargon, which may deter deviations in the early rounds of the game. At the very last round, however, even this threat becomes ineffective: the game will finish and the players will never meet again. At this point,

the Potential Deviant has no selfish incentive to stick to the convention that has evolved thus far.

Hume's game is quite different from the situations that are typically analysed by game theorists. It is, to begin with, a complex game obtained by combining two different strategic settings.[3] There is radical uncertainty about the rounds that are going to be played, and players must rely on speculative conjectures about the future. Although they are difficult to analyse using the standard tools of game theory, complex games are very common in real life. Social life is neither an infinitely repeated Prisoner's dilemma nor a pure coordination game, but probably a combination of both. We rarely know for sure what stage-game we are facing, and even when we think that we know it, the equilibria are often too many or too few to allow the formulation of precise game-theoretic predictions.

Following a custom – a regularity that has evolved in the past, in a class of similar games – provides a cheap solution to this predicament. This idea is very much in the spirit of the Schelling-Lewis approach: if conventions help simplifying decisions when there are too many equilibria, they may do the same trick when there are none. Humean theories in fact belong to a broader tradition in economics and social science that views institutions as means to facilitate coordination and reduce transaction costs (North 1984, 1990). The convention of paying a one percent fee on house sales for example can save us a lot of time and bargaining hassle, if I don't know the costs of my estate agent and he doesn't know my reservation price. Or consider the classic problem of enforcing contracts: payment and delivery of a good or service do not always take place simultaneously, so the buyers need to trust that the goods will be delivered, the sellers that payment will follow.

To deal with these problems, successful societies have developed institutions like courts, merchants' trades, watchdogs and arbitrators. But informal customs and norms also contribute substantially to reduce uncertainty and transaction costs. Successful conventions however must be resilient – robust to the threat of temptation highlighted by Hume. While external factors (for example, policing institutions that reduce the incentive to deviate) certainly play a role, there may also be *internal* cognitive mechanisms that enhance the resilience of conventions and facilitate sociality in complex games.

3. Scanning the Humean brain

Cognitive neuroscience is a branch of psychology devoted to studying the neurological substrates of mental mechanisms. Over the last decade it has made quick progress thanks to the improvement of hemodynamic imaging techniques, such as fMRI and PET, which give the opportunity to identify brain areas that are significantly active during the performance of cognitive tasks. Functional Magnetic Resonance Imaging (fMRI) allows the measurement of blood flowing in the brain of human and animal subjects, via the detection of so-called BOLD (Blood Oxygenated Level Dependent) signals. Oxygenated blood feeds neurons, and moves towards those areas of the brain that are most active at a given time. Because of the lag between the firing of neurons and blood flow (4-5 seconds on average), fMRI scanning provides a relatively rough temporal map of brain events, which is however compensated by higher spatial precision than can be achieved by any other existing technology.[4]

[3] Complex games are remarkably understudied – but see Zollman (2008) for an exception.

[4] Other technologies, such as electroencephalography, have the opposite trade off: high temporal but low spatial precision. Notice that imaging data are not the only source of insight in cognitive neuroscience; other kinds of

Armed with imaging technology we can look at Hume's game in the experimental laboratory. The experimental setting is the following: two individuals try to coordinate by pressing one of two buttons named "Left" and "Right". The players interact anonymously via a computer network, and earn money whenever they converge on the same option. The subjects are told that the game will be played for twenty-one rounds and that cumulative earnings will be paid at the end of the experiment. They are also warned that some unspecified rounds may have a "special" payoff structure, which may or may not be revealed before the round is played.

Right at the start one of the subjects is selected randomly and invited to lie in the fMRI scanner. The screen of her PC is reflected via a mirror, and her decisions are transmitted using two buttons on a remote control. The subject's brain is scanned repeatedly before she makes her decisions (the "decision period"), and when she receives feedback about the other player's moves (the "outcome period").

Our version of Hume's game features four surprise Temptation stage-games at rounds 9, 13, 17, and 21. The other seventeen rounds are pure coordination games such as the one of Table 1. Deviation rates in Temptation rounds range between 40 and 53 percent. There is an increasing – but statistically insignificant – tendency to breach the convention in later rounds, perhaps because realise that the payoffs from future coordination decline as the end of the game is approaching.[5] Still, almost fifty percent of Potential Deviants cooperate even in the last round, when the "shadow of the future" has completely disappeared.[6]

So conventions tend to induce conformity in spite of individualistic incentives to deviate. Several experimental subjects are willing to give up some material gains and stick to the rule of conduct that evolved in the early part of the game. The interesting question, then, is *how* this happens: what proximate mechanisms make conventions robust to disturbances and even changes in the structure of payoffs?

4. Trembling hands and moral disgust

Subjects typically reach coordination between the fourth and the eighth round. Once a convention is established, most of them keep choosing it unproblematically, and make money by simply sticking to the rule. They apply the inductive principle that – other things being equal – the other player will continue to behave as she has done until now, and she expects us to do the same. Even before the first Temptation round, however, coordination is disrupted by occasional "trembling hand" deviations. Figure 3a shows activations in so-called "outcome periods" for the other player, when she realizes that a trembling-hand deviation has just occurred. The coloured areas represent regions of the brain where a statistically significant increase in the level of blood has been detected, compared with the control condition (in this case, outcome periods after successful coordination).

---

evidence, such as lesion studies, give additional information concerning causal relations between brain, cognition, and behaviour (we shall see some examples later in the paper).

[5] It is well known from experimental game theory that subjects have difficulties applying backward induction reasoning thoroughly (e.g. Johnson et al. 2002).

[6] This replicates previous experimental results, such as e.g. Guala and Mittone (2010).

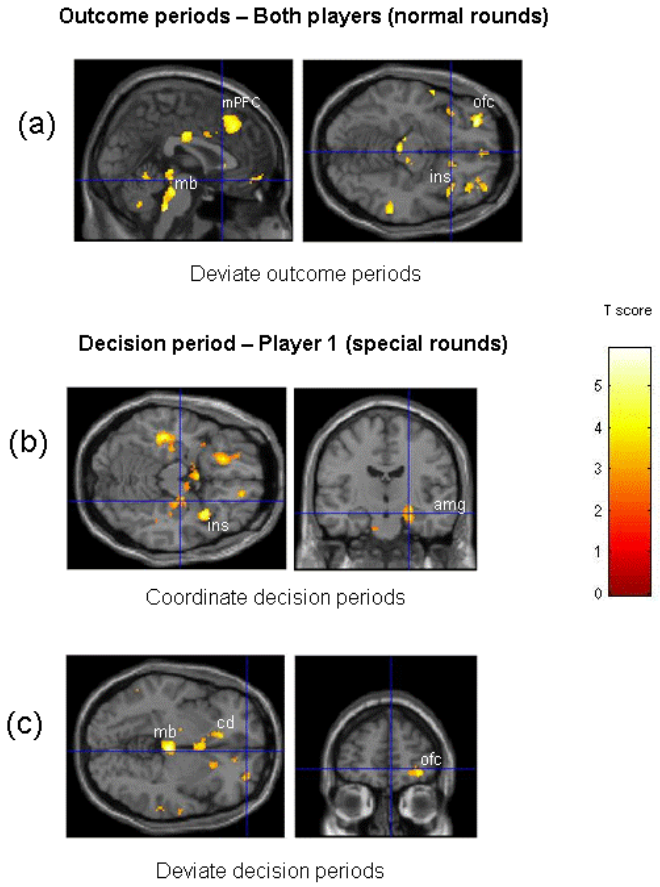

Figure 3

Trembling-hand deviations from a convention elicit activations in the *substantia nigra*, which is part of the *basal ganglia* in the human mid-brain (mb). The substantia nigra is associated (among other things) with the processing of surprising events. In this case it probably reflects the surprised reaction of experimental subjects when their partners breach the convention without apparent reason. This surprise, however, is charged with negative emotion. Indicated by the arrow is the anterior *insular cortex* (ins) A substantial body of previous research has mapped insula activation onto a set of negative feelings – such as pain – and emotions – including anger, fear, and disgust.

In a much-publicized experiment, Sanfey et al. (2003) have discovered that subjects who receive unfair offers in a simple bargaining game (the Ultimatum game) display disproportionate activation of the anterior insula. The Ultimatum game is a two-stage sequential game which gives all the bargaining power to the first mover. Two players receive a sum of money (say, 10$) from the experimenter, and must decide how to split it. Player 1 (the Proposer) makes an offer, which Player 2 (the Responder) can only accept or reject. If she accepts, they each get what the Proposer offered; if she rejects, they walk out with nothing. The rational equilibrium of this game is a very unequal split: Player 1 offers the minimum amount possible (say, 1$) and Player 2 accepts because it's better than nothing.

In practice, things do not work out this way. Experiments in Western societies – where fairness and equality are important values – report mean offers of around 40% of the cake, and modal offers at 50%. Offers of 20% or less are rejected about half of the time. The fMRI

data of Sanfey and his colleagues confirm an old hypothesis in behavioural game theory: emotions may sustain pro-social behaviour when rational deliberation fails (Hirshleifer 1987, Frank 1988). Anger and moral disgust force the rejection of unfair offers. The other players know this, and do not try to exploit their bargaining power to their full advantage.

So it seems that exploratory behaviour in repeated coordination games is received pretty much in the same way as the violation of norms of equality in the Ultimatum game: breaches of convention hurt. But why do Potential Deviants care? Disappointing fellow players may have negative consequences in the complex games that we play in real life. For if I disrupt others' expectations

> they will be surprised, and they will tend to explain my conduct discreditably. The poor opinions they form of me, and their reproaches, punishment, and distrust are the unfavourable responses I have evoked by my failure to conform to the convention. (Lewis 1969: 99)

This account can be generalized to cover a wider class of situations than pure coordination games (Sugden 1998). The crucial mechanism is a hypothetical human tendency to please others: as social animals, human beings have likely acquired through evolution a "basic desire" to keep the good will of the members of their group (Sugden 1986: 156). This basic desire makes us feel uneasy when we become the object of resentment, as for instance when we breach a convention. Add to this fact that *we* resent it when other people frustrate our expectations, and we have a plausible explanation of why conventions tend to turn into social norms.

5. The cautious brain

We also looked at decision periods before Temptations rounds, in the few seconds that precede decision to conform. Two areas are involved in particular: the *amygdala*, and the *orbitofrontal cortex*. What we know about these regions of the brain suggests an interesting picture of decision-making in complex coordination games.

Conformist players display increased activation in the amygdala, a limbic region of the brain integrated in the dopaminergic pathway (Phelps 2006, Seymour and Dolan 2008). The amygdala is involved in the processing of positive and negative emotions, in particular fear, and it is known to interact with frontal areas of the brain during decision-making. While subjects who stick to the convention have greater amygdala activation during the decision periods, those who deviate in Hume's game display more activation in the orbitofrontal cortex (OFC) (Figure 3b,c).

Previous studies suggest that the OFC and the amygdala work together to learn associations between stimulus and reward. Hampton et al. (2007) report fMRI scanning of two patients with amygdala lesions, while engaged in a switch/stay task that is in many ways similar to our experiment. Subjects had to choose one of two buttons, each delivering a monetary gain or loss with a given (but unknown) probability. The probabilities were not stable throughout the game. On the basis of observed outcomes, the subjects had to learn to switch buttons when the probabilities changed (i.e. when pressing a button became on average more profitable).

Normal subjects displayed greater activity in the OFC during switch trials compared with stay trials.[7] Interestingly, subjects with amygdala damage in the experiment of Hampton and colleagues had anomalous (enhanced) OFC activity compared to normal subjects. Behaviourally, the amygdala patients had problems "staying", and switched buttons too frequently. This suggests that in normal subjects the amygdala damps the OFC impulse to deviate from a rule that has been followed until now. While the OFC seeks new opportunities, the amygdala acts as the "moderator" or conservative advisor in our brain.

The causal relations between OFC, amygdala, and behaviour have been studied by surgical intervention in monkeys. In an instrumental learning experiment (Rudebeck and Murray 2008) monkeys with amygdala lesions switched more frequently to a recently rewarded option following negative feedback (they persevered *less* in following a learned rule), while monkeys with OFC lesions continued to choose previously rewarded options regardless of negative outcomes (they persevered *more*). This is the same tendency reported by Damasio (1994) in experiments with the "Iowa test", where OFC-impaired human subjects cannot switch to more profitable decks of cards as normal subjects do.[8]

Notice that there is no social learning in these experiments – subjects are reasoning about the properties of natural (non-intentional) systems. So it is intriguing that amygdala and OFC activations are inversely correlated in Hume's game too. Collectively, these data support the hypothesis that OFC and amygdala play "opportunistic" and "moderating" functions across a variety of decision and learning tasks, of social and non-social nature alike.

It is well established for example that the amygdala is involved in the recognition of trustworthy faces (Adolphs et al. 2000) and the perception of racial difference (Phelps et al. 2000). Clinical studies confirm that the amygdala has a "gateway" function, screening between safe and potentially threatening social situations. Subjects with "Williams Syndrome" manifest excessive friendliness towards strangers and reduced capacity to identify threatening social stimuli (such as angry facial expressions). Post-mortem and fMRI studies indicate that Williams syndrome patients suffer from anomalies at the level of amygdala-OFC interaction that impair their social inhibition mechanisms (they are too friendly and trustworthy, especially with strangers – see Meyer-Lindenberg et al. 2006).

Finally, high amygdala activity is associated with low transfers in economic trust games,[9] and its suppression using the synthetic neuropeptide Oxytocin causes an increase of trusting behaviour (Baumgartner et al. 2008). Interestingly, Oxytocin does not change subjects' beliefs in the probability of a positive outcome, nor does it simply prompt a desire to benefit others. Damping amygdala activity, rather, seems to reduce subjects' anxiety regarding social betrayal.

6. Punishment and uncertainty

The amygdala has a regulatory function on social behaviour. In particular, it acts as an alarm system that signals potentially dangerous situations. In the context of Hume's game, this

[7] This is consistent with previous evidence that individuals with OFC lesions find it difficult to reverse an association once it has been learned (e.g. Fellows and Farah 2005).
[8] For a critical analysis of Damasio's classic interpretation of the Iowa Gambling Task, see Colombetti (2008).
[9] In a trust game one player (the investor) sends a sum of money to another player (the entrepreneur); the money is doubled by the experimenter, and the entrepreneur has the opportunity of sending back part, none, or all of the money to the investor.

raises the question of *what* exactly the danger may be. Or, in folk-psychological terms, what are conformist subjects worried about?

Breaches of conventions may be associated (consciously or subconsciously) with sanctions. Sanctions are an important mechanism for the enforcement of social norms (Sober and Wilson 1998) and come in many varieties – from gossip, verbal reproaches and ostracism, to material and physical punishment (Boehm 1999). It is unlikely however that conformity is caused by an internalized fear of material punishment. A recent study (Li et al. 2009) reports differential brain activations in a trust game played with and without punishment. When punishment is available, typical reward areas such as the parietal cortex are activated in the trustee's brain. In the absence of punishment threat, in contrast, there is activation of the very same regions associated with conformist behaviour in Hume's game – including amygdala and OFC. Conformism without punishment exploits different neural mechanisms and is not merely an internalization of punishment threat.

If the amygdala is involved in detecting social threat, we must conceive the latter in much broader terms than the punishment theory does. Although we can only speculate at this stage, we propose an interpretation that combines evidence coming from both social and non-social experimental tasks.

The amygdala is almost certainly involved in the evaluation of probabilistic prospects. As in standard decision theory, we must distinguish between *risk* and *uncertainty*: current work in cognitive neuroscience suggests that the evaluation of risky prospects (where *objective* probabilities are involved) takes place in more evolved, "higher" regions of the brain such as the medial prefrontal cortex (mPFC) (Knutson et al. 2005). In contrast, the evaluation of ambiguous and uncertain prospects (where objective chances are unknown and actors must rely on their *subjective* estimates) relies on visceral signals from the limbic brain. Hsu et al. (2005) for example report significant increases in amygdala-OFC activation in tasks with ambiguous and uncertain prospects (e.g. the number of winning cards in a deck is unknown) compared with risky tasks (where the number of cards is known).[10]

Notice that there is little to calculate in Hume's game: there are no objective probabilities upon which one can build an objective estimate of future rewards, and the reaction of the other players adds an extra element of uncertainty.[11] So from a neural point of view following a social rule has less to do with the belief-updating of a Bayesian calculator than with the trial-and-error attempts of an animal who has to cope with uncertainty. The main reason why conformist subjects stick to conventions is that they are *averse to uncertainty*, a braking mechanism that prevents the exploitation of short-term advantages at the expense of long-term stability of reward.[12]

[10] See also de Martino et al. (2006).

[11] Moreover, there is no difference in mPFC activation between conformist and deviant subjects in Hume's game. The only mPFC significant activations take place during outcome periods *after* Temptation rounds, when a convention has been breached. They probably reflect the attempt to make sense of the deviant's behaviour by "simulating" their decision on the basis of the unusual payoff structure of Temptation rounds.

[12] There is also a small cluster of activation (about 10 voxel) in the left dorsolateral prefrontal cortex (DLPFC) of conformist players (not displayed in Figure 3). The DLPFC mediates between emotive impulses and long-term, abstract, or impersonal rewards (Greene et al. 2004, Hare et al. 2009). Conformity thus seems to result, at least in part, from inhibition – resisting temptation in social games. We have strong evidence for causal inference here: when the DLPFC is temporarily "knocked down" using transcranial stimulation, subjects playing the Responder in the ultimatum game accept unfair offers more frequently than subjects with fully functioning DLPFC (Knoch et al. 2006a, 2008). There is also evidence that DLPFC controls behaviour by

From an evolutionary point of view, it is possible that our attitude towards uncertainty has been shaped by the pressure of social selection. According to the "Social Brain" hypothesis (Dunbar 1998) many characteristic cognitive functions of homo sapiens evolved to cope with the complexity of coordination in relatively large social groups. If the Social Brain hypothesis is true, in other words, we may be averse to uncertainty because following customs is advantageous in social decision-making.

The sensitivity of the amygdala to racial differences (Phelps et al. 2000) finds a natural explanation in this context. Social group boundaries (including race) signal that we are entering a morally ambiguous zone where the usual customs and rules do not necessarily apply. The transgression of social conventions – even relatively recent and transient rules such as those that we observe in Hume's game – probably triggers the same alarm system: the amygdala limits our explorations into socially uncertain terrain.

7. Social norms and predictability

The breakdown of social customs is an extremely important source of uncertainty for homo sapiens – the *most* important one, perhaps, in terms of fitness. Coordination required the recruitment of an alarm system in the human brain, the amygdala, devoted to the detection and inhibition of ambiguous situations. Social cognition then may result from a compromise between two desiderata: maximizing the advantages of Machiavellian reasoning and exploratory behaviour for the exploitation of new opportunities; but also maximizing reliability and predictability for the sake of coordination. These desiderata often pull in opposite directions, and behaviour in Hume's game reflects different ways of coping with this tension. Conformity with social norms and conventions is enhanced by a neural "braking system" that guarantees a degree of stability in spite of changes in incentives and uncertainties in the payoff structure.

These claims are largely consistent with the emphasis on pro-social emotions that pervades contemporary social neuroscience (Adolphs 2003, Singer and Fehr 2005, Frith 2007). Our theory departs from the latter, however, in one important respect. While we agree with evolutionary and neuro-psychologists that higher cognitive functions located in the human cortex provide crucial skills – such as the interpretation and anticipation of other people's intentional states – that allowed our ancestors to develop complex forms of sociality (e.g. Tomasello et al 2005, Amodio and Frith 2006), in our view they do not explain the very human capacity to preserve social cooperation in spite of incentives to defect from the social contract. In fact, sociality is to some extent made possible by mechanisms that *override* our strategizing capacities and stabilize behaviour in complex environments.

Stable, predictable behaviour has obvious advantages. Contrary to formal game theory, in real life we are never entirely sure what kind of games we are playing. We can never monitor exactly the payoff structures or the information that is available to our opponents, to begin with. Moreover, we often play several games on several tables at once, and bystanders rely on our moves to predict what we shall do in analogous (but rarely identical) future games (Sugden 1986: 155-157; Ross 2005: 282-289). Customs simplify decisions enormously in such circumstances, eliminating the requirement of constantly monitoring incentives, and

---

limiting exposure to uncertain prospects. Inhibition of DLPFC increases risk-taking behaviour (Knoch et al. 2006b); excitation using direct current stimulation, in contrast, makes subjects abnormally averse to uncertainty in gambling tasks (Fecteau et al. 2007).

grouping entire classes of games under the same heading. Signalling conformism is one way of saying that you are reliable – a predictable agent with whom to play (complex) coordination and cooperation games in the future.

References

Adolphs R., Tranel D. and Denburg N. (2000) "Impaired emotional declarative memory following unilateral amygdala damage". *Learning and Memory* 7, 180-186.

Adolphs R. (2003) "Cognitive neuroscience and human social behaviour". *Nature Reviews Neuroscience* 4, 165-178.

Amodio D.M. and Frith C.D. (2006) "Meeting of minds: the medial frontal cortex and social cognition", *Nature Reviews Neuroscience* 7, 268-277.

Bacharach M. and Bernasconi M. (1997) "The variable frame theory of focal points: an experimental study". *Games and Economic Behavior* 19, 1-45.

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, 639-650.

Binmore K. (1998) *Game Theory and the Social Contract, Vol. 2: Just Playing*. Cambridge, Mass.: MIT Press.

Binmore K. (2006) *Natural Justice*. Oxford: Oxford University Press.

Boehm C. (1999) *Hierarchy in the Forest*. Cambridge, Mass: Harvard University Press.

Boyd R. and Richerson P.J. (2005) *The Origins and Evolution of Cultures*. Oxford: Oxford University Press.

Colombetti, G. (2008) "The somatic marker hypothesis, and what the Iowa gambling task does and does not show". *British Journal for the Philosophy of Science* 59: 51-71.

Damasio, A. (1994) *Descartes' Error: Emotion, Reason, and the Human Brain*. London: Vintage.

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

Dunbar R.I.M. (1998) "The social brain hypothesis". *Evolutionary Anthropology* 6, 178–90.

Fecteau S, Pascual-Leone A, Zald DH, Liguori P, Théoret H, Boggio PS and Fregni F (2007) "Activation of prefrontal cortex by transcranial direct current stimulation reduces appetite for risk during ambiguous decision making", *Journal of Neuroscience* 27: 6212-6218.

Fellows, L.K. and Farah, M.J. (2005) “Different underlying impairments in decision-making following ventromedial and dorsolateral frontal lobe damage in humans”. *Cerebral Cortex* 15, 58-63.

Frank R.H. (1988) *Passions within Reason.* New York: Norton.

Frith C.D. (2007) “The social brain?” *Philosophical Transactions of the Royal Society B* 362, 671-678.

Gauthier D. (1979) “David Hume, contractarian”. *Philosophical Review* 88, 3-38.

Gintis H. (2009) *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton: Princeton University Press.

Greene J., Nystrom L., Engell A., Darley J. and Cohen J. (2004) “The neural bases of cognitive conflict and control in moral judgment”. *Neuron* 44, 389-400.

Guala F. and Mittone L. (2010) “An experimental study of conventions and norms”, *Journal of Economic Psychology,* forthcoming.

Hampton A.N., Adolphs R., Tyszka M.J. and O’Doherty J.P. (2007) “Contributions of the amygdala to reward expectancy and choice signals in human prefrontal cortex”. *Neuron* 55, 545-555.

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

Hirshleifer J. (1987) “On the emotions as guarantors of threats and promises”, in J. Dupré (ed.) *The Latest on the Best*. Cambridge, Mass.: MIT Press.

Hsu M., Bhatt M., Adolphs R., Tranel D. and Camerer C.F. (2005) **“**Neural systems responding to degrees of uncertainty in human decision-making”. *Science* 310, 1680–1683.

Hume D. (1740) *A Treatise of Human Nature*. Oxford: Oxford University Press.

Hume D. (1777) *Essays: Moral, Political and Literary*. Oxford: Oxford University Press.

Johnson E.J., Camerer C., Sen S. and Rymon T. (2002) “Detecting failures of backward induction: monitoring information search in sequential bargaining”. *Journal of Economic Theory* 104, 16-47.

Knoch D.., Pascual-Leone A., Meyer K. and Fehr E. (2006a) “Diminishing reciprocal fairness by disrupting the right prefrontal cortex”. *Science* 314, 829-832.

Knoch D., Gianotti L.R., Pasqual-Leone A., Treyer V., Regard M., Hohmann M. and Brugger P. (2006b) “Disruption of right prefrontal cortex by low-frequencey repeated transcranial magnetic stimulation induces risk-taking behaviour”, *Journal of Neuroscience* 26, 6469-6472.

Knoch D., Nitsche M.A., Fischbacher U., Eisenegger C., Pascual-Leone A. and Fehr E. (2008) "Studying the neurobiology of social interaction with transcranial direct current stimulation – the example of punishing unfairness". *Cerebral Cortex* 18, 1987-90.

Knutson B., Taylor J., Kaufman M., Peterson R. and Glover G. (2005) "Distributed neural representation of expected value". *Journal of Neuroscience* 25, 4806-4812.

Lewis D.K. (1969) *Convention: A Philosophical Study*. Cambridge, Mass.: Harvard University Press.

Li J., Xiao E., Houser D. and Montague P.R. (2009) "Neural responses to sanction threats in two-party economic exchange". *Proceedings of the National Academy of Sciences*, in press.

Mehta J., Starmer C. and Sugden R. (1994)" The nature of salience: an experimental investigation of pure coordination games". *American Economic Review* 84, 658-673.

Meyer-Lindenberg A., Mervis C.B. and Berman K.F. (2006) "Neural mechanisms in Williams syndrome: a unique window to genetic influences on cognition and behaviour". *Nature* 7, 380-393.

North D.C. (1984) "Transaction Costs, Institutions, and Economic History", *Journal of Institutional and Theoretical Economics* 140, 7-17.

North, D.C. (1990) *Institutions, Institutional Change, and Economic Performance*. New York: Cambridge University Press.

Phelps E.A., O'Connor K.J., Cunningham W.A., Funayama E.S., Gatenby J.C., Gore J.C. and Banaji M.R. (2000) "Performance on indirect measures of race evaluation predicts amygdala activation". *Journal of Cognitive Neuroscience* 12, 729-738.

Phelps E.A. (2006) "Emotion and cognition: insights from studies of the human amygdala". *Annual Review of Psychology* 57, 27-53.

Ross D. (2005) *Economic Theory and Cognitive Science: Microexplanation*. Cambridge, Mass.: MIT Press.

Rudebeck P.H. and Murray E.A. (2008) "Amygdala and orbitofrontal cortex lesions differentially influence choices during object reversal learning". *Journal of Neuroscience* 28, 8338-8343.

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

Schelling, T. (1960) *The Strategy of Conflict*. Cambridge, Mass.: Harvard University Press.

Seymour B. and Dolan R. (2008) "Emotion, decision making, and the amygdala". *Neuron* 58, 662-671.

Singer T. and Fehr E. (2005) "The neuroeconomics of mind reading and empathy". American *Economic Review* 95, 340-345.

Skyrms, B. (1996) *Evolution of the Social Contract*. Cambridge: Cambridge University Press.

Skyrms B. (2004) *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.

Sober E. and Wilson D.S. (1998) *Unto Others: The Evolution and Psychology of Unselfish Behaviour*. Cambridge, Mass.: Harvard University Press.

Spitzer M., Fischbacher U., Herrnberger B., Gron G. and Fehr E. (2007) "The neural signature of norm compliance". *Neuron* 56, 185-196.

Sugden R. (1986) *The Economics of Rights, Cooperation and Welfare*. Oxford: Blackwell.

Sugden R. (1998) "Normative expectations: the simultaneous evolution of institutions and norms", in *Economics, Values, and Organization*, edited by A. Ben -Ner and L. Putterman. New York: Cambridge University Press.

Tomasello M., Carpenter M., Call J., Behne T. and Moll H. (2005) "Understanding and sharing intentions: The origins of cultural cognition", *Behavioural and Brain Sciences* 28: 675-691.

Zollman K.J.S. (2008) "Explaining fairness in complex environments", *Politics, Philosophy and Economics* 7: 81-98.