Neuro-cognitive Mechanisms Mediating the Impact of Social Distance on Human Coordination

PhD student: Gabriele Chierchia

Advisor: Giorgio Coricelli

Doctoral School in Cognitive and Brain Sciences

Center for Mind/Brain Sciences (CIMeC), University of Trento

The structure of this thesis

This thesis is divided in 2 broad sections, both of which broadly focus on decision making, with a particular focus on strategic interactions (i.e. inspired from the behavioral game theory approach). The 1st section, "Neuro-Cognitive Mechanisms Mediating the Impact of Social Closeness on Coordination", constitutes the main body of the thesis and regards the project I personally most worked on during my PhD. It contains 2 articles: a behavioral study (Chierchia&Coricelli, under revision) (comprised of 3 experiments) and an fMRI study (Chierchia et al., in preparation). In these articles I attempt to connect a "hard problem" of game theory, namely coordination games, to notions from social psychology, sociology and social neuroscience/neuroeconomics within a simple framework. Experimentally, the studies in this section adopt one-shot coordination games and thus focus on how social information affects initial expectations and outcomes of strategic interactions. The second section contains 2 additional works. The first one is an fMRI study I contributed to (Fouragnan et al., 2013), in which we extend the previous line of inquiry from static to dynamic interactions. Here we were interested in how reliable social information (reputational priors) can constrain updating in repeated trust games, and how this learning mechanism is articulated in the brain. The last essay is a review chapter I wrote (Chierchia&Coricelli, 2011), which critically evaluates the capability of a strong "dual vs. unitary" dichotomy of cognition to account for recent findings in neuroeconomics.

Section 1 Neuro-cognitive Mechanisms Mediating the Impact of Social Distance on Coordination

Abstract

To model strategic interactions standard game theory assumes that agents have common knowledge of rationality. This allows agents to use deduction to form expectations on the behavior of their counterparts. In coordination games however, deduction fundamentally fails to prescribe a unique solution to agents, raising a "matching" problem in game theory. The question is, when deduction is of no use, how are agents to match or decouple their choices? The thesis explored here is inspired from the recent finding that humans recruit the same neural structures to reason about themselves and similar but not dissimilar others, and of friends but not strangers; a finding which has led some investigators to speak of self-referential mentalizing. This meshes nicely with the widely-established cross-species observation that social beings usually exhibit a preference for similar others; as well as with the well-known observation that social closeness fosters cooperation. However, as investigated by nearly all previous experiments, cooperation critically required agents to match their choices. My work adds to the experimental literature in several respects: i) it shows that both objective social closeness (friendship) and psychological closeness (perceived lab-induced similarity) can have an opposite effect on strategic interactions, depending on whether they require to match or decouple choices; ii) that this behavior is best explained by synergistic contributions of expected reciprocity and altruism, which we show to be dissociable both in behavior and the brain. From a neural perspective, expected reciprocity relies on the ventromedial prefrontal cortex, an area previously implicated in reward, interpersonal similarity and depth of reasoning; while the temporo-parietal junction is particularly important for altruism.

Taken together, our results provide novel insight into the neuro-cognitive mechanisms that facilitate social cohesion.

Introduction

The strategic mechanisms of social cohesion

More than 2 centuries ago Jean-Jacques Rousseau (1754) illustrated a dilemma, which today re-emerges as "the most difficult problem of game theory" (Camerer, 2003). The dilemma is this: 2 hunters are trying to catch a stag. To do so, it is critical that each one keep his post on separate sides of the hunting grounds, because the stag cannot be caught alone. However, there is some likelihood of a hare occasionally passing by and a hunter can catch the hare alone, with certainty. The hare is worth less than the stag, but it is there[1], and this could tempt a hunter to abandon his post to catch it. In which case, the other hunter will go home with nothing[2]. Crucially, both hunters are aware of this risk, so even in the absence of temptation (i.e. the hare), a hunter could defect from cooperation only for fear of the defection of his counterpart. Furthermore, hunter A could also defect because he believes hunter B believes A will defect, and this fear can *reverberate* further. Indeed, in lab experiments for real money and in large-scale economies, this type of "win-win" cooperation has often been shown to fail. This story captures an essential aspect of cooperation: we all know that we can often achieve more by working together, however, cooperation forces us to rely on others, which makes it intrinsically risky. In brief, with sociality comes exposure. So the question is: should we live in safe isolation, and catch our hares separately, or should we stick together, and catch a stag?

[1] "If it was a matter of hunting a deer, everyone well realized that he must remain faithful to his post; but if a hare happened to pass within reach of one of them, we cannot doubt

[2] To make this more concrete, think that the stag is worth $1000 to each hunter, and the hare only $500. The stag can be caught with certainty, but *only if* both hunters cooperate, the hare can be caught alone.

We actually know the end of this story. Humans, and other animals, "decided", through evolution, to live together, that is, to live close to one another, in a group. However, we are not uniformly distributed within this macro-group, but we are *clustered* in subgroups of nations, provinces, neighborhoods, gradually extending families and households; of culture and language, of ethnicities, religions, clans, clubs and political parties. The rationale of this thesis is that, to study what may have been the (neuro-cognitive) processes that enabled humans to "decide" to live close to one another and cooperate, we can try to study what mechanisms differentiate the interactions of "closer" and "farther" others in terms of such naturally-occurring subgroups. In other words, this thesis investigates the *strategic* mechanisms of social cohesion.

With the stag hunt example however, it seems we have told only "half" of the story. Suppose we decided to live close together, to reduce the commute and go hunting every day (and catch many stags). A new problem emerges: stag hunts offer the possibility of mutually profiting from an interaction; but how will we avoid conflict over resources that aren't divisible? It is in fact clear that the world doesn't only offer "win-win" (albeit risky) situations, like stags. It offers equally many profit opportunities that can only be consumed in isolation. However, this may pave the road to conflict. For instance, if we live together and a particularly attractive mate walks by, we'll be more likely to engage competition (in monogamous cultures at least, mating privileges cannot be shared). Similarly, we know we can't all talk at the same time during an argument, or no one will understand anything (attentional and perceptual resources also have a cap and risk to be over-crowded); we also know that if we all take too much water from the same well, or take all our cattle to graze on the same field, we'll finish both water and the grass (which are common dilemmas in the economic literature); that if we all take our preferred route

home (i.e. the freeway) at same time there will be a traffic jam. In all such cases, sociality thus leads us to 2 possible inefficiencies: either we raise our voice, honk our horns, shove our way to the field or well, hoping others don't do the same, or we wait in line. In economies, this known to happen (i.e. Ochs, 1999) when agents would like to enter a particular market. They know that if too many of them do (and everyone starts producing the same brand of the same product) there will be a price-war and all investors lose. For this reason, such situations have been called "entry games".

Clearly, all of the situations described above (both, stag hunts and entry games) could be detailed in so many ways that could matter (i.e. can agents talk things over? Is it a one-shot or a repeated interaction? What do they know about how much they each value the same goods? Etc.). However, they all have at least 1 thing in common, that is, they all involve multiple agents and incentives. This means we can, as a preliminary simplification, focus only on these and see what we can say. When we do so, the interactions start to look like games. Below (Fig. 1), we show such a simplified representation of both of the situations described above (in typical game-theoretic notation matrix – see caption).

STAG HUNT	**RISK (stag)**	**SAFE (hare)**
RISK (stag)	1 , 1	0 , SP
SAFE (hare)	SP , 0	SP , SP

ENTRY GAME	**RISK (enter)**	**SAFE (wait)**
RISK (enter)	0 , 0	1 , SP

SAFE (wait)	SP , 1	SP , SP

Fig. 1. Stag hunt games (SHs) and entry games (EGs) in matrix notation: 2 agents (one choosing a column, the second a row) make simultaneous choices without communicating. They choose between the same pair of options: a safe payoff (SP) that can be obtained with certainty and in isolation, and a potentially higher paying but risky one (here worth 1, with 0<SP<1), which depends on the choices of others. For each combination of choices the payoff to each of the players is shown, the payoff on the left refers to the player choosing between rows, the one on the right to column player's payoff. In SHs agents would prefer to match their choices. In EGs, they would prefer to decouple them.

Very ideally, we have represented a simplified version of the 2 halves of what we called the sociality problem. Now, we can turn back to our question ("safe isolation or risky interdependence?") and ask what would be needed in order to make the choice of sociality dominant relative to the choice of safe isolation. For social cohesion to be an optimal (evolutionary) solution, a (neuro-cognitive) mechanism should realize 2 objectives: it should facilitate the successful exploitation of situations requiring joint effort (i.e. cooperating to catch stags) - thus earning an advantage over safe isolation - while simultaneously optimizing traffic over limited common resources. Within an economics and game -theoretic framework, understanding what guides choices in either of these situations - even when much simplified relative to their real-word analogues - has resulted to be extremely challenging. Both of the depicted scenarios can in fact be represented by *coordination games*, which have been said to "constitute the most difficult problem of game theory" (Camerer, 2003), as they apparently involve a difficult matching problem. Our general proposal, which I will try to articulate throughout this thesis, is that a correspondence between social network closeness, psychological closeness and neural closeness, could afford precisely this, that is, propensity for cooperation, and aversion to

conflict.

Chapter 1 Behavioral Game Theory and Common Knowledge

1.1. On deduction and incentives

Our focus is on the *multi-agent decision problems*, that is, on decisions that depend on the decisions of others. In other terms, we focus on strategic interactions. Our departing point is the work from Von Morgenstern and Neumann (1947), on the one hand, and John Nash (1950) on the other. The former founded expected utility theory, which provided a mathematical framework for prescribing economically "rational" choices. To do so, the theory proposed a formal tool to predict the behavior of "toy" agents. Such agents followed simple rules: they need things (that is to say, they have preferences, i.e. food/water, occasions to reproduce, money etc.) and they will always make the choice that maximizes their chances of obtaining them, given the information/beliefs they have. Typically, to mimic the fact that, in the real world, events are uncertain, such agents were viewed as always choosing between lotteries. For instance, a "rational" agent, as depicted above, should prefer a sure payoff of 60 to a 50-50 bet of 100 or nothing, because the expected value of accepting the bet it 50, which is less than 60. However, how should such agents behave when the outcomes of their choices depended not on lotteries, but on the choices of other agents? The solution was a logical follow-up: agents expect other agents to behave exactly like they would, that is, "rationally"; they also know that everyone knows this, and that everyone knows that everyone knows etc. This recursion in forming beliefs about beliefs (about beliefs…) is the peculiarity of strategic interactions and it constitutes the critical additional assumption in the passage from individual decision making to inter-individual/strategic decision problems. Game theorists call this the

assumption of "common knowledge of rationality". Furthermore, as the theory was mostly intended to give advice in economic decisions, which were often competitive (i.e. finding the "right" compromise in bargaining), agents were also assumed to be strictly self-interested.

Given such assumptions, John Nash completed a demonstration that, as long as agents have a finite set of options, then *all thinkable* interactions have an intriguing mathematical feature called a fixed-point, or, as we will refer to it from now on, a Nash equilibrium (NE). That is to say that, out of all the combinations of possible choices by players, there is only a subset of them (which can be mathematically derived) in which all players are choosing their best option (they are maximizing their utility/payoff), given the choices of all the others. In such situations, since no agent has an incentive to unilaterally deviate from his current choice, everyone is in equilibrium. Since then, a fundamental question of the whole enterprise of behavioral game theory and experimental economics (Camerer, 2003) has been whether such a mathematical model of deductive rationality had any empirical bite, that is, whether it could yield valid predictions of the outcomes of strategic interactions. Intriguingly, in many instances it does.

One example I find compelling is the following game, first alluded to by Keynes (General Theory of Employment Interest and Money, 1936), then formalized and empirically tested by Nagel (1995). To describe the type of reasoning agents in the stock market often go through, Keynes spoke of a hypothetical newspaper contest, in which readers were to try and guess which of the 6 depicted faces, was the most beautiful. Of this he said:

"It is not a case of choosing those [faces] that, to the best of one's judgment, are really the

prettiest, nor even those that average opinion genuinely thinks the prettiest. We have reached the third degree where we devote our intelligences to anticipating what average opinion expects the average opinion to be. And there are some, I believe, who practice the fourth, fifth and higher degrees." (Keynes, General Theory of Employment Interest and Money, 1936).

In Nagel's version (1995) of the "beauty contest" *n* agents simultaneously pick a number x_i, with $0 \leq x_i \leq 100$. Their objective is to choose the number that comes closest to a target number. This target number is equal to the average of all chosen numbers (hence the guessing what others choose), multiplied by a parameter k (i.e. ½). The winner obtains a dollar amount X, the others earn 0 (and X is split in case of ties). When k<1, this game has a single NE, which is 0. To understand why, consider the game with k=1/2. The highest possible target number for this game is 50, which would only occur if everyone chose 100. So, any choice above 50 is "strictly dominated" by any choice below it, that is, it is pays off less no matter what numbers the other players choose. Critically however, if others get this (and since they are assumed to be rational, they will), they'll be choosing numbers of 50 or below, which lowers the mean and thus the target number, to at least 25 or lower. The same reasoning is iterated until it is impossible to do so further, that is, at 0. This game shows many interesting characteristics in action. One of them is that deductive rationality may not always be an optimal strategy. Here for instance, agents choosing 0 can lose (as Camerer puts it, "they're smart and poor"), if the other players are choosing higher numbers. However, when this game is repeated, and players are told the outcome of each successive round, they quickly, as a group, lower their choices, until almost everyone is choosing 0 (Fig. 2).

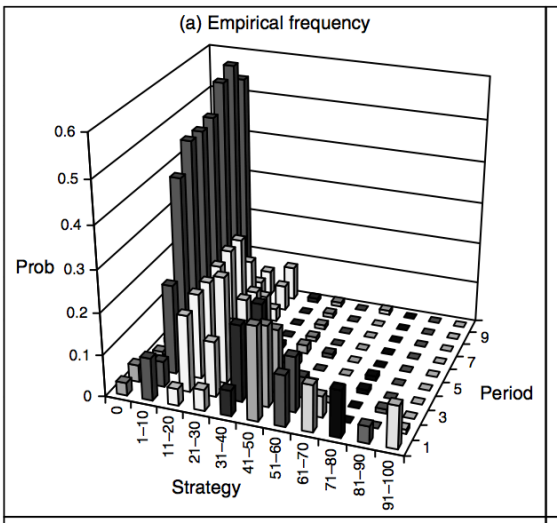

Fig. 2. Frequency of choices in a beauty contest (with parameter k=2/3): in period 1, choices are rather randomly dispersed, by period 9, nearly everyone is choosing the NE.

It follows that, while in the short run a game-theoretic approach may not always do too well (not in this game at least) – though, even in the 1st round, some agents are already selecting the theoretic solution of the game -, in the longer run, it seems to capture a fundamental aspect the game, namely, the *direction* of its unraveling: indeed (at least in beauty contests) eventually, subjects do tend to equilibrium[3]. The interesting thing is that incentives may be doing much of the work: though some subjects immediately play at equilibrium, or may learn to do so through introspection (Weber, 2003), not all subject need to "be rational to behave rationally". For instance, they needn't all learn to reason in greater depth – thinking about what others think they think etc. - (indeed, Nagel provides some evidence that they don't); rather, it is sufficient that they be sensible to gains and losses – as in reinforcement learning – to notice that high numbers don't pay, and thus

[3] Whereas critics have often stressed failures or "rationality" in one-shot decisions or interactions, I find that its potential pragmatic/predictive appeal is stronger under repeated conditions, that is, in the long run. Camerer says, "in the modern view, equilibrium should be thought of as the limiting outcome of an unspecified learning or evolutionary process that unfolds over time" (Camerer et al., 2001). Apparently Nash had similar ideas: "In his thesis proposing a concept of equilibrium, Nash himself suggested that equilibrium might arise from some 'mass action' that adapted over time".

gradually reunite with the "higher level" thinkers in equilibrium (though, for a discussion of this see Erev&Roth, 1998). As Aumann puts it:

"One of the simplest, yet most fundamental ideas in bounded rationality - indeed, in game theory as a whole - is that no rationality at all is required to arrive at a Nash equilibrium; insects and even flowers can and do arrive at Nash equilibria, perhaps more reliably than human beings." (Aumann, 1997)

For instance, in a matching pennies game, 2 agents have to decide between 2 options, say "heads" or "tails". One agent gets paid if both player match their choices, the other if they mismatch. In such a competitive (0-sum) game any "pure strategy" (i.e. "always choose heads", or "always choose tails") can be exploited by one's opponent. Indeed, in this game, there is no equilibrium in pure strategies. There is however an equilibrium in *mixed strategies* which dictates that both agents should mix between options with a given probability, specifically, with probability, p=0.5, which is roughly what occurs. In line with Aumann's words however, this doesn't require rationality, but only that agents adapt their choices. Incidentally, non-human primates have been shown to behave closer to Nash in a subset of interactions, such as "matching pennies games" (Martin et al., 2012) and Ultimatum Games (Jensen et al., 2007; Sanfey et al., 2003).

Thus, in some empirical cases, incentives and deductive rationality seem to work synergistically in constraining choices towards equilibrium: where some agents may "understand" the solution by introspection, and adopt it, the others should, as interactions unravel, get "hammered into it" by the payoff structure itself. Later on (in the discussion on entry games), we will also see how game theoretic predictions can be accurate even in

the short run for some games, in which subjects reach equilibration, immediately, without communication, or feedback. "To a psychologist", Daniel Kahneman said, "it looks like magic" (Kahneman, 1988).

1.2. The failures of deduction

In many instances, however, game-theoretic predictions and deductive rationality are grossly off. One particularly strident example is the "centipede game" (Rosenthal, 1981) (Fig 3). Here, 2 players have in front of them 2 unequal piles of dollars; lets say, $4.00 in one pile, and a single $1.00 bill in the other. They then take turns choosing whether to "stop" or "continue". Player 1 starts. If he stops he keeps the bigger pile, while player 2 gets the smaller pile. If he continues, the 2 piles are doubled and passed to player 2 (who thus receives 2 piles of $8.00 and $2.00) who is in turn to decide whether stop (and keep the bigger pile), or pass back to 1 (getting both piles doubled again). Whoever stops gets the bigger pile, which is however doubled at each continuation. If at round 6, the last round, player 2 continues, then player 1 obtains $256.00 and player 2 gets $64.00.

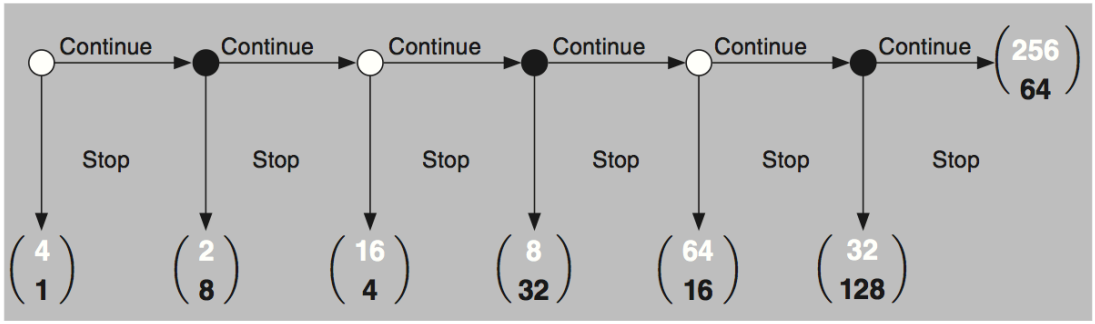

Fig 3. The centipede game (Rosenthal, 1981)

The paradox of this game is that both players should prefer the last-period outcome to the 1st period one. However, game theory sais that if they're rational, they'll never get to the last round, because the unique NE of this game is for the 1st player to stop in the 1st round.

The reasoning is that in sequential games like this one - in which agents take turns in making their choices, rather than choosing simultaneously - game theoretic analysis prescribes player 1 to use "backward induction" (Gibbons, 1992) and start examining the game by its final stage, when player 2 is deciding whether to pass or continue. Player 1 should in fact notice that, in his last decision, player 2 would essentially be choosing between stopping, and getting $128.00 or continuing and getting $64.00. So, if player 2 is rational, he will never continue in the last stage. Thus, knowing that the last stage is "lost", player 1 should stop in the 2nd to last stage, to earn $64.00, rather than $32.00 (his second-best option). However, again, if player 1 thinks player 2 is rational, then player 2 should anticipate that player 1 will stop in the 2nd to last stage and should thus not give him the possibility to do so, by stopping at the 3rd to last stage. This correct reasoning unravels backwards towards an apparently incorrect ending, where player 1 stops at the 1st turn and earns $4.00. Such a result is so strongly counterintuitive that for many players, Aumann said (1992) "if this is rationality, they want none of it". Indeed, though learning has an effect in this game as well, NE-play is still rarely observed (Palacios-Huerta&Volij, 2008).

This paradox has led to a tremendous amount of work, and several accounts have emerged as to why these sorts of inconsistencies emerge. We talk about them here, not so much for this specific game – since, usually, if any "refinement" has received attention it is because it accommodates deviations from NE in many games - rather, to show what type of adjustments seem necessary, if we want the "toy" agents of EUT and game theory, to look a little more like us.

One of the largest classes of such explanations involve "social preferences" (i.e.

Fehr&Camerer, 2007). What such models have in common is that they relax the game-theoretic assumption of self-interest: players are no longer only motivated by their own payoffs, but they can "care" about others (i.e. about others' payoffs); where such "caring" can be declined in many different ways: a competitive subject will be happy to see a counterpart lose, an altruistic one, to see another win, others still could be sensitive to the ratio, the difference or other relations between payoffs (i.e. they make choices that maximize fairness) (Fehr&Schmidt, 1999). In all such models, players remain for the rest "rational", that is, they keep maximizing the utility, however, what has changed is that their payoffs become interdependent. In the centipede game for instance, the mere possibility that altruists exist can change the structure of the game, and its NE (McKelvey& Palfrey, 1992). Since an altruist would place a positive weight on my own payoff, he could choose to "continue" just to benefit me. In this case, a rational self-interested player could continue in turn, proportionally to his estimated probability (his belief) of having to do with an altruist. If both players were mutually altruistic, this

On the other hand, very different types of models argue that social-preferences may have nothing to do with the paradox, which may instead emerge from the fact that game theory implicitly assumes its agents to have unlimited computational abilities, which is generally not true for humans (i.e. Simon, 1957). Indeed, games like the centipede game, which involve backward induction, and games like the beauty contest, involving iterated beliefs, can undoubtedly be computationally demanding (i.e. in terms of working memory). Indeed, mouse-lab studies have shown that agents that fail to converge to NE often don't apprehend the necessary information to do the required backward induction (i.e. they don't even look at what happens in hypothetical stages of the game – such as the last stage of centipede games) (Johnson et al., 2002; Carrillo et al.,2008). Recent models of NE

("quantum response equilibria") relax the assumption of unlimited cognitive resources by allowing agents to make small mistakes in their decision process (McKelvey&Palfrey, 1995). Importantly, such mistakes are not random, but they will be attracted towards choices that can yield a higher payoff, as is choosing to "continue" in centipede games. Indeed, such models do predict that the probability of agents "stopping" in centipede games increases, as the game progressively approaches its end (McKelvey&Palfrey, 1998), and they also yield better predictions in a number of other cases in which behavior apparently departed from NE. Critically, such models remain equilibrium models, they simply account for the fact that approach towards NE is probabilistic, rather than deterministic[4].

Finally, an intriguing literature on the "epistemic conditions for NE" (Aumann&Brandenburger, 1995) calls into question what it means to have common knowledge of rationality, thus what players know about each other, their intentions and beliefs. In contrast to the idea out-of-equilibrium choices in centipede games are due to subjects hitting the roof of their cognitive limitations, such models suggest that an important role could be played by the uncertainty that subjects may have about the "limitations" of others. An interesting study by Palacios-Huerta&Volij (2009) showed that agents that have no problems with backwards induction, such as chess players, choose the NE in centipede games 70% of the time when playing with other chess players, and by the 5th repetition of the game, they have learned to pick it always. If the players were grand masters, they chose it 100% of the time on their first shot. These percentages however go down when chess players played with students. On the other hand, when college students

[4] Camerer (2003) says that if Nash had been a statistician, rather than a mathematician, he would have invented quantal response equilibrium.

play against other college students (the typical pool of behavioral economics studies), NE play was as low as 3%, with no sign of convergence over repetition. The interesting thing however is that such a percentage was multiplied 3-fold - that is students picked the NE immediately 30% of the time – when students played against chess players, and 70% of the time when they had a chance to learn. This is a clear indication that knowledge on the rationality of others can strongly affect strategizing and NE-play in interactions. Though human abilities are limited, and these certainly play a role in many "irrational" behaviors, it may be the case that, more often than we think, humans seek to adapt their "rationality" and strategic behavior to what they perceive of the rationality of others. In a centipede game, if I believe my counterpart is rational, I could still have the doubt that he thinks I'm not. In which case, my counterpart could have the temptation to "continue", which gives me some incentive to "continue" as well. It follows that, on top of talent or practice, an important factor in inducing the recursive thinking that unravels towards Nash, as Palacios-Huerta&Volij (2009) put it - and Aumann&Brandenburger (1995) formalize - is neither that you are rational, nor that I am, and not even that we are both rational, but common knowledge of rationality is the key.

In the examples we showed so far however we focused on games with a specific characteristic: they have a unique NE. This gives the "common knowledge" assumption a specifically deductive connotation (i.e. students and chess players in the centipede game were interested in how "smart" their counterparts were likely to be); that is, the only way subjects had to infer the choices of others was to believe they were rational and, so to speak, deduction was the only "common language". However, what happens when deduction seems to lead nowhere, how are agents supposed to infer beliefs then? This problem emerges in games with multiple equilibria, in which, with no deductive common

knowledge available, agents must find some other common form of knowledge that may allow them to coordinate their choices and infer their beliefs.

1.3. Coordination games

For the games we described so far, we said that NE does prescribe a unique solution, though agents may or may not "get it" (i.e. because they have limited computational abilities), or they may think others don't get it (or that others think they won't get it). In games with multiple equilibria it's often the contrary: deductive rationality says nothing, and, if anything, humans are often able to tacitly *coordinate* their choices. The 2 games we opened with, SHs and EGs, are games of this type. So to answer our question about social cohesion, we must understand what mechanisms can guide coordination. In line with the assumption of game theory, we will argue that common knowledge is critical for this, and we will begin to highlight the apparently intricate connections between common knowledge and similarity.

1.3.1. Coordination and intuition

Coordination in its purest, and the clear failure of deduction, can be demonstrated by "pure matching games" (Schelling, 1960). Consider a game in which 2 agents choose between "heads" and "tails". If they match they win (i.e. $100), if they don't, they lose ($100). This game has 2 equilibria in pure strategies: if both players choose A, neither has incentive to deviate; the same holds if both choose B (and there's also a mixed strategy equilibria in which both players randomize with probability p=1/2). That's all economic deduction can say, since the payoffs alone do not make any distinction between strategies. However, the labels do, and subjects have been shown, to pick up on this very easily, by

choosing “heads” 87% of the time (Mehta et al., 1994), by choosing “Everest” among the set of mountains and Ford among the set of cars. The failure of deductive rationality in such games is clear, as Schelling pointed out:

“One cannot, without empirical evidence, deduce what understandings can be perceived in a nonzero-sum game of maneuver any more than one can prove, by purely formal deduction, that a particular joke is bound to be funny" (Schelling, 1960, p. 140).

What is remarkable of such games, is the flexibility and resilience with which humans find novel dimensions of stimuli to use as coordination devices. For instance, in the example above, suppose we wanted to try and *break* coordination by calling the labels “heads” and “heads”, rather than “heads” and “tails”. Now not even the labels differentiate between strategies. Yet, subjects will still be able to coordinate on the top-left strategy, because “upper-leftness” now appears psychologically prominent – at least in western cultures that read from left to right. One could try further, by removing even the graphical display and presenting the matching problem orally (and with the same names for both strategies). Still, I think subjects would coordinate on the 1st of the 2 proposed options, because temporal antecedence is salient (the “logic” of “finders keepers”). Similarly, a game requiring to match numbers, out of the set of numbers has infinite NE, however the number “1” was chosen 29% of the times. Schelling calls these characteristics “focal” (1960), where the ability to establish focality is clearly far from the type of deduction we spoke of in the games with a single NE seen above.

1.3.2. Coordination and common knowledge

However, focality is fundamentally different from our ability to establish salience. It is about understanding what is likely to be *mutually understood* as salient. A rather clear example: a matching game in which agents are to coordinate their choices by choosing a common year has also infinite NE. However, when this experiment was ran, in 1990 (Mehta et al., 1994), 61% of subjects chose 1990. Indeed, the present is another strong element of focality, in that it is shared by all. However, when the same subjects were asked to simply pick a year, without the objective of coordinating, 43 different years were chosen: 1971 was picked 8% of the time, plausibly since most of the subjects in that pool were born that year, and the percentage of people picking 1990 dropped to 6.8%. What this says is that salience is personal, focality is shared, and it is focality that drives coordination: it isn't what subjects personally prefer, but what they think they could all *agree* on preferring. Again, it follows that common knowledge is key, however, in contrast with games with a single NE, it cannot be based on rationality in strictly deductive sense. *The ability to coordinate relies on the extent in which we believe others to perceive the world as we do*, that is, in the extent to which there is common knowledge.

For instance, people coordinate on "heads" rather than tails, but is it intrinsically more rational than tails? Clearly not. A different cultural system, with a different established convention of expressing the notion of "heads or tails" (i.e. "tails or heads"), could have as successfully coordinated on tails. Furthermore, were 2 agents of 2 such different (hypothetical) cultures to interact, we could predict that they would miscoordinate, and conflict could emerge. This was studied by Weber&Camerer (2003). In their experiment subjects looked at a photograph depicting many objects. One subject, the speaker, had a list of such objects and had to tell the other to point at the corresponding objects in the picture with a penalty for being slow. At the beginning of the experiment it could take the

speaker as long as 30s to direct attention towards a man gesticulating with his hands in front of the desk. By the end of the experiment, the speaker readily said "Macarena" and this was immediately understood. When subjects from different pairs were mixed, response times shot back up again, because speakers kept using their previously established conventions with listeners who had developed different ones. Both agents of the new dysfunctional firm also blamed one another, rather than the contingent difficulties of "inter-cultural" mediation. Though some conventions may seem smart, in coordination games, it isn't deductive rationality that leads to agreement, it is agreement that makes outcomes rational. Camerer appropriately cites Pascal, "Why do we follow old laws and old opinions? Because they are better? No, but they are unique, and remove sources of diversity".

1.3.3. Games with Pareto-ranked equilibria: an efficiency problem

This doesn't seem to be exactly the case for games with pareto-ranked equilibria (see below), such as SHs. Agreeing on a convention would be uncontroversial if agents are completely indifferent between the convention alternatives. However, in games with pareto-ranked equilibria the alternatives left open by deduction *are* different: in the SH case, one of the 2 equilibria is (Pareto-) efficient, as both players earn more, the other is not[5]. In spite of this, coordination failure was found to be extremely frequent[6]. This was first observed by Cooper et al. 1990. Below (Fig. 4) we show the matrix of the group's

[5] When I was first told the SH payoffs, I found it completely trivial: I was going to go for the high payoff, and I was virtually 100% sure that my counterpart would have done the same. I clearly noticed that, on paper, the higher payoff was also risky, but I held it to be a risk under our complete control.

[6] By coordination failures, accepting the suggestion of Devetag&Ortmann, we intend failures to coordinate on the pareto-efficient outcome, though it could also mean failure to coordinate on any of the 2 equilibria.

replication in 1992, the column/row labels are here modified to align with the stag/risk story[7].

STAG HUNT	**Stag (risk)**	**Hare (safe)**
Stag	**1000, 1000**	0, 800
Hare	800, 0	**800, 800**

Fig. 4. Payoffs of one of the first incentive-compatible laboratory stag hunt games conducted by Cooper et al. (1990, 1992). Points were probability points, with 1000 points leading to $1.00 with certainty. The game has 2 equilibria in pure strategies: if both players are choosing to risk, neither has an incentive to deviate, as occurs if they both choose to stay safe, thus the strategy profiles (that is, the combinations of choices by all players) {stag, stag} and {hare, hare} are the 2 pure-strategy NE of the game. There is also a very counterintuitive mixed strategy equilibria which we don't talk about here (and which doesn't seem to work at all).

In that study (Cooper et al., 1992), when collapsing over the last 11 periods of the game, 97% of subjects (out of 165) chose the secure option. Nearly simultaneously, Van Huyck et al. (1990, 1991) showed that similar inefficient outcomes occurred when the multiple-equilibria were more than 2, such as in "weak-link" or "median" games. This was unsettling, as games with pareto-ranked equilibria appeared to have a 2-fold problem: not only game theory appeared unequipped to analyze them (like for the other coordination games) (though, for a notable exception, see the global games framework, Carlssen&Van Damme, 1993); but outcomes were also inefficient (and rather unsocial). Indeed, such

[7] Here, we call the available options with the name of the stag story, though we will often call the stag option the "risky" one, and the hare option the "safe" one. In other experiments, the stag option is often treated as "effort" levels. This terminology is incorrect in economics, for which the term "risk" uniquely indicates forms of uncertainty with known probabilities. Harsanyi&Selten (1988) do speak of risk in the context of SHs, but their definition regards the equilibria, not the strategies. The stag option should be called the Pareto-dominant, Pareto-superior or efficient equilibrium. However, for non-economists, I believe the term "risk" is more appealing to describe the tension between safety and efficiency which is typically observed in order-statistic games.

findings sparked a wave of follow-up studies (reviewed in Devetag&Ortman), the majority of which probed the reasons for such inefficiency.

A first important idea seemed to justify indifference between conventions. Indeed, an important realization is that inefficiency doesn't always occur: one can easily imagine that as the SP value approaches 0, agents should be more and more willing to risk. Harsanyi&Selten (1988) formalized this concept and suggested the existence of 2 criteria for equilibrium selection in SHs: payoff dominance and risk dominance. In SHs, the {stag, stag} outcome always constitutes the payoff dominant equilibrium, while the risk-dominant equilibria depends on the value of the hare, that is, of SP: if the value of risking is sufficiently higher than the value of security, then both criteria coincide, and the theory predicts that subjects should choose the efficient outcome; if, on the other hand, one can be only marginally profit by risking, then, faced with the failure of deduction, agents may choose the risk-dominant SP. For our symmetric SH games (Fig. 1), where the stag was worth 1, risk-dominance predicts that agents should risk when SP<1/2, mix strategies when SP=1/2, and stop risking for values of SP>1/2, which is (only very) roughly what is observed (i.e. Rankin et al., 2000). This meant that, in some cases, agents could behave *as if* they were playing against a chance[8]. It remained disturbing that they should be unable to behaviorally agree on an outcome that is favorable to both and that they tended to choose security when they had to expect from one another anything more than random behavior (that is when SP>1/2).

Rankin, Huyck&Battalio (2000) made an interesting alternative suggestion: perhaps the

[8] As Aumann&Dreze (2004) put it, "… games against nature and strategic games are in principle quite similar, and can –perhaps should– be treated similarly".

problem had to do with the fact that players were "getting stuck" in consistent but unfavorable conventions. Indeed, a feature that seems to stick out in SHs is that, whereas learning often leads to increased "rationality" in some games with unique NE (as we saw earlier for beauty contexts, and to some extent for the centipede game), it seemed to do the contrary in SHs: it appeared to drive towards inefficiency. While the majority of subjects did not attempt to coordinate on the secure options in early trials, during the last ones, safety appeared by far to be the prominent strategy (Battalio, 1997). Rankin et al. had the intriguing intuition that this was due to the fact that miscoordination on early trials may have been anchored to non-strategic details of the game presentation, such as specific payoffs or strategy labels. They thus attempted to "perturb learning" by pseudo-randomly changing SH payoffs and scrambling labels, to see what convention would emerge when the only commonality between games was their "strategic similarity", that is, the fact that they all involved a choice between security and efficiency. To allow for different possible culture-specific conventions to emerge, subjects played in segregated groups of 8, and were randomly re-matched, within their group, at each round. Group-conventions did emerge, and they all tended towards efficiency: even when the SP was larger than ½, subjects kept cooperating. The main difference between this SH setting and others appears to be that here there were many games with rather low SPs. Its as if subjects had to learn how to rely on one another in situations in which coordination failure was less costly (i.e. learning how to swim in shallow water). Then, seeing that they were able to do so successfully created a common history or convention for cooperation, which was then more easily extendable to games with greater risk. Similar results were shown by Brandts&Cooper (2004). They showed that groups that had got "stuck" in inefficient equilibria/conventions could get "unstuck" by a sudden increase in the incentive to cooperate. Once such incentives were subsequently decreased, changing back

to the way they were, agents didn't change back with them. All this suggests that the tradeoff between social efficiency and security is indeed fragile, but that agents do not seem indifferent between these 2 conventions, they also all seem to prefer the same one. However, though they seem to require some common history of successes in order to realize this.

1.3.4. SHs and communication

The strongest suggestion that there is common knowledge on whether the efficient or secure convention is preferable in SHs is that agents quickly agree when they are allowed to communicate, as shown by Cooper et al. (1992). Indeed, in their experiment, relative to the normal/tacit SH, in which successful coordination *never* occurred, when 1-way communication was allowed (so that 1 player could announce his/her announce decision to the other), 53% of the agents reached efficient coordination. However, when 2-way communication was allowed, coordination was almost at roof, at 91%. SHs are very particular in this sense.

For instance, consider a close relative of the SH also involving potential cooperation: the notorious prisoner's dilemma (PD)[9].

Prisoner's Dilemma	**Cooperate**	**Defect**

[9] Like the SH, PDs have a story: 2 suspects are being held in custody for interrogation in separate rooms and they both have 2 options. If they cooperate by not confessing they get 3 years of prison each (because of insufficient proof). If they both confess they get 4 years. If however, one confesses and the other doesn't the one who does gets a "get out jail free card" (0 years of prison). If we now change the payoffs from losses (time in jail) to gains (i.e. by simply adding 10 to all the payoffs), we can preserve the structure of the PD, and confront it to the SH.

Cooperate	7 , 7	0 , 10
Defect	10 , 0	**6 , 6**

Stag Hunt	**Cooperate**	**Defect**
Cooperate	**7 , 7**	0 , 6
Defect	6 , 0	**6 , 6**

Battle of the Sexes	**Box**	**Ballet**
Box	**10 , 6**	0 , 0
Ballet	0 , 0	**6 , 10**

The critical difference is that, in contrast to SHs, PDs have a unique NE (in pure strategies), which is to defect (defecting strictly dominates cooperating, because it pays off more whatever one's counterpart does). The apparently modest change in payoffs can have an effect on communication because to the extent that interactions are competitive they can generate an incentive to lie (i.e. bluffing in poker or price shading). Indeed, though communication does have a positive effect in PDs, meta-analytic reviews suggest that the increase in cooperation is around 40% (Sally, 1995; though see Balliet 2009 for moderating effects), far less than in SHs. Indeed, communication seems to even add some drama to the game and a number of TV game shows – i.e. "golden balls" – had participants talk their choices over for large monetary stakes in what were actually PD variants. It was probably considered entertaining to watch many participants solemnly give their word that they would cooperate, only to subsequently defect, either alone or together. This is also suggested by the different subjective impact lies have once they are revealed: in SHs,

if one announces to cooperate and then doesn't, he/she might simply be viewed as obtuse or excessively fearful. Lying in PDs on the other hand could be seen as the most manipulative and spiteful betrayal, since agents benefit from convincing lies. [10] For similar reasons economists call non-binding communication "cheap-talk". The reason why cheap-talk seems to work so well in SHs, is that SHs are the "building blocks of strategic complementarities" (Camerer, 2003): situations in which increased effort/risk/action on the part of one party generates incentive for another party to do the same (I.e. a company increasing production and sales in cars, generates a another company to increase production of fuel), with no incentive to defect. Indeed, by introducing "half-way" differences in agents' goals, the effects of communication change as well. In the battle of the sexes game (BOS), 2 agents, John and Mary, would both like to go out together, and this is their main objective. However, (to invert the standard gender-stereotype) John would prefer to go see a ballet show, while Mary prefers the boxing match. Like for the SH, BOS has 2 equilibria in pure strategies (and 1 mixed strategy NE) in which both agents either choose "box" or both choose "ballet". Thus, agents have a common set of goals. However, critically, they disagree on which of the two should be chosen. Uncertainty in such situations is very high, as one should concur by trying to make a "rational" choice in the payoff matrix shown above (Fig.). In line with this, choice percentages observed by Cooper et al., 1994 were close to the mixed equilibria of the game: 59% chance of miscoordination. Like for the SHs, Cooper et al. confronted the effects of 1-way and 2-way communication on the BOSs. The revealed pattern was different: like for SHs, 1-way

[10] Interestingly however, the same reverberant fear of SHs, could easily be playing a role in PDs. For instance, maybe agent are not lying when they profess their intentions to cooperate but they get scared at the last moment, thinking that others may have lied, or, because they think the other may doubt their trustworthiness at the last minute. Indeed, there are many commonalities between SHs and PDs. For instance, Camerer (2003) notes that when PDs are repeated under certain conditions (with high enough discounting or altruism parameters) they mathematically become SHs.

communication facilitated coordination, agents tended to announce their preferred equilibria (i.e. the John-players announced they would be going to ballet) and their counterparts (the "Marys") tended to accommodate, yielding a coordination rate of 95%. Interestingly however, in contrast to SHs, adding 2-way communication didn't further better coordination rates, instead, it brought them back down to roughly where it was before (Cooper et al. 1992,). It seems that 1-way communication worked as a "tie-breaker" (Camerer, 2003) offering some way to coordinate choices, however 2-way communication simply re-instantiated the conflict. In few games, the cumulative effect of common knowledge, such as in 1-way and 2-way communication, reduces uncertainty as in SHs.

Often, people tend to not think much of uncertainty in strategic interactions because they think, "well, that's what language is for": its easy to know/predict the intentions of others because we talk about them all the time. We already saw that 1 fundamental flaw of this argument is that it doesn't hold in competitive situations, where communication and mutually increased predictability (i.e. common knowledge) seem to lead to an increase rather than an increase in uncertainty. A second flaw is that there are so many situations in which agents logistically cannot all talk to one another (indeed, all large scale market interactions are of this type). However, the counter-argument that I find most intriguing, is that communication too seems to require common knowledge. The question here becomes: how can we be fully sure of what others understand of our utterances, unless we admit some common back ground knowledge. Think of teaching: if conveying knowledge was simply about communicating, then teaching would always be snap, while it is not. Indeed, this is plausibly why its success is evaluated by counter-interrogation of the teacher to the student (i.e. tests). Indeed, it appears that 2-way and not 1-way communication is necessary to establish successful communication; if doubts still remain,

more communication and "testing" may be necessary, and we may well have to go back and forth several times, to *make sure* we mean the same thing by a given proposition. This back and forth assuring however is a burden in terms of time. Non-linguistic "mind-reading" on the other hand would be costless, and, in a sense, there seems to be nothing mysterious about it: it could easily occur to the extent to which we assume that others know what we do. For instance, a doctor would probably take much less to explain a medical problem to another doctor than to a non-doctor (i.e. a patient). This "making sure" is what many feel is the crux of SHs, which also called the "assurance game" (Camerer, 2003). On one hand, the SH payoff matrix (that is, the mere incentives of the game) would seem to speak for itself, it could be thus be held to parallel the utterance, however others may not interpret the matrix/utterance in the exact same way we do. In strategic interactions, this seed of potential diversity, may then generate what Hofstadter (1985) calls "reverberant doubt" and mutual suspicion, which we opened with; and this, we will argue, may be one of the reasons why we generally "like" similar others. I would next like to conclude this section on common knowledge and behavioral game theory with a last philosophical argument, which I believe makes the same point from the opposite perspective: not that similarity can favor communication, but that communication cannot emerge from complete diversity.

In what has been said (Wright, 1999) to be one of the most discussed arguments of this century, Quine made the following example: suppose we were confronted with an unknown utterance by an indigenous speaker, who, pointing at a rabbit, said "Gavagai". We would be strongly tempted to conclude that what the speaker means by "gavagai" is what we mean by "rabbit", and be happy with the idea that we've obtained a successful translation, a first step towards communication and perhaps future cooperation. However,

the meanings of words are derived from their context, and here there is no such mutually accepted context. Thus, given the evidence at hand, we could equally suppose that the native means "food", "lets go hunting", "a rabbit ear", "it'll rain tonight" (if he's superstitious), "look, the ground" (if he doesn't care about the rabbit at all) and, under complete diversity, the utterance could virtually mean anything. However, to the degree in which we think the stranger perceptually divides the world in similar chunks as ourselves (i.e. making a division between rabbit and background more likely between particular portions of the rabbit's ear), to the degree in which we believe that he, like us, could consider the rabbit as a possible form of sustenance, in short, to the degree in which we assume similarity and some fundamental likemindedness, we can proportionally restrict the possible ways in which to translate the word "Gavagai", excluding at least, the wildest ones. Indeed, Quine's problem appears very similar to the one of multiple NE in games, as Gibson puts it, the problem for Quine "is not that successful translation is impossible, but that it is multiply possible". Our idea is that, to the degree in which we can refer the beliefs of others to our own system of beliefs, we can restrict the amount of multiple possibilities, and thus reduce our uncertainty. In brief, my take on Quine's story is that, if there is no pre-existing commonly accepted bundle of knowledge, then mutual understanding appears to be logically indeterminable.

1.4. Synthesis

We opened by posing a rather ominous question which we use as a general guideline: what type of (neuro-cognitive) mechanism could make cooperation efficient and competition difficult? We thus turned to a game theoretic analysis of strategic interactions, because of its formalized framework, and because it sometimes seems to

capture something of how people actually behave (or end up behaving in the long run).

The theory tells us that an assumption of common knowledge is fundamental when passing from individual to inter-individual or interdependent decision-making. It enables agents to recursively form beliefs (about beliefs about beliefs…) of what others will do and thus best-respond to their beliefs, which is the central aspect of strategizing. In other words, common knowledge is what seems to allow predictability. However, (at least in game theory) it does so rigorously through deduction.

In games with multiple NE this logic breaks down as deduction leaves open multiple viable alternatives. Rational agents must then rely on some other form of common knowledge in order to coordinate their choices. One obvious solution is to *create* such common knowledge, by founding arbitrary conventions. Perhaps as a result of such arbitrariness, we saw examples (i.e. Weber&Camerer, 2003) of how different conventions can emerge in different groups, seemingly on the basis of situational contingencies or "historical accidents". However, in other cases, groups seem to establish less-arbitrary conventions, in that they all seem to flow in the same direction. For instance, we saw that when shallow learning is perturbed (Rankin et al., 2000), most groups tend to coordinate on an efficient, rather than a secure equilibrium; similarly, that increasing and decreasing economic incentives to cooperate doesn't seem to affect cooperation in a symmetric way (Brandts&Cooper, 2004); and that certain characteristics (i.e. temporal antecedence, the present etc.) have a special property of focality, that is, they seem salient to all. In the first case, common knowledge of shared historical accidents seems to generate behavioral similarity and efficiency (as well as potential inter-cultural clashes), in the latter it seems that actual similarity may engender a sense of common knowledge. In both cases,

whatever the origin, similarity and common knowledge seem to be tightly bound.

Finally, we considered language as a potential coordination device. Indeed, language is meant precisely to "share" knowledge. Furthermore, we showed that verbal communication can potentially comply to our purported dual function of a "cohesion mechanism", in that it has been shown to drastically increase efficient (cooperative) coordination and to potentially muddle competition. However, with the example on teaching, we hinted at the suspicion that language may too require common knowledge (Cooper et al., 1992). I will review some empirical literature (this time from psychology and sociology), further strengthening the link between similarity and common knowledge.

Chapter 2 *Homophily*

Similarity has been said to be the fundamental laws of interpersonal attraction (Byrne, 1971): generally speaking, it is commonly accepted that, we like and tend to approach, interact and form ties with people who we feel similar to ourselves (Morry, 2007). However, a problem immediately emerges.

"Any event in the history of the organism is, in a sense, unique. Consequently, recognition, learning, and judgment presuppose an ability to categorize stimuli and classify situations by similarity. As Quine (1969) puts it: "There is nothing more basic to thought and language than our sense of similarity; our sorting of things into kinds". (Tversky&Gati, 1978).

The apparently "good" thing about similarity is that it is a very intuitive concept that can be applied to nearly all levels of perception and cognition. This is also its fundamental limitation, that is, that it seems to apply to everything. For instance, it appears that no 2 elements are so diverse, that no similarity at all could be found between them. Then, to make similarity of any possible use, some separate mechanism would seem necessary to restrict among the possible inputs to a putative similarity-computing device. However, if such a mechanism existed, then it remains unclear why the same mechanism couldn't sort out the similarity outputs as well, thus eliminating similarity all together. In their edition of articles, Sloman&Rips (1998) frame the problem as one regarding an apparent contraposition between 2 general explanations of cognition: those based on similarity (association) and those based on rules. Suspiciously similar contrapositions appear open in neuroscience - such as the one between the procedural vs. representational nature of

computations in the prefrontal cortex (Wood et al., 2003) – as well as classical debates in epistemology - such as the one between empiricism (i.e. Mach) and rationalism (i.e. Planck) (Fuller, 2005). I hope to be able to steer clear of such debates in this thesis, though I probably won't fully be able to do so. In the attempt, I start from a rather different track, involving similarity in attraction, rather than categorization.

2.1. Similarity, repetition and attraction in non-social domains

"People love those who are like themselves" (Aristotle)

"Similarity begets friendship" (Plato)

In classic reinforcement learning organisms "repeat" those behaviors that led to rewards (and "stop repeating" those that led to punishments). It seems likely then that they might tend to "invert the causal chain", not limiting themselves to repeating what is "good" but also taking what is repeated to *be* "good". Indeed, repeated exposure and reward seem to share an intimate connection. One possible declination of this phenomenon is perhaps captured in the mere exposition effect (Zajonc, 2001), which consists in the observation (we give an example below) that simple repeated exposure to previously neutral stimuli increases their perceived attractiveness. In lay words, we prefer the familiar to the unfamiliar. Similarity then might result rewarding because it logically implicates a form of repetition: i.e. to recognize that 2 or more entities are similar we must be repeatedly exposed to the common feature that, by definition, they share.

The mere exposure effect is very well rooted as it has been observed across an array of domains and in different species (Zajonc, 2001). It has even been shown to generate

approach behavior prenatally. For example, newly hatched chicks would move towards a tone that they had prenatally been exposed to, but not towards novel tones (Rajecki, 1974). A study on humans for instance (Monahan et al., 2000), subliminally exposed 2 groups of subjects to either 5 *repetitions* of 5 previously neutral Chinese ideographs, or of 25 different ideographs. The "repetition" group subsequently rated more positively (with respect to the non-repetition group) the observed ideographs. This effect appears to tap on affect rather than cognition, as it is stronger when subliminal. Moreover, it is rather unspecific, as it contaminates easily to "similar" but novel stimuli (similar to the repeated ones), and to dissimilar but still adjacent stimuli.

Furthermore, there is evidence showing that not only we like what we've been repeatedly exposed to, but we also feel that what we like is familiar, as if we had already been exposed to it (Monin, 2003). This phenomenon is partially explained by prototypicality. For instance, faces that have been artificially generated as the geometric mean of a number of faces are rated to be both more attractive and more familiar (Langlois&Roddman, 1990), where similar results hold for non-social stimuli such as birds or watches (Halberstadt&Rhodes, 2000). Prototypical objects are the more representative exemplars of their category, they constitute the "average exemplar", and as such they are likely to be somehow *similar* to the majority of exemplars of their category, to contain that mix of features that we seem to find "repeated" in all of them. Indeed, even if subjects haven't been exposed to a prototypical object, but only to non-prototypical samples of it, they feel it to be familiar, possibly because we spontaneously generate prototypes (Rosch, 1978) and mistake this for prior exposure (Strauss, 1979).

How exactly the above phenomena are connected and why we should *like* prototypical

objects, familiarity and repetition is still a matter of debate: some investigators suggest that perceptual fluency may underlie them all (Bornstein&D'Agostino, 1994); Zajonc argues (Zajonc, 2001) that there is adaptive value in approaching repeated stimuli, as they imply the absence of punishments, signaling "safety". In decision theory it has long been known that uncertainty and one of its declinations, ambiguity (Ellsberg, 1961), decreases the perceived utility of prospects.

2.2. *Similarity and attraction in the social domain*

The folk-psychological rule that similarity generates interpersonal attraction has been widely observed at both macro and micro levels of social ties, below we briefly address them separately.

2.2.1. Similarity in the large: "homophily"

The notion of repetition recurs in McPherson et al.'s (2001) definition of homophily, which refers to the principle that contact between similar others occurs at *higher rates* than between dissimilar others. An important implication of this is that any type of information (genetic, cultural, behavioral or material) that flows through networks will tend to be localized in both geographic and network distance ("the number of relationships a piece of information has to travel to connect two individuals" (McPherson et al., 2001). A voluminous sociological literature has proved this empirical pattern along a number of relation types and similarity dimensions such as ethnicity, gender, age, religion, education, occupation; and value based similarity, such as that based on shared attitudes, behaviors, beliefs or similar tastes. For instance, at a national probability sample only 8% of adults discussed "important matters" with someone of another race, 1/7th of what

would be expected on the basis of random extraction given relative group sizes (Marsden, 1987). Lu et al. (2012) refer that gender and age related homophily has been observed in zebras and dolphins and that meerkats assort depending on common attributes of dominance or foraging networks. Infants as old as 12 months have a notorious knack for imitation and they also prefer others that imitate them (Meltzoff, 2007). The commonly quoted evolutionary reason for this was proposed by Hamilton (1964) in terms of kinship selection. The idea is simply that agents may have an incentive to benefit others proportionally to their relatedness because, by doing so, they promote the survival of the portion of genes they share with them. Indeed, phenotypic matching (Porter, 1987), that is, the implicit evaluation of relatedness based on phenotypic similarity, has been observed in ground squirrels (Holmes&Sherman, 1982), baboons (Alberts, 1999), rhesus monkeys and a number of other species. Interestingly, even genetic homophily has been reported in humans (Fowler et al., 2012), such that friends are more likely to share given genes, plausibly precisely as a consequence of homophilic assortment. DeBruine (2002) showed economic trust was increased when human agents played with a fictive player who's face had been morphed to resemble themselves. Homophily was also predictive of cooperation in hunter-gatherer populations of Tanzania (Apicella et al., 2012), where social distance was shown to be as important as genetic relatedness in predicting assortment. A rather fascinating account of genotypic similarity was provided by Ghirlanda&Vallortigara (2004). The authors are concerned with the fact that selection pressures on the individual cannot explain the fact that, at the population level, the great majority of vertebrates show lateralization in proportions that are different from ½ (i.e. humans being more frequently right-handed). They make the intriguing proposal (with a model) that this might emerge as an evolutionary stable strategy when asymmetric organisms must coordinate their behavior with other asymmetric organisms. In line with

this argument, one of my main proposals is that similarity is not only about attraction, but about the impact it has on interactions. In brief, similarity appears to be a particularly powerful mechanism for social cohesion, in both phylogeny and onthogeny.

2.2.2. Similarity in the small: proxemics and propinquity

The network distance above however has a physical counterpart, which is object of a branch of social psychology called proxemics. Argyle&Dean (1965) proposed the principle that intimacy predicts greater physical closeness and that this distance represented an equilibrium that agents actively took effort to maintain when perturbed by external factors. For example, in a laboratory, they observed that eye contact was reduced as the chairs of seated strangers were moved closer. Mehrabian (1969) found that the degree of liking and the physical separation and gaze avoidance between two people were negatively related. Burgess (1983) analyzed people as they walked through a mall and discovered that companions were nearer to each other than to strangers, and that as the density of the crowd increased, the companion groups compressed so as to maintain spacing from strangers.

Furthermore, as in the non-social cases seen above, not only does liking predict interpersonal closeness but closeness predicts liking (as in the mere exposure effect): Priest and Sawyer (1967) tracked the friendships formed within a new dormitory at the University of Chicago. They found that proximity was positively related to both recognition and liking of other students: roommates were liked more frequently than neighbors, who were liked more than floor mates, etc. Similar patterns were observed for factors such as street arrangements (Hampton&Wellman, 2000) and legislative seating

(Calderia&Patterson, 1987).

In synthesis, we argued that non-social similarity/familiarity in the perceptual domain and reward could share some intricate connections and how this could spill over into social domain. Indeed, similarity and proximity appear to be fundamental factors of interpersonal attraction. However, though this may be sufficient to trigger a number of affect-laden processes leading to social approach, it seems unlikely that reward/motivation alone could be capable of guiding the complex cognitive processes that occur during the social interactions that follow. Such interactions are likely to be what matters. If homophily induced approach to malfunctioning/inefficient interactions (or even only neutral ones) it would seem implausible that it emerged as such a widespread phenomena. On the other hand, for homophilic approaches to actually work, they should be reinforced by the interactions that follow them. In other words, we're here interested in the possible strategic components of homophily.

Throughout this thesis we've stressed how one of the core assumptions of strategizing is common knowledge, in that it enables agents to form beliefs and expectations of what others will do. What we're contemplating here is the potential role of homophily as a "belief-correlating device", which could precisely promote efficient coordination in situations where deduction is of no use. In the next section, we will review evidence supporting the idea that similarity can indeed affect the way agents form beliefs on the intentions of one another, that is, on how it can affect "mentalizing" (Premack&Woodroof, 1978).

2.3. Similarity and common knowledge

Similarity supports self-referential mentalizing strategies

One of the long-standing open challenges of philosophy of mind and cognitive science is to explain how we come to infer the mental states of others. A simulationist account of mentalizing (Goldman, 2005) supports the view that we have direct access to our own mind and that we use such knowledge to reason about others. A different account however, called Theory-Theory (Carruthers&Smith, 1996), assigns a less important role to self-knowledge. Rather, it stresses the importance of our ability to build theories in general; the processes of forming which (for instance via the tacit generation of abstract and flexible rules or representations) shouldn't be too different whether they are meant for making inferences on mechanistic processes (humans excluded) or intentional ones. Proponents of this view alike children to "child scientists", who test their homemade theories of mind, or folk-psychological rules (Gopnik, 1996).

The self-referential approach appears to be particularly compliant in accommodating a series of experimental observations regarding egocentric biases. Many of which have to do with the well-documented observation that, often, social inferences are "contaminated" by one's first person views and perspective. Children have been long held to exhibit such egocentric beliefs (Piaget, 1957). Seminal studies using the Sally-Anne task (Wimmer&Perner, 1983; Baron-Cohen et al., 1985) showed that younger children (i.e. ages 1-3) have difficulties suspending their privileged knowledge when making attributions to uninformed others. Adults too however tend to do the same, usually, but not always in a self-serving way: by assuming their intentions and beliefs are more transparent to others than they actually are (Gilovich et al., 1998) and by overestimating

the degree to which others attend such intentions (Gilovich et al., 2000) or share their beliefs/thoughts (Keysar, 1994), the well-established "false consensus effect" (Ross et al., 1977; Marks&Miller, 1987; Krueger&Clement, 1994). As Goldman notes (2006), this form of self-projection seems to occur across domains, from knowledge, to valuations to feelings.

For instance, following Goldman (2006), Keysar (2003) had 2 agents, a "director" and a "follower", face a same grid with several objects on it. In front of the follower, but not the director, one of the objects, a roll of scotch tape, was hid in an opaque bag and placed on the grid. The follower thus knew that the director did *not* know what was in the bag. The director was then to orally "direct" the follower to shift some of the objects around the grid. One of such objects was a videotape. When the director said "move the tape" the follower should have realized that he couldn't mean the scotch tape, since the director didn't know about it. Nonetheless, agents often moved this, which indicates some difficulty in disanchoring themselves from their own privileged information. Similarly, Newton (1990) had "tappers" tap the theme of 25 famous songs so to allow listeners to guess what song it was. Tappers professed to nicely hear the songs in their heads while tapping and predicted that roughly 50% of listeners would guess what songs they were, though only 3% actually did. Camerer, Weber & Lowenstein note that asymmetric information of this sort is frequent in economic interactions (i.e. sellers know more about their product than buyers, employees know more about their abilities than employers). For instance, in their experiment (Camerer, Weber & Lowenstein, 1989), a group of subjects made forecasts on the earnings of given firms. A second group was then told the actual earnings of said firms and asked to trade assets that paid a liquidation dividend equal to the forecasts of the 1st group. To make accurate forecasts, the second group of participants should have

completely neglected their privileged knowledge to make optimal trades, which they didn't, thus making sub-optimal choices. Furthermore, economists have suggested that egocentric perspectives may play a role in adverse selection problems (and that "cursed knowledge" is one of the plausible explanations for "adverse selection" problems (Brocas et al., 2011; Carrillo&Palfrey, 2008). For instance, in the notorious "market of the lemons" (Akerlof, 1970), an agent is to make an offer for a firm, the true worth of which is known privately only to the seller. What the buyer knows is that i) the firm is worth some value q, with $0<q<100$, and ii) that it will be worth more in his hands than the current owner's, say $3/2q$. The (Bayesian) NE of this game is for no trade to occur[11]. Unsurprisingly, very few buyers reach this conclusion in the lab (), plausibly because they unprofitably generalize their ignorance, rather than their knowledge. Critically, such egocentric based-errors do not always occur (Epley et al. 2004), nor all errors go in the direction predicted by simulation. For instance, Ruffman (1996) ran an experiment involving 4-year olds and an observer. In front of both, there was a round dish, containing red and green beads, and a square dish, containing only yellow beads. Hidden to the view of the observer, but accessible to the children, a green bead was taken from the round dish and placed in an opaque bag. The children were then asked to say what bead the observer thought had been taken. Most of them didn't answer "green", as would have been predicted by self-projection, rather most of them answered "red". This suggests that the source of the inference was not the self, but, plausibly, a general "psychological rule", such as: "if you

[11] The reasoning is the following: suppose the buyer was considering to make an offer for the average expected value of the firm, that is 50. Taking the perspective of the seller, who knows the true worth of the firm, the buyer should realize that there is no world in which the seller would accept an offer of 50 if the firm is worth $q>50$. It follows that, if the firm is sold, it will be worth anywhere between 0 and 50, the expected value of which is 25. Under the buyer's new administration this would amount to $3/2*25 = 37.5$, however the buyer spent 50, thus prospectively losing 12.5. Since the same reasoning occurs for any positive offer, the buyer should make no offer at all.

don't know, you get it wrong".

Indeed, the debate between simulation and theory-theory is still very open (Goldman, 2006), yet an increasing numbers of cognitive scientists seem to support "hybrid views" of mentalizing, in which self-related and rule-based inferences interact to support social cognition. One interesting hybrid perspective suggests that both cognitive strategies may play important roles and that contexts may cue how much weight to give to each one (Goldman, 2006; Mitchell, 2005; Meltzoff, 2007; Epley et al., 2004; Ames, 2004;). This latter hypothesis is fundamentally based on similarity. It makes the argument that the self-related knowledge may be adopted to predict someone's behavior ("what I would do in her place?"), in the extent to which agents assume that their target is similar to themselves. Meltzoff (2007) says, "The bedrock on which social cognition is built is the perception that others are 'like me'."

Interestingly, though not expressed in terms of simulation vs. theory-theory, social psychological accounts of social judgment have converged on a somewhat similar distinction. On one side, similar to a rule-based approach to social cognition, a voluminous literature has focused on stereotypes (i.e. "How smart is Joe the football player?") (Fiske&Neuberg, 1990; Hamilton&Sherman, 1994) or prototypes (Karniol, 2003). On the other, much research has focused on the self (Krueger, 2000; Ross et al., 1977; Allport, 1931), accounting for the pervasive egocentric/false consensus biases succinctly reviewed above. In line with hybrid approach, one idea is that these 2 strategies of social inference may not be exclusive, but complementary, and may be moderated by social categorization. Indeed, a particularly consistent finding is that projection is greater for ingroup members than outgroup members (for a meta-analysis, see Krueger, 2000). Krueger (2000) says,

"The surest way to eliminate projection is to ask people to estimate social consensus for a group to which they do not belong. . . . It is as if people treat members of out-groups as members of different species". Ames (2004) took this as an indication that similarity may be the main modulator for adopting self-referential vs. stereotype based strategies (Fig. 5). For instance, "I found this movie great and, since I feel all proper cinephiles are like me, they'll like it too".

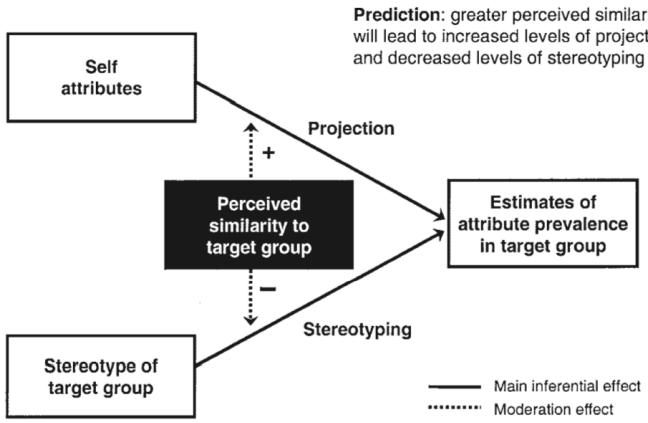

Fig. 5. Ames (2004): a similarity-contingency model accounts for the observation that self-projection and stereotyping appear to be inversely modulated by one's perceived similarity to a target group.

Critically, a number of studies in social psychology have shown that, depending on context, self-referential inferences can be either helpful (Hoch, 1987) or detrimental (Birch&Bloom, 2004). On the "helpful" side, similarity has been repeatedly shown to ease mentalizing: we are more empathetic towards others believed to have similar personalities to our own (Stotland, 1964), as we are better at decoding mental states from faces of ingroup as opposed to outgroup targets (Adams, 2010; Elfenbein&Ambady, 2002), where outgroup individuals are also believed to have a simpler, less rich, mental life than our own, both at the cognitive and affective level (Leyens et al., 2000). Possibly as a result of this, "group-membership similarity" has been widely shown to generate "ingroup

favoritism and discrimination towards the outgroup" (as Tajfel&Turner (1979) summarize 15 years of experiments on social identity theory). The "Minimum group" paradigm is particularly interesting because it shows that the specific characteristics identifying a group are relatively unimportant in driving discrimination, as agents will aggregate on virtually any trivial characteristic (such as pertaining to the group that overestimated, rather the underestimated the number of dots on a screen, or having preferred a painting by Klimt rather than Kandinsky); something which suggests that similarity effects are occasionally driven by the recognition of similarity itself, rather than its object. Furthermore, social closeness across a number of dimensions has been shown to benefit economic interactions in a wide range of games (which we review in study 1).

On the "detrimental" of similarity and self-referential mentalizing, egocentrism (Royzman et al., 2003) and ingroup favoritism has been shown go to go hand in hand with misunderstanding and conflict (Ross&Ward, 1995; Weber&Camerer, 2003). Todd (2010) and colleagues devised an interesting paradigm to show how similarity can occasionally interfere with perspective taking. In the 1st section we spoke about non-social similarity and its tendency to be unspecific. This enabled Todd to induce a focus on social similarity with non-social similarity. He did so, by having separate groups of subjects compare the same pair of images and either to "look for similarities" or "look for dissimilarities", before taking part in a series of social tasks. As we mentioned earlier, similarity facilitates mentalizing when others are likely to share our own perspectives (i.e as might occur if they are part of our same cultural group). However, dissimilarity may help take the perspective of others, when these differ from our own. Todd (2010) demonstrates this point at both the perceptual and conceptual level through 5 experiments, which we briefly review. 1) Perceptual perspective taking: after the similarity/dissimilarity priming,

subjects viewed a picture of a man facing them and sitting at a table. At the left of the depicted man there was a bottle (thus to the "right", according to the subjects perspective). Subjects were then simply asked, among several "distractor" questions, on what side of the table the bottle was. Subjects in the dissimilarity mindset more frequently said it was on the left, thus taking the perspective of the man in the picture, while the similarity-focused subjects tended to say the bottle was on the right, thus remaining anchored to their own perspective. 2) Deciphering ambiguous communication: subjects were told about "John's dinner out...". John had asked his wife to suggest him a restaurant to go to with his parents and both food and service turned out to be terrible. After dinner he sent an email to his wife, saying, "the dinner was marvelous, just marvelous". After hearing this story subjects were asked to rate how likely it was for Micheal's wife to have understood his sarcasm. Subjects in the dissimilar mindset thought it was more unlikely than those in the similar mindset. They were thus able to better detach themselves from their privileged knowledge, which indeed Micheal's wife was unlikely to have. 3) False belief attribution: subjects were shown comic strips depicting a girl who finishes playing the violin, puts it in a blue container (out of 4 differently colored containers) and leaves the room; the strip then depicts Vicky's sister, who comes into the room and moves the violin from the blue to the red container and leaves. Subjects were then asked to write down the probabilities of Vicky looking for her violin in each container. As before, subjects in the similarity mindset were relatively less able to take Vicky's naïve perspective and placed higher probabilities on the red container. 4) Inter-group false-belief attribution: this experiment was identical to the former with the exception that i) there was no search task to induce focus on similarities/dissimilarities, and ii) the between group manipulation regarded simply the fact that, for one of the 2 groups, Vicky and her sister had foreign names, while for the other they had German names (the same nationality of

the subjects). Subjects in the "foreign" name group behaved in the same way as the difference mindset group had in experiment 3, that is, they correctly attributed lower probabilities to the fact that the girl would look for her violin in the red container. This shows that common group characteristics, and similarity, can easily induce subjects to assume that others share their same beliefs. 5) Finally, minimum-groups were formed between subject pairs. They were then placed one opposite to the other and each of them took turns in i) either being blindfolded and orally guided through a virtual maze by the other participant, or ii) being the guide, who could only pronounce 4 directions (up, down, left, right). Outgroup pairs took less time to complete the maze, indicating that their perspective taking had been facilitated, and as a consequence, their coordination rates enhanced. In short, these results show nicely how both i) a focus on similarity and ii) ingroup membership, can be of obstacle in mentalizing, specifically in the case in which perspectives differ (Santiesteban et al., 2012).

2.4. Synthesis

The findings reported above can be well-integrated with what Fiske (2006) proposes to be the "universals" of social cognition, by which humans across cultures have been shown to differentiate amongst one another. She proposes a space delimited by 2 axes: warmth and competence (Fig. 6). The "self" can be considered to be in the top-right quadrant. Then perceived similarity to the self implies a shift in an affective/attraction-related and a cognitive/belief-related one. We not only like social closer others, but we also tend to assume that they perceive the world like ourselves and that they are thus more likely to behave like us.

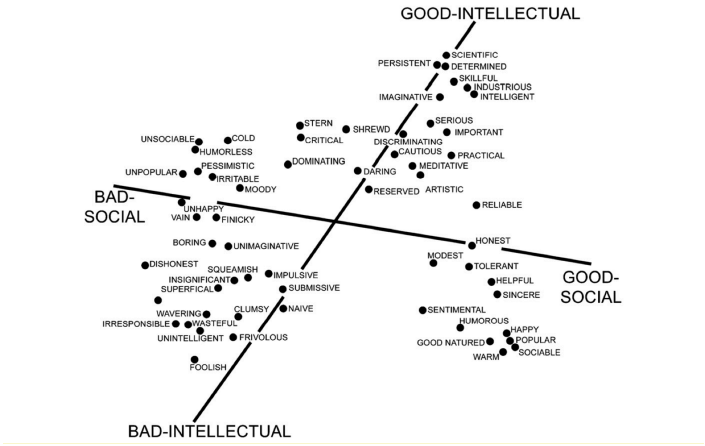

Fig. 6. Fiske's "social evaluation" space along dimensions of "warmth" and "competence".

We showed that the impact this can have on interactions is strictly task-dependent. Specifically, if perspective coincide, then projection can help, but if they differ, it can lead to increased misunderstanding and sub-optimal behavior. In our studies, we aimed to link these findings to strategic interactions. Specifically, if common knowledge based on deduction fails in coordination games, then, perhaps closeness based common knowledge could vicariate. Indeed, a sense of likemindedness could play the role we needed of promoting cooperation and disfavoring competition, by respectively decreasing uncertainty when choices are to be matched, but increasing it when they are to be decoupled.

Chapter 3 Two Roads to Social Cohesion: Propensity for Cooperation and Aversion to Conflict

Gabriele Chierchia[1] and Giorgio Coricelli[1,2]

[1]Center for Mind/Brain Science, University of Trento and [2] Economics Department, University of Southern California

Abstract

Humans are often faced with a tradeoff between safe isolation and potentially beneficial, but risky, interdependence. To account for such decisions, standard economics typically focuses on their economics incentives while relatively neglecting that social species are accustomed to differentiating between one another along various dimensions, such as their degree of relatedness, their similarities or their group membership. Indeed, such forms of "interpersonal closeness" are known to foster cooperation. However, no study has investigated how closeness may play out in more competitive environments. We report 2 novel findings: playing with actual friends (vs. strangers) or with similar (vs. dissimilar) others raises risk-rates and can resolve notorious coordination problems involving cooperation; the same forms of closeness however decrease risky behavior when choices offset one another, for instance when competing over limited common resources. Interestingly, both effects increase group payoffs, thus shedding light on their potential selective advantages.

Introduction

Sally (2001) on Adam Smith:

"Smith suggested a geometry of human relations: we perceive a space in which our self is the origin and other people are arrayed at recognizable positions and at a calculable distance from the origin. Our ability to change places in fancy with another declines as the other moves further away from the self; accordingly, sympathy is an inverse function of distance."

As social beings, it is extremely frequent that the outcomes of our choices depend on the choices of others. This requires us to peer upon the minds of others, while they peer into ours, to optimize behavior. Such "mutual mind-reading" in decision-making can give rise to "strategic uncertainty" (Van Huyck et al., 1990).

For instance, consider the decision of joining a rebellion. All know that the chances of victory are proportional to the number of those joining, and that if enough people do, all of those people will be better off. However joining the rebellion is risky because one would never want to do so alone. Alternatively, the tribulation one may experience upon deciding to enter the freeway around rush-hour may be similar to the one experienced by agents who would like to enter particular markets (Camerer&Fehr, 2006): in both cases, agents know that if too many enter, there will be a traffic jam on the free way, or a price war in the market, thus everyone loses.

In rebellion-style cases potential profits are aligned (if "we all risk, we all profit") and agents would thus prefer to match their choices (i.e. "we either all risk" or "no one does", but isolated risks are costly). In traffic-style situations agents can only profit alone and

would thus prefer to *decouple* their choices (such that "either I take the freeway", or "you do", but "we shouldn't take it together"). In the first case, strategies are said to be complements, and they can foster cooperation, while in the latter they are *substitutes*, and are typical of competition (Bulow et al., 1985; Fehr&Camerer, 2006). Both cases however, require agents to coordinate their choices without communicating.

Understanding what mediates strategic uncertainty in coordination problems of this sort has perplexed economists and game theorists for decades (Schelling, 1960). The problem is that, whereas, in many interactions, incentives alone can guide "rational"/deductive agents to optimize their behavior (i.e. in tic-tac-toe, a rational first-mover will never lose), coordination dilemmas have multiple deductively valid solutions (i.e. multiple Nash equilibria), leaving agents with the problem of finding some other form of "tacit agreement". Indeed, coordination games have been said to constitute "the biggest problem of game theory" (Camerer, 2003).

However, standard economics typically assumes a parsimonious "social void" of undifferentiated agents (Charness et al., 2007), whereas sociobiology, social psychology, anthropology and sociology tend to explain behavior precisely by assuming that agents differentiate between one another along various important dimensions, such as their degree of relatedness (Hamilton, 1964), their degree of interindividual similarities (McPherson, 2001) - which could be a proxy for relatedness (Fowler et al., 2009)- or in terms of their common group membership (Tajfel&Turner, 1979), among others (Akerlof&Kranton, 2000). Such factors suggest the notion of a social space (Fiske et al., 2006; Jones&Raichlin, 2006), in which tuning behavior to the "social distance" of others appears to be the rule, rather than the exception of interactions. Humans are in fact

accustomed to do so since the earliest ages, preverbally (age 1 and younger) exhibiting affiliative/cooperative behavior towards parents but not strangers (Lamb, 1977), evaluating others (Hamlin et al., 2007), and progressively expanding their social network beyond genetically related others, to include friends, with which they share more than with strangers (Olson et al., 2007). Similar social tuning has also been observed in a number of non-human species including spider monkeys (Pastor-Nieto et al., 2001).

It is perhaps less surprising then that "closeness"-based behaviors recur in the economic interactions of adults. Indeed, many different forms of decreased social distance - in terms of social network distance (Apicella et al., 2012; Harrison et al., 2011), artificial group membership (Charness et al., 2007; Chen&Chen, 2011), natural group membership (Berhard et al., 2006; De Cremer&Van Vugt, 1999), common culture (Efferson et al., 2008), social identification (Bohnet&Frey, 1999; Hoffman et al., 1996), pre/post play communication (Fehr&Gachter, 1999), social distance (Charness et al., 2003), motor synchronization (Wilthermuth et al., 2010), facial resemblance (DeBruin, 2002), demographic similarities (Cole&Teboul, 2004) and friendship (Haan et al., 2007; Reuben et al., 2008; Yamagishi&Sato, 1986; Thompson et al., 1998) - have all been associated with increased cooperation.

However, no study to our knowledge has investigated how closeness may play out in the frequent cases when there is no such possibility, rather, as incentives are in conflict, choices offset one another and agents must decide whether to engage competition. Here we investigated the hypothesis that interpersonal closeness would decrease uncertainty when choices are to be matched, but that it would increase it when they are to be decoupled. As an illustrative metaphor, one may think of the paradigmatic case of

interpersonal closeness, namely, that of monozigotic twins: though twins should (and indeed do, Segal&Herchberger, 2009) take more risk when it comes to matching choices to maximize mutual gains, the same processes that mediate such an effect could exacerbate conflict when they are to outsmart one another, thus raising their uncertainty when strategies are substitutes. Our study's goal is to fill this gap by systematically contrasting the effect of social distance in cooperative and competitive coordination games.

The current research

To do so, we transformed the above examples (i.e. rebellions, traffic) in economic games, in which matched counterparts received real monetary payoffs depending on their ability to tacitly coordinate their choices in games with strategic complements and substitutes (see below). Using such games, we measured strategic uncertainty under 3 conditions of social distance. Study 1 attempted to maximize social distance and served as a baseline. In this first study, in keeping with standard behavioral economics experiments (Camerer, 2003), agents knew nothing of one another. Furthermore, as choices in monetary interactions often carry a moral value, and subjects have been shown to be sensitive to the potential judgment of experimenters (Hoffman et al., 1996), a double-blind setting was adopted (for specifics, see Methodological Details S1, in the Supplemental online material - SOM, available online). In study 2 we relaxed anonymity and investigated "objective" social distance by assessing how actual friends, as opposed to strangers, coordinated their choices in cooperative and competitive environments; in study 3, we restored anonymity and induced "perceived" closeness by manipulating one of its most important predictors, namely, perceived similarity (McPherson, 2001), matching subjects with other (unfamiliar) agents perceived as similar in terms of specified personality traits. In the next section we illustrate the games and procedures that were common to all studies. Then, in

the three subsequent sections, we describe and motivate the specifics of each experiment and present their results in turn.

The games and procedures

In all 3 experiments we used the same 2 types of (2-player) coordination games: "Stag Hunts" (SHs) and "Entry Games" (EGs), which have been extensively studied, both in theory and experimental settings (see Camerer, 2003 for a review). In our versions - adapted from Heinemann et al. (2008) - we attempted to keep the superficial aspects of the 2 games as similar as possible, so that any behavioral difference would be due to their structural (incentive-related) differences. The games were as follows: in both games, 2 agents had to choose between the same two options: a high paying "risky" option, always worth $/€15.00 or 0, and a lower paying but safe payoff (SP) of a given $/€ amount (SP ≤ 15.00). Both games capture a frequent scenario, namely, that low gains can be obtained in isolation. Indeed, if the SP was chosen, it was obtained for sure, regardless the choice of one's counterpart. High paying outcomes on the other hand required coordination and risk: in SHs, $/€15.00 were obtained, by both players, only if both risked, while if only one risked, he obtained 0. In EGs, on the other hand, the high gain could *only* be obtained in isolation, thus if both risked, both obtained 0. Then, by varying the value of the SP and having participants choose at each (randomized) step we obtained a measure of their uncertainty in the 2 games. Importantly, since initial coordination patterns usually determine the outcome of their repeated versions (Heinemann et al., 2004), and since we were here interested in the way social distance biases choices rather than how it may bias learning (Fouragnan et al., 2013), no feedback on the outcomes of decisions was provided until the end of the experiment, when one choice was extracted at random and paid (in addition to a flat "show up fee" of $/€5.00).

Finally, to control for the potential impact of inter-individual differences in (non-social) risk attitudes, participants took part in lotteries. Here, they were to choose between the same options as in the strategic games, however, in contrast to those, if one chose the risky option the probability of winning was fixed at p=0.5 and depended on a blind extraction from an actual urn (containing 1 "winning" ball and 1 "losing" ball), thus completely independent from the choice of others.

Each study took place in behavioral economics labs. Participants interacted in groups of approximately 16, and interacted via computers from shielded cubicles. All dependent variables were analyzed with generalized linear mixed effects models (GLMMs), as implemented in the lme4 package (Bates & Sarkar, 2006), in the R environment (R Development Core Team, 2006) (for specifics see Methodological Details S2 in SOM-R). All procedures were approved by local ethical committees.

3. 1. Experiment 1: strategic uncertainty under anonymity

The first study was conducted at the CEEL lab (Cognitive and Experimental Economics Laboratory) of the University of Trento (Italy). In 4 experimental sessions, 75 participants took part in the economic coordination tasks, implemented in zTree software (Fischbacher, 2007).

Results

Collapsing across SPs, agents were shown to risk in SHs (68% of the time) and to avoid risk in EGs (67% of the time). However, in SHs, for SP values above roughly 2/3 of what

agents could receive by cooperating (threshold: SP=10.18) (Fig. 1), agents tend to stop risking. Indeed, as expected in SHs, GLMMs revealed that increasing SPs decreased likelihood of risking (b=-0.94, s.e.=0.08, p<2e-16) increased miscoordination (coefficient=0.02, s.e.=0.002, p<0.001) and lowered expected payoffs (coefficient=0.25, s.e.=0.03, p<0.001), thus posing an efficiency dilemma of how to maintain coordination when SPs increase. In EGs too, SPs decreased likelihood of risking (b=0.54, s.e.=0.05, p<2e-16), however, in contrast to SHs, coordination failures were especially high at low SPs (p<0.05), rather than high ones. Response time (RT) analysis further confirmed that uncertainty was highest at opposite SP ranges, and for opposite choices in the 2 games (game*choice, and game*SP interactions both significant at the 0.01 level): in EGs agents appeared to reluctantly engage competition (taking longer to do so) when the alternative option was particularly unappealing, that is, at low SPs (coefficient=-0.02, SE=0.006, p<0.001), and when ultimately choosing to risk, rather than staying safe (mean difference=0.2s, SE=0.05, p<0.01). Indeed, risking generally lowered payoff (p<0.05). Conversely, in SHs, decision times (i.e. fear of miscoordination) grew with increasing SPs (coefficient=0.008, SE=0.003, p<0.05), and agents took the longest when finally interrupting cooperation by choosing the safe option (mean difference=0.16, SE=0.04, p<0.001). Finally, we found that across subjects mean risk rates in the individual decision-making domain (lotteries) correlated with the homologous scores in the strategic games, for both SHs (Pearson's r = 0.27, p<0.05) and EGs (r=0.4, p<0.001). Thus participants who risked more in the lotteries domain risked more in the two games (Fig. 7). .

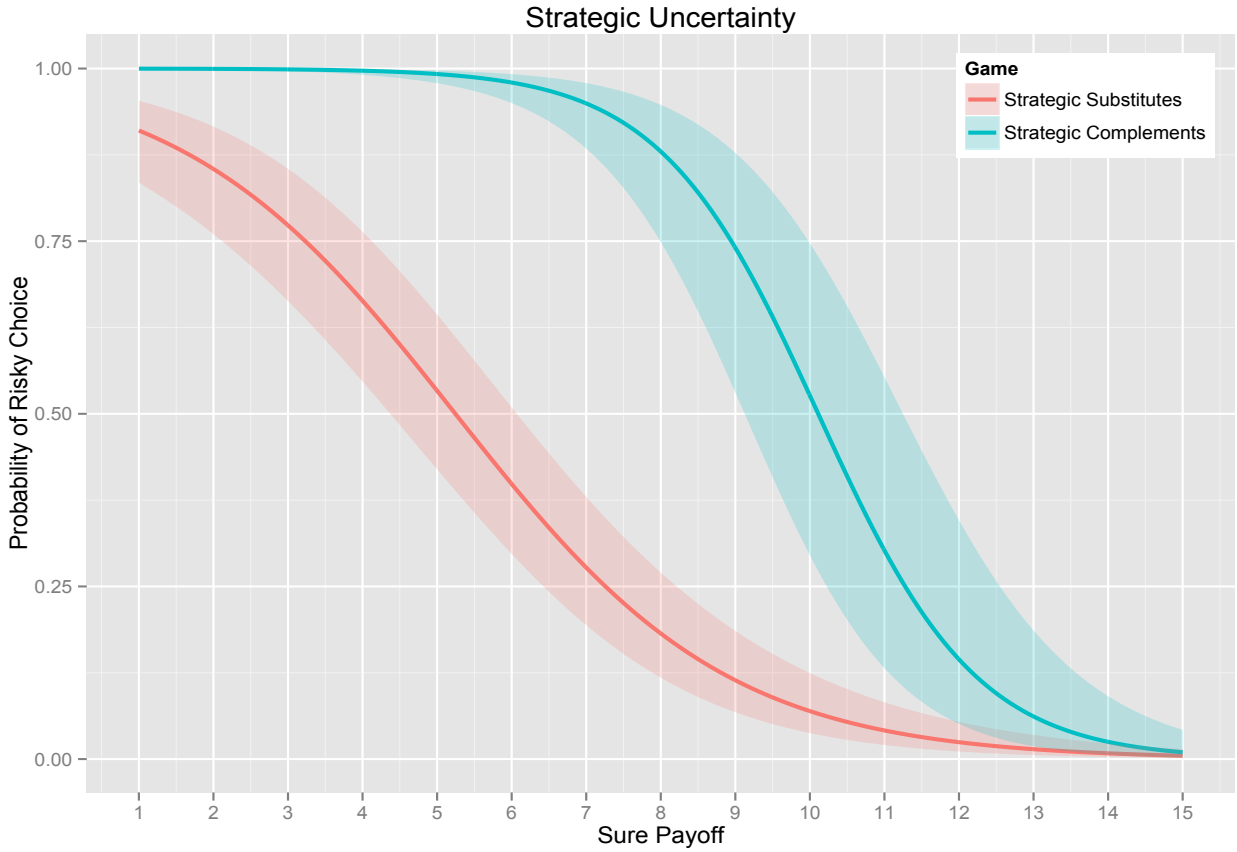

Fig 7. Probability of risky choice (y-axis) (estimated with a generalized logistic mixed model, with 95% confidence bands in grey), when mutually anonymous counterparts interacted in one-shot coordination games, given increasing sure payoffs (x-axis) and the incentive to either match (strategic complements, dashed line) or decouple (strategic substitutes, full line) their choices.

3.2. Experiment 2. Objective social distance: friendship

The study was conducted at the Los Angeles Behavioral Economics Lab (LABEL, University of Southern California). In 4 sessions of approximately 20 participants, 78 students (44 females, mean age = 20.5, s.d.=3) took part in economic coordination tasks, implemented in Qualtrics software (Qualtrics, Provo, UT). Each participant was required to bring a non-romantic friend to the experimental session and was to take part in our games of interest twice, once with a stranger and once with their friend. In both cases, agents were mutually informed of whom they were playing with. After taking part in the

lotteries, while payoffs were being calculated, participants also took part in a brief questionnaire (see SOM-R S3 for details).

Results

GLMMs revealed a 2-way interaction between game and friendship (b=-1.73, s.e.=0.14, p< 2e- 16). Indeed, restricted models showed that, in SHs, risky behavior was drastically reduced when agents played with their friends as opposed to strangers (b=-3.43, s.e.=0.41, p<2e-16): even when friends could only break even or lose by risking (that is, when SP=15), more than 40% of them still kept doing so. In contrast to this, the already high uncertainty of mutually offsetting choices in EGs appeared aggravated by friendship. Here GLMM logistic fits revealed an effect of friendship (b=0.73, s.e.=0.28, p<0.01), as well as a significant interaction between SP and friendship (b=0.05, SE=0.02, p<0.05), suggesting that, in EGs, friendship lowered likelihood of risking at low SPs (i.e. when competition is highest) (see Fig. 2). In line with this, and with experiment 1, decision time analyses revealed a marginally significant 4-way interaction between friendship, game, choice and SP (b=0.11, s.e.=0.06, p=0.06): while in SHs, at high SPs (i.e. SP>7), friends hesitated more than strangers when interrupting cooperation (by choosing the safe option) (b=-0.15, s.e.=0.05, p<0.05); in EGs, at the lowest SP (SP=0), they appeared most uncertain when risking collision (b=0.23, s.e.=0.12, p=0.06). Furthermore, the main effects of friendship in this study and in experiment 1, where no social comparison was salient (because participants only played with strangers), were undistinguishable (SHs and EGs, both ps>0.3), thus ruling out potential experimental "demand effects". Also, the correlations we observed in experiment 1 between risk attitudes in lotteries and strategic games completely broke down in SHs (friends: Pearson's r=0.03, p=0.54; strangers: r=0.11, p=0.33), and were maintained in EGs (friends: r=0.30, p=0.006; strangers: r=0.56,

p<0.001).

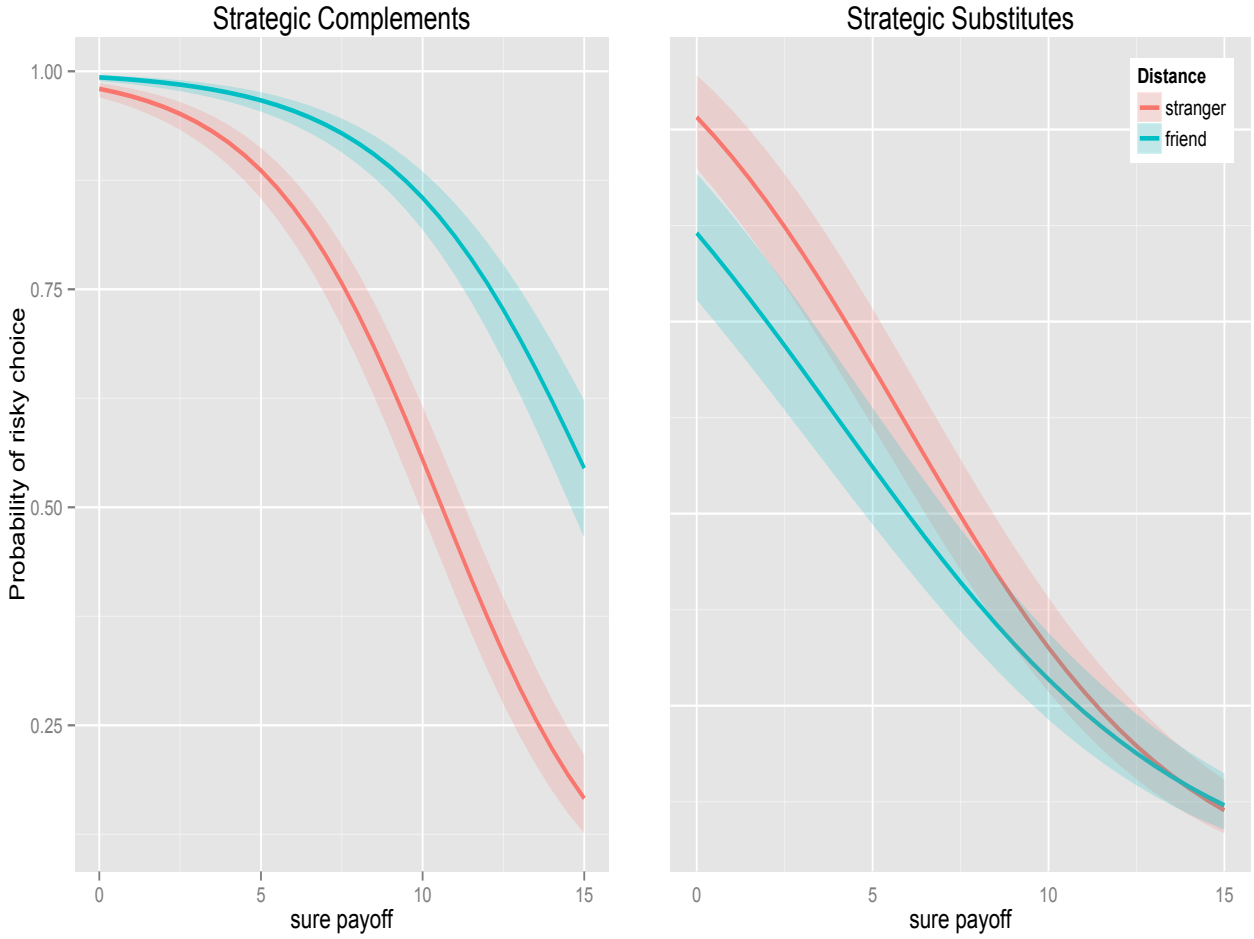

Fig. 8. Probability of risky choice (y-axis) (estimated with a generalized logistic mixed model, with 95% confidence bands), when friends (green) or strangers (red) interacted in one-shot coordination games, given increasing sure payoffs (x-axis) and the incentive to either match (strategic complements, left panel) or decouple (strategic substitutes, right panel) their choices.

Notably, in analyzing the consequences of such behaviors, we found that both of the effects were associated with increased coordination rates - that is, same choices in SHs or opposite ones in EGs (both p_s<0.05) - as well as payoffs (both p_s<0.05, though in EGs, mirroring the behavioral distinction above, this was detected as a slope difference). Interestingly, both of which were unchanged (p_s > 0.4) when choices of a player were

matched to a random participant in the "friend" condition, rather than to their actual friends, thus suggesting that such beneficial effects of closeness were not driven by the fact that friends were able to better anticipate each others' idiosyncratic choices - though they may still have believed so - so much as by an unspecific sense of "friendliness".

In such "random matching" procedure, payoff's were computed given the percentage of agents risking in the same condition. For instance, if a player risked in an EG with SP=7.5, in the friendship condition, and 50% of the other players risked in this exact same condition (target's friend excluded), then the participant's payoff in this trial was equal, in expected value, to the probability of being matched to someone who did not risk (thus 1-0.5=0.5), multiplied by the earning in case of success (15.00), that is, 0.5*15=7.5. Furthermore, if this specific percentage occurs when SP=7.5, then no player has an incentive to change his/her choice, as even 49% risking would generate incentive for an extra player to risk as well - since 0.51*15 is larger that the SP of 7.5. Indeed, for each SP, there exists such a percentage, which coincides with the mixed strategy solution of the game. Notably, it is a well-documented observation (Camerer&Fehr, 2006) that humans approximate such equilibrium points remarkably well in EGs, without communication or trial and error. "To a psychologist", Daniel Kahneman said, "it looks like magic" (Kahneman, 1988). Indeed, equilibration was approximated in our study as well, as is clear from Fig. 3 (albeit with some under-entry). In this figure, which here only shows the higher competition SP range, SP≤7 (for full figure see supplementary online material, SOM-R S4) one can appreciate several aspects: i) as described above, friends (filled circles) risk less than strangers in EG (squares); ii) by doing so, they are foregoing more profit opportunities, because they are playing farther from equilibria (empty circles); however, iii) since many equilibria in EGs, though economically "rational", yield low

collective payoffs, as a group, friends are ultimately better off than strangers (they are closer to the social optimum, that is, the peak of each curve). Such payoff dynamics thus shed light on how closeness-induced uncertainty could ultimately be reinforced, and thus sustained in competitive environments.

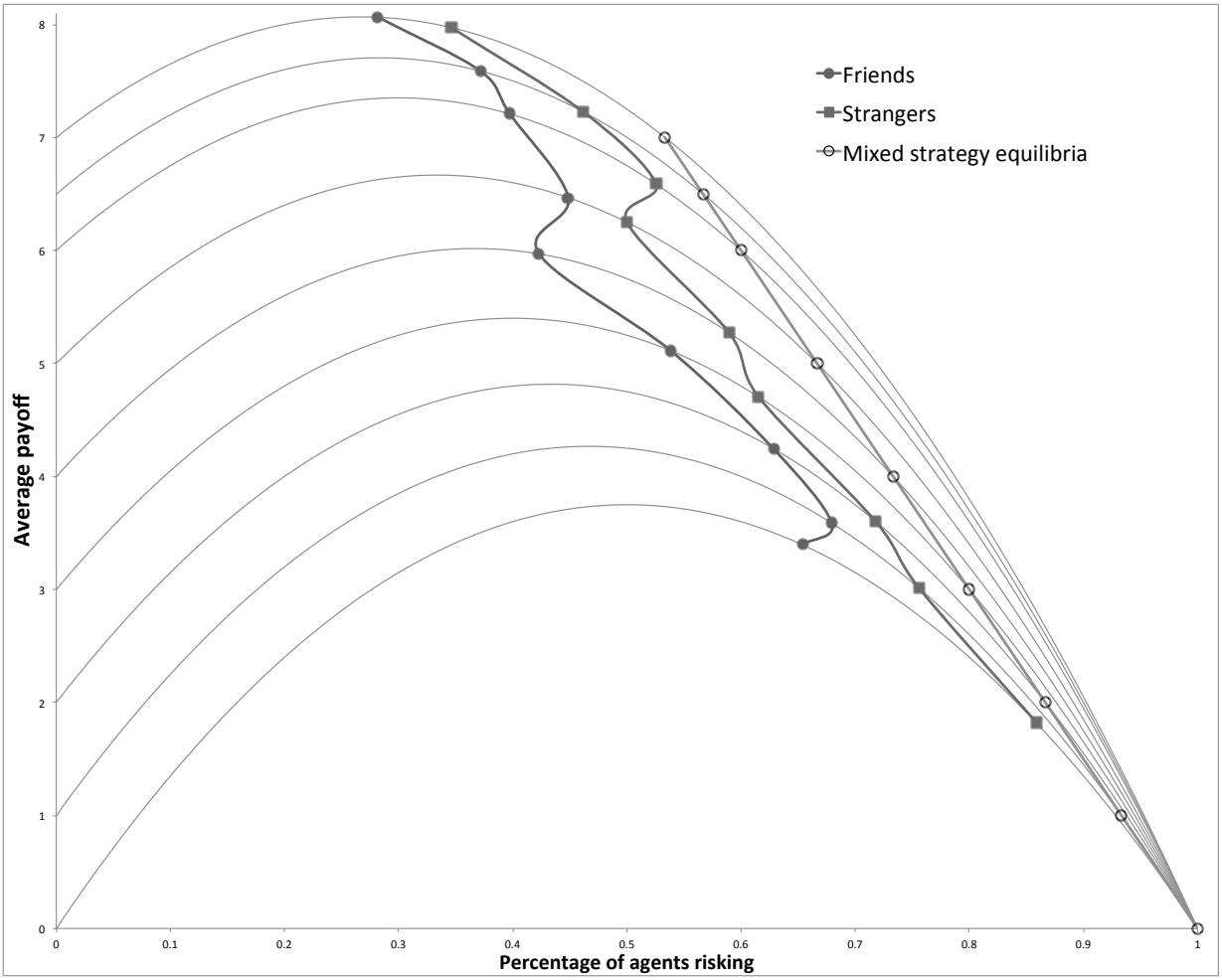

Fig.9. Payoff dynamics in coordination games with strategic substitutes ("Entry games"). The series of inverse-U curves represent average payoffs (y-axis) given a hypothetical proportion of agents that "enter" (risk) (x-axis). Each one represents an EG with a specific sure payoff (SP), which are differentiable by examining the average payoff value (y) when no one enters (i.e. when x=0). The indicators represent the observed proportion of agents entering in the friend (full circles) vs. stranger (squares) conditions. Friendship induced participants to risk less, thus play farther from equilibrium, however, by doing so, they earn more than strangers as a group.

3. 3. Experiment 3. Pychological social distance: similarity and liking

Studying the impact of perceived social distance through friendship raised a few issues:

1) Though friendship decreases social distance, it remained unclear what aspects of this could drive the observed effects. For instance, a questionnaire from the previous study showed that friends believed they were, relative to strangers, characterized by the same personality traits - though, interestingly, they disagreed on which ones. This is in line with the observation that similarity (i.e. homophily) along a vast range of dimensions is one of the best predictors of social ties (McPherson, 2001), such as friendship (Morry, 2007). From a psychological perspective, interpersonal closeness has been shown to induce agents to project (Goldman, 2006) their own thoughts and preferences to others (Todd et al., 2012; Robbins&Krueger, 2005 Mitchell et al., 2005; Krienen et al., 2010), which could indeed decrease uncertainty when choices are to be matched, and increase it when they are to be decoupled. However, an alternative hypothesis is that, independently of similarity, agents who are "close" also generally *like* one another, and may thus try and benefit one another by cooperating more and competing less. This led us to ask whether similarity or liking alone were sufficient to induce the polar effects on coordination observed above.

2) A second issue is that closeness was not the only factor distinguishing friends from strangers. Importantly, friends knew they would have seen each other after the experiment thus possibly raising reputational concerns (Fehr&Fischbacher, 2003).

In study 3 we thus attempted to overcome both of these intrinsic limitations, by re-instantiating anonymity and simultaneously trying to disentangle between similarity and

liking.

Material and Methods

The experiment was carried out at LABEL (University of Southern California). In 3 sessions, 40 anonymous participants (29 males, mean age = 21**,** s.d= 3) took part in the 2 coordination tasks, implemented in zTree (Fischbacher, 2007). As soon as participants entered the lab they rated (on a scale from 1 to 7) each of 100 personality traits. They did so twice: once, indicating how much they identified in a given trait, and the second, how much they liked the same traits. Then, for each participant, an algorithm clustered the traits in 4 groups, of 3 traits each: 1) traits that one liked and also identified in (i.e. maximizing both liking and identity scores), 2) traits that one identified in but did not like (i.e. maximizing identity while minimizing liking), 3) traits that one liked but did not identify in, and, 4) traits that one did not identify in and did not like (one can imagine a 2-dimensional space, with a "liking" and "identity" axes). Finally, participants played the games while being mutually informed whether their counterparts similarly or dissimilarly identified in a given triplet of personality traits, for each of the 4 trait triplets above. As the decisions were many, we adopted the strategy method (Selten 1967; Brandts&Charness, 2000), in which agents view all options, for a given game, trait cluster and similarity condition, on a single page, rather than sequentially (see fig. 4).

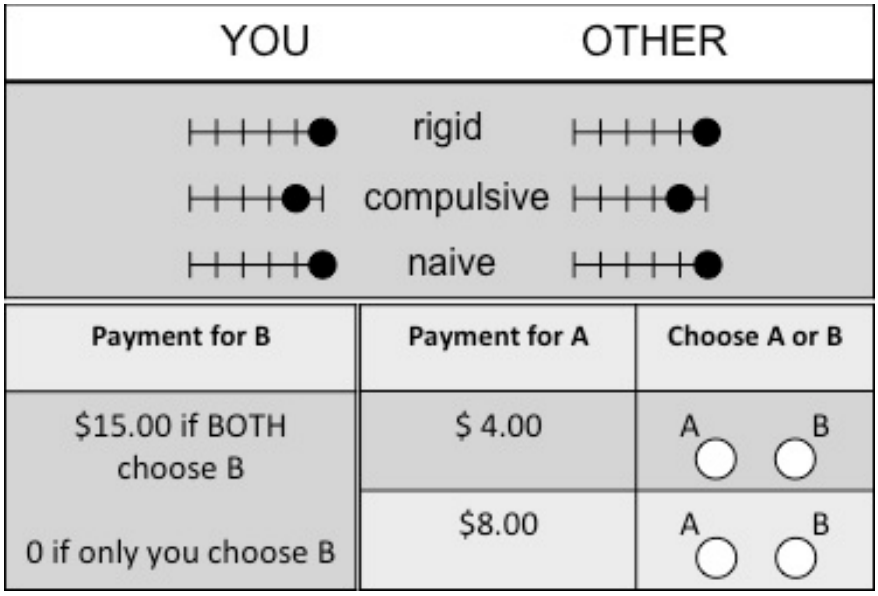

Fig. 10. Stimuli for testing the impact of similarity and liking on coordination games. Top: an example from the "high identity, low liking" trait cluster for a given participant. In a similarity condition matched counterparts identified in the same traits, in a dissimilarity condition (not shown) they oppositely identified in the same traits (black circles for "other" were shifted to the left). Bottom left: in this example the screenshot for SHs. In EGs (not shown) the "15.00" and "0" were simply inverted. Bottom right: in the strategy method agents view all sure payoff options, in randomized order (for a given game, trait cluster and similarity condition) on a single page.

Results

GLMMs revealed a significant interaction between game and similarity (b=0.75, s.e.=0.22, 38p<0.01), which further interacted (in 3-way interactions) with both liking (b=0.68, s.e.=0.25, p<0.01) and, marginally, with identity (b=0.45, s.e.=0.25, p=0.08). Indeed, restricting the model to only the cluster of traits that subjects identified with and liked - the condition which most resembles the case of friendship - we replicated the findings of experiment 2, albeit to a lesser degree: similar agents were more likely to risk than dissimilar ones in SHs (b=6.11, s.e.=1.33, p<0.0001), while the logistic fit to EGs revealed an opposite marginal effect of similarity (b=-1.63, s.e.=0.9, p=0.08), and a SP*similarity interaction (b=0.25, s.e.=0.11, p<0.05). Importantly, restricting the model to the any of the other trait clusters revealed that similarity ceased to induce

this dual pattern when shared traits were disliked, or when liked traits were not identified in. This suggests that liking, identity and similarity interact to induce the differential effects of interest and that, in addition to objective network distance, psychological distance alone (i.e. perceived similarity and liking) can foster cooperation and discourage conflict.

3.4. Discussion

It has long been known that interpersonal closeness fosters cooperation: in treating moral sentiments Adam Smith (Smith, 1759) recognized that "sympathy" declined with social distance, from "brothers and sisters" to "the children of brothers and sisters", to "the children of cousins" and "the affection gradually diminishes as the relation grows more and more remote". Similarly, Haldane (1939) provocatively remarked, "I would lay down my life for two brothers or eight cousins". Indeed, from an evolutionary perspective (Hamilton, 1964), it can make sense to incur costs to benefit genetically closer others, because, by doing so, individuals may indirectly promote the survival of the portion of genes they share with them. Paralleling this, and perhaps relatedly, in non-kin relations, homophily (the tendency to associate with similar others) is held to be one of the fundamental mechanisms underlying social ties such as friendship (McPherson, 2001). Incidentally, and possibly linking the two phenomena, friends have been shown to exhibit correlated genotypes, plausibly as a consequence of the fact that they actively seek others with similar/dissimilar phenotypes (such as similar cognitive skills, preferences or physical traits) (Fowler et al., 2011). Here we thus sought to investigate how social closeness, in terms of friendship, or interpersonal similarities, can play out on actual economic interactions. We demonstrated that both can potentially resolve one of "the biggest problems of game theory", namely coordination: indeed, when notorious "Stag hunt" dilemmas (with their typically observed inefficiencies) are grounded in social contexts mediating closeness, the dilemma nearly dissolves, together with agents' uncertainty. However, we show that closeness can also increase uncertainty if choices offset one another. Here, for instance, a perception of similarity can deter agents from

collectively over-exploiting common resources, thus increasing collective profits. It follows that 2 roads of decreased and increased uncertainty can lead groups whose members feel close to maintain coordination advantages over groups whose members are "strangers".

Finally, our experiment on similarity demonstrates that the effect that closeness has on social interactions isn't necessarily only about "liking". Instead, and in synthesis, we suggest that one of the reasons homophily works, as a social attractor, is that it may resolve fundamental coordination problems: if others are more "like us" they are more likely to pursue our own goals and thus reciprocate our efforts. Consequently we may cooperate with ingroup others, not only to benefit them, but also because we think cooperation with them is more likely to be successful. For the same reasons however, we may avoid conflict with ingroups not only to avoid harming them, but also because conflict with them would be more harmful than with others, paralleling the adage that “there is no worse conflict than with one’s self”.

Chapter 4 *Neural Mechanisms Mediating the Impact of Social Distance on Human coordination*

Introduction

In the preceding sections we showed how homophily and social closeness can serve 2 probably related but conceptually distinct functions: i) Along a seemingly reward/motivation-related dimension (i.e. Fiske's (2007) "warmth" dimension) they can induce attraction; ii) along a more cognitive/informational dimension (i.e. Fiske's "competence" dimension) they can foster a sense of likemindedness, mutual understanding, or "common knowledge". We also provided evidence (Chierchia&Coricelli, under revision) that both actual closeness (friendship) and perceived closeness (i.e. similarity) support a qualitatively similar 2-fold role in mediating coordination: they lead agents to risk more when trying to match their choices, and to risk less when trying to decouple them, and that both of these effects generate an economic advantage to groups whose members feel "close". Here we aimed to further articulate these effects in relation to the 2 dimensions illustrated above, and to attempt to dissociate between them at both the behavioral and neural level.

Indeed, a recent debate has sparked in behavioral economics concerning the nature of the traditionally observed effects of social distance on cooperation (Cox, 2004; Guala et al., 2012; De Cremer, 1999, Bohnet&Frey, 1999). Two broad classes of approaches have emerged: one argues that decreased social distance affects preferences, such that "closer" counterparts come to positively regard one another's welfare (Fehr&Camerer, 2007; Fehr, 2003). We call this an "altruism hypothesis" (AH). Importantly, an AH posits that altruistic agents will attempt to benefit others, regardless how they expect them to behave. Indeed,

altruism can benefit active and passive recipients alike (i.e. as charity) and is hence a possible account of social but not (inherently) strategic behavior. Conversely, a second approach has focused precisely on expected reciprocity (RH) (Rabin, 1993; Charness&Dufwenberg, 2006) and it takes social distance to be "degree of reciprocity that subjects believe exist in an interaction" (Hoffman et al., 1996). A RH assumes particular relevance in light of the frequent presence of "conditional reciprocators", who are willing to cooperate with others conditional on their beliefs that others will do the same (Camerer&Fehr, 2006; Suzuki et al., 2011; Chang et al., 2011). Indeed, both AH and RH could, in principle, contribute to the observed polar effect of social distance on coordination: reasoning by extremes, a completely selfless/altruistic agent would always risk in SHs, and never do so in EGs, as these strategies maximally benefit one's counterpart. On the other hand, the belief that others will reciprocate one's actions would also alleviate uncertainty when actions are to be matched (as in SHs), but aggravate it when they are to be decoupled (EGs). In this study we thus aimed to weigh the potentially separate contributions of non-strategic altruism and strategic reciprocity.

Intriguingly, evidence from social neuroscience and neuroeconomics appeared to trace a similar functional division, within the "default/mentalizing network" (Amodio&Frith, 2006; Buckner et al., 2008), between the medial prefrontal cortex (mPFC) - in particular its ventral portion (vmPFC) - and the temporo-parietal junction (TPJ). Specifically, the v/mPFC is held to be critical for the formation of recursive beliefs in strategic interactions (Hampton, 2006; Coricelli&Nagel, 2009; Yoshida et al., 2006). Furthermore, fortifying this view of self-projection to closer others, the vmPFC has been shown to be recruited when agents think about themselves and similar others (Mitchell et. al., 2005; Jenkins et al., 2008), or friends (Fareri et al., 2012; Krienen et al., 2010), but not dissimilar others or

strangers. On the other hand, imaging studies on altruism have consistently implicated the TPJ (Tankersley et al., 2006; Morishima et al., 2012; Tricomi et al., 2010; Young et al., 2010). This raised the possibility of dissociable contributions of altruism and expected reciprocity to the polar effect of social distance on coordination, both in behavior and the brain.

4. 1. Closeness in the brain

4.1. 1. Common beliefs

Above, we reviewed evidence that social closeness in terms of common social categorizations (i.e. such as common group membership, friendship, interpersonal similarity) can induce agents to attribute their own beliefs and mental world to others, apparently drawing on self-knowledge to mentalize about them (i.e. Krueger, 2008; Ames, 2004; Goldman, 2005; Morry, 2007). Now, if we rely on related information structures to mentalize on ourselves and similar, but not dissimilar others, and related information structures tend to be co-localized in the brain, this affords a prediction: we should observe overlapping neural regions to be recruited when subjects mentalize about themselves and similar others, and less for dissimilar others. This was observed in a study by Mitchell et al. (2006). We briefly illustrate.

In a first study, Mitchell et al. (2006) told participants that the aim of the experiment was to assess how well they could use a small set of information to infer a number of characteristics of 2 target strangers. One of the targets was described as stereotypically conservative and the other as liberal, such that, supposedly, each participant

spontaneously felt more similar to one than the other. Subsequently, in the scanner, subjects were asked to answer the same set of unrelated questions (i.e. "Does target X like to go back home for thankgiving?") about i) themselves, ii) the conservative other and iii) the liberal other. After scanning subjects provided a behavioral measure of how much they implicitly associated with either the liberal or conservative other, in turn enabling the investigators to divide them into "similar to liberal" and "dissimilar to liberal" subjects. Mitchell reported that answering questions about one's self (self-mentalizing) recruited the ventromedial prefrontal cortex (vmPFC) in both groups. This region has consistently been implicated in a number of tasks tapping on self-referential knowledge (for a meta-analyses, see Northoff et al., 2006). Importantly, the same region was recruited when answering questions i) about liberal targets for the "similar to liberal" group only and ii) about conservative targets for the "dissimilar to liberal group only. Furthermore, for both groups, a distinct region in the dorsomedial PFC (dmPFC) was recruited for the respective dissimilar others (Fig. 11).

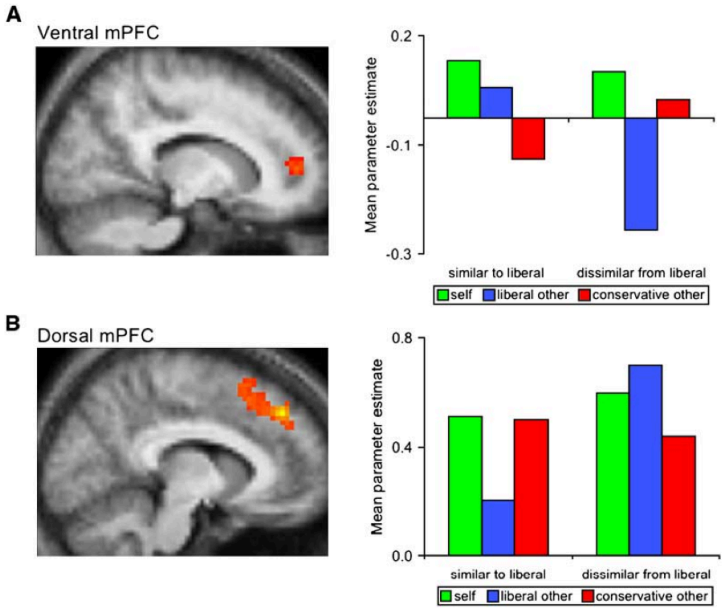

Fig. 12. Mitchell et al. (2006) show that the vmPFC is similarly recruited to answer questions about one's self and similar but not dissimilar others. The dmPFC exhibits the opposite pattern.

A second study (Jenkins et al. 2008) replicated these results with repetition suppression. In the latter study a behavioral measure of self-projection was observed: subjects tended to attribute to similar others the same answers they gave for themselves. Mitchell takes his results to suggest that similarity has a moderating role on self-referential mentalizing. Krienen et al. (2010) showed that similar effects on the brain also take place when subjects answer questions about themselves and their friends, independently of similarity (see discussion). Indeed, the idea of an overlap between other-related and self-related processing has spurred a long list of studies. In a recent meta-analyses (Denny et al., 2012) 107 imaging studies involving self vs. other related contrasts were compared. The results confirmed Mitchell's findings in that they revealed a gradient along the medial wall of the PFC, such that, relative to other-related information, self-related information seems to more commonly recruit a more ventral portion of the mPFC (Fig. 13)

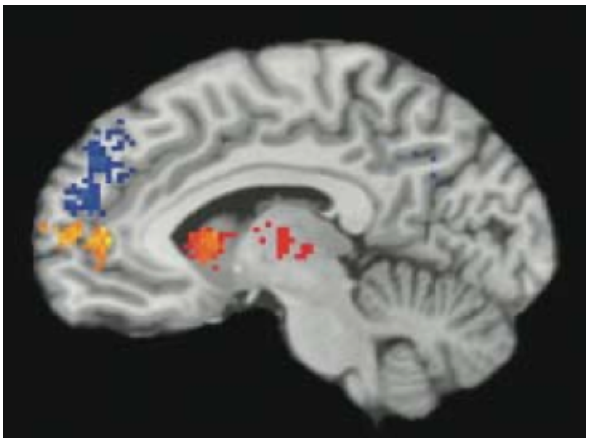

Fig. 13. Results of a meta-analysis (Denny et al. 2012) on 107 imaging studies involving contrasts of self vs. other-related information. Evidence for a gradient along the medial wall of the mPFC: ventral portions are more frequently recruited for self, rather than other-related information, while the opposite contrasts often show activity in more dorsal portions.

The question of why the vmPFC may contribute to the formation of beliefs for self and closer others remains open. Indeed, what makes self-related information specific? One possibility is that self-relevant information is more accessible: it is usually retrieved faster (Kuiper&Rogers, 1979) and with more confidence (Bower&Gilligan, 1979) than any other type of information. Related to this, self-relevant information may tend to be richer than other types of information. In this perspective, vmPFC activity could be a neural signature for information that is being processed in greater depth. For instance, Coricelli&Nagel (2009) showed that it was recruited by subjects that adopted higher degrees of recursion in economic games.

4. 1. 2. Common preferences

Interpersonal attraction and reward have a common behavioral manifestation, namely, approach. Indeed, the vmPFC is part of the dopaminergic mesolimbic pathways, which have been strongly implicated in mediating a value signal (cfr. Schultz, 2000; Haber&Knutson, 2011) and its diminished functioning has been associated with a number of pathologies related to decision making (for a review see Chierchia&Coricelli, 2011). In social contexts, the vmPFC has been consistently implicated in mentalizing (Mitchell et al., 2006), moral reasoning (Moll et al., 2005), empathy (Jackson et al., 2005; Shamay-Tsoory et al., 2003), cooperation (Rilling, 2000) and, in brief, notions related to a social value signal (Fehr&Camerer, 2007). Its reduced functional connectivity with other regions has been observed in patients with autistic spectrum disorder (Kennedy&Courchesne, 2008), and even its sheer volume has been associated to social competences and size of social networks (Lewis et al., 2011). However, with regards to the link between this region and interpersonal attraction, one study explicitly focused on the perception and subsequent

approach behavior towards *liked* others (Güroğlu et al. 2008) - which was, together with similarity, a necessary factor in affecting coordination rates in our experiment. In their study members of an orchestra rated how much they liked/disliked i) each member of the group, ii) a series of celebrities and iii) a number of objects. During scanning, they performed a social interaction simulation task: a figure representing the subjects was placed in the center of a screen, which they were asked to imagine as a room. At the top of the screen was represented either a picture of an orchestra member, a celebrity or an object. Subjects simply had to move a lever "up" towards the picture to indicate approach, "down" to indicate avoidance and "left" to indicate neutrality. Clearly, subjects tended to approach liked elements (of either of the 3 classes), however, interestingly for our purposes, the vmPFC was recruited preferentially for liked peers, relative to liked celebrities and objects, implicating this region in interpersonal attraction. In Mitchell's experiment as well (2006), positivity IAT scores, a measure of the association between liberal/conservative others and positively/negatively valenced words correlated with vmPFC activity when inferring the preferences of similar and liked others. Apparently, the tight correlation between liking and similarity is hard to tease a part, even at the neural level. One attempt was by Mobbs et al. (2009). In their task, subjects viewed 2 fictive game-show contestants who, before taking part in the game, answered questions about themselves in such a way to appear either likeable or unlikeable. Subjects rated both how much they liked each of the participants and how much they felt similar to them, indeed the 2 correlated. In the scanner, subjects than watched the contestants win/lose monetary awards. The ventral striatum and the vmPFC expressed both higher activity and connectivity when subjects either won in the 1st person, or observed liked others winning, as opposed to disliked others winning. Intriguingly, the strength of the coupling between the 2 regions scaled not with liking-rating but with the difference between the similarity

and liking ratings. This led the authors to speak of a key role for similarity in vicarious reward. Similarly, Fareri et al., 2012, found that ventral striatum and the vmPFC were more active when subjects won monetary prizes to be shared with their friends, rather than strangers.

Such results raise an important question. What drives effects of social closeness on coordination? The fact that closer others have better mutual understanding, that they can project to one another's minds with greater ease and depth (Coricelli&Nagel, 2009)? Or rather the fact that they care for one another and intimally share one another's rewards and punishments? Plausibly, both dimensions of social closeness contribute to interactions, however, no study has systematically attempted to disentangle between the two. One way to do so is to force friendly and stranger counterparts to randomize over their choices. If subjects still exhibit increased cooperation for friends as opposed to strangers, it would have meant that non-strategic motivational components of closeness are sufficient to drive cooperation. In any case, this idea promised to dissociate between behavioral and neural signatures of closeness-based altruism and mutual understanding, the former aligning better with positivity/liking, the second with similarity.

4. 2. Material and methods

The games

To investigate the effect of social closeness on coordination, we re-adapted the 2 games from our previous study (Chierchia&Coricelli, under revision) - which in turn had been adapted from Heinemann&Nagel (2006): one game involved strategic complements (Stag hunts – SHs), in which agents have an incentive to match their choices; the other involved

strategic substitutes (Entry games – EGs), which should lead agents to attempt to decouple their choices. As in our previous studies, we attempted to keep the superficial aspects of the 2 games as similar as possible, so that any behavioral difference would be due to their structural (incentive-related) differences. The games were as follows: in both games, 2 agents had to choose between the same two options: a high paying "risky" option, always worth €15.00 or 0, and a lower paying but safe payoff (SP) of a given € amount (SP ≤ 15.00). Both games capture a frequent scenario, namely, that low gains can be obtained in isolation. Indeed, if the SP was chosen, it was obtained for sure, regardless the choice of one's counterpart. High paying outcomes on the other hand required coordination and risk: in SHs, $/€15.00 were obtained, by both players, only if both risked, while if only one risked, he obtained 0. In EGs, on the other hand, the high gain could *only* be obtained in isolation, thus if both risked, both obtained 0. Then, by varying the value of the SP and having participants choose at each (randomized) step we obtained a behavioral measure of uncertainty in the 2 games. Importantly, since initial coordination patterns usually determine the outcome of their repeated versions (Heinemann et al., 2004), and since we were here interested in the way social distance biases choices rather than how it may bias learning (Fouragnan et al., 2013), no feedback on the outcomes of decisions was provided until the end of the experiment, when one choice was extracted at random and paid (in addition to a flat "show up fee" of $/€5.00).

Social distance

To assess the impact of social closeness in such games, participants played each of the same games twice (with identical payoffs), once with a friend who came to the experimental session with them, and once with a anonymous stranger, who had

previously taken part in the experiment[12].

Non-strategic conditions

To disentangle between alternative explanations of closeness-based cooperation (illustrated above), we included an additional "non-strategic" condition to the standard strategic games. Here, players were informed that their counterparts' choices would have been determined by a coin flip. Strategic and non-strategic trials were visually only differentiated by the word "flip" or "choice" (Fig.) and were identical in their payoff schedules. For instance, in a non-strategic SH, 1 participant makes a deliberate choice between risking or choosing the SP. The other participant has 50-50 of being assigned either one or the other. If the choosing player takes the risky option, there is a 0.5% chance of both players earning the maximum. With the complementary probability the choosing player will earn 0, but the non-deliberating player will still earn the SP. It follows that in non-strategic flip trials, deliberating players retained the possibility to benefit their non-strategic counterparts, by risking more in SHs and less in EGs, however they can't do so based on expected reciprocation. Since it is possible that some closeness-based cooperation is non-strategic, we were interested in comparing the 2 cases. An identical polar effect of friendship on risk in strategic and non-strategic conditions would have suggested that even the simplest preference-based models (i.e. 1st order altruism) could be sufficient to explain the impact of social distance on interactions in SHs and EGs. On the other hand, observing differential effects in strategic and non-strategic conditions (in terms of an interaction between friendship and strategy) would have successfully

[12] For the first pair of friends we were ready to ask them to come back for payment, however a "friend condition" was extracted so we were able to pay them immediately.

dissociated and quantified the respective contributions of the 2 mechanisms to closeness-based interactions. This yielded a 2 (game: SHs vs. EGs) x 2 (social distance: friend (F) vs. strangers (S) x 2 (strategy: strategic (s) vs. non-strategic (ns)) full factorial within subject design; in addition to which, to control for the effect of social distance, agents also played each game with a computer (C) that made random selections among options, thus adding 2 experimental cells, for a total of 10 experimental cells.

Stimuli

For each of the 10 experimental cells, participants made 31 decisions between a risky option (always worth €15.00 or nothing) and a safe option (SP), which varied 31 times between $0 \leq SP \leq 15$, in steps of 0.5, for a total of 310 decisions. The order of the SPs and the order of the conditions was pseudo-randomized between subjects. Furthermore, pilot studies had shown switching between SHs and EGs to be confusing. Thus, to minimize task-switching costs, the 2 games were played in separate runs (2 consecutive runs for each game), the order of which was counterbalanced across participants. No feedback on decisions was given until the end of the experiment. Participants had 6 seconds to make each choice. As soon as they did, a white frame highlighted their selected option for 500ms. After this, a fixation cross appeared, the duration of which covered the remainder of the 6 s decision time, plus a jittered interval (min 1.5s, max 6s, log-normally distributed) (Fig. 14). Stimuli were prepared and administered in psychtoolboox (Brainard, 1997; Pelli, 1997).

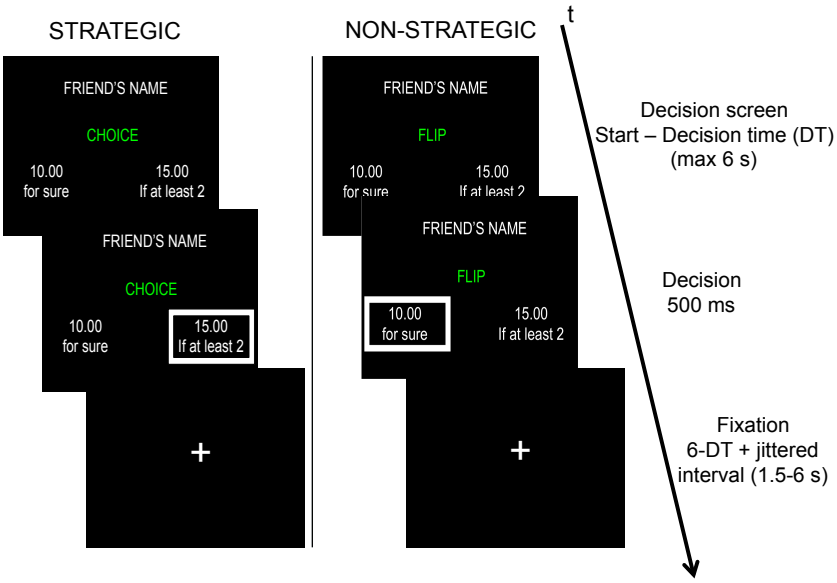

Fig. 14. Task structure. Participants played 2 types of coordination games: stag hunts (SHs) (shown above) and entry games (EGs) (in which the "if at least 2" label became "if at most 1"). In both, they chose between a generally low paying sure payoff (SP) and potentially higher paying but risky one (worth always €15.00). In SHs, the high payment was obtained only if both players chose to risk, thus matching their choices. If one player risked and the other didn't, he earned 0. In EGs, a player would only earn the high payoff if she was the only one choosing it, thus players were to decouple their choices. If both chose the risky option, both earned 0. Above a SH is shown, in EGs the risky option was labeled "it at most 1". Participants played each of the games repeatedly without feedback, for different values of the SP. They did so twice: once with a friend playing outside the scanner, once with an anonymous participant ("stranger"). To dissociate between altruistic (non-strategic) motives and reciprocity (strategic motives), in "flip" trials, counterparts' choices were determined at random. As a control, in "computer" conditions, participants made choices with a randomizing computer.

Participants and procedure.

19 participants (14 female[13]) of the University of Regensburg were required to bring a same-sex non-romantic friend to the experimental session (mean age = 22.3, s.d.=0.6). One participant per friend-pair underwent *f*MRI, while the other took part in the same task

[13] Our previous studies had shown that, at least with regards to behavior, there is no detectable gender effect of closeness in SHs and EGs.

outside the scanner. Most of the 38 participants (26) were in the human sciences (for the larger part psychology), none were economists. 14 participants had "ever heard" of game theory - those who had, had studied it, on average, less than 4 hours in total (s.d.=3.7). On average, friends had known each other for 6-12 months (17 for more than 1 year, none for less than a month). When asked how frequently they had regularly seen one another in the past 6 months (on a scale from 1 to 5: 1=every day – 5 = once a month), the average score was 2.1 (s.d.=1.3). As soon as friend pairs entered the room, they were separated and given written instructions. They both then took part in a questionnaire that probed their understanding of the task. Only upon correct completion did we proceed, in which case, both participants were informed that both them and the other participants who previously took part in the experiment had correctly answered the questionnaire and were thus clear on how the games worked. Participants were informed that they wouldn't have known the outcome of any of their interactions before the end of the experiment.

Incentive-compatibility

Participants were informed of the following. At the end of the experiment, 1 trial would have been selected at random and paid according to actual choices. If a stranger condition was selected, payoff depended on the corresponding choice (for the exact payoff matrix) of a randomly designated participant (friend excluded), who also had played in "stranger" matching, and who was awaiting payoff. If a non-strategic condition was extracted, then 2 subsequent 50-50 coin flips determined payoffs: 1 selecting which of the 2 participants was to be the "deliberating" one, and a second determining the "choice" of the non-deliberating participant.

Questionnaire

After taking part in the tasks of interest, participants responded to a questionnaire consisting of several parts: 1) the two "McGill Friendship Questionnaires for late adolescents and young adults" (Mendelson&Aboud, 1999), one tapping an "affective" component (MFQ-A) of friendship (i.e. how much friends "like" each other), the other on the fulfillment of certain friendship functions (MFQ-F) (i.e. "reliable alliance" of friends); 2) The “social value orientation task” (SVO, Murphy et al. 2011). In this task, participants are to choose between different predetermined ways in which to distribute hypothetical monetary sums to one’s self and another person. Some of the choices are rigged to as to allow for competitive behavior to emerge (i.e. such that one would be willing to lose money to decrease the payoff of a stranger), rather than altruistic behavior (i.e. in participants, to different extents have been shown to be willing to incur costs to benefit others). In synthesis, the SVO is held to yield a measure of pro-sociality. We had participants respond to it twice, once deliberating on allocations to strangers, the other to their friend; 3) a novel "perceived similarity" questionnaire, in which participants had to score, relative to 0 (the average University of Regensburg), how much i) they identified in each of 20 positively valenced personality traits, and ii) how much they thought each of the same traits represented their friends; 4) a risk-attitude questionnaire, in which we assessed subjectively accessible beliefs about i) the risk propensities of one's friend and ii) those of the average USC student; 5) In a reduced version of the same task games they had just concluded, we asked participants to state their *beliefs* about either a stranger’s or their friend’s choices for given SPs (when playing respectively with other strangers or their friends, that is, the subjects themselves). Finally, 6) we asked participants whether it made a difference for them to play with their friend or a stranger, and why.

Behavioral analysis

We analyzed dependent variables with generalized linear mixed effects models (GLMMs), as implemented in the lme4 package (Bates & Sarkar, 2006), in the R environment (R Development Core Team, 2006). We would start from the most complex models including all terms of interest and allowing all their possible interactions. Nuisance variables and covariates (gender, run, trial order, block order) were also included in the models, though they were not left free to interact with our interest variables, since this made the models computationally intractable[14]. We then proceeded as follows: non-significant higher-level interaction terms were progressively excluded, by comparing the nested models via likelihood ratio tests (Baayen, 2008). We did this until no factor could be justifiably excluded. In the main text, we always report effect sizes and p-values of the effects of interest from such best-fitting models. Furthermore, since all our designs involved repeated measures, all models included 1 random-intercept term (Baayen, 2008), thus accounting for the fact that responses provided by a same subject were not independent. In the experiments involving friends, an additional random effect term was incorporated to nest friends within friend pairs, so to account for eventual dependencies among friends. As additional controls, once a satisfactory model was found, random slopes were also progressively added to ensure that none of the observed effects of interest was due to inter-individual variability in the susceptibility to our factors of interest. Finally, extreme residuals (larger than 2.5 standard deviations) of the final models were excluded, and the model re-ran so as to control that none of the reported effects was due to extreme values.

With this statistical procedure, we ran the following models, all of which included our basic terms of interest and their interactions: game, the SP covariate, friendship and

[14] In other words, our reported conclusions control for direct effects of nuisance variables on the dependent variable, however they do not exclude that these may have indirectly affected them, i.e. through interaction with some other term.

strategy. In a 1st behavioral model on choice (BMC1) we dropped the computer control condition, in order to have a balanced 2 (player: friend/stranger) x 2 (game: entry/stag) x 2 (strategy: strategic/non strategic). This model was also implemented to investigate effects on response times (BRT1). In a second model on choice (BMC2), to allow comparison with the computer condition, strategic cells were dropped, thus yielding a 3 (player: friend/stranger/computer) x 2 (game). In a follow-up RT model (BRT2), the term "choice" (whether participants had chosen the safe or risky option) was also added. Since however "choice" is theoretically a non-manipulable, though we treat it here as an independent variable, we nested this term within the subject random effect term, so that different risk rates between subjects could be accounted for as random slopes, and their contributions to the main effects consequently discounted (http://www.ats.ucla.edu/stat/r/pages/mesimulation.htm).

MRI acquisition

Imaging was performed on a 3T head-only scanner (Siemens Allegra, Siemens, Erlangen, Germany) equipped with a single-channel head coil. 4 functional runs of 400 volumes consisting of 34 axial slices were acquired with a standard T2*-weighted echo-planar imaging sequence (repetition time TR = 2 s, echo time TE = 30 ms, flip angle FA = 90 °, 64 x 64 matrix, in-plane resolution 3 x 3 mm^2, slice thickness including gap 3.45 mm). To allow for saturation of the magnetic field the first 10 (can't remember, there are a few dummy scans at the scanner anyway and we excluded the first four saved volumes?) volumes of each run were discarded from preprocessing and analysis. Between the 2nd and the 3rd functional run an anatomical T1-weighted volume with 160 saggittal slices was measured using an MP-RAGE sequence (TR = 2250 ms, TE = 2.6 ms, FA = 9 °, 240 x 256 matrix, voxel size 1 x 1 x 1 mm^3).

Preprocessing

Preprocessing and fMRI data analysis were conducted with SPM8 (Wellcome Department of Imaging Neuroscience, London, UK) running under Matlab 7.5 (Mathworks, Natick, MA, USA). In the first step individual structural volumes and corresponding functional series were reoriented parallel to the AC-PC line (a line running through the anterior and posterior commissures) and the intrahemispheric fissure with the origin set to the anterior commissure, according to definitions of the Talairach space (Talairach & Tournoux, 1988). Functional volumes were then slice-time corrected with the temporally middle slice serving as a reference and realigned to the mean volume of the four runs. After coregistration onto the functional mean the structural volume was segmented and normalized into MNI standard space using unified segmentation (Ashburner, & Friston, 2005). Corresponding normalization parameters were reapplied onto the functional series, which were resampled to a voxel resolution of 2 x 2 x 2 mm^3 and spatially smoothed by an isotropic Gaussian kernel of 8 mm FWHM. Data quality and subjects' motion was checked with ArtRepair V4 toolbox (Mazaika, Whitfield, & Cooper, 2005). Fast head motion, defined as scan-to-scan motion above 0.5 mm in more than xyz % of the volumes, lead to exclusion of two subjects.

MRI analysis

General linear models (GLM) were used to fit hemodynamic signal to our experimental design. GLM1 included 11 in each run, 5 for our conditions of interest (Fs = strategic friend, Ss = strategic stranger, Fn = non-strategic friend, Sn = non-strategic stranger and Cmp = computer) and 6 motion realignment parameters (the 4 additional regressors corresponded to the average run-specific average signal). Corresponding boxcar functions

defined by trial onsets and trial-specific duration were convolved with the implemented first-order canonical hemodynamic response function. When necessary, an additional regressor was added to model trials in which subjects did not respond within the response window of 6 s. Slow signal drifts and temporal correlations between the residual errors were removed with a high-pass filter of 1/128 Hz and an auto-regressive AR(1) model. From GLM1, 2 analyses were carried out, GLM1_A1 and GLM1_A2.

GLM1_A1. Paralleling our behavioral analysis, a first analysis focused human conditions only (thus dropping computer conditions) in order to have a 2 (player: friend/stranger) x 2 (strategy: strategic/non-strategic) x 2 (game: EG/Sh) within subject design. For each of these "factors" differential contrast images of the two levels were generated on single-subject level using t contrasts. Contrast images corresponding to the three two-way interactions and the three-way interaction were built in a similar way. Group analyses were then conducted with one-sample t-tests based on the contrast images.

The initial voxel threshold was set to puncorr < .001 with an extent threshold of k > 100 voxels [depends on your preferences, choose something between say 100 and 150] for analysis 1 and a cluster threshold of pcorr < .05 (FWE) for analysis 2. The statistical threshold for post-hoc tests was set to .001, reflecting the initial voxel threshold. Anatomical labels were derived from the Anatomical Automatic Labeling (AAL) toolbox (Tzourio-Mazoyer et al., 2002), reported coordinates are in MNI space. Activation maps were mapped onto the population-average landmark- and surface-based (PALS) standard brain (Van Essen, 2005) with Caret 5.6 (Van Essen et al., 2001; http://brainvis.wustl.edu/wiki/index.php/Caret).

GLM1_A2. To incorporate the computer conditions, a second analysis focused on non-strategic conditions only. Contrast images for the 6 non-strategic conditions (3 player: friend/stranger/computer x 2 games) were calculated on single-subject level in SPM, averaging across the two runs of a game, and subjected to a random-effects 2x3 ANOVA with factors game and player. The group statistics were estimated with GLM Flex (http://nmr.mgh.harvard.edu/harvardagingbrain/People/AaronSchultz/GLM_Flex.html), a collection of scripts that allows for statistically valid within-subject ANOVAs using partitioned error terms (see McLaren et al., 2011). Significant clusters in F contrasts were further analyzed with post-hoc tests. For this purpose beta estimates averaged across the cluster were extracted individually for the different levels using MarsBaR 0.42 (Brett et al., 2002) and then tested with paired t-tests in SPSS ().

2 further GLMs were conducted to analyze the impact of 2 further regressors of interest, both of which used the same regressors as GLM1 but added cell-specific parametric regressors representing i) SPs (GLM2) and ii) payoffs (GLM3). Monetary payoffs were calculated as follows: when a participant chose the SP, the given SP was attributed to him, while when he risked, earnings were computed in expected value (EV), given random matching to any of the other participants for that specific trial, according to the following formula:

$$EV^{SH}(Risk_j)=15*p(Risk_{-j})$$

$$EV^{EG}(Risk_j)=15*[1-p(Risk_{-j})]$$

This says that the EV of an agent when risking, is a positive (in SHs) or negative (in EGs) linear function of the relative frequency (*p*) with which the other agents (*-j*) also chose to risk for that specific SP, in that specific condition. The 2 regressors were modeled in separate GLMs, because presented correlations.

4. 3. Behavioral results

BMC1 (see behavioral analysis section) on the total 38 participants revealed a 3-way interaction between game, social distance and strategy ($p<0.01$) as well as a 4-way interaction that additionally included SPs ($p<0.01$). Replicating our previous findings (Chierchia&Coricelli, under revision), friends were estimated to be more likely to risk than strangers in SHs (mean difference=11.8, s.e.=3, $p<0.0001$), while, they were less likely to risk at low SPs in EGs (mean slope difference=0.05, SE=0.02, $p<0.05$) (Fig 15). Our addition of non-strategic trials proved to be effective, as playing with randomizing others decreased risk rates in SHs ($p<0.001$) and increased them in EGs ($p<0.001$). Interestingly, even when counterparts' choices were random, friends still risked more in SHs ($p<0.01$) and less in EGs ($p<0.05$). However, this difference was amplified in SHs ($p<0.05$). In EGs on the other hand, though the interaction between social distance and strategy did not reach significance in term of risk rates, it did in terms of response times ($p<0.01$) (see BRT2 in behavioral analysis section above). Indeed, in EGs, participants were quicker to choose the safe option with friends in comparison to strangers ($p<0.05$), though this only occurred when interactions were strategic. This suggests that different processes may have driven the similar choice patterns in strategic and non-strategic EGs.

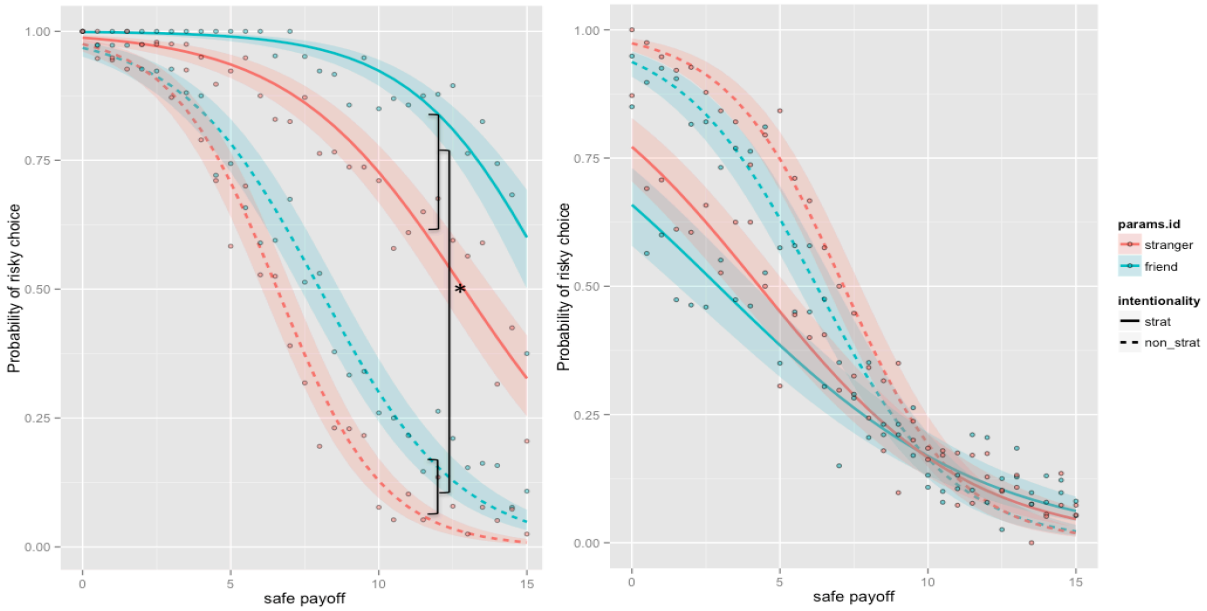

Fig. 15 A GLMM (BMC1, see behavioral analysis section) estimated probability of risking (y-axis) in SHs (left panel) and EGs (right panel), given SPs (x-axis), when playing with friends (green) and strangers (red) in strategic (full line) and non-strategic (dashed line) condition. Error bands represent 95% confidence intervals. Points are observed percentages of risky choices, for each SP. In non-strategic conditions, counterparts' choices were determined by a coin flip. The graph shows that when payoffs are aligned (in SHs), then agents risk more when they interact with a motivated counterpart, than when they play against chance (full vs. dashed lines, left panel). Conversely, when incentives are in conflict (as in EGs), then players risk less with a motivated counterpart, than against chance (full vs. dashed lines, right panel). Within both strategic and non-strategic conditions, the same pattern differentiates play with friends and strangers: friends risk more in SHs, and less in EGs. The square brackets in the left panel indicate the significant interaction between strategy and social distance in SHs ($p<0.05$). All curves are differentiable at $p<0.05$.

Finally, to incorporate our non-social computer condition, and to control for residual effects of social closeness, we ran the above model while focusing on non-strategic conditions only (since computer counterparts always randomized) (BMC2, see behavioral analysis section above). The model confirmed and extended the previous results: in SHs, players risked more when the outcomes of their decisions affected their friends, rather than both strangers and computers ($p_s<0.001$). In EGs, the opposite was observed, such

that friends risked less with friends than strangers or computers (p_s<0.001). However, interestingly, in both games, players risked at non-dissociable rates when their choices impacted the payoff of a stranger, or that of a computer (all p_s>0.2) (Fig. 16).

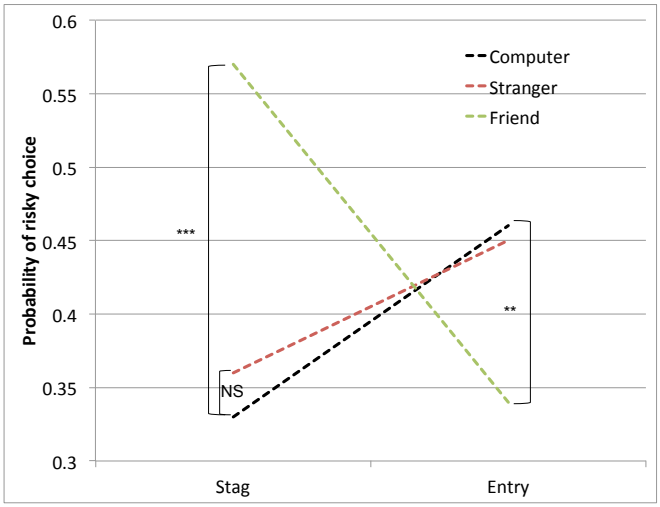

Fig. 16. Comparison with non-social "computer" control: a GLMM (BMC2) estimated probability of risky choice in non-strategic conditions. In stag hunts, players risked more when their choices affected their friends, rather than strangers or computers. In entry games, the opposite occurred. In both games, players risked at non-dissociable rates when their choices affected strangers or computers. ***p<0.001; **p<0.01; NS=not significant.

When restricting the above choice model (BMC1) to only the participants who underwent MRI (n=16), the interaction of strategy and game on probability of risking remained strongly significant (b=2.77, se=0.26, z=10.51, p< 2e-16) (Fig. 17).

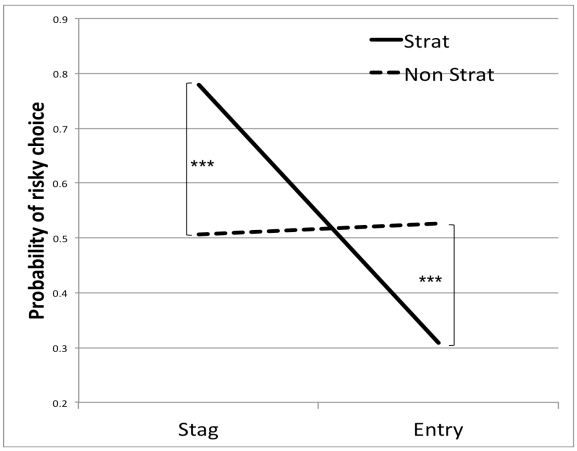

Fig. 17. Behavior in the scanner: interaction of strategy and game (BMC1). Interacting with a strategic, as opposed to a non-strategic (random) counterpart, raises risk rates when payoffs are aligned (Stag hunts), but decreases them in competitive environments (Entry games).

As did the interaction between social distance and game (b=-1.11, se=0.25, z=-4.300, p<1.71e-05): friends risked more than strangers in SHs (b=0.87, se=0.3, z=2.843, p<0.01), however the opposite effect was no longer significant in EGs (p<0.3)[15]. This latter finding wasn't too surprising, since our previous research (Chierchia&Coricelli, under revision) showed that, relative to the effects of social closeness in SHs, the corresponding effects of social closeness in EGs tend to be smaller, and usually require at least 20 subjects to emerge. In line with this interpretation of a power problem, when restricting the above model to "counterpart" participants (those who didn't undergo fMRI), we observed a quantitatively similar results as for when the model was restricted to MRI participants alone. Furthermore, the full including all participants showed that that there was no significant difference of choice behavior between MRI participants and their counterparts outside the scanner (p>0.8).

[15] Friend-pliant choices in non-strategic conditions were also eliminated in SHs, though they remained marginally significant in EGs (p=0.05).

Finally, RT analyses showed that behavior of MRI subjects alone had been affected by our manipulations in the predicted direction. A GLMM revealed a significant 3-way interaction between game, strategy and player (b=0.12, se=0.03, t= 3.11, p<0.01). Specifically, in both games, players were faster to reach a decision when interacting with their friends as opposed to strangers, provided the condition was strategic. Importantly however, players were not always facilitated when interacting with friends, rather, they were only facilitated to make opposite choices in opposite games. This was suggested by an additional GLMM that controlled for choice (risk or safe). This model revealed a significant 4-way interaction between game, friendship, strategy and choice (risk or safe) (t=3.79, p<0.001). What this interaction says, in synthesis, is i) that friendship facilitated *risking* in SHs (b=-0.18, se=0.02, t=-8.77, p<0.0001), while facilitating safe choices in EGs (b=-0.09, se=0.02, t=-4.111, p<0.01), and ii) that this pattern was disrupted in non-strategic conditions (both p_s >0.4) (Fig. 18).

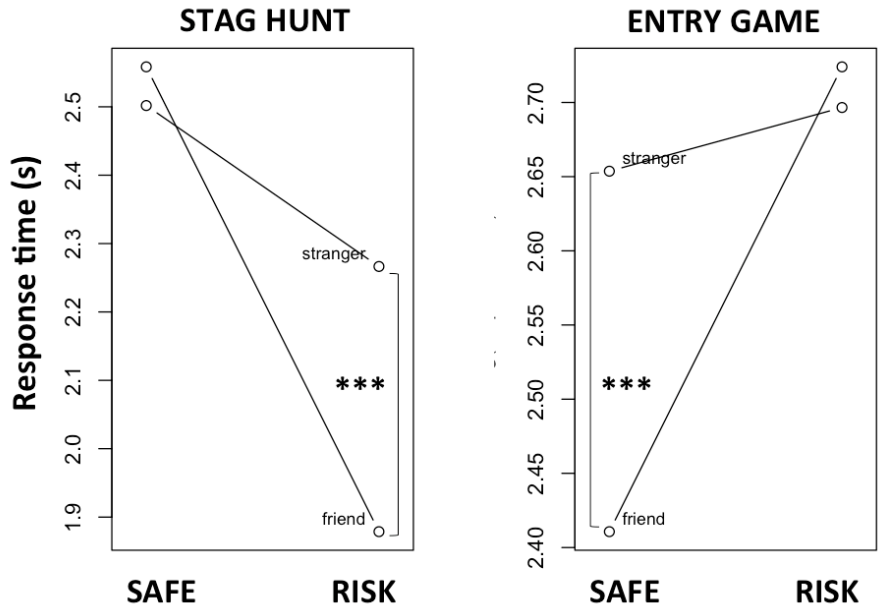

Fig. 18. Behavior in scanner (BRT2, see behavioral analysis section): interaction of strategy, game, choice and player. Collapsing across choices, a previous model (BRT1) showed that players were faster to reach a

decision when interacting with friends rather than strangers. The addition of choice however (BRT2) shows that this such a facilitation is completely due to opposite choices in the 2 games: in EGs, participants take less when they choose the safe option with friends, in SHs, they take less to risk. This pattern only occurs in strategic conditions.

4. 4. Imaging results

Analysis revealed neural areas that track social distance, independently of the presence of a strategy (main effect of social distance), areas that were more sensitive to the strategic vs. non-strategic nature of the interaction (main effect of strategy), and importantly, areas that were sensitive to the interaction of these factors. We report such results in turn.

The main effect of social distance was similar for both GLM1_A1 and GLM2_A2. We thus report effects for GLM1_A2, since this enabled comparison with the computer conditions. This model focused only on non-strategic conditions, and revealed a number of neural regions (Fig., Table 1). Of particular interest, bilateral TPJ ($t=743$, $p<0.001$) and dmPFC ($t=6.83$, $p<0.001$), nicely tracked social distance, independently of whether counterparts randomized or deliberated over their choices. Post-hoc-tests further revealed that a right-sided bias in such sensitivity, such that right but not left TPJ significantly differentiated between computer counterparts, strangers and friends (all $p_s<0.001$), as did the superior medial frontal cortex (Fig. 19).

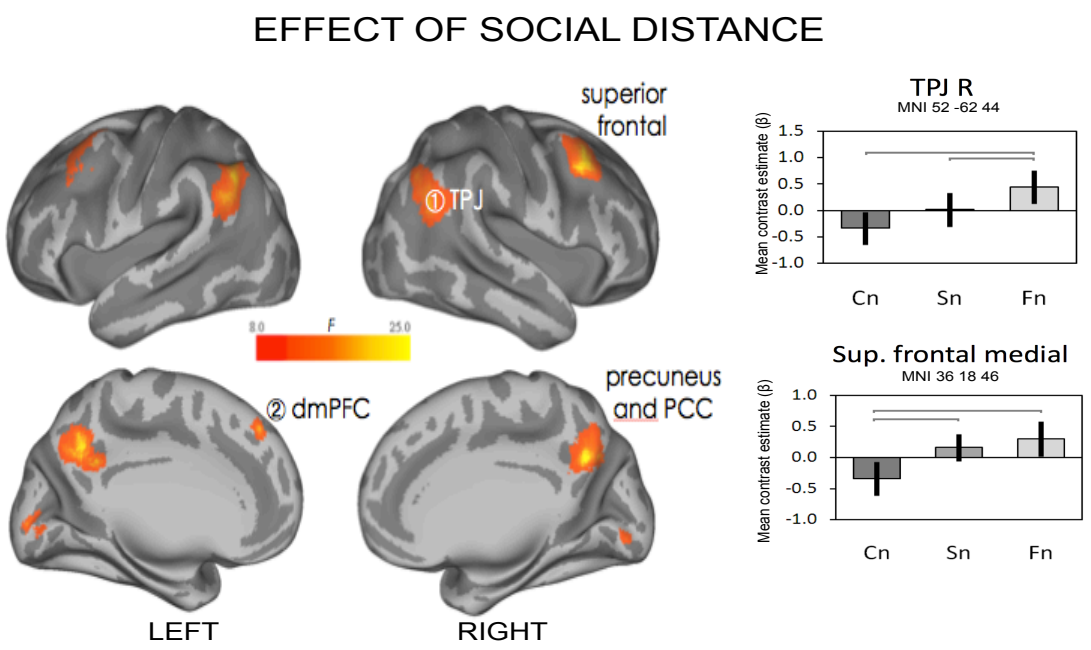

Fig. 19. Effect of social distance in non-strategic conditions (GLM1_A2). F-test, initial voxel threshold p_{uncorr} = .001, cluster threshold p_{corr} = .05 (FWE). Parameter estimates for selected clusters are based on peak voxels. Error bars correspond to 90% C.I

Anatomical	**Side**	**Label**	**x**	**y**	**z**	**k**	p_{corr}	***T***
Angular gyrus, inferior parietal, supramarginal gyrus, middle temporal	R	TPJ R	52	-62	44	1917	<.001	7.43
Middle temporal	L	STS L	-56	-34	-10	365	.002	7.09
Middle frontal, superior frontal	R		36	18	46	1 374	<.001	6.83
Precuneus	L/R		-4	-50	30	2 097	<.001	6.80
Middle temporal	R	STS R	62	-18	-16	340	.002	6.62
Inferior parietal, middle temporal, angular gyrus	L	TPJ L	-58	-58	28	1 153	<.001	6.07

Table. Main effect of social distance in non-strategic conditions

On the other hand, a rather distinct set of neural regions was revealed to be sensitive to our strategic vs. non-strategic manipulation (Table 2). GLM1_A1 showed that, in addition to the previously reported regions that tracked social distance, nearly the whole medial wall of the PFC (t=7.56, p<0.001) was recruited when counterparts deliberated as opposed to randomized between choices (Fig. 20)

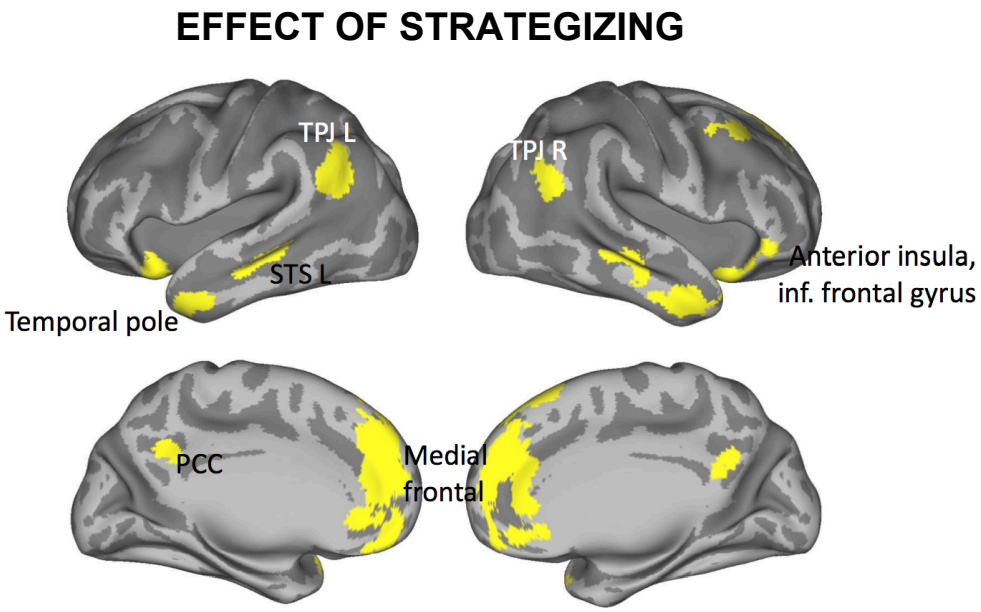

Fig. 20. Main effect of strategizing (GLM1_A2). F-test, initial voxel threshold p_{uncorr} = .001, cluster threshold p_{corr} = .05 (FWE).

Anatomical	Side	Label	x	y	z	k	p_{corr}	*T*
Temporal pole, middle temporal, inferior temporal	L		-52	6	-32	187	.040	7.74
Superior frontal medial, ACC	L/R		-6	46	36	4661	<.001	7.56
Angular gyrus, middle temporal	L	TPJ L	-50	-58	26	610	<.001	7.41

Middle frontal	R		36	22	42	211	.025	6.64
Inferior temporal, middle temporal, temporal pole	R		50	2	-38	279	.007	6.51
Inferior frontal pars orbitalis, insula	R		30	14	-20	381	.001	6.51
Middle temporal, inferior temporal	R		56	-22	-18	464	<.001	6.18
Angular gyrus, superior temporal, middle temporal	R	TPJ R	52	-58	26	429	.001	5.97
Middle temporal	L		-60	-18	-14	456	<.001	5.81
Insula, temporal pole, inferior frontal pars orbitalis	L		-32	16	-24	283	.007	5.52
Precuneus, PCC	L/R		6	-54	28	252	.012	5.25

Table 2. Neural regions recruited for strategizing

Finally, GLM1_A1 revealed that only a subset of regions of this network was sensitive to the interaction between social distance and strategy. In particular, the vmPFC (t=6.33, p=0.01) differentiated between friends and strangers only if the interaction was strategic (Fig. 21).

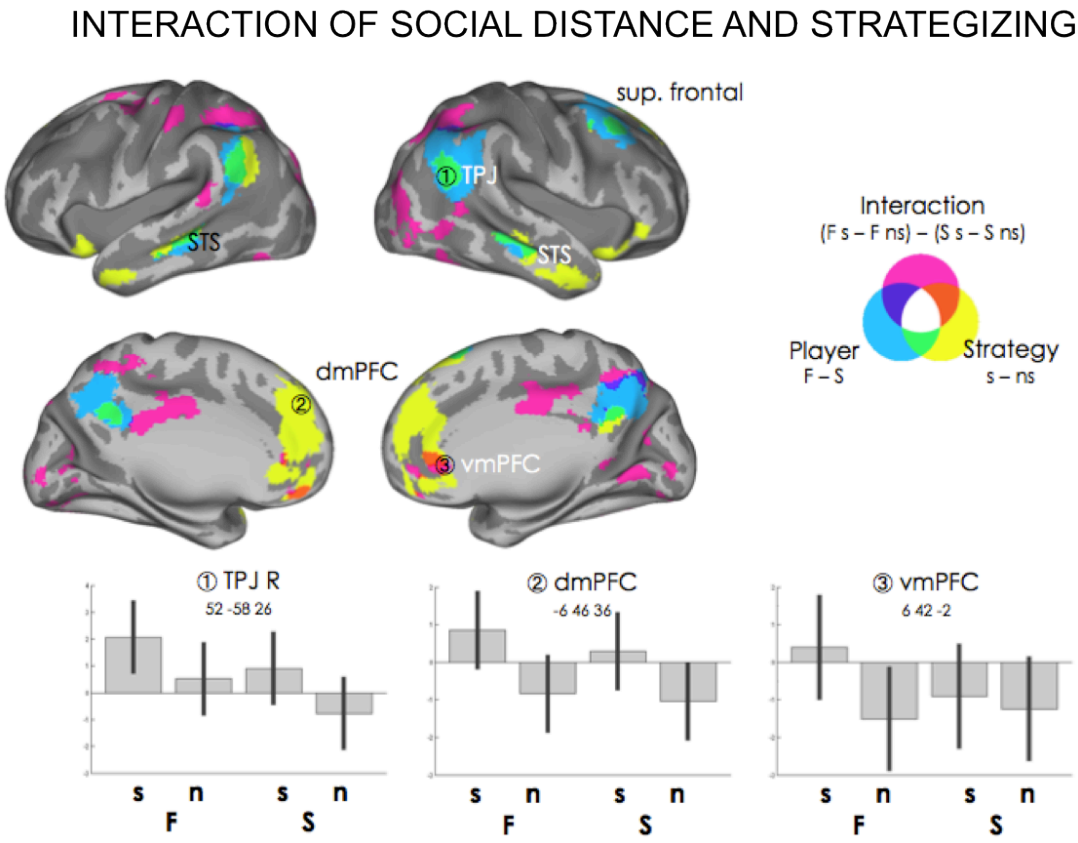

Fig. 21. Effects of Player, Strategy and corresponding interaction. The TPJ responds preferentially to social distance, dorsal regions of the mPFC to strategy, while vmPFC, middle cingulate cortex and PCC respond to their interaction. Results are based on t-tests, initial voxel threshold p_{uncorr} = .001, cluster threshold p_{corr} = .05 (FWE). Parameter estimates for selected clusters, are based on peak voxels. Error bars correspond to 90% C.I. F = Friend, S = Stranger, s = strategic, ns = non-strategic.

Anatomical	Side	x	y	z	k	p_{corr}	*T*
Parietal lobe	L/R	18	-66	42	4272	<.001	7.40
Middle occipital, cerebellum, middle temporal	R	28	-62	-44	1694	<.001	7.02
Lingual gyrus R, calcarine sulcus R/L, cerebellum R/L	R/L	20	-74	-44	1690	<.001	6.59

Rectus L/R, medial frontal pars orbitalis L	L/R	-8	54	-20	202	.011	6.33
Superior temporal	L	-58	-40	12	209	.009	6.15
Middle occipital, superior occipital	L	-26	-82	12	310	.001	6.02
Cerebellum, fusiform gyrus	L	-38	-68	-22	445	<.001	5.65
Paracentral lobule, SMA	L	-14	-22	60	146	.043	5.38
Middle frontal, superior frontal, precentral gyrus	R	36	2	44	174	.021	5.34
ACC L/R, medial frontal pars orbitalis R	R/L	4	40	-2	337	.001	5.28
Cerebellum, lingual gyrus	L	-14	-84	-18	326	.001	4.84

Table 3. Brain regions sensitive to the interaction of social distance and strategizing.

4. 5. Discussion

Social closeness has long been known to foster cooperation (Smith, 1759). Humans have been shown to cooperate more frequently with other humans, than computers (Kiesler et al., 1996), with friends more than strangers (Yamagishi&Sato, 1986), with similar rather than dissimilar others (Chierchia&Coricelli, under revision) and with ingroup rather than outgroup members (Charness et al., 2007; Chen&Chen, 200). In all of such cases, similarities between self and other are stressed. Such perceived closeness can favor

cooperation in two different ways: it can lead to "liking" others, so that they will be willing to altruistically incur more risk/costs, to benefit one another, or it can increase expected reciprocity, which decreases the perceived risk that others may defect. In both cases, agents may will be cooperating, but they will be doing so for fundamentally different reasons. Here, we aimed to weigh the respective contributions of such mechanisms. Replicating our previous results (Chierchia&Coricelli, under revision) we show that, in a game that requires agents to match choices, subjects risk more if they are playing with their friends, rather than strangers; while if the game requires to decouple choices, friends risk less. Here we extended those results by showing that even when subjects know that both friendly and stranger counterparts are randomizing their choices, they keep exhibiting this behavior, albeit to a lesser degree. This suggests however, that both expected reciprocity and altruism could work synergistically in promoting cooperation. Interestingly, we find the brain to rather clearly distinguish between these two components of closeness. A first important finding is that of an apparent gradient within the so-called mentalizing network (Saxe&Kanwisher, 2003; Van Overwalle, 2011), with nearly the whole medial prefrontal cortex being recruited for social interactions to a much greater extent when they are strategic (i.e. when choices are interdependent) than when they are not (that is when other people may be affected by our decisions, though we're not affected by theirs), and the TPJ being more finely tuned to social distance. The mPFC, especially its dorsal component has been linked to performance monitoring and uncertainty in many forms of abstract goal directed behaviors (i.e. Ridderinkoff, 2004). Through its connection with the hippocampus and dlPFC (Kim&Whalen, 2009) it is frequently considered to favor more cognitively mediated forms of perspective taking, in opposition to the vmPFC, which through its connections with the amygdala, accumbens and insula, is held to mediate more affect laden forms of mentalizing (Vollm et al., 2006;

Shamaay-Tsoory, 2009). One recent account (Venkatramman&Huettel, 2012) has implicated the dmPFC specifically in strategic control. Indeed, false-belief tasks, which require subjects to adopt a perspective that differs from their own - and are thus, by definition, more mediated - regularly recruit this area (Buckner, 2008), as well as the TPJ (Saxe&Kanwisher, 2003). One interesting addition of our study to this literature is that it shows that the involvement of the d/mPFC by strategizing isn't necessarily a mere matter of difficulty or uncertainty, which often may co-occur with strategizing. Indeed, strategizing in EGs raised difficulty (as suggested by the increased RTs) and decreased uncertainty (as subjects risked less) relative to its non-strategic analog, while the opposite occurred in SHs. In spite of these differences, contrasting strategic and non-strategic interactions in both of these games resulted in dmPFC activation.

The TPJ, together with the mPFC has been overwhelmingly involved in mentalizing (see Van Overwalle, 2012, for a review). At lower levels, it is has been consistently implicated in attention reorientation (i.e. Posner tasks) and stimulus driven control: when attention is focused on some external object (i.e. reading) and a novel event occurs (i.e. an unusual noise) that reorients our attention, the TPJ passes from deactivated to transiently activated. Interestingly however, it doesn't do so for just any event, for instance, it has been shown to not get "distracted" (activated) by highly salient but task irrelevant stimuli (Indovina&Macaluso, 2007). This involvement of the TPJ in tracking "relevant" environmental differences that require attention reorientation has been proposed to be at the basis of its involvement in reading the mind of others and perspective taking (Corbetta et al., 2008), especially when they differ from one's own. In line with this, there is a vast literature linking the TPJ to self-other differentiation and control (i.e. Decety&Lamm, 2007). A fascinating series of studies by Santiesteban et al. (2012) showed how excitatory

stimulation via tDCS on TPJ improved social abilities in a number of context involving co-representations of self and other. In one task, participants had to respond to a finger movement seen on a screen, with an incongruent finger movement (i.e. if the index moved, subjects had to move their middle pinky and viceversa). To better resist the tendency to imitate, they had to thus inhibit the other-related representation in favor of their own perspective. Stimulation on TPJ (relative to sham) increased accuracy and speed. Conversely, in another task, subjects viewed a matrix-like cupboard with a series of objects in a subset of its cubicles. Some of such objects (i.e. a small candle) could be seen from both the subject and a "director" on the other side of the cupboard; others (i.e. a big salient candle) could only be seen for the subject. The director then gave instructions to the subject of which objects to take. Contrary to before, in this task, subjects had to inhibit their own perspective and amplify that of the director. Again, TPJ excitation resulted in better performance. Finally, in a last important task, subjects were asked (i.e. as in Mitchell et al. 2005 or Krienen's tasks) to answer mental (on preferences) or physical questions about either themselves or others, so self and other representations were not co-represented. In a later memory retrieval task subjects exhibited typical self-referential biases (i.e. they were faster and better at remembering self-judgments) that were not been altered by TPJ stimulation. From this, it would almost appear that vmPFC and TPJ serve complimentary functions of, respectively, integrating vs. segregating self-other representations, which could, in turn lead respectively to self-projection (i.e. egocentric biases) or perspective taking. In line with this, and the previous claims on dmPFC, a meta-analysis by Denny (107 studies) (2012) showed that while the vmPFC was more frequently reported in self vs. other contrasts, both dmPFC and TPJ have been found more frequently observed in other vs. self contrasts. Finally, Tankersley et al (2007) had the intriguing idea that the TPJs involvement in agency detection, the ability to understand

behaviors as motivated (Castelli et al., 2000), could constitute a low-level determinant of altruism. In their task subjects either viewed others play a reaction time game, or played the game themselves. Indeed, TPJ was more active for when they watched others play, rather then when they played in first person. The magnitude of this effect predicted subsequent scores in altruism questionnaires. This very low-level role for TPJ in altruism was further neurobiologically grounded by a recent study by Morishima et al. (2012), which showed that inter-individual differences in altruistic tendencies were predicted by grey matter volume in the TPJ. For our task, we had hypothesized that regions that tracked social distance in a way that wasn't modulated by strategy would have been apt neural mediators of altruistic motives. The region we found is precisely the TPJ.

Finally, with regards to Mitchell's idea linking the vmPFC to similarity. Krienen et al. (2010) put it to the test with 4 fMRI experiments, in all 4, subjects answered questions about themselves and given target others. The first served to define a localizer in which ROIs were defined that preferentially responded for questions about one's self vs. president Bush. The second experiment found that previously observed vmPFC region was preferentially active for friends rather than strangers, regardless whether they were similar or dissimilar. The third dropped the friendship factor in order to assure that the previous failure of similarity to recruit vmPFC wasn't due to a "rescaling" of the similarity perception due to the highly salient similar friend. The fourth took the 2 extremes: dissimilar friends and similar strangers, and showed that vmPFC was much more sensitive to the former than the latter. The interpretation the authors give is that rather than similarity, behavioral relevance is the important factor. In Mitchell's defense - though the authors themselves admit this – their similarity manipulation used here was not effective. The authors had run into the same problem I found for my similarity experiment, that is,

to try to separate similarity from liking. In the attempt to provide similar targets that had only mildly likeable traits, they probably ended up making the similarity characteristics irrelevant or uninteresting. Indeed, it is clear that, as a feature, of similarity per se cannot mean much, because, virtually, there will always be some element by which 2 subjects are similar. Plausibly, context will determine which ones are likely to matter. For instance, if a drunk man comes towards me on the street with hostile intentions, I might not even notice that he's wearing the same shirt as I am, though in a different circumstance, i.e. at a dinner, the same shirt could be a conversation starter. What is interesting about social categorizations (i.e. minimal group paradigms) and similarity is not that they can be made irrelevant, rather the ease with which it they often spontaneously adopted by participants in search for social (and non-social) navigation devices. When we are first struck by a given similarity with someone, i.e. we both liked a given movie, and we feel that sense of approach, I don't believe it is because we are interested in object of similarity per se, rather I believe we are interested in the fact that the common object of interest promises other similarities and other common interests. Indeed, were we to follow up with questions (i.e. what scenes of the movie were liked? Or maybe probe the other's reaction to our favorite quote from the movie), we would feel pleasure to see our expectations confirmed. That these expectations are formed in the first place, and that they drive our questions and predictions seems to be characteristic of similarity and analogical reasoning (Vosniadou&Ortony, 1989).

Krienen et al. (2010) thus attempted to replace similarity with "behavioral relevance", though it isn't clear whether this was successful. A particularly interesting study by Nicolle et al. (2012) proposes to replace the "self" with behavioral relevance. They had participants take part in a typical time discounting task, in which they chose over varying

magnitudes of reward "now", over some larger reward in the future. Subjects were pre-tested and had known discount functions, which described their idiosyncratic preferences in regards to time discounting. They were then asked to learn the preferences of another person, who had different (but not anti-correlated) function. Then, in the scanner, they were to make choices for either themselves or said others. As predicted by a self/other distinction along the vm/dmPFC, while responding for themselves, vmPFC tracked their own subjectively discounted values, while simultaneously tracking, in the dmPFC, the preferred values for their counterparts. However, even more interestingly, when subjects were asked to answer for their counterparts, this pattern completely flipped over: now the vmPFC was tracking their counterparts' preferences and not their own. The authors' conclusion is that of an "agent independent axis" in the mPFC. I believe they make an excellent point. However, were subjects not "themselves" when responding for the other person? Were they not responding to their own contingent objective of complying to experimental demands? Are we not ourselves when we role-play? If we are, then Nicolle's study paradoxically seems to chain the self to the vmPFC, rather than freeing it.

From an overview our own results, and the existing literature, the perspective that emerges is certainly *not* that the vmPFC is selective for similarity (Mitchell et al. 2005, 2008). Rather, its activity can be usually taken to indicate that diverse functions, such depth of reasoning (Coricelli&Nagel, 2009) and reward-related signals (Schultz, 2000) are synergistically working to drive decisions. It is this particular mixture of features, if anything, that makes vmPFC somewhat specific in the brain. What many studies may thus pick up on then is the fact that "self"-related stimuli are *usually* processed in greater depth (Symon&Johnson, 1997), or are liked (Ferguson et al., 1983). Consequently, the same is probably true for stimuli that make us see ourselves in others, that is homophily. The

"coincidence" that we propose is that homophily or social closeness seem to bare a quite similar distinction: they not only motivate people to like and approach one another, but they also carry important information which can be often behaviorally relevant in interactions. Our results further qualify what this "behavioral relevance" (Nicolle et al., 2012) means with regards to important building blocks of strategic interactions (such as games with strategic complementarities and substitutability): the relevance of social closeness in situations requiring joint effort, will drift towards cooperation and risk, while its relevance in situations involving conflicting incentives will drift towards isolation and security. We show that, while agents exhibit this behavior in interactions, the vmPFC – likely through integration of information from different areas, such as the dmPFC and the TPJ - distinguishes between aspects of closeness related to altruism and reciprocity.

6. Conclusion

We opened with Rousseau's Stag Hunt, a metaphor, for characterizing the potential benefits, but also the risks of sociality: agents can achieve more by working together, however, to do so they need to depend on one another, which makes cooperation risky. Moreover, by living together, agents will incur more frequent competition over limited common resources, thus creating an additional potential cost of social life. With hindsight, humans "chose" to live socially. Our question was, in a typical "cost-benefit" analysis, what (neuro-cognitive) mechanisms could have tilted the balance towards sociality?

To illustrate this point we proceeded with 3 sections, which attempted to capture the problem at 3 apparently different levels and approaches: the game-theoretic level, the psychological level, and the level of neuroscience. In brief these sections were as follows:

Behavioral game theory and common knowledge

Behavioral game theory meshes the typical *deductive approach* of game theory, with empirical observation. To illustrate this, in 1.1., we very briefly described basic applications of game-theoretic concepts to strategic interactions and confronted its predictions to games played by actual players, for real money, in labs. In some of such games, game theoretic-predictions seem to do remarkably well, in others - exemplified in 1.2. - they drastically fail. To account for the latter findings, we briefly went through recent classes of proposals on which of game theory's assumptions should be relaxed. For instance, standard game theory assumes strict self-interest and unbounded computational abilities, either or both of which could be unrealistic. However, we will stress the importance of the "common knowledge of rationality" assumption. In 1.2. we passed to

different particular class of games, *coordination games*. We will show why the game theoretic-concepts illustrated above fundamentally fail to capture the essence of coordination. In such situations, inferences cannot be made by deliberation or deduction and an important "matching" problem emerges; here, common knowledge must rely on common intuitions (focality), common conventions or communication. In some cases, agents seem to develop conventions on the basis of "historical accidents", which can remain "culture/group-specific" and generate inter-group conflict; in other cases, the conventions seem to emerge from actual similarities among agents, as all groups seem to take the same ones. In both cases however, when incentives are aligned, common knowledge will bring agents to behave similarly.

Homophily

One of the strongest patterns that sociologists and biologists have found to predict network closeness in social species is homophily, the preference for similar others. We will discuss why similarity and reward/motivation apparently share an intimate connection. However, the literature suggests that similarity is not merely about reward, but that it also serves the purpose of aligning beliefs, that is, it re-instantiates common knowledge. The suspicion that emerges is that if similarity is so successful as a social attractor it isn't merely about motivation/attraction, but also (and perhaps critically) about its potential benefits on the actual strategic interactions that follow. Separating this motivational component of closeness and similarity from a potential "knowledge" component will be one of the principle objectives of my experimental work. A different question however is what impact such "psychological closeness" could have on interactions. Previous experimental evidence shows that it can be helpful if perspectives are aligned, but detrimental when they are not, though no study had focused on how this

could oppositely impact games in which choices should be matched as opposed to decoupled.

Novel experimental findings

Throughout this thesis we reported the following findings i) relative to complete chance (p=0.5), such as placing bets on a coin flip – for real monetary outcomes -, agents risk much more frequently (and take less time do so) when they are required to match their choices with another anonymous but motivated/intentional agents; conversely, if mutually anonymous agents are to decouple their choices, then they risk less than when they play against chance (taking longer to do so) (study 1, experiment 1); ii) a similar behavioral pattern is observed when friends, as opposed to strangers, interact in the same scenarios: friends more quickly and more frequently accept the risks of cooperation relative to strangers; however, when choices are to be decoupled, friends quickly avert from risk (study 1, experiment 2); iii) a similar behavioral pattern occurs when subjects *don't* know each other but perceive certain similarities between them. The pattern holds only when features are not only shared but also liked. However, it also doesn't hold if features are liked but not shared (study 1, experiment 3); iv) Two mechanisms could explain these results: expected reciprocity and altruism. In an fMRI experiment we show that, even when agents know that their counterparts (friends or strangers) are randomizing over choice, a similar "polar" effect on risk is observed. This is consistent with altruism. However, the interaction between "social distance" (i.e. friendship) and strategy (whether counterparts randomized or not) was significant both in behavior and the brain: the vmPFC – an area previously implicated in interpersonal similarity, depth of reasoning and reward - was recruited to a greater extent for friends, rather than strangers, especially in strategic conditions; the dmPFC was preferentially active for

strategic conditions, independently of counterpart; while the TPJ differentiated between counterparts independently of strategy. These results are discussed in the respective sections (Study 2).

Section 2

Study 1 Reputational priors magnify striatal responses to violations of trust

Elsa Fouragnan,[1] Gabriele Chierchia,[1] Susanne Greiner,[2] Remi Neveu,[3] Paolo Avesani, [2] and Giorgio Coricelli [1, 3, 4]

[1]Interdepartmental Centre for Mind/Brain Sciences (CIMeC), University of Trento, 38060 Trento, Italy; [2]NeuroInformatics Laboratory (NILab) of Bruno Kessler Foundation, Neuroimaging Laboratory (LNIF) of CIMeC, University of Trento, 38060 Trento, Italy; [3]National Scientific Research Center (CNRS), UMR5292, University of Lyon, 69003 Lyon, France ; [4]Department of Economics, University of Southern California, Los Angeles, CA 90089-0253, U.S.A.

Abstract

Humans learn to trust each other by evaluating the outcomes of repeated interpersonal interactions. However, available prior information on the reputation of traders may alter the way outcomes affect learning. Our functional magnetic resonance imaging (fMRI) study is the first to allow the direct comparison of interaction-based and prior-based learning. Twenty participants played repeated trust games with anonymous counterparts. We manipulated two experimental conditions: whether or not reputational priors were provided, and whether counterparts were generally trustworthy or untrustworthy. When no prior information is available our results are consistent with previous studies in showing that striatal activation patterns correlate with behaviorally estimated reinforcement learning measures. However, our study additionally shows that this correlation is disrupted when reputational priors on counterparts are provided. Indeed participants continue to rely on priors even when experience sheds doubt on their accuracy. Notably, violations of trust from a cooperative counterpart elicited stronger caudate deactivations when priors were available than when they were not. However, tolerance to such violations appeared to be mediated by prior-enhanced connectivity between the caudate nucleus and ventro-lateral Prefrontal Cortex (vLPFC) which anti-correlated with retaliation rates. Moreover, on top of affecting learning mechanisms, priors also clearly oriented initial decisions to trust, reflected in medial prefrontal cortex activity.

Introduction

Trusting others involves risk and uncertainty: people invest a form of good (i.e. money, work, time etc.) in interactions that can yield a profit or a loss, depending on whether others hold to their end of the bargain (Coleman, 1994). Critically, when others are not contractually committed to doing so, they may be untrustworthy for their own benefit and harm the person that initially placed trust in them (Berg et al., 1995). In financial transactions, investors should then either anticipate this, and not invest money to begin with, or develop efficient strategies to estimate the trustworthiness of others (Camerer and Weigelt, 1988).

Experiments with repeated Trust Games (RTGs) allow to empirically observe trust-based dynamics (Chang et al., 2010). Neuroimaging studies employing RTGs have shown that, when no prior information on transaction partners is available, the brain's reward circuitry is involved in learning about their type (i.e. their level of trustworthiness), based on the outcomes of previous trust-based interactions (King-Casas et al., 2005). Indeed, reward-related brain regions have been found to respond positively to trustworthiness and negatively to violations of trust (Krueger et al., 2007; Phan et al., 2010; Long et al., 2012). We refer to this as "interaction-based" learning.

However, a second important alternative for investors to efficiently engage in financial decisions is to rely on priors provided by a third-party. Such priors may affect the way agents evaluate the outcomes of transactions and thus how they learn about the type of their counterparts. We refer to this as "prior-based" learning. For example, in web-based transactions, which are increasingly used, investors interact with complete strangers and rely on available reputation priors (e.g., reports on previous transactions, customer reviews etc.) to predict expected returns and potential risks associated with

investments (Kim, 2009). However, while the neural correlates of interaction-based learning to trust have been largely explored, only few studies have investigated the neural bases of trust when reputation priors are provided (Delgado et al. 2005; Stanley et al. 2012). No studies on date have directly compared the two forms of trust-based decision making within the same experiment.

To confront this issue, we conducted a functional magnetic resonance imaging (fMRI) experiment in the attempt to characterize the neural activation patterns related to trust-based decisions during RTGs. Two situations were analyzed and compared, one in which we provided information about the social attitude of counterparts (i.e. reputational priors), and one in which no such information was provided. Furthermore, in contrast to a previous neuroimaging study on the same issue (Delgado et al., 2005), we also manipulated the actual level of trustworthiness demonstrated by counterparts during an RTG, such as to make it consistent with the provided priors. Finally, we used standard fMRI analysis, model-free and model-based reinforcement learning (RL) models to approach the problem of social learning and reputation effects. Our main goal was to assess whether and how reputation priors affect RL mechanisms at both the behavioral and neural level.

Materials and methods

Participants

Twenty male participants (mean age, 29.5 ± 3.53 years) took part in the fMRI experiment; two were removed from the analysis for excessive head movement (See fMRI analysis). All of them were healthy, gave written informed consent, had normal or corrected-to-normal vision without any history of psychiatric, neurological, or major medical problems, and free of psychoactive medications at the time of the study. Participants were told that the

experiment aimed at studying decision making in a social context, that they would receive a compensation of 15 Euros/hour and that the money gained in ten randomly extracted trials would be added to their compensation. The study was approved by the local institutional ethical committee of the University of Trento.

Task

The experimental task was based on the Trust Game (TG) (Berg et al., 1995). In one round of our task, each participant played as "investor" with an anonymous counterpart as "trustee". Both players were endowed with 1 euro before starting a round composed of 2 stages (see Fig. 1A): in stage 1 the participant decided whether or not to share his euro with the trustee. If he decided to share, the euro was multiplied by 3 by the experimenter before being allotted to the trustee. In stage 2 the response of the trustee could be to either equally share his money with the investor (1/2 of 4 euros = 2 euros) or keep his money and return nothing. It follows that if the investor invested and the trustee reciprocated, both players were better off than if the interaction has not occurred at all. However, investing was risky, as if a trustee returned nothing, the investor incurred a loss.

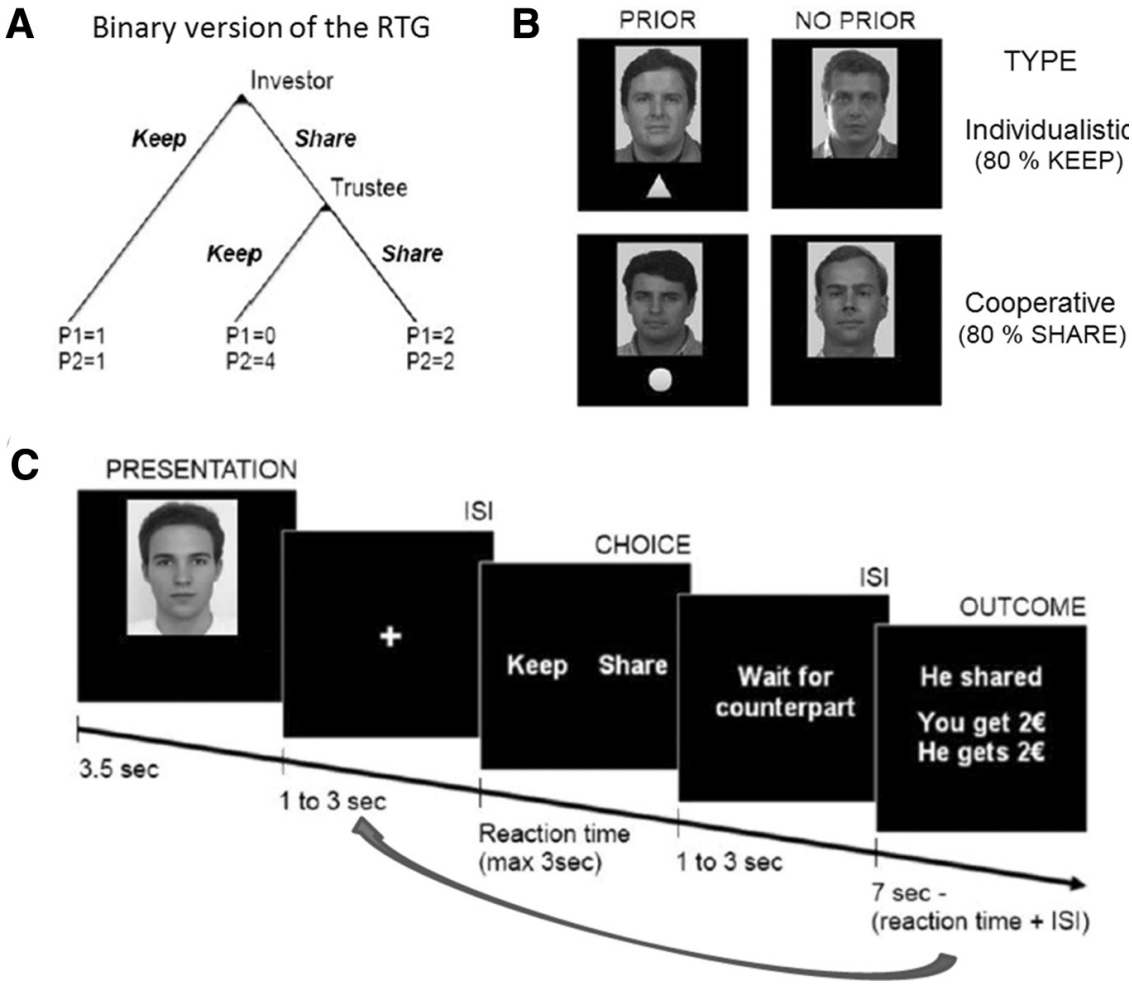

Figure 1. Experimental design. A. One round of the two-player repeated trust game (RTG). P1 is the payoff of the participant, who always plays as investor; P2 is the payoff of his counterpart, who plays as trustee. Before each round both players are endowed with 1 euro. The participant moves first and chooses either to "keep" or "share". If he keeps, both players maintain their initial endowments. If he shares the participant's endowment is multiplied by 3 and passed to the counterpart. The trustee then decides whether to share in turn (by returning 2 euros), or to keep (by returning nothing). RTGs consisted of several consecutive rounds with a same counterpart. Participants played with many different counterparts and were told that their counterparts had already made their choices. **B. Experimental conditions.** Two conditions were adopted: 1) the "type" of

counterpart, and 2) the presence vs. absence of "reputational priors". Types: counterparts could be either "cooperative" or "individualistic" in their (simulated) behavior in RTGs; the former shared and the latter kept in 80% of RTG rounds. Reputational priors: participants were told that cues indicated whether the current counterpart had obtained a high or low score in a social orientation task (triangles indicated low scores, circles indicated high scores). Such priors reliably differentiated between the 2 counterpart types. **C. Timeline of the first RTG round**. **Presentation**: Face of the counterpart (with a prior or no-prior) was displayed for 3.5 s, and only presented for the first round of an RTG. **Fixation**: Fixation cross was presented during a jittered inter-stimulus interval (ISI). **Choice**: Participants made their choice by pressing "Keep" or "Share". **Delay**: ISI corresponding to the (simulated) decision of the counterpart. **Outcome**: Outcome of the game and the payoffs of each player.

We used a repeated version of this TG (RTG), which consisted in a series of consecutive TG rounds with a same counterpart. However, this alters the nature of the single-shot TG, as RTGs allow for additional strategic maneuver. For instance, investors tend to invest more (and trustees to reciprocate) in initial rounds of RTGs, than in final rounds or single shot games (Isaac et al., 1985). For similar reasons, both parties may strategically punish (by not investing) if they believe this might incentivize uncooperative counterparts to review their strategies in future rounds.

Our study intended to minimize the strategic component of trust-related behavior; hence our version of the game differentiated from the typical repeated TG in few but important respects.

(i) Subjects were informed that trustees had already made their choices, which thus wouldn't have been affected by those of the participant. In other words, participants knew

that counterparts were not interactive. This feature should have eradicated any strategic component usually present in RTGs. In reality, the trustees were computer simulations and they reciprocated an investment with fixed probabilities unknown to participants.

(ii) Another feature was also adopted to make learning independent on participants' actions. In traditional RTGs, when an investor does not trust, the round ends and nothing is learned about the behavior of counterparts. In our study, on the other hand, participants learned about the trustees' choices even when they invested nothing. This adjustment enabled to keep the amount of feedback fixed (regardless the choice of participants), thus allowing us to compare learning mechanisms between conditions.

(iii) Finally, to further reduce strategic reasoning, participants did not know how many games composed each RTG with a given trustee but only that RTGs were consecutive and if they were not paired with the same trustee twice in a row, then they would have never encountered the counterpart again. Specifically, we fixed a constant probability of 1/3 to continue the game with a same counterpart; this resulted in a minimum of one and a maximum of eight games with a same trustee.

Then, each trustee was introduced with a picture of his face before a RTG began (see **Fig. 1B**). The association between pictures and RTGs was randomized, as was the order of RTGs. To reduce facial information extraction and gender attraction, we assembled a database of colored pictures from 20 to 60 years old Caucasian men (mean age: 34.05 ± 11.19) controlled for attractiveness, emotion and racial traits. 128 pictures were selected and used with authorization from the FERET database of facial images collected under the FERET program (Phillips et al., 2000). The words "trust" or "trustworthy" were never mentioned during the training session and the experiment.

Experimental conditions

A first key manipulation was that trustees were divided into 2 predefined types: they could be either "cooperative" or "individualistic". Cooperative trustees would reciprocate 80% of the times, while individualistic counterparts would defect 80% of the times (though participants were not informed of such contingencies). The distinction between types furthermore allowed confronting the cases in which trustees behaved consistently ("Cons") or inconsistently ("Incons") with their types.

The second key feature of our study was whether or not a reputation prior was provided (see **Fig. 1B**). In the prior-condition, half of the cooperative and half of the individualistic trustees were flagged, respectively by a circle and a triangle. These cues signalled their "reputation". Specifically, participants took part in the Social Valuation Orientation (SVO) (Messick and McClintock, 1966; Van Lange et al., 1999) and were told that the distinct cues were based on the trustees' scores for the same task. This task distinguishes between different types of social value orientations (e.g., cooperative or individualistic). The main difference between each category is the extent to which one cares about own payoffs and that of the others in social dilemma situations. Finally, for the remaining half of the counterparts, no prior information was provided (no-prior condition).

In order to insure no difference in learning scheme in each of the four conditions (Prior/Cooperative, Prior/Individualistic, No-Prior/Cooperative, No-Prior/Individualistic), RTG length and share/keep schedules within each RTG were counterbalanced.

Procedure

Training

Participants received written instructions, took part in a simplified version of the SVO task and completed a 20 minutes RTG practice session (20 trials). The experiment was implemented using Presentation® software (version 0.70).

Inside the MRI

In the scanner, subjects completed 356 trials (89 for each condition: Prior/Cooperative, Prior/Individualistic, No-Prior/Cooperative, No-Prior/Individualistic), divided in 4 runs of 20 minutes. **Figure 1C** shows the timeline of the first trial of a RTG. Each RTG started with a 3.5 s display of the face of the trustee (which, only in "prior" conditions, was flagged with a reputational cue). This was followed by a fixation cross and then by a "decision-screen", which required participants to choose between 1 of 2 option, labeled "share" or "keep". After making their choice, participants waited a jittered interval before an "outcome screen" appeared, displaying the trustee's choice and the corresponding payoffs to both players. For those trials in which participants chose to keep, the outcome screen was still shown.

Analysis

Behavioral data analysis

Behavioral data were analyzed using Stata© Statistical Software version 9.2 and the R environment (Development Core Team 2008). A two-way repeated measure ANOVA was performed to identify differences between conditions for each variable of interest (e.g., decision to trust, payoffs made in each condition). Next, we computed regression analyses

using mixed-effects linear models (MEL), in which participants were treated as random effects and hence were allowed to have individually varying intercepts. Parameter estimates (*b*), standard error (*se*), *t*-values and *p*-values were reported.

RL models

Model 1: Model-free TD learning

We first used a "model-free" temporal-difference (TD) (model 1) learning algorithm (Rummery and Niranjan, 1994; Sutton and Barto, 1998), which assumes that agents are initially unaffected by the presence of priors, but that, as trials with a counterpart unravel, they may update reward values differently when priors are available as opposed to when they were not available. Participants would sample the reward probability of two choices ("Keep" or "Share") in the Cooperative and Individualistic conditions. We then hypothesized that participants would obtain reliable expectation of these conditions updating the estimated value of each choice with a discounted "step-size". Thus the stochastic prediction error δ, based on the Rescorla-Wagner learning rule (Rescorla and Wagner, 1972) was computed as follows:

$$\delta_t = r_t - Q(_{C,t}) \qquad (1)$$

where r is the payoff obtained at time t, when choosing an option C at time t or $t+1$, Q is the value of each choice "Share" or "Keep" in each trial. In addition to this, the following learning rule differentially updated the stochastic prediction error in the Prior (P) and No Prior (NP) conditions:

$$Q(_{C,t+1}) = Q(_{C,t}) + \alpha^{P} \cdot \delta^{P}(_{C,t}) + \alpha^{NP} \cdot \delta^{NP}(_{C,t}) \qquad (2)$$

The degrees in which δ^{P} and δ^{NP} influence the new action value are weighted by two learning rates α^{P} and α^{NP} where $0 < \alpha^{P}, \alpha^{NP} < 1$.

Model 2: Model with separate expectations for positive or negative priors

Additionally, we hypothesized that, in the Prior condition, participants may have "optimistic" or "pessimistic" expectations, at the beginning of the game due to the presence of a positive (P^+) or negative Prior (P^-), respectively (Biele et al. 2011; Wittmann, 2008) (model 2). Thus, the values of initial choices when playing with a Cooperative or Individualistic counterpart in the prior condition were computed as:

$$Q^{P+}(_{C,0}) = g^{P+}.\mu\theta_{P+}.N \qquad (3)$$

$$Q^{P-}(_{C,0}) = g^{P-}.\mu\theta_{P-}.N \qquad (4)$$

where g^{P+} g^{P-} are equal to 1 when playing with a counterpart with a positive or negative prior, respectively; and 0 otherwise. θ_{P+} and θ_{P-} are free parameters capturing the optimistic or pessimistic impact of the priors expectation, μ is the expected payoff from choosing randomly among all options, which serves as a normalization constant (in our case μ = 1), and N is the number of trials experienced in the learning condition, which is a scaling factor, allowing for the comparison between an expected value decision and the outcome of the decision. On the other hand, in the no prior condition, only one parameter weighted the initial expected value of choices, $Q^{NP}(_{C,0})$.

The Softmax function was then used for the two models to determine the probability of choosing a given choice option given the learned values:

$$p_1(t) = \frac{\exp[V_1(t)/\beta]}{\exp[V_1(t)/\beta] + \exp[V_2(t)/\beta]} \qquad (5)$$

where β is called a temperature parameter. For high values of β, all actions have almost the same probability (i.e. choices are random), while for low βs the probability of choosing the action with the highest expected reward ($Q_1 > Q_2$) is close to 1.

In order to generate model-based regressors for the imaging analysis, both learning models were simulated using each subject's actual sequence of rewards and choices to produce per-trial, per-subject estimates of the initial values Q_t and error signals δ_t (Morris et al., 2006; Wittmann et al, 2008). All parameters of interest were implemented in Matlab R2009 and were estimated using the negative log likelihood of trial-by-trial choice prediction. Model comparisons were performed with the Bayesian Information Criterion, the pseudo r^2 value using the Log likelihood of a random distribution and tested with the likelihood ratio test.

fMRI method

fMRI data acquisition

A 4T Bruker MedSpec Biospin MR scanner (CiMEC, Trento - Italy) and an 8-channel birdcage head coil were used to acquire both high-resolution T1-weighted anatomical MRI using a 3D MPRAGE with a resolution of 1 mm^3 voxel and T2*-weighted Echo planar imaging (EPI). The parameters of the acquisition were the following: 34 slices, acquired in ascending interleaved order, the in-plane resolution was 3 mm^3 voxels, the repetition time 2 sec and the echo time was 33ms. For the main experiment, each participant completed 4 runs of 608 volumes each. An additional scan was performed in between two different runs in order to determine the point-spread function (PSF) that was then used to correct the known distortion in a high-field MR system.

Preprocessing

The first five volumes were discarded from the analyses to allow for stabilization of the MR signal. The data were analyzed with Statistical Parametric Mapping 8 software (SPM8®, Welcome Department of Cognitive Neurology, London, UK) implemented in Matlab R2009 (Mathworks, Sherborn, MA). We used SPM8® for the preprocessing steps. Head motions were corrected using the realignment program of SPM8®. Following realignment, the volumes were normalized to the Montreal Neurological Institute (MNI) space using a transformation matrix obtained from the normalization process of the first EPI image of each individual subject to the EPI template. The normalized fMRI data were spatially smoothed with a Gaussian kernel of 8 mm (full-width at half-maximum) in the (x, y, z) axes. Imaging data for participants with head motions exceeding one voxel (3mm) in transition and 3° in rotation were discarded (Eddy et al., 1996). We also used the xjView package and MRICron to create the pictures presented in the results (version 1.39, Build 4).

fMRI analysis

GLM 1a and b. Our first analysis considered the main effect of the presence or absence of reputation priors when a new counterpart is presented for the first time. We used a general linear model (GLM), estimated in three steps: 1) first, individual BOLD signal was modeled by a series of events convolved with a canonical hemodynamic response function. The regressors representing the events of interest were modeled as a boxcar function with onsets at the beginning of each RTG ("Pre") and durations of 3.5sec. For **GLM1a,** regressors represented trials in which i) priors were provided ("Prior_Pre") and ii) no priors were provided ("NoPrior_Pre"). For **GLM1b,** regressors represented trials in which i) priors were provided for a cooperative counterpart ("Prior+_Pre"), ii) priors were

provided for individualistic counterparts ("Prior-_Pre") and iii) no priors were provided ("NoPrior_Pre"). For *t*-contrasts, we then computed first-level one-sample *t*-tests comparing trials with and without priors on the basis of the GLM1a. 2) We then analyzed second-level group contrasts. Our fMRI results were initially thresholded at $p < 0.001$ uncorrected and were subsequently cluster-thresholded at $p < 0.05$ FWE. All reported coordinates (x, y, z) are in MNI space. Anatomical localizations were performed by overlaying the resulting maps on a normalized structural image averaged across subjects, and with reference to an anatomical atlas. Finally, 3), we used the Marsbar toolbox from SPM8® to perform functionally defined (based on the averaged parameter estimates in the cluster found with GLM 1b) region of interest analysis (ROI) and compute percentage signal changes.

GLM 2 Model-based fMRI analysis. A second GLM model still focused on the distinction between prior and no prior conditions but additionally separated between two phases of the RTG: the decision phase and the outcome phase. This allowed to assess how the impact on BOLD signal of priors was parametrically modulated by two behaviorally estimated learning measures (from model 2): 1) at time of choice, the parameter *Qt,* weighted the value of options, on a trial to trial basis, depending on RTG history; 2) while δ_t scaled outcomes on the basis of their estimated prediction error. We performed this analysis at the individual level and ran group statistics, taking individual participants as random effects. We then focused on a subset of our resulting brain regions on the basis of effect strength ($P < 0.05$ FWE corrected). Specifically, averaged parameter estimates were extracted from bilateral caudate (MNI coordinates: (-14, 20, 2) and (12, 18, 6)), separating between prior vs. no prior contexts.

GLM 3. Violation of trust. In a third GLM we differentiated between consistent ("Cons") and inconsistent ("Incons") outcomes. We classified consistent outcomes as those rounds in which either i) participants had kept with individualistic counterparts that defected ("Cons-") (distribution of trials: M = 57 ± 3) or ii) they had shared with a cooperative counterpart that reciprocated ("Cons+") (M = 56 ± 4 trials); inconsistent outcomes, on the other hand, occurred when either iii) participants had kept with an individualistic counterpart that reciprocated ("Incons-") (M = 14 ± 4 trials) or iv) they shared with a cooperative counterpart that defected ("Incons+") (M = 15 ± 4 trials), and who thus "violated" their trust.

Functional connectivity analysis (PPI). To explore the interplay between the caudate and other brain regions following violations of trust (Incons+), we assessed functional connectivity using psychophysiological analysis (**PPI**: Friston 1994; Cohen et al. 2005) that compares the pattern of activity of a seed region to every other regions of the brain. We took the bilateral caudate resulting from the reported **GLM3** ("Cons" > "Incons") as seed regions, as these areas showed highest sensitivity to violations of trust ($t = 6.78$, $p < 0.05$, FWE). Then, we created three regressors: 1) the caudate time course (physiological regressor); 2) an event related regressor that distinguished between violations of trust in the prior and no prior conditions (with a boxcar function ranging from the beginning of the outcome phase until the end of the ISI) and 3) the interaction term. Additionally, we also conducted a correlation analysis between the retaliation rate for each subject (measured by the percentage of choices to keep after violation of trust when playing with a cooperative partner) and the parameter estimates in left ventro-lateral Prefrontal Cortex (vLPFC) (MNI -40, 42, 4) across subjects. Finally, to examine how striatal responses to

violations of trust were related to learning, we plotted individual parameter estimates against the individual learning rates (estimated with model 2 described above).

Results

Behavioral results

Our main goal was to determine whether reputation priors influence initial expectations and decisions in the games, and subsequent learning mechanisms. A repeated measure two-way ANOVA was performed using type of counterpart (cooperative or individualistic) and prior condition (prior or no-prior) as within participant factors. The percentage of decisions to share was significantly higher with cooperative counterparts (M = 71.77, SE ± 4.03) than with individualistic counterparts (M = 27.34, SE ± 3.71; $F_{1, 17}$ = 174.01, p < 0.001). The results also showed a significant interaction effect of prior with type of counterpart ($F_{2, 35}$ = 30.87, p < 0.001). *Post hoc* tests (t-tests Bonferroni corrected) indicated that participants decided to share with cooperative partners more when provided with a prior (M = 81.09, SE ± 4.78) than when priors weren't provided (M = 62.45, SE ± 5.81; t = 5.89, p < 0.001), whereas they decided to share with individualistic counterparts less in the prior (M = 18.37, SE ± 4.66) than in the no prior condition (M = 36.3, SE ± 5.05; t = 4.23, p < 0.002, see **Fig. 2A**). When payoffs are analyzed with type of counterparts and prior condition as within-subject variables, we found that payoffs were significantly higher when playing with cooperative counterparts (M = 1.43, SE ± 0.13) than individualistic counterparts (M = 0.94, SE ± 0.11; $F_{1, 17}$ = 138.32, p < 0.001) and significantly higher in the prior condition (M = 1.20, SE ± 0.10) than the no prior condition (M = 1.08, SE ± 0.06; $F_{1, 17}$ = 28.98, p < 0.001, see **Fig. 2C**).

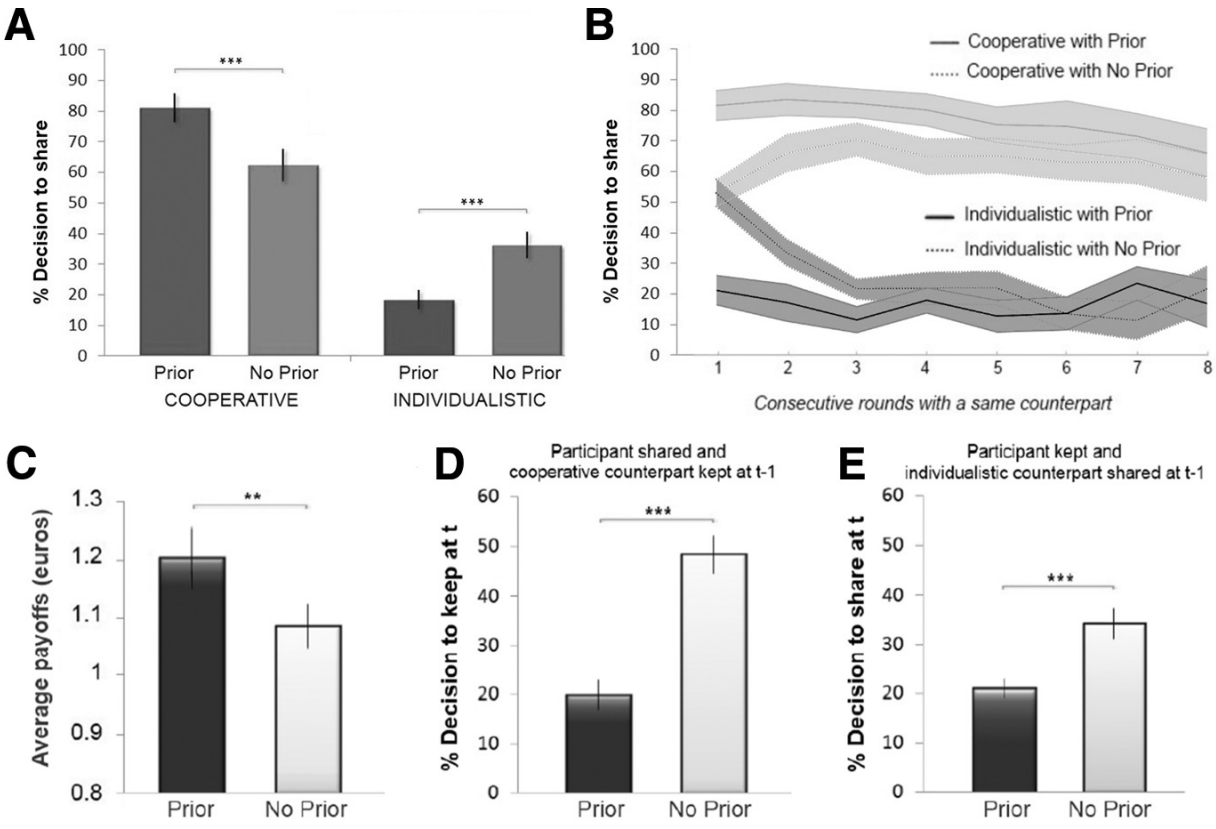

Figure 2. Behavioral results. A. Average percentage of decision to trust across conditions. Mean ± standard error of participants' decision to trust (share) are broken down for trustee's type (Cooperative/Individualistic) and prior condition (Prior/No Prior); *** $p<0.001$. Priors enabled participants to match (on average) their choices with the counterpart's level of trustworthiness. **B. Learning dynamics across RTG rounds.** Average percentage of the decision to trust for each round when playing with a "cooperative" vs. "individualistic" counterpart, and when priors were present vs. absent. When participants know nothing of their counterparts they tend to randomize between trusting and not trusting during initial rounds and adjust their choices to their counterparts' type in succeeding rounds. On the other hand, when priors are present,

participants tend to rely on them already from early rounds. Shaded areas above and below the curves are standard errors. **C. Average payoffs in the Prior and No-Prior conditions.** Average payoffs ± standard error (in euros) in Prior/No Prior conditions. When priors are available, participants significantly earn more when they adjust their choices to counterparts' types; ** p<0.01. **D. Choices following unexpected behavior of cooperative and individualistic counterparts.** Average (± standard error) of percentage of "keep" choices in prior vs. no prior condition at time *t*, following rounds in which participants shared and a cooperative counterpart violated their trust by deciding to keep (at *t*-1). Decisions to Keep at time *t* (i.e., retaliation) was less frequent when priors were available. **E.** Percentage of "share" choices (at *t*) following rounds in which participants had kept and an individualistic counterpart has shared (at *t*-1).

In order to examine the effect of the prior condition, the trustees' type, the order of the repeated game and the interactions of such factors on the decision to share (binary dependent variable), we performed regression analyses using mixed-effects linear (MEL) models. The results revealed that participants shared with cooperative counterparts more often as compared to individualistic counterparts (b = 1.29 (SE ± 0.08), t = 15.8, p < 0.001); shared less when they did not receive priors (b = - 1.09 (SE ± 0.09), t = - 12.1, p < 0.001); and shared less over time (b = - 0.12 (SE ± 0.02), t = - 6.81, p < 0.001). These results suggest that participants took into account reputation priors and played according to the counterpart's level of trustworthiness. Instead, when priors were not available, participants learned counterparts' types on the basis of their actions. Interestingly, we found an interaction effect between the trustees' type and the prior condition (b = 2.27 (SE ± 0.13), t = 17.39, p < 0.001). These results indicate that the difference between prior and no prior conditions was greater when playing with a cooperative than with an

individualistic counterpart. Furthermore, even though participants in the no prior condition adjusted their decisions to their counterparts' type over rounds, they still shared with cooperative counterparts less than when they had priors (see **Fig. 2B**). Post hoc *t*-test revealed that, in the no prior condition, in rounds when cooperative counterparts kept, participants subsequently kept more (*Mean percentage of decisions to keep* = 0.48, *SE* ± 0.019), whereas they persisted in sharing in the prior condition (*M* = 0.2, *SE* ± 0.015; t_{17} = - 4.99, p < 0.001), (see **Fig. 2D**). Similarly, when individualistic counterparts shared in a round, participants subsequently shared more when not provided with a prior (*Mean percentage of decisions to share* = 0.34, *SE* ± 0.015) than when given a prior (*M* = 0.21, *SE* ± 0.009; t_{17} = - 4.783, p < 0.001, see **Fig. 2E**).

Results from learning models

A likelihood ratio test revealed that the Prior model (model 2) with separated expectations for cooperative and individualistic counterparts (Prior mode) performed better than the classical TD learning model (model 1) (p < 0.001) (Additional statistics are reported in **Table 1**). The best-fitting parameters are shown in **Table 2**. For these parameters, we found that the average learning rate estimated from trials in the No Prior condition, $\boldsymbol{\alpha_{NP}}$, was significantly higher than the average learning rate estimated from trials in the Prior condition $\boldsymbol{\alpha_{P}}$ (t_{17} = 2.29; p < 0.05). We also found that the initial value in the Cooperative Prior condition, $\boldsymbol{Q_{P+}(0)}$ was significantly higher than the initial value in the No Prior condition $\boldsymbol{Q_{NP}(0)}$ (t_{17} = - 2.82; p < 0.001), and the initial value in the Individualistic Prior condition, $\boldsymbol{Q_{P-}(0)}$ (t_{17} = - 3.07; p < 0.001). There was no significant difference between the initial value in the Individualistic Prior condition, $\boldsymbol{Q_{P-}(0)}$ and the initial value in the No Prior condition $\boldsymbol{Q_{NP}(0)}$ (t = 0.46). Finally, we found that the average

learning rates estimated for each participant when they kept was higher (M = 0.46, SE ± 0.04) than when they shared (M = 0.38, SE ± 0.048; t_{17} = - 2.27, p < 0.05, see **Table2**).

Table 1. Learning model comparison

Learning model comparison		
	Classical model-free TD learning model	Prior $^{+}$ and Prior $^{-}$ expectations RL learning model
BIC	7619	6460
Log likelihood (random model $=$ -4442)	-3809	-3230
Pseudo r^2	0.14	0.273

Bayesian information criterion value (BIC), Log likelihood, and the pseudo r^2 suggest that the Prior $^{+}$ and Prior $^{-}$ expectations TD learning model fits the observed behavior better the other TD learning models.

Table 2. Averaged best-fitting parameter estimates (across subjects) SE

Parameter estimate for best behavioral model, depicted as mean $\pm$ SE		
	Mean	SE
Learning rate Prior condition α_P	0.3373	$\pm$ 0.0456
Estimates for Cooperative counterparts	0.327	$\pm$ 0.0424
Estimates for Individualistic counterparts	0.3475	$\pm$ 0.0398
Learning rate No Prior condition α_{NP}	0.5075	$\pm$ 0.0689
Estimates for Cooperative counterparts	0.4686	$\pm$ 0.0701
Estimates for Individualistic counterparts	0.539	$\pm$ 0.0599
Estimates learning rates for Invest trials (participants shared)	0.3845	$\pm$ 0.0459
Estimates learning rates for Non-Invest trials (participants kept)	0.4603	$\pm$ 0.0476
Softmax inv. Temp Betha β	4.7769	$\pm$ 0.3149
Initial value Cooperative Prior condition, $Q_{P+}(0)$	1.3814	$\pm$ 0.1031
Initial value Individualistic Prior condition, $Q_{P-}(0)$	0.9838	$\pm$ 0.1055
Initial value No Prior condition $Q_{NP}(0)$	1.0641	$\pm$ 0.126

fMRI results

Effect of prior at time of counterpart presentation

The contrast ("Prior_Pre" > "NoPrior_Pre") (see Methods, **GLM1a**) revealed differential activity in the mPFC (0, 62, 31), to the presence vs. absence of any priors when new counterparts were presented (t = 8.26; p < 0.05 FWE cor.) (See **Fig. 3A** and **Table 3**). Further functional ROI analysis, based on **GLM1b**, qualified this activation pattern as responding with increased activity to the presence of priors, regardless of their nature (positive or negative), and decreased activity to their absence (see **Fig. 3B**). The opposite contrast ("NoPrior_Pre" > "Prior_Pre") revealed activity in bilateral anterior insula (-36, -4, 15), t = 3.91; p < 0.001 unc., and (38, 3, 10), t = 3.45; p < 0.002 unc.).

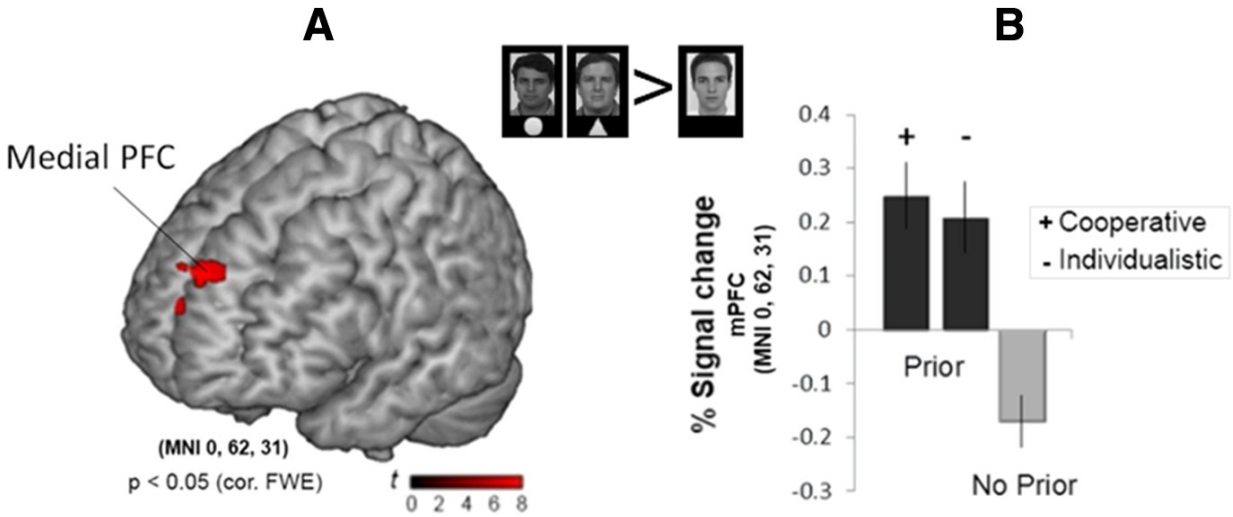

Figure 3. mPFC encodes reputational priors when a new counterpart is first presented. A. Random Effect Analysis. When contrasting (Prior) > (NoPrior) conditions at time of counterpart presentation, activity in the medial prefrontal cortex survived FWE correction, p < 0.05. **B. Functional ROI Analysis in mPFC.** Functional ROI analyses

further revealed percentage signal changes in the medial prefrontal cortex MNI (0, 62, 31). The figure shows an increased activity when priors were present, regardless of their type, and decreased activity when there were no priors.

Table 3. Activations correlated with contrasts of interest

Analysis/Location	BA	Side	Cluster size	T	*p* value FWE cor.	MNI coordinates (mm)		
						X	*Y*	*Z*
Prior > No Prior (GLM 1)								
mPFC	10		95	8.26	6.8×10^{-06}	0	62	31
VTA		—	14	3.177	0.0032 unc.	0	−1	−5
No Prior > Prior (GLM 1)								
Anterior insula	44	Left	106	3.912	0.0009 unc.	−36	−4	15
Anterior insula	44	Right	55	3.450	0.0017 unc.	38	3	10
Parametric regression of Choice (GLM 2)								
mPFC	10	—	87	6.562	2.7×10^{-06}	−2	64	10
Lateral PFC	46	Left	122	5.987	7.8×10^{-05}	−38	38	32
Lateral PFC	46	Right	109	6.342	2.1×10^{-06}	30	38	34
Superior parietal lobule	48	Left	43	5.01	6.7×10^{-04}	−38	6	24
Parametric regression at Outcome for the No Prior condition (GLM 2)								
Caudate nucleus	—	Left	77	7.091	8.9×10^{-06}	−14	20	2
Caudate nucleus	—	Right	56	8.298	7.9×10^{-06}	12	16	8
Violation of rust in the Prior condition (GLM 3, Cons > Incons)								
Caudate nucleus	—	Left	82	6.78	2.8×10^{-06}	−10	18	11
Caudate nucleus	—	Right	56	6.34	2.4×10^{-06}	12	21	5

Note: BA, Brodmann area; mPFC, medial prefrontal cortex; VTA, ventral tegmental area.

Effect of prior at RTG choice

Applying parametric analysis (see **GLM 2**) to the functional MRI data, we focused on trial-to-trial weights on decision values as represented by per-trial *Qt* estimate amplitude. We found that decision value estimates were correlated with neural activity in a network consisting of the mPFC (-2, 64, 10) and the dorsolateral prefrontal cortex (dLPFC) (-38, 38, 32), surviving $p < 0.05$ FWE corrected (see **Fig. 4A** and **Table 3**). These two regions reflected the contributions of prior's valence (positive or negative) to the pattern of activity related to the decision to trust (see **Fig. 4B**). Moreover, the difference at a neural level between prior and no prior condition was greater when playing with a cooperative counterpart compared with an individualistic counterpart. This is consistent with the observed behavioral asymmetry of the effect of priors between cooperative and individualistic conditions (see **Fig 2B**).

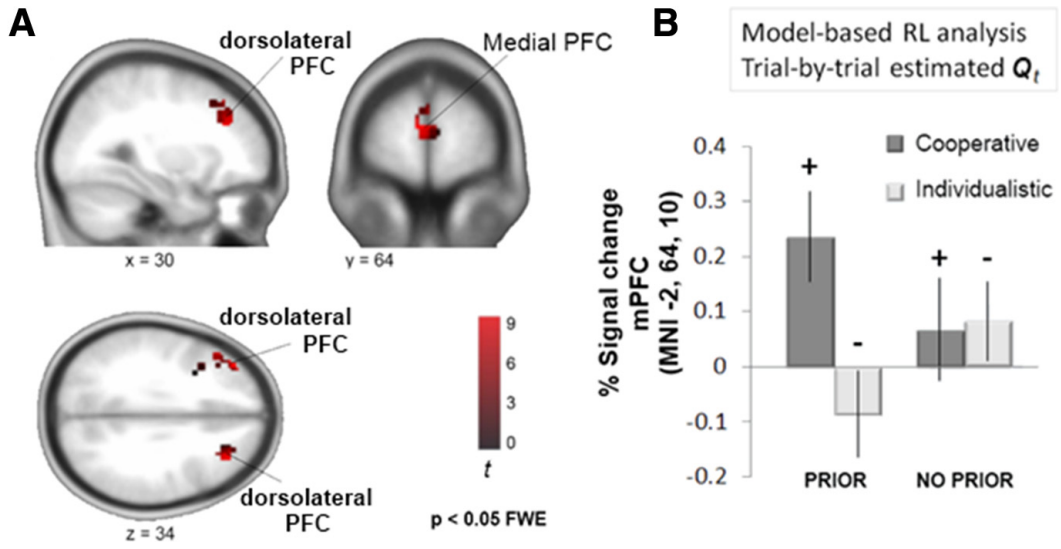

Figure 4. Brain regions parametrically correlated with the estimated "optimistic" and "pessimistic" decision value from the Prior model. A. Random effect fMRI analysis. To look for neural correlates of value signals (*Qt*) at time of choice, we entered the trial-by-trial estimates of the values of the two stimuli ("Share" and "Keep") into a regression analysis against the fMRI data. We found enhanced activation in mPFC and dLPFC, surviving FWE correction, p<0.05. **B. Functional ROI analysis in mPFC.** Percent signal change by condition in the mPFC area represented in (A). Similar pattern of activity was found in the dlPFC (not reported). These regions encoded prior valence (positive and negative) that guided decision to trust at time of choice. Error bars represent *SE*.

Effect of prior at RTG outcome

Across all RTGs, during the outcome phase of the game (see **GLM2**), individually estimated trial-wise prediction errors (positive and negative combined) correlated significantly with BOLD responses in the bilateral caudate in the No Prior trials only (p < 0.05 FWE), (see **Fig. 5A** and **Table 3**). On the other hand, striatal activity appeared to track estimated prediction errors in a more blunted fashion when priors were provided (see **Fig. 5A**).

Moreover, from a direct comparison between the no prior and prior conditions, we found higher activity in the left caudate for the no prior condition compare to the prior condition with a group peak MNI coordinates at -12, 20, 8 (see **Fig. 5B**).

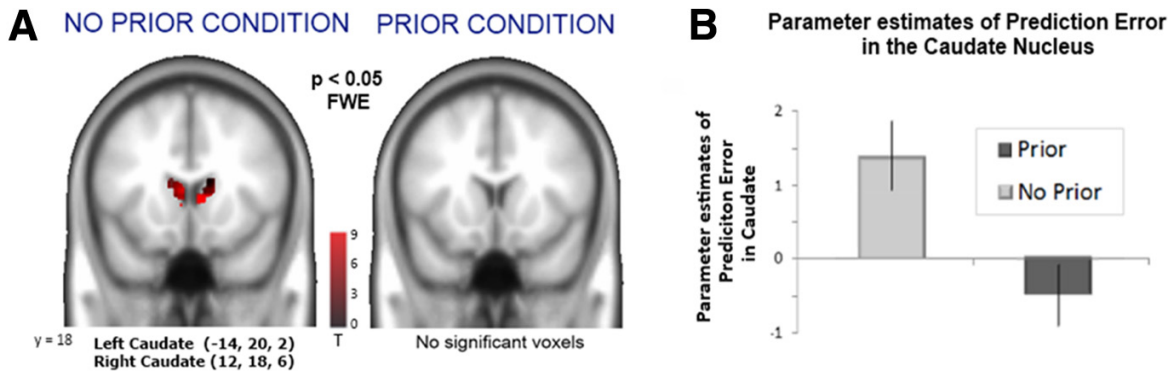

Figure 5. Brain regions parametrically correlated with the estimated Prediction Error of the best fitting RL model. A. Random effect fMRI analysis: Activity of the caudate showed significant correlation to the estimated PE signal in the no prior condition ($p < 0.05$ FWE cor.). Such activities were not observed in this brain area in the prior condition. Peak coordinates are given in MNI space. Colour bars indicate T-values. B. Parameter estimates were extracted from the left caudate (-12, 20, 8) for the direct comparison between prior and no-prior conditions. Caudate activity correlates with PE in the no-prior condition only.

<u>Pattern of activity related to violation of trust: functional connectivity analysis</u>

Finally, we specified the changes in activity in the caudate related to the effects of violation of trust (e.g. the decisions to keep of a cooperative counterpart in response to a decision to trust of a participant) in the prior and no-prior condition (analysis from **GLM3**, **Table 3**). This analysis showed a stronger deactivation of the caudate in the prior condition compared to the no-prior condition ($t = 6.78$; see **Fig. 6A** and **Fig. 6C**). However,

in contrast with the no prior condition, striatal deactivations to violation of trust were not reflected in the behavior of our participants. Indeed, the pattern of striatal activity related to violation of trust did correlate with individual learning rates only in the no prior condition (from the **model 2**: r = - 0.687, p < 0.001; see **Fig. 6D**). No such correlation was found in the Prior condition (see **Fig. 6D**).

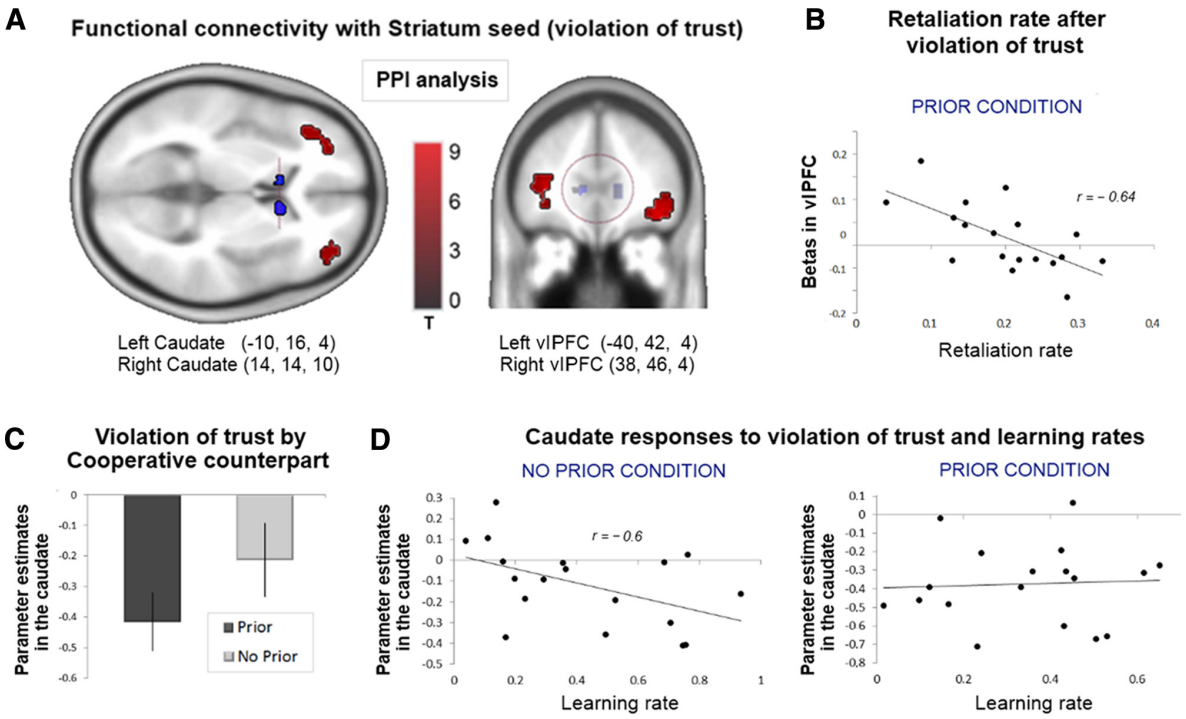

Figure 6. Functional connectivity between the caudate nucleus and vLPFC correlates with the choice to retaliate after violation of trust in the prior condition. A. PPI analysis. With a caudate seed, bilateral vLPFC shows stronger connectivity with this region in the prior compare to the no prior conditions. **B. vLPFC prevents retaliation to violation of trust in the prior condition.** vLPFC anti-correlates with retaliation rate in the Prior condition after participants experimented violation of trust from a cooperative counterpart. Spearman r = -0.6, p < 0.009. **C. Reputational priors magnify striatal**

response to violation of trust. The caudate shows a stronger deactivation to violation of trust from a cooperative counterpart in the prior condition compare to the no prior condition. **D. Striatal responses to violation of trust and learning rates.** The correlation between caudate and learning rates is significant only in the no-prior condition, thus striatal responses to violation of trust in the prior condition are not reflected in learning.

We used Functional connectivity analysis to search (see Methods **Connectivity analysis**) for brain areas that could have mediated such striatal responses when reputation priors were provided. We found that left and right vLPFC showed strong functional connectivity with the caudate seed region after violation of trust in the Prior compared to no prior conditions; vLPFC, left (-40, 42, 4), $t = 3.73$; and; right (38, 46, 4), $t = 6.37$, $p < 0.05$ cor. SVC; see **Fig. 6A** and **Table 3**). Finally, we found that the strength of connectivity between caudate-vLPFC was anti-correlated with participants' decisions to keep following violation of trust (Spearman correlation $r = - 0.67$, $p < 0.001$). Moreover, we found that the activity in the vLPFC was inversely correlated with individual retaliation rates (computed as the percentage of Keep over Share choices) after violations of trust ($r = - 0.6$, $p < 0.009$, see **Fig. 6B**).

Discussion

Reputation-based social decision-making has been investigated both by theoretical and empirical studies (Boero et al., 2009; Camerer & Weigelt, 1988; Fudenberg et al., 1990), however research on its neurocognitive bases is still in its infancy. Though it is rather unlikely that, in daily decisions, people possess absolutely no prior/contextual information on who they interact with, the growing literature using RTGs in fMRI focused

mainly on situations in which strictly no priors are available (McCabe et al., 2001; King-Casas et al., 2005; Krueger et al., 2007). Only two recent fMRI studies investigated how social priors (i.e. the moral character of their counterparts) affect the way people engage in RTGs (Delgado et al., 2005; Fareri et al., 2012). These studies however did not completely isolate the effect of priors on trust (prior-based trust) by confronting them to identical conditions with no priors (interaction-based trust). Our experimental setting is the first to allow this direct comparison. The main goal of our study was to determine whether, and how, reliable reputational priors affect initial decisions and subsequent learning mechanisms at both the behavioral and neural level.

From a behavioral point of view, we show that priors affect decisions to trust in at least 2 ways: 1) in initial stages of the interaction, participants clearly chose to trust or distrust according to the positive or negative reputation of their counterparts; furthermore, 2) players tend to keep relying on reputation priors, even when their counterpart's behavior was inconsistent with it. As a consequence, and since priors were accurate predictors of trustworthiness in our study, players earned more when reputational cues were available than when they were not.

mPFC encodes reputational priors

From a neural point of view, our fMRI results revealed that the presentation of a new counterpart yielded enhanced activation in the mPFC when accompanied by a prior (irrespective of it signaling a positive or negative reputation). We suggest that the enhanced mPFC activity may reflect the fact the prior information reduced the uncertainty about the behavior of the other faced by participants when beginning a new RTG. Indeed, this region has been previously implicated in uncertainty resolution in interactive contexts (Yoshida & Ishii, 2006). This is furthermore consistent with the inverse

activation pattern observed in the insula, which showed stronger activity when priors were *not* available, consistently with previous findings reporting a role for this region in tracking increased uncertainty (Preuschoff et al., 2008).

mPFC and dlPFC encode the value of reputation priors

At time of choice, the valence of priors elicited dissociable activation patterns when integrated with the behaviorally estimated (from the prior-based RL model) option values (Q_t). Specifically, the mPFC and dlPFC differentially responded to cooperative vs. individualistic counterparts, however only when priors were available. As reported in previous studies, our results suggest that this brain network keeps track of contextually-modulated decision values over trials, and doing so improves participants' performance (Wunderlich et al. 2009).

As reputational priors conveyed information on the social attitudes of counterparts in our study, this activation is also consistent with a well-established role of the mPFC in ascribing attitudes to others (Mitchell, 2009), and anticipating their choices (Coricelli and Nagel, 2009; Hampton et al., 2008; Krueger et al., 2007). Thus the mPFC is encoding a first response to reputational priors as well as the effect of priors during subsequent interactions. This is in accordance with findings from humans (Hampton et al., 2008; Rilling et al., 2002) and non-human primates (Barraclough, Conroy, & Lee, 2004) on the role of the PFC in encoding value-related signals in repeated interactions.

Caudate nucleus encodes reward PE only when prior information is not provided

Consistent with previous studies, trial-by-trial prediction errors estimated by RL models correlated with activity in the striatum (Bunge et al., 2003, McClure et al., 2003; O'Doherty et al., 2004; King-Casas et al., 2005; Schonberg et al., 2007) but, critically, only when no

priors were available. This confirms a role for the caudate in tracking the difference between expected and obtained outcomes in RTGs, triggering learning. However, when priors were available they appeared to prevent participants from reinforcement-based learning, which was reflected in the reduced covariance between caudate responses and estimated prediction errors.

Priors magnify reward-prediction error signals in the caudate nucleus

As regards the striatal activation patterns, these are well-aligned with an established role of the striatum in tracking reward contingencies, in both non-social (O'Doherty et al., 2004) and social domains (Delgado et al., 2005; King-Casas et al., 2005; Jones et al., 2011). More specifically, the observed patterns are consistent with the idea that the caudate mediates the neural computation of reward prediction error (RPE). Indeed, we observed RPE-pliant signals in the caudate only when no priors were provided, while the same signals appeared blunted when priors were available. Previous studies on non-social tasks (Doll et al., 2009; Doll et al. 2011; Li et al., 2010) and social tasks (Biele et al., 2011; Delgado et al., 2005; Fareri et al., 2012) have shown that, when priors are available, participants tended to hinge on to them, and to relatively discount the impact of the outcomes of their past decisions.

However, in addition to the previous studies, our results show that the presence of priors magnifies striatal deactivation to violations of trust (i.e. when a counterpart with positive reputation, as opposed to no reputation, violated trust), rather than blunting their response. Why previous studies didn't find such magnified response due to violation of priors requires further investigation, though several hypotheses are possible. For instance, two studies (Delgado et al., 2005; Fareri et al., 2012) focused on the subset of unreliable priors, that is, on priors that carried no information on trustees' actual choices;

it is likely that, in such a scenario, participants were gradually learning to disregard such priors, converging towards their extinction rather than exploitation. On the other hand, the opposite may have occurred in a more recent study on the non-social domain (Li et al., 2010), in which priors were perhaps too reliable. Indeed, in that study, agents were explicitly instructed on the precise probabilities of outcomes, which may have reduced their surprise when infrequent, though anticipated losses occurred. In both these previous studies, the space for learning via priors may have been reduced, as the actual prior-to-reward contingencies appeared either non-existent (Delgado et al., 2005; Fareri et al., 2012), or already completely exploited (Li et al., 2010). It is also possible that the different methods used to instil priors tapped on different neural mechanisms: Delgado and colleagues (2005) provided short descriptions of the "moral character" of counterparts, whereas Fareri and colleagues (2012) used direct evidence from previous experience (i.e. playing a ball task). Such methods of instilling priors may have also made them more salient or intuitive and, as a result, harder to extinguish in spite of conflicting evidence. On the other hand, our task reported on characteristics of counterparts that were possibly more directly linked to the main task (i.e. the priors were based on results indicating the extent to which one cares about his own payoffs and that of others - SVO task). Further investigation specifically manipulating prior reliability should clarify some of the points of divergence. Until then, the open question in our study regarded the reason as to why striatal deactivations to trust violations were not leading to behavioral adjustments when priors were available.

VLPFC - Caudate stronger functional connectivity preventing retaliation

On the other hand, when priors were present, we suggest that the impact on learning of the striatal deactivations to violations of trust may have been disrupted by

other brain areas. Our results are in line with attributing this role to the vLPFC, which we found to functionally correlate with such striatal deactivations. In particular, the strength of connectivity between caudate and vLPFC was stronger in the prior compared to the no prior condition. We thus propose that the vLPFC contributes in maintaining choices aligned with the reliable prior beliefs, when beliefs momentarily conflict with observations. This might occur by compensating for the relatively automatic behavioral changes to reward prediction error signals. In line with this interpretation previous literature has implicated the vLPFC in top-down cognitive control by biasing processing in other brain regions towards contextually appropriate representations (Cohen et al., 1990; Millet et al., 2001). Furthermore, not only the vLPFC plays a role in modulating bottom-up fashion cognition processes, but this area has also been found to play a role in goal-directed behavior (Souza et al., 2009; Valentin et al., 2007).

In conclusion, our study integrates theories and methods from cognitive neuroscience, economics, and reinforcement learning to gain a greater understanding of how reputation priors are encoded in the brain and how they affect learning to trust anonymous others. Our findings suggest that priors influence both initial decisions to trust and the following learning mechanisms involved in repeated interactions. Specifically, the present study showed that reputational priors magnify striatal responses to violations of trust. However, when such priors are reliable, other phylogenetically younger brain regions involved in higher cognition may contribute to keep decisions anchored to those priors, thus relatively discounting the weight of conflicting evidence. The interplay between striatum and lateral orbitofrontal cortex may prevent unnecessary retaliation when others violate our trust, and thus constitute an important neuro-cognive mechanism that favors social stability.

Book Chapter The Neuroeconomics of Cognitive Control

Gabriele Chierchia and Giorgio Coricelli

Introduction

In cognitive (neuro)science cognitive control broadly refers to our capacity to go beyond relatively reflexive reactions to salient stimuli in accordance to internal often far-removed goals (Miller, 2000). This idea is of interest to economics: even economist Vilfredo Pareto, among the first and strongest advocates of the separation of economics and psychology (a position similar to that held today by many economists with regard to neuroscience (Gul&Pesendorfer, 2007)), felt the need to distinguish between choice-guided versus routine-guided behaviors, such as "a man removing his hat whenever he enters a drawing room or [perhaps provocatively] a Catholic who regularly attends mass." (Bruni&Sugden, 2007).

Indeed, cognitive control is deeply connected to decision making, and at least three related lines of evidence suggest this: (1) The factors that are held to trigger cognitive control are also implicated in the economic notion of utility; (2) cognitive skills correlate with decision-making tendencies; and (3) cognitive "loads" can impact on decision making. Let us briefly illustrate these points separately.

1. There are many ways to think of and subdivide the environmental or cognitive factors that recruit cognitive control. Norman and Shallice (Shallice, 2000) propose that there are five general classes of them; among which are novelty or complexity of the environments or tasks, performance error, and uncertainty and conflict. Ridderinkhof (2004)

synthesizes these well as situations in which actions need to be adjusted to goals. In what follows we give some examples of why the same factors are fundamental in decision-making and how, in particular, they appear to be connected to the economic notion of utility.

For instance, regrding conflict and uncertainty, imagine we were offered to choose between the following options: a) $10 dollars for sure, or b) a bet on a fair coin flip such that you win $11 for “heads” and $0 otherwise. It probably wouldn't take us much to decide and our responses would be rather automatic (fast) and stereotyped (constant in time, within and between subjects). However, imagine now that we changed the value of the uncertain payoff to $21, keeping the sure payoff fixed at $10, This would probably elicit much more variability in responses and slower response times, which are one of the behavioral signs of cognitive control. Specifically, what happened between the two decision proposals is that we modulated the desirability, henceforth, the utility, of one of the options, thus generating higher conflict and uncertainty; two factors, which in turn, signal that higher cognitive control is required.

Similar reasoning holds for other factors proposed by Norman and Shallice, such as complexity. To give one example of a topic that will be treated later in this chapter, economists have shown that if an option is presented in an ambiguous manner this will decrease its perceived utility (see Theme 4). To keep the example above, in which we win $21 if heads comes out on a coin flip, lets imagine now that we were offered the same option but we are also told that the coin actually isn’t fair, and is unbalanced towards either heads or tails. Though it is intuitive how this might decrease the appeal of the coin-flip bet, relative to the 10$ sure payoff, it isn’t clear that committing to this impression would be the best choice. Indeed, from what we know, the coin has equal chances to be

unbalanced towards the winning outcome (heads) as it does towards the losing one. Thus, ultimately, the expected value of the betting option is the same as before, when we knew the outcome probabilities. What changed is the fact that resolving ambiguity requires second-order probability estimations (inferring the probability of outcome probabilities), which are clearly more complex than reasoning on established outcome probabilities. Thus, aversion to ambiguity in economic decision-making could easily have to do with the fact that increased complexity of ambiguously described options requires more cognitive control, which is costly, and subjects might then avoid ambiguity to not incur such costs.

Though we don't go over all of Norman and Shallice's factors for reasons of space, the above examples should give an idea of how such factors share intricate connections with those thought to shape the utility of options. The following two points provide two general lines of empirical support for this idea.

2. There is behavioral evidence linking cognitive skills to decision preferences. For instance, early studies showed that children who are better at postponing an immediate gratification for a later larger one (Theme 3) are more likely to develop better social and cognitive competences as adolescents (Mischel et al., 1989). Along the same line, adults who obtain higher scores on IQ tests are also less susceptible to risk when deciding in uncertain contexts (Benjamin et al., 2006) where, on average, it has been shown that risk affects people more than it should (Binswanger, 1980) (Theme 2)). Subjects with higher IQ scores are also more patient in postponing gratification and, in social decision contexts, are more generous and cooperative, as well as readier to retaliate if counterparts fail to reciprocate (Rustichini, 2008).

3. Another eloquent example of how (value-unrelated) cognitive processes are linked to decision preferences is given in “cake versus fruit experiments.” (Shiv&Fedorkhin, 1999). In such experiments, subjects were divided into two groups, one of which was asked to remember seven digits, the other only two. Both groups were subsequently asked to choose between a slice of chocolate cake and a bowl of fruit (both equally priced). The “seven-digit group” was shown to more frequently choose the unhealthier but probably more gratifying chocolate cake. These data were taken to suggest that both memorization and decisions tap the related cognitive processes. In other words, the seven-digit group had to process a larger “cognitive load,” leaving it with less cognitive resources to resist the more tempting option.

In summary, cognitive control and decision-making processes are deeply entangled areas of cognition, thus one way to recruit, and study, cognitive control is to make decisions harder, or as some say more interesting (Rustichini, 2008). To do so,in turn, we need to manipulate the factors that determine the utility of options. In what follows we illustrate several of them in respective themes (Themes 1–5), first giving an example and then a definition. We will then proceed by attempting to draw the borders between broad opposing tendencies in the neurocognitive explanations of cognitive control in decision making. In particular, we focus on a dual versus unitary framework (Rustichini, 2008) which has been very prevalent in neuroeconomic research. The “battlegrounds” of such opposing views are the neuroeconomic data, which we illustrate using the factors mentioned in Themes 1 through 5. Throughout, we will argue that, at present, none of such broad models fully accounts for the growing corpus of neuroeconomic data. Finally, we discuss recent neuroeconomic studies that stress a more interdependent nature of controlled and controlling processes in the brain.

Theme 1: Loss Aversion (Kahneman&Tversky, 1979)

Imagine you are invited to either accept or reject the following coin flip bet: heads you win $50, tails you lose $30. If you feel some struggle, that's the grip of loss aversion. Theoretically, winning $50 should attract you more than losing the same amount scares you. However, in a number of experimental settings, people (as well as young children (Harbaugh et al., 2001) and nonhuman primates (Chen et al., 2006)) tend to refuse similarly structured bets. Normally, they require that potential gains nearly double potential losses to take the risk. To account for this, it has thus been proposed that losses are weighted differently from gains.

Theme 2: Risk Aversion (Bernoulli, 1954)

Imagine being proposed the following choice between (a) $100 for sure, or (b) $200 if heads comes up on a fair coin flip. If you choose b, you are susceptible to risk. One definition of risk is variance of outcomes; the two gambles here, indeed, have the same mean, or expected value, but different variance. On average, people are risk averse (Binswanger, 1980) (they tend to go for option a); however, there is also much interindividual variability.

Theme 3: Temporal Discounting (Samuelson, 1937)

Do you prefer (a) $10 right now or (b) $11 next month? If you chose b—that is, you are patient—try increasing the time of payoff receipt in b by, say, an extra month. If you keep repeating this, at some point you are likely to pick option a, no matter how patient you are. Indeed, even though we may expect different people to give different answers on options with specific values, the tendency remains: people (and nonhuman animals, from pigeons to macaques), appear to discount the value of goods, in our case money, as the time to

their receipt increases. Much of this behavior can be accounted for by exponential discounting, which decreases the value of goods constantly across time. However, choose between the following: (c) $10 in 12 months, or (d) $11 in 13 months. This choice should be similar to the former, as we only added one year to both options (a) and (b). However, subjects tend to switch their preferences, from the nearer payoff to the farther one when both far and near payoffs are delayed. If we were to discount goods in a constant fashion (e.g., exponentially), such reversals shouldn't occur. One way to account for this behavior is to hypothesize that discounting is stronger when immediate payoffs are involved, whereas it decreases when there is no possibility to act immediately.

Theme 4: Ambiguity Effect (Ellsberg, 1961)

Imagine a game host offers you two extraction-type lotteries, presented as two boxes, to bet on. For either box, you win $50 if a red ball is extracted. In box 1 is one red ball and one blue ball. In box 2, the game host initially put two red balls and two blue balls and subsequently extracted two balls but didn't show their colors. Thus, in box 2 there could be either two balls of the same color (either red or blue) or one ball of each color. Which box do you prefer to bet on? If you choose box 1 you are susceptible to ambiguity. Indeed, the two boxes offer the same chances of winning (the same expected value). The simplest definition of ambiguity is that outcome probabilities are unknown to the subject.

Theme 5: Framing Effects (Kahneman&Tversky, 1958; Tversky&Kahneman, 1981)

You are offered 100 euros to make two separate choices, 50 prior to each: in choice 1, you are offered to decide between (A) keeping 20 of your 50 euros and (B) betting everything on a "wheel of fortune" type lottery with a 65% chance to keep all and a 35% chance to lose all. Now, suppose you are offered decision 2, between (C) losing 30 of your 50 euros

and (D) betting everything on the same lottery above. If you chose A and D, you are in line with the majority of subjects; alternatively, you might have realized that the two decisions are equivalent. In fact, B = D, but also A = C, since in one case you keep 20, in the other you lose 30 from the originally endowed 50 euros. Indeed, it all boils down to preferring a half empty glass or a half full one: the two glasses refer to the same object, that is, they are extensionally equivalent, as are the preceding prospects; however, subjects tend to reverse their choices according to how the options are framed.

Cognitive Control and Emotions in Economic Decision Making

Let us think in extremes: perhaps the largest doubt one can have about cognitive control is whether it exists at all as a dissociable anatomical and functional system. At the opposite extreme, cognitive control could be completely integrated with other structures/functions unrelated to control, perhaps functionally emerging from a more distributed network. This schematization lends itself to a very broad and yet open-ended debate in cognitive (neuro)science regarding the relatively dualistic or unitary nature of decision processes. This issue is more specific to economics and decision making, as similar debates in economics (i.e., regarding the impact of emotions on decisions) predate brain studies, to the point that some hope that neuroscience could help resolve some of the lingering problems of economics (Camerer et al., 2007).

Dual models stress the relative "independence" of "decision subsystems," that is, systems that can independently generate a decision, "as if" we had different "selves" competing for different options (Laibson, 1997). The unitary approach, on the other hand, also predicts the involvement of a number of "subsystems," however, none of these can generate an

independent decision. From a neuroscientific viewpoint, dual models predict that a dissociable neuroanatomical network subserves cognitive control, whereas in a unitary framework there is no need for a functionally or anatomically distinguishable control system.

The distinction between different subsystems apparent in the dual models often runs parallel to the one between emotional and deliberative processes (Ochsner&Gross, 2005) (or between variously labeled fast and frugal, automatic/effortless, intuitive, experiential or hot processes, on one hand, versus effortful, analytic, rule-based, verbal, cool, or rational processes, on the other) (Mukherjee, 2010).

Broadly speaking, both unitary and dual frameworks have apparent strong and weak points. For instance, it is nearly a truism that the unitary approach is simpler, as it explains decision phenomena with one rather than two systems. The dual system on the other hand appears particularly appealing for explaining "inconsistencies" observed in decision behavior. In what follows, we explain why this is so, reviewing the neuroeconomic literature that has focused on a number of such behavioral inconsistencies (see Themes). We show, however, that in many cases unitary frameworks can also accommodate the data. Throughout, we argue that both models fail to capture some important aspects of how the brain processes decisions.

Loss Aversion (Theme 1)

There is an intuitive appeal in hypothesizing that the different impacts that gains and losses have on behavior could be explained by different underlying neurocognitive systems. In particular, it would be consistent with a dual approach to predict that losses might have a greater impact on behavior as a result of their being processed in more

emotion-related cortical regions. An alternative explanation more consistent with the unitary approach, however, is that the same neural network is differentially recruited by the processing of both gains and losses.

Neuroimaging evidence on healthy decision makers appeared to support the dual systems hypothesis, as the anticipation and experience of economic losses has been repeatedly associated with activity in structures strongly associated with affective and autonomic processing, such as the amygdala and the anterior insula (Knutson&Bossaerts, 2007). With some exceptions (Smith et al., 2009; Yacubian et al., 2006), the same regions were not sensitive to gains, which have instead been shown to recruit a system centered on the midbrain and the striatum, branching to various regions of the PFC (Schultz, 2006). Only one study, by Tom and colleagues (Tom et al., 2007), showed that increasing potential losses and gains recruited a same network, which was activated by gains and deactivated by losses. However, a study that included a task very similar to that used by Tom and colleagues was unable to replicate their results (Canessea et al., 2013). Recently, De Martino and colleagues (2010) showed that patients with circumscribed damage to the amygdala clearly dissociated from their matched controls, as they didn't exhibit loss aversion. Overall, though studies employing different tasks show that the amygdala is sensitive to both positively and negatively valenced cue, (Hamann&Mao, 2002) studies specifically focusing on loss aversion seem to tilt in the direction of a dual view.

Risk (Theme 2)

In a dual view, risk attitudes could be the result of emotions (and emotion-related cortices), which would be modulated by cognitive control in risk-neutral subjects. There is evidence that corroborates this hypothesis. Patients with lesions in areas thought to

integrate emotion and cognition, such as the orbitofrontal cortex (OFC) (Damasio, 1994) exhibit risk-neutral behavior, (Kable&Glimcher, 2007) paradoxically, as do high-scoring subjects on IQ tests (Benjamin, 2006). Moreover, imaging studies revealed that areas previously associated with cognitive control, such as the lateral prefrontal cortex lPFC (Tobler et al., 2006) (in particular, the ventrolateral PFC) play a role in mediating aversion to risk (Tobler et al., 2009). These findings are consistent with transcranial magnetic stimulation (TMS) studies showing the causal regulatory link between the inferior frontal gyrus (IFG) and risky behavior, by which interference with IFG activity using repetitive TMS (rTMS) decreases risk aversion (Knowch et al., 2006). The authors of the latter study propose that, when facing choices between options with different levels of risk (i.e., choose between (a) winning $20 with an 80% chance or lose -$20 otherwise, and (b) winning $80 with a 20% chance or lose -$80 otherwise), the risky option is more salient and attractive, as it usually features a greater outcome. This automatic attraction toward higher-paying outcomes would require the intervention of control processes, which, in turn, would support a more analytical assessment of the options, that is, enabling one to weigh the higher-paying option by its probability, making it overall less attractive. However, Rustichini (2008) stresses that the same data are compatible with a unitary view, as the IFG may subserve general information processing, thus its disruption leads to the failure of integrating reward magnitude and probability. Moreover, although some subcortical and PFC regions appear to code risk and expected value separately (Phelps, 1968; Seymour et al., 2007), others dissociate between the measures through distinct temporal dynamics, rather than regional segregation (such as dopamine neurons – Fiorillo et al., 2003).

Temporal Discounting (Theme 3)

Models to account for preference reversals have been proposed within both (a) dual and (b) unitary models (Rustichini, 2008). Dual type explanations hinge on the idea that competition for guiding behavior occurs between an "impulsive" and a "patient" system. To represent this, Phelps and Pollak proposed (Phelps&Pollak, 1968) a model that employs two parameters in a temporal discounting function. One, "delta," discounts evenly across different time points—and is consistent with the previous exponential discounting (see Theme 3)—the other, "beta," gives the function a steep curvature for immediate rewards. In contrast to this, supporters of the unitary view have often taken from psychophysics, stressing parallelisms with better-understood perceptual systems (Rustichini, 2008). A third line of research has proposed that hyperbolic discounting (i.e., preference reversals) can be explained by a logarithmic perception of time and exponential time discounting (Takahashi, 2005). Incidentally, this seems to be supported by a neuropsychological study showing that ventromedial PFC (vmPFC) patients behaved comparably to controls on intertemporal decisions but were impaired in a task that assessed their ability to consistently focus on different time horizons (Fellows&Farah, 2005). Therefore, even if discounting behaviors can be described as a result of two processes (i.e., patient vs. impatient), they seem to presently leave open a number of possible subfunction combinations.

Taking the "unitary versus dual" dispute into the brain doesn't simplify the scenario foreshadowed by the preceding behavioral debates. A first study by McClure and colleagues (McCLure et al., 2004) was able to dissociate between beta- and delta-pliant systems; a second study by Kable and Glimcher (Kable&Glimcher, 2007) however, showed a unitary set of reward-related regions modulated by near and far rewards, and a third

one by Ballard and Knutson (2009) was partially consistent with both studies. Overall, while there appear to be some "dualisms" in the brain, they don't align well to those of a typical dual model. For instance, dual models predict that a neural system would be preferentially activated by immediate as opposed to future rewards; however, Ballard and Knutson's study suggests that a key dissociation might be between reward magnitude and reward delay, which is compatible with Kable and Glimcher's results. Overall, the most consistent result appears that of an lPFC involvement in the processing of the delay of rewards, as this is confirmed by two of the preceding studies (Ballard&Knutson; McClure et al., 2004) an electrophysiological study on monkeys (Kim et al., 2008), and several patient and imaging studies in different but related tasks (Knoch&Fehr, 2007).Moreover, this idea is not in conflict with Kable and Glimcher's findings, as this could not differentiate well between reward magnitude and delay (Ballard&Knutons, 2009). The lPFC's involvement for processing rewards that are delayed in time is consistent with the notion that this region is needed to override prepotent responses such as those that could derive from the temptation to accept immediate payoffs.

Decisions under Ambiguity (Theme 4)

It could be tempting to explain ambiguity aversion within a dual framework. Not knowing the contingencies of our decision environments could easily "frighten" us, perhaps so quickly and automatically that we don't give ourselves the time to consider the possible situations and make a balanced choice. The first neuroimaging research by Huettel and colleagues (Huetttel et al., 2006) to directly confront neural responses to risk versus ambiguity showed that subjects that chose the ambiguous lotteries more often (see Theme 4) exhibited enhanced inferior frontal gyrus (IFG) activity in response to ambiguity. Such

activity was interpreted to be a signature of cognitive control, which could override the impulsive decision of automatically avoiding ambiguity and plausibly mobilize cognitive resources to explore the ambiguous scenario (i.e., considering the various alternatives underlying the ambiguously described probabilities). A second study, by Hsu et al. (Hsu et al., 2005), was particularly consistent with dual models, as it showed that emotion-related cortices, among which, the amygdala and the OFC, responded preferentially to ambiguity and that striatal responses were more sensitive to risk. The two types of responses also differed in timing, as the amygdala was activated seconds earlier than the striatum. Moreover, the causal role of the OFC in ambiguity processing was demonstrated by the observation that patients with lesions in this area were less sensitive, and even became prone to both ambiguity and risk, relative to their matched controls. Together, the functional magnetic resonance imaging (fMRI) and lesion data leaded the authors to speak of an amygdala-OFC centered vigilance-evaluation system (requiring regulation, via the dorsomedial PFC, or dmPFC) that quickly tracks salient aspects of the stimuli that carry uncertainty-related information (i.e., signaling that information is missing).

Though Hsu and colleagues' results seem to support the idea that risk and ambiguity are processed by distinct mechanisms in the brain, their neuropsychological results also suggested that ambiguity and risk tendencies are connected, as they seemed to correlate in both the control and patient samples (which is consistent with a previous study linking ambiguity and risk in healthy subjects (Boassaerts et all., 2010). In line with this, and closer to a unitary perspective, a study by Levy et al. (Levy et al., 2010) found that the activity in the set of regions, including the medial PFC, striatum, amygdala, and posterior cingulate cortex (PCC) covaried with subjective value in both risky and ambiguous decisions. There was moreover evidence for differential activation patterns (rather than

segregation), as connectivity analysis suggested that connection "weights" are stronger between the amygdala and the striatum under ambiguous than risky choices.

It is hard to argue that these results answer the question of whether dual or unitary systems underlie ambiguity.

Framing Effects (Theme 5)

Consistently with a dual systems approach, it has been proposed that emotional processes may underlie subjects' susceptibility to choices framed either as losses or gains. Such a model would predict that frame-driven behavior would correlate with activity in emotion-related regions and that behavioral consistency across frames (the "rational" behavior) would elicit activity in areas associated with cognitive control, since consistent behavior across different contexts is costly. In line with this, a study by De Martino and colleagues (2006) showed that amygdala activity correlated with risk-averse behavior in "gain frames" and risk-seeking behavior in games framed negatively, which is consistent with the idea that this limbic structure amplifies risk-related biases by processing contextual cues. In contrast, when subjects "resisted" frames, the anterior cingulate cortex (ACC) was preferentially recruited in a subregion later associated with strategic control (Venkatraman et al., 2009). Moreover, the authors obtained individual "rationality" indexes from behavior (a measure of their subjects' degree of susceptibility to frames) that correlated with medial OFC (mOFC) activity. The OFC is considered to integrate emotional valence and goal-oriented behavior (Damasio, 1994), and as such the authors suggested that subjects who chose more "rationally" had richer representations of their own emotional biases, enabling them to better modify their behavior.

Interplay between Emotions and Cognitive Control: A Paradigmatic Example

Coricelli et al (2005) measured brain activity using fMRI while subjects participated in a simple gambling task. The experimental task required subjects to choose between two gambles, each having different probabilities and different expected outcomes. Regret was induced by providing information regarding the outcome of the unchosen gamble. Increasing regret was correlated with enhanced activity in the medial orbitofrontal region, the dorsal ACC and anterior hippocampus. This hippocampal activity is consistent with the idea that a cognitive-based declarative process of regret is engaged by the task. This supports a modulation of declarative (consciously accessible) memory (Eichenbaum, 2004; Steidl et al., 2006) such that after a bad outcome the lesson to be learned is: "In the future pay more attention to the potential consequences of your choice." Furthermore, Coricelli et al. (2005) showed that activity in response to experiencing regret (OFC/ACC/medial temporal cortex) is distinct from activity seen with mere outcome evaluation (ventral striatum), and in response to disappointment elicited by the mismatch between actual and expected outcome of choice. Indeed, the magnitude of disappointment correlated with enhanced activity in middle temporal gyrus and dorsal brainstem, including periaqueductal gray matter, a region implicated in processing aversive signal such as pain. This suggests distinctive neural substrates in reward processing, and that the OFC and medial temporal cortex areas can bias basic dopamine-mediated reward responses (Eichenbaum, 2004).

Coricelli et al. (2005) reported that, across their fMRI experiment subjects became increasingly regret aversive, a cumulative effect reflected in enhanced activity within ventromedial orbitofrontal cortex and amygdala. Under these circumstances, the same pattern of activity that was expressed with the experience of regret was also expressed

just prior to choice, suggesting the same neural circuitry mediates both direct experience of regret and its anticipation. Thus, the OFC and the amygdala contribute to this form of high-level learning based on past emotional experience, in a manner that mirrors the role of these structures in acquisition of value in low-level learning contexts (Gottfried et al., 2003).

Moreover, and of particular interest for our current discussion, affective consequences of choice can induce specific mechanisms of cognitive control (Yarkoni et al., 2005). Coricelli et al. (Coricelli et al., 2005) observed enhanced responses in right dorsolateral prefrontal cortex, right lateral OFC, and inferior parietal lobule during a choice phase after the experience of regret (Coricelli et al., 2005) ,where subsequent choice processes induced reinforcement, or avoidance of, the experienced behavior (Clark et al., 2004) Corroborating results from Simon-Thomas et al,. (2005) show that negative emotions can recruit "cognitive" right hemisphere responses. Thus, negative affective consequences (regret) induce specific mechanisms of cognitive control on subsequent choices. These data suggest a mechanism through which comparing choice outcome with its alternatives (fictive error), and the associated feeling of regret, promotes behavioral flexibility and exploratory strategies in dynamic environments so as to minimize the likelihood of emotionally negative outcomes. These studies stress a more interdependent nature of controlled and controlling processes in the brain.

Brief Discussion and Synthesis

One of the problems with treating cognitive control in economic decision making is that there is a resilient idea that control makes behavior rational and that emotions make it

irrational. A line of literature coming from neuropsychological observations supports this idea: Patients with lesions in the amygdala do not exhibit loss aversion (De Martino et al., 2010) patients with lesions in the OFC/vmPFC are less risk and ambiguity averse, and are close to neutrality in both domains (Hsu et al., 2005), they are also utilitarian in moral decision making (Greene, 2007) and are less influenced by regret in economic decisions (Camille et al., 2004); similarly, subjects with autistic syndromes are less susceptible to framing effects (De Martino et al., 2007). All these pathologies are thus associated with increased "economic rationality" in a number of contexts.

This interpretation, however, ignores the most prominent and consequential behavioral feature of these patients; that is, they are also severely impaired in everyday decision making. In experimental tasks, this is suggested by vmPFC/OFC patients' inability to learn from negative decision outcomes (Damasion, 1994), their impairments in reversal learning (Fellows&Farah, 2005), their violations of preference transitivity (Fellows&Farah, 2007) (i.e., they are more likely to exhibit inconsistent preferences of the type A > B, B > C, but A < C) and abnormal decision making in a number of interactive choice contexts (Van den Bos&Guroglu, 2009) Thus, overall, emotions take part in inconsistent and consistent/adaptive decisions.

This has implications for the dual versus unitary discussion. We suggest that there is a "strong" interpretation of dual models and a "weaker" one. The weak version makes only the first of the following two claims, the strong one makes both: (1) that there are two relatively distinct broad systems in the brain, one that preferentially takes part in fast, effortless, emotional, and context-related processes, another that is preferentially activated in situations requiring control and deliberation; and (2) that these two systems

make separate contributions to, respectively, ”rational” and “irrational” economic decision making. The stronger version appears at odds with current neuroscientific evidence. Our review of neuroimaging evidence further stresses and complicates this point: Even within economic categorizations of behavior, which depend on the factors manipulated in the decision environment (Themes 1–5), the brain is capable of responding either as a unitary or as a dual system, plausibly according to specific differences in task designs that should gradually be disentangled. In none of the individual factors we examined do imaging studies uniquely support either a unitary or dual view: in some designs, the two putative neural systems do not dissociate whereas in others they do. Thus, under a strict falsificationism, both theories are falsified.

Our impression is that the reviewed results appear less odd outside a strict opposition between a dual and unitary framework; although it is hard to deny that there are cortices more related to bodily/emotional processes and others more related to analytical ones (something close to the weaker claim above), results ultimately stress the flexibility with which the two systems seem to interact, thus the different effects cognitive control and emotions can have on behavior.

Outstanding Questions

- What is the relationship between cognitive control and the reward system?
- What is the role of cognitive control in the computations underlying social interaction?
- To what extent do we need cognitive control to behave optimally?

Further Reading

Koechlin E, Hyafil A. 2007. Anterior prefrontal function and the limits of human decision-making. Science 318:594–598. In scenarios in which goals do not match expectations, one of the big problems a cognitive agent faces is that of analyzing and confronting a number of possible plans of actions. However, the lPFC is functionally limited, and only serially represented plans can be processed, as in a bottleneck. The authors suggest that "branching" is the function that counters the bottleneck problem in the lPFC. It is attributed to the FPC (frontopolar cortex, BA 10) and would enable the exploration/execution of a target task, while maintaining a previously selected task in a pending state for subsequent automatic retrieval and execution.

Venkatraman V, Alexandra GR, Taran AA, Huettel SA. 2009. Resolving response, decision, and strategic control: Evidence for a functional topography in dorsomedial prefrontal cortex. J Neurosci 29:13158–13164. Several studies have investigated further functional dissociations within the pmPFC, reporting a ventral-dorsal gradient for emotional versus more cognitive processes as well social relevance. This recent fMRI study further qualifies anatomofunctional specialization of cognitive control in the mPFC.

References

Akerlof, G. and R. Kranton, Economics and Identity, Quarterly Journal of Economics CVX (3), (August 2000), pp. 715–753.

Ballard K, Knutson B. (2009). Dissociable neural representations of future reward magnitude and delay during temporal discounting. Neuroimage 45: 143–150.

Barraclough, D. J., Conroy, M. L., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. Nature neuroscience, 7(4), 404–410.

Bates, D., & Sarkar, D. (2006). lme4: Linear mixed-effect models using S4 classes

Bernhard, F, Fehr E, Fischbacher U. (2006). Group Affiliation and Altruistic Norm Enforcement. American Economic Review, 96 (2): 217–221.

Benjamin DJ, Brown SA, Shapiro JM. (2006). Who is "behavioral"? Cognitive ability and anomalous preferences. Unpublished working paper.

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

Bernoulli D. (1954). 1738. Exposition of a new theory on the measurement of risk (Translation of Speciment theoriae novae de mensura sortis). Econometrica 22: 23–36.

Biele G, Rieskamp J, Krugel LK, Heekeren HR (2011) The Neural Basis of Following Advice. PLoS Biol 9: e1001089.

Binswager HP. (1980). Attitudes toward risk: experimental measurement in rural India. Am J Agric Econ 62: 395–407. Boero R, Bravo G, Castellani M, Squazzoni F (2009) Reputational cues in repeated trust games. The Journal of Socio-Economics 38:871-877.

Bohnet I & Frey B (1999). Social Distance and other-regarding Behavior in Dictator Games. Comment. American Economic Review.

Bossaerts P, Ghirardato P, Guarnaschelli S, Zame WR. (2010). Ambiguity in asset markets: theory and experiment. Rev Financ Stud 23: 1325–1359.

Brandts J & Charness G. (2000). Hot vs. Cold: Sequential Responses and Preference Stability in Experimental Games. Experimental Economics.

Bruni L, Sugden R. (2007). The road not taken: how psychology was removed from economics, and how it might be brought back. Econ J 117: 146–173.

Bulow J, Geanakoplos J, Klemperer P (1985). Mulitmarket oligopoly: Strategic substitutes and complements. Journal of Political Economy

Bunge SA (2004) How we use rules to select actions: a review of evidence from cognitive neuroscience. Cogn Affect Behav Neurosci 4:564-579.

Camerer CF, Loewenstein G, Prelec D. (2005). Neuroeconomics: how neuroscience can inform economics. J Econ Lit 34: 9–64.

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

Canessa N, Chierchia G, Motterlini M, Baud-Bovy G, Tettamanti M, Cappa S. (2013). Distinct neural correlates for the processing of magnitude, probability and uncertainty of potential monetary wins and losses.

Camerer C. (2003). Behavioral Game Theory. Princeton University Press

Camerer C & Fehr E (2006). When does economic man dominate social behavior? Science

Camerer C, Ho T, Chong J (2004). Behavioral Game Theory: Thinking, Learning and Teaching." Nobel Symposium on Behavioral and Experimental Economics, Stockholm, December 2001.

Camerer C, Weigelt K (1988) Experimental Tests of a Sequential Equilibrium Reputation Model. Econometrica 56:1-36.

Charness G, Haruvy E, Sonsino D (2003). Social Distance and Reciprocity: An Internet Experiment. J. of Economic Behavior and Organization.

Charness G, Rigotti L, Rustichini A. (2007). Individual behavior and group membership. American Economic Review, 97: 1340-1352.

Chen MK, Lakshminaryanan V, Santos LR. (2006). The evolution of our preferences: evidence from capuchin monkey trading behavior. J Polit Econ 114: 517–537.

Chen R & Chen Y (2011). The Potential of Social Identity for Equilibrium Selection. American Economic Review

Clark L, Cools R, Robbins TW. (2004). The neuropsychology of ventral prefrontal cortex: decision-making and reversal learning. Brain Cogn 55: 41–53.

Cohen, M. X., Heller, A. S., & Ranganath, C. (2005). Functional connectivity with anterior cingulate and orbitofrontal cortices during decision-making, Brain Res Cogn Brain Res. 23(1).

Cohen JD, Dunbar K, McClelland JL. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. Psychol Rev. 97:332–61.

Coleman JS (1994) Foundations of Social Theory: Harvard University Press.

Coleman TF, Li Y (1996) An Interior Trust Region Approach for Nonlinear Minimization Subject to Bounds. SIAM Journal on Optimization 6:418.

Coricelli G, Critchley HD, Joffily M, O'Doherty JP, Sirigu A, Dolan RJ. (2005). Regret and its avoidance: a neuroimaging study of choice behavior. Nat Neurosci 8: 1255–1262.

Coricelli G, Nagel R (2009) Neural correlates of depth of strategic reasoning in medial prefrontal cortex. Proc Natl Acad Sci U S A 106:9163-9168.

Damasio AR. (1994). Decartes' Error: Emotion, Reason, and the Human Brain. New York: Putnam Publishing.

De Cremer & Van Vugt (1999). Social identification effects in social dilemmas: a transformation of motives. European Journal of Social Psychology.

De Martino B, Camerer CF, Adolphs R. (2010). Amygdala damage eliminates monetary loss aversion. Proc Natl Acad Sci USA 107: 3788–3792.

De Martino B, Harrison NA, Knafo S, Bird G, Dolan RJ. (2007). Explaining enhanced logical consistency during decision making in autism. J Neurosci 28: 10746–10750.

De Martino B, Kumaran D, Seymour B, Dolan RJ. (2006). Frames, biases, and rational decision-making in the human brain. Science 313: 684–687.

Delgado MR, Frank RH, Phelps EA (2005) Perceptions of moral character modulate the neural systems of reward during the trust game. Nat Neurosci 8:1611-1618.

Doll B. B., Hutchison K. E., Frank M. J. (2011). Dopaminergic genes predict individual differences in susceptibility to confirmation bias. The Journal of neuroscience 31, 6188–6198.

Doll B. B., Jacobs W. J., Sanfey A. G., Frank M. J. (2009). Instructional control of reinforcement learning: a behavioral and neurocomputational investigation. Brain Res. 1299, 74–94.

Eddy WF, Fitzgerald M, Genovese CR, Mockus A, Noll DC. (1996). Functional image analysis software - computational olio. In: Prat A, editor. Proceedings in Computational Statistics. Heidelberg: Physica-Verlag;. pp. 39–49.

Efferson C, Lalive R, Fehr E (2008). The Coevolution of Cultural Groups and Ingroup Favoritism. Science.

Eichenbaum H. (2004). Hippocampus: cognitive processes and neural representations that underlie declarative memory. Neuron 44: 109–120.

Ellsberg D. (1961). Risk, ambiguity, and the savage axioms. Q J Econ 75: 643–699.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. The quarterly journal of economics, 114(3), 817-868.

Fellows LK, Farah MJ. (2005). Different underlying impairments in decision-making following ventromedial and dorsolateral frontal lobe damage in humans. Cereb Cortex 15: 58–63.

Fellows LK, Farah MJ. (2005). Dissociable elements of human foresight: a role for the ventromedial frontal lobes in framing the future, but not in discounting future rewards. Neuropsychologia 43: 1214–1221.

Fellows LK, Farah MJ. (2007). The role of ventromedial prefrontal cortex in decision making: judgment under uncertainty or judgment per se? Cereb Cortex 17: 2669–2674.

Fareri D. S., Chang L. J., & Delgado M. R.. (2012). Effects of Direct Social Experience on Trust Decisions and Neural. Front Neurosci., 6: 148.

Fehr E & Fischbacher U (2003). The nature of human altruism. Nature

Fiorillo CD, Tobler PN, Schultz W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. Science 299: 1898–1902.

Fiske S, Cuddy A, Glick P (2006). Universal dimensions of social cognition: warmth and competence. TICS.

Fouragnan E, Chierchia G, Greiner S, Neveu R, Avesani P, Coricelli G (2013). Reputational Priors Magnify Stratal Responses to Violations of Trust. J. of Neuroscience.

Fowler JH, Dawes CT, Christakis NA (2009) Model of genetic variation in human social networks. PNAS

Friston, K. J., Buechel, C., Fink, G. R., Morris, J., Rolls, E., & Dolan, R. J. (1997). Psychophysiological and modulatory interactions in neuroimaging. NeuroImage, 6(3), 218–229.

Fudenberg D, Kreps DM, Maskin ES (1990) Repeated Games with Long-Run and Short-Run Players. The Review of Economic Studies 57:555.

Fuller, S. (2005). Philosophical History Of Our Times, A. Orient Blackswan.

Gachter G & Fehr E (1999). Collective action as a social exchange. J. of Economic Behavior and Organization.

Goldman A (2006). Simulating minds: The philosophy, psychology, and neuroscience of mindreading. Oxford University Press, USA

Gottfried JA, O'Doherty J, Dolan RJ. (2003). Encoding predictive reward value in human amygdala and orbitofrontal cortex. Science 301: 1104–1107.

Greene JD. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. Trends Cogn Sci 11: 322–323.

Guala F, Mittone L, Ploner M (2012). Group membership, team preferences and expectations. J. of Economic Behavior and Organization.

Gul F, Pesendorfer W. (2007). The case for mindless economics. In: Handbook of Economic Methodologies (Caplin A, Schotter A, eds). New York: Oxford University Press.

Haan M, Kooreman P, Riemersma T (2006) Friendship in a Public Good Experiment. IZL Discussion Paper series No. 2108. Bonn, Germany: Institute for the Study of Labor.

Hamann S, Mao H. 2002. Positive and negative emotional verbal stimuli elicit activity in the left amygdala. Neuroreport 13: 15–19. Hamilton WD (1964) The genetical evolution of social behaviour I & II. J Theor Biol 7: 1–52.

Hamlin J., Wynn K, Bloom P (2007). Social evaluation by preverbal infants. Nature.

Hampton AN, Bossaerts P, O'Doherty JP (2008) Neural Correlates of Mentalizing-Related Computations During Strategic Interactions in Humans. Proceedings of the National Academy of Sciences 105:6741-6746.

Harbaugh WT, Krause K, Vesterlund L. (2001). Are adults better behaved than children? Age, experience, and the endowment effect. Econ Lett 70: 175–181.

Harrison F, Sciberras J, James R (2011). Strength of Social Tie Predicts Cooperative Investment in a Human Social Network. Plos1.

Heinemann F, Nagel R, Ockenfels P (2008). Measuring Strategic Uncertainty in Coordination Games. Review of Economic Studies.

Heinemann F, Nagel R, Ockenfels P (2004). Global Games on Test: Experimental Analysis of Coordination Games with Public and Private Information. Econometrica.

Hoffman E., McCabe K, Smith V (1996). Social distance and other-regarding behavior in dictator games. American Economic Review.

Hsu M, Bhatt M, Adolphs R, Tranel D, Camerer CF. (2005). Neural systems responding to degrees of uncertainty in human decision-making. Science 310: 1680–1683.

Huettel SA, Stowe CJ, Gordon EM, Warner BT, Platt ML. (2006). Neural signatures of economic preferences for risk and ambiguity. Neuron 49: 765–775.

Isaac, R M, McCue K, Plott C (1985) Public goods provision in an experimental environment. Journal of Public Economics. 26: 51–74.

Jones B, Raichlin H (2006) Social discounting. Psychological Science.

Jones R. M., Somerville L. H., Li J., Ruberry E. J., Libby V., Glover G., et al. (2011). Behavioral and neural properties of social reinforcement learning. The Journal of neuroscience 31, 13039–13045.

Kable JW, Glimcher PW. (2007). The neural correlates of subjective value during intertemporal choice. Nat Neurosci 10: 1625–1633.

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

Kim HH (2009) Market Uncertainty and Socially Embedded Reputation. American Journal of Economics and Sociology 68:679-701.

Kim S, Hwang J, Lee D. (2008). Prefrontal coding of temporally discounted values during intertemporal choice. Neuron 59: 161–172.

King-Casas B, Tomlin D, Anen C, Camerer CF, Quartz SR, Montague PR (2005) Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange. Science 308:78-83.

Knoch D, Fehr E. (2007). Resisting the power of temptations: the right prefrontal cortex and self-control. Ann N Y Acad Sci 1104: 123–134.

Knoch D, Gianotti LR, Pascual-Leone A, Treyer V, Regard M, Hohmann M, Brugger P. (2006). Disruption of right prefrontal cortex by low-frequency repetitive transcranial magnetic stimulation induces risk-taking behavior. J Neurosci 26: 6469–6472.

Knutson B, Bossaerts P. (2007). Neural antecedents of financial decisions. J Neurosci 27: 8174–8177.

Krienen F, Tu P, Buckner R (2010). Clan Mentality: Evidence That the Medial Prefrontal Cortex Responds to Close Others. J. of Neuroscience.

Krueger F, McCabe K, Moll J, Kriegeskorte N, Zahn R, Strenziok M, Heinecke A, Grafman J (2007) Neural Correlates of Trust. Proceedings of the National Academy of Sciences 104:20084-20089.

Laibson D. 1997. Golden eggs and hyperbolic discounting. Q J Econ 112: 443–477. Li, J., Delgado, M. R., & Phelps, E. A. (2011). How instructed knowledge modulates the neural systems of reward learning. Proceedings of the National Academy of Sciences, 108(1), 55–60.

Lamb M (1977). Father-Infant and Mother-Infant Interaction in the First Year of Life. Child development.

Levy I, Snell J, Nelson A, Rustichini A, Glimcher P. (2010). Neural representation of subjective value under risk and ambiguity. J Neurophysiol 103: 1036–1047.

Long Y, Jiang X, Zhou X (2012) To believe or not to believe: trust choice modulates brain responses in outcome evaluation. Neuroscience 200:50-58.

Martin, C. F., Bhui, R., Bossaerts, P., Matsuzawa, T., Camerer, C., & Camerer, C. Experienced chimpanzees behave more game‐‐theoretically than humans in simple competitive interactions.

McCabe K, Houser D, Ryan L, Smith V, Trouard T (2001) A Functional Imaging Study of Cooperation in Two-Person Reciprocal Exchange. Proceedings of the National Academy of Sciences 98:11832-11835.

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal Prediction Errors in a Passive Learning Task Activate Human Striatum. Neuron, 38(2), 339–346.

McClure S.M., Laibson D. I., Loewenstein G, Cohen JD. (2004). Separate neural systems value immediate and delayed monetary rewards. Science 306: 503–507.

Messick, D.M., & McClintock, C.G. (1968). Motivational basis of choice in experimental game. Journal of Experimental Social Psychology, 4, 1-25.

Miller EK. (2000). The prefrontal cortex and cognitive control. Nat Rev Neurosci 1: 59–65.

Miller EK, Cohen JD (2001). An integrative theory of prefrontal cortex function. Annu Rev Neurosci.; 24:167–202.

Mischel W, Shoda Y, Rodriguez MI. (1989). Delay of gratification in children. Science 244: 933–938.

Mitchell JP (2009) Social psychology as a natural kind. Trends Cogn Sci 13:246-251.

McPherson M, Smith-Lovin L, Cook JM. (2001). Birds of a feather. Homophily in social networks. Annual Review of Sociology.

Mitchell JP, Macrae CN, Banaji MR. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. Neuron.

Morris, J. S., Frith, C. D., Perrett, D. I., Rowland, D., Young, A. W., Calder, A. J., & Dolan, R. J. (1996). A differential neural response in the human amygdala to fearful and happy facial expressions. Nature, 383, 383(6603), 812–815.

Morry M. (2007). Relationship satisfaction as a predictor of perceived similarity among cross-sex friends: A test of the attraction-similarity model. Journal of Social and Personal Relationships.

Mukherjee A. (2010). Dual system model of preferences under risk. Psychol Rev 117: 243–255.

Norman D. A., Shallice T. (2000). Attention to action: willed and automatic control of behaviour. In: CHIP Report 99. San Diego: University of California.

O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ (2004). Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. Science 304:452-454.

Ochsner KN, Gross JJ. (2005). The cognitive control of emotion. Trends Cogn Sci 9: 242–249.

Olson K. & Spelke E (2008). Foundations of cooperation in young children. Cognition.

Pastor-Nieto R (2001). Grooming, kinship, and co-feeding in captive spider monkeys (Ateles geoffroyi). Zoo Biology 20: 293–303.

Phan KL, Sripada CS, Angstadt M, McCabe K (2010). Reputation for Reciprocity Engages the Brain Reward Center. Proceedings of the National Academy of Sciences 107:13099-13104.

Phelps ES, Pollak RA. (1968). On second-best national saving and game-equilibrium growth. Rev Econ Stud 35: 185–199.

Phillips PJ, Moon H, Rizvi SA, Rauss PJ (2000). The FERET evaluation methodology for face-recognition algorithms. Pattern Analysis and Machine Intelligence, IEEE 22:1090 - 1104.

K., Quartz S.R. , Bossaerts P. (2008). Human Insula Activation Reflects Risk Prediction Errors As Well As Risk. The Journal of Neuroscience 28:2745-2752.

Ridderinkhof KR, Ullsperger M, Crone E.A., Nieuwenhuis S. (2004). The role of the medial frontal cortex in cognitive control. Science 306: 443–447.

Rescorla RA, Wagner AW, Black AH, Prokasy WF (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Classical Conditioning II: Current Research and Theory, pp 64-99: Appleton-Century-Crofts.

Reuben E, Winden F. (2008). Social ties and coordination on negative reciprocity: The role of affect. Journal of Public Economics 92 34–53

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

Rosenthal, R. (1981). "Games of Perfect Information, Predatory Pricing, and the Chain Store". Journal of Economic Theory 25 (1): 92–100.

Robbins J, Krueger J (2005). Social Projection to Ingroups and Outgroups: A Review and Meta-Analysis. Personality and social psychology review.

Rummery, G. A., & Niranjan, M. (1994). On-Line Q-Learning Using Connectionist Systems.

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

Samuelson PA. (1937). A note on measurement of utility. Rev Econ Stud 4: 155–161.

Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. The Journal of neuroscience 27(47), 12860–12867.

Schultz W. 2006. Behavioral theories and the neurophysiology of reward. Annu Rev Psychol 57: 87–115. Sally D. (2001). On sympahty and games. Journal of Economic Behavior & Organization. Vol. 44 1–30.

Schelling Thomas. (1960). The Strategy of Conflict. Cambridge, MA: Harvard University Press.

Segal N & Hershberger S (1999). Cooperation and Competition Between Twins: Findings from a Prisoner's Dilemma Game. Evolution and Human Behavior.

Selten, R. (1967). "Die Strategiemethode zur Erforschung des eingeschränkt rationalen

Verhaltens im Rahmen eines Oligopolexperiments," in: Sauermann, H. (Ed.), Beiträhe zur experimentellen Wirtschaftsforschung, Tübingen: J.C.B. Mohr, 136-168.

Seymour B, Daw ND, Dayan P, Singer T, Dolan RJ. (2007). Differential encoding of losses and gains in the human striatum. J Neurosci 27: 4826–4831.

Shiv B, Fedorkhin A. (1999). Heart and mind in conflict: the interplay of affect and cognition in consumer decision making. J Consum Res 26: 278–292.

Simon-Thomas ER, Role KO, Knight RT. (2005). Behavioral and electrophysiological evidence of a right hemisphere bias for the influence of negative emotion on higher cognition. J Cogn Neurosci 17: 518–529.

Smith BW, Mitchell DG, Hardin MG, Jazbec S, Fridberg D, Blair RJ, Ernst M. (2009). Neural substrates of reward magnitude, probability, and risk during a wheel of fortune decision-making task. Neuroimage 44: 600–609.

Souza MJ, Donohue SE, Bunge SA (2009) Controlled retrieval and selection of action-relevant knowledge mediated by partially overlapping regions in left ventrolateral prefrontal cortex. NeuroImage 46:299-307.

Stanley DA, Sokol-Hessner P, Fareri DS, Perino MT, Delgado MR, Banaji MR, Phelps EA (2012). Race and Reputation: Perceived Racial Group Trustworthiness Influences the Neural Correlates of Trust Decisions. Philosophical Transactions of the Royal Society B: Biological Sciences 367:744-753.

Steidl S, Mohi-uddin S, Anderson AK. (2006). Effects of emotional arousal on multiple memory systems: evidence from declarative and procedural learning. Learn Mem 13: 650–658.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. Psychological Review, 88(2), 135–170.

Tajfel H, Turner J. (1979). An Integrative Theory of Intergroup Conflict. In Stephen Worchel and William Austin, eds., The Social Psychology of Intergroup Relations, Monterey, CA: Brooks/Cole.

Takahashi T. 2005. Loss of self-control in intertemporal choice may be attributable to logarithmic time perception. Med Hypotheses 65: 691–693.

Tobler PN, Christopoulos GI, O'Doherty JP, Dolan RJ, Schultz W. (2009). Risk-dependent reward value signal in human prefrontal cortex. Proc Natl Acad Sci USA 106: 7185–7190.

Tobler PN, O'Doherty JP, Dolan RJ, Schultz W. (2007). Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. J Neurophysiol 97: 1621–1632.

Todd AR, Hanko K, Galinsky AD, Mussweiler T. (2010). When Focusing on Differences Leads to Similar Perspectives. Psychological Science 2011 22: 134.

Tom SM, Fox CR, Trepel C, Poldrack RA. (2007). The neural basis of loss aversion in decision-making under risk. Science 315: 515–518.

Tversky A, Kahneman D. (1981). The framing of decisions and the psychology of choice. Science 211: 453–458.

Valentin VV, Dickinson A, O'Doherty JP (2007) Determining the Neural Substrates of Goal-Directed Learning in the Human Brain. The Journal of Neuroscience 27:4019-4026.

Van den Bos W, Güroglu B. (2009). The role of the ventral medial prefrontal cortex in social decision making. J Neurosci 29: 7631–7632.

Van Huyck JB, Battalio RC, Biel RO. (1990). Tacit coordination games, strategic uncertainty and coordination failure. American Economic Review.

Van Lange PA The pursuit of joint outcomes and equality in outcomes : An integrative model of social value orientation. Journal of personality and social psychology 77:337-349.

Venkatraman V, Alexandra GR, Taran AA, Huettel SA. (2009). Resolving response, decision, and strategic control: evidence for a functional topography in dorsomedial prefrontal cortex. J Neurosci 29: 13158–13164.

West S, Diggle S, Buckling, Gardner A, Griffin A (2007). The social lives of microbes. Annual Review of Ecology, Evolution and Systematics.

Wiltermuth, S.S., & Heath, C. (2009). Synchrony and cooperation. Psychological Science

Wittmann, B. C., Daw, N. D., Seymour, B., & Dolan, R. J. (2008). Striatal activity underlies novelty-based choice in humans. Neuron, 58(6), 967–973.

Wood, J. N., & Grafman, J. (2003). Human prefrontal cortex: processing and representational perspectives. Nature Reviews Neuroscience, 4(2), 139-147.

Wunderlich, K., Rangel, A., & O'Doherty, J. P. (2009). Neural computations underlying action-based decision making in the human brain. Proceedings of the National Academy of Sciences, 106(40), 17199–17204.

Yacubian J, Gläscher J, Schroeder K, Sommer T, Braus DF, Büchel C. (2006). Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain. J Neurosci 26: 9530–9537.

Yarkoni T, Gray JR, Chrastil ER, Brach DM, Green L, Braver TS. (2005). Sustained neural activity associated with cognitive control during temporally extended decision making. Brain Res Cogn Brain Res 23: 71–84.

Yamagishi T, Cook KS (1993). Generalized exchange and social dilemmas. Soc Psychol Quart 56: 235–248.

Yoshida, W., & Ishii, S. (2006). Resolution of uncertainty in prefrontal cortex. Neuron, 50(5), 781–789.