In brains we trust: How neuroeconomists stylize trust, the brain, and the social world

Klaassen, P.

Chapter 5

Passive connections created: Neuroeconomic facts about trust

the signal of resistance opposing free, arbitrary thinking is called a *fact*
Fleck (1979, p.101)

5.1 Introduction

Now that we are introduced to neuroeconomics, we can descend into the esoteric circle of the endeavor and excavate what facts concerning trust are established in it. In the present chapter, this will be done rather clinical, reserving the epistemological, ontological and social questions pertaining to the neuroeconomic stylization of trust for chapters 6 and 7. Before going there, however, I first provide a brief ground plan and some pointers.

Oxytocin and functional localization

One of the cornerstones of the neuroeconomics of trust is its research on the neuropeptide oxytocin. We have encountered this already in section 1.1 and in the previous chapter, but in sections 5.2 and 5.3 I will described this in somewhat more detail. Surely oxytocin is on its way to achieving iconic status. However, neuroeconomics, since its days of incipience, shares in the brilliance of one of our contemporary icons of scientific progress—the whole-brain images produced through positron emission tomography (PET) and, especially, functional magnetic resonance imaging (fMRI). Four years before the publication by Kosfeld

et al. on the role of oxytocin in human trust, the first ever "behavioral economics in the scanner" experiment was reported on by Kevin McCabe and colleagues (2001), and this experiment crucially involved the Trust Game. Their work constitutes the first publication regarding the neuroimaging of trust.[1]

When McCabe et al. decided to align behavioral economics, psychology and fMRI neuroimaging all in one experimental protocol, this must have looked groundbreaking from the point of view of behavioral and experimental economists. From the point of view of (cognitive) neuroscientists, however, it probably appeared to be simply one among many logical options of what to do next. For even though it had been less than ten years since fMRI had been welcomed into the family of neuroscientific research technologies, it had already become one of its most prominent members, especially in the context of cognitive neuroscience.[2]

Figure 5.1 illustrates fMRI's fast growth: by the time of the publication of McCabe et al.'s publication, fMRI had already worked its way into about a quarter of all publications in the *Journal of Cognitive Neuroscience*. Since then its share in neuroscientific research has only continued to grow.

In conformity with the general trend, further imaging research on or relating to the neural basis of trust followed after McCabe et al.'s publication. Anticipating the analysis of the neuroeconomics of trust that will follow in chapter 6, I can say that such research confirms that neuroeconomics is coherently stylized, even though variation in the choice of research technologies is possible. Expressly, neuroimaging studies of trust by and large mobilize much the same network of linkages and are, accordingly, argumentatively structured in a way very similar to neuroeconomic research on trust involving neuroendocrinology (i.e., to neuroeconomic research focused on the role of oxytocin in trust).

Levels of neuroeconomic analysis

Different as neuroeconomic investigations of trust based in either neuroendocrinological or neuroimaging research may be, they both promise a comparable return in terms of passive linkages—that is, facts about the neurobiology of trust and economic or social decision making and facts concerning such social phenomena that can (only) be established as a consequence of this biological knowledge. Such facts involve either specific neural correlates or the workings of neuropeptides such as oxytocin.[3]

As we will see, neuroimaging studies concerning trust are comparably networked as neuroendocrinological studies of trust. The neuroscientific part of all such investigations concern the level of implementation, to put it in terms analogous to those of David Marr's well-known model, which was developed in the context of the neuroscience of perception, and in terms resembling those used in the historical introduction to neuroeconomics

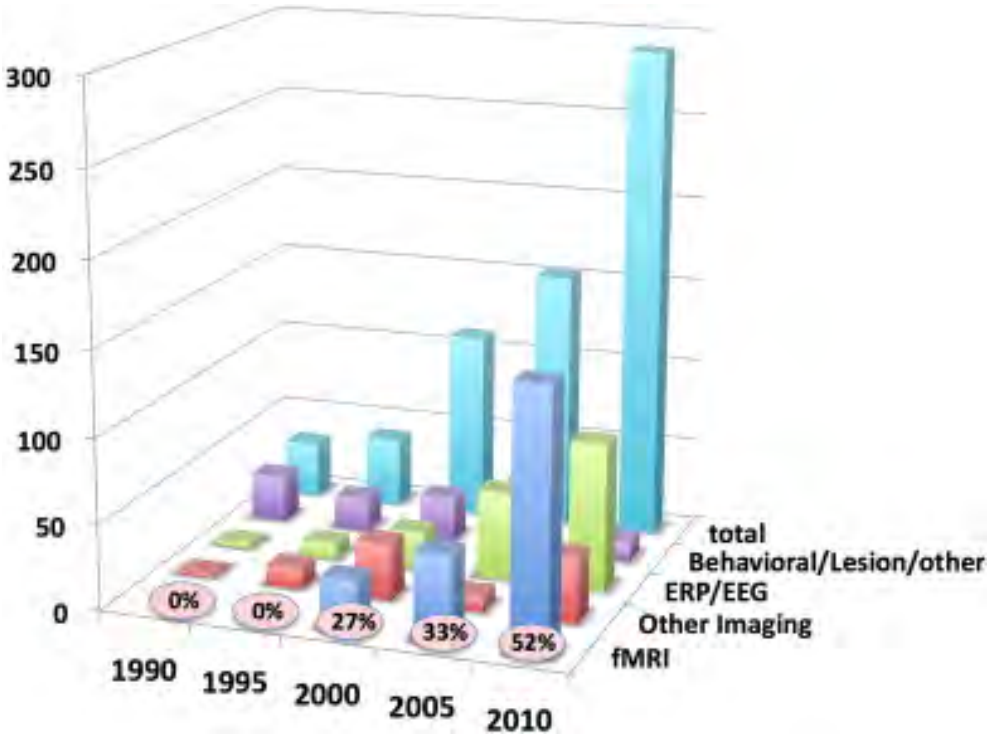

Figure 5.1: An overview of the use of various research technologies in papers published in the *Journal of Cognitive Neuroscience*, 1990–2010.
Adapted from Rosen (2012, p.1320)

presented in the previous chapter. This level exists next to two other levels: the levels of meaning or function, on the one hand, and the level of algorithms, on the other. The level of function at issue here is explicated in (behavioral) economic terms, in which social or economic decision making has been modeled and where behavior is the relevant measure. The algorithms in question are the domain of cognitive psychology. Here the distinctive measures are the data from questionnaires, which are part of most if not all neuroeconomic investigations, and here there are "theoretical entities" too—namely, the psychological interpretations given to shed light on both observed behavior (function/ meaning) and neural activation patterns (implementation).

5.2 Neuroendocrinological neuroeconomics

Oxytocin increases trust in humans. This is the fact established by Kosfeld et al. which we saw readers of *Nature* were "prepared for" by Damasio's introductory article (2005). In the next chapter I will examine the active linkages at work in the establishment of this fact; here, however, I will present a preparatory overview of the ground covered in Kosfeld et al.'s article (2005) in order to provide a better understanding of what it means to state that oxytocin increases trust. To begin with, then, one must know that this fact is established using an experimental setup composed of two components: a specific adaptation of a game previously developed in behavioral economics for probing strategic or social interactions between individuals, and a component built on the branch of neuroscience that goes by the name of neuroendocrinology and deals with endocrine secretion in the brain and with the workings of neuroendocrine substances. As for the first component, experimental subjects are asked to play the Trust Game—the very same game subjects played in the experiment already briefly presented in chapter 1. As for the second component, the focus here is again on oxytocin.

Trust Game

The Trust Game is played between two different people, where Player One is often called the "investor" and Player Two the "trustee." Many variations are possible and can be found in the literature, but in the version of the game at issue here, graphically displayed in figure 5.2, the game was played as an anonymous one-shot, one-way Trust Game. That it is a one-shot game means that investor and trustee played together only once. This is to ensure that learning effects or built-up reputations do not interfere with the behavior of the players. That it is a one-way game means that each player could make one move only. As will become clear, this entails that trust is only potentially visible in one player—the investor.

At the start of each round of play, both the investor and the trustee received twelve monetary units. After that, investors had four discrete options for action: transferring zero, four, eight or twelve monetary units to the trustees. On the way from the investors to the trustees, the number of monetary units sent would be tripled by the experimenters, and this was common knowledge to both investors and trustees. Hence, trustees could receive zero, twelve, twenty-four or thirty-six monetary units, and consequently, prior to their move they would have at their disposal twelve, twenty-four, thirty-six or forty-eight monetary units. Therefore, they could choose to return any number of monetary units to the investors they had at their disposal.

All in all, both players could potentially end up with higher payoffs then they started out

with. This only happens, though, if the investors send at least some monetary units to the trustees, and if the trustees subsequently return more monetary units to the investors than had been initially sent to them. That this is possible is of course due to the tripling of monetary units that takes place in between investor and trustee. If both players profit, then, it is thought to be because the investor *trusts* the trustee to be trustworthy (i.e., to return more monetary units than the investor had sent). Thus, we see that trust is operationalized in the Trust Game as a specific behavior: investors sending monetary units to trustees.

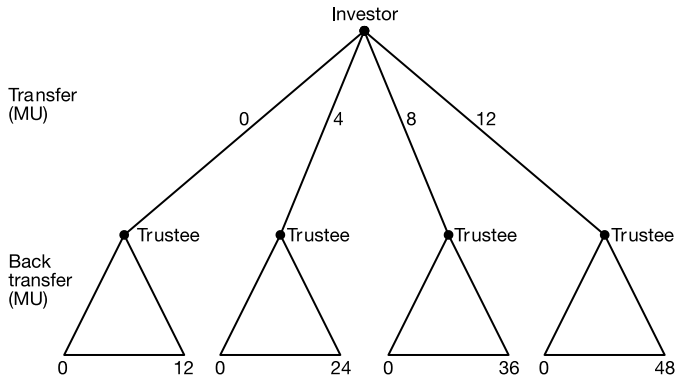

Figure 5.2: A schematic of the Trust Game as played in the experiment by Kosfeld et al.*Reproduced from Kosfeld et al. (2005).*

Now it is clear why this is properly called a one-way Trust Game: only the move by the investor can be characterized as revealing or involving trust. For only the investor is in a situation in which he or she can take a costly action that might, in the future, be repaid by an exchange partner; only the investor runs a risk, and only the investor's final outcome is fully dependent on his or her exchange partner's decision. That a risk is involved in the behavior of the first mover is crucial, as will be discussed more extensively in chapter 6.

What would you do in a game like this—or perhaps better: What *should* you do? To start with, let us consider what trustees should do. Trustees aiming for maximized personal gain would never return anything: Investor and trustee play together only once during the experiment, so trustees have no reason whatsoever to honor the trust placed in them by investors. As rational choice theorists tend to put it, any incentive to honor the trust that one's exchange partner gives is lacking here (see e.g. Hardin 2002).

Investors, for their part, can use backward inference in their decisions. Accordingly, investors can preemptively decide not to transfer anything to the trustees. This, then, is the noncooperative equilibrium of the Trust Game predicted by rational choice theory: The

investors do not send anything to the trustees, and the trustees do not send anything back. No one gains from the agreed upon tripling of the sum transferred by the investor to the trustee, and thus both players end up with less than they potentially could have had.

Oxytocin

However, as behavioral game theory had already established prior to the work of Kosfeld et al. (see, for example, Berg et al. 1995), people in fact tend not to act as predicted by rational choice theory in games such as these. This was affirmed in Kosfeld et al.'s study, which, however, aimed at something in addition to simply replicating previous results from behavioral economics. That is to say, in their experiment we find the coupling of behavioral economics with something more complex, namely, the experimental investigation of the precise role a specific neuromodulator plays in trust and social interaction by way of its effect on the human nervous system. The guiding hypothesis behind this experiment is that oxytocin is an important part of the biological basis of trust. More specifically, it is hypothesized that the intranasal administration of exogenously produced oxytocin will increase the amount of trust investors put in trustees. And indeed, according to Kosfeld et al., the main result of their experiment was the confirmation of this hypothesis.

Thus we have arrived at the neuroendocrinological component of the experiment, and the confirmation that oxytocin increases trust in humans. I will discuss this in detail in the next chapter; for now I simply emphasize that, first, oxytocin's effect on trust could not have been established if it were not for the use of a control condition in which subjects play the Risk Game I described in section 4.4; second, in oxytocin, we find one of neuroeconomics' most widely recognized beacons. That is to say, possibly most of the attention neuroeconomics has attracted ever since its inception, has been concerned with investigations involving oxytocin. To a large extent this can likely be attributed to the persistent efforts of the molecule's "ambassador," Paul Zak, director of Claremont Graduate University's Center for Neuroeconomic Studies and coauthor of the Kosfeld et al. article. Zak has variably dubbed oxytocin the "moral molecule" the "love molecule" and the "trust molecule," and his message has proven to have not been lost on the public, judging by his innumerable media appearances and the many reviews, cover stories and blog posts dedicated to his research on oxytocin.[4]

That the coverage in the mass media does not increase our understanding of the investigations at issue should be clear. But it does give an indication of what promise such work holds: a molecular-biology of everything we hold dear in public as well as private life, to phrase it in keeping with the conventional rhetoric.

5.3 *This* is where you trust

The biological reality of social trust caught in an image, part I: *or*, Trust in the subcortical brain

So far we know that neuroeconomic trust is behaviorally defined as a specific move in the Trust Game, that it can be promoted through the intranasal administration of the neuropeptide oxytocin and that it can be distinguished from nonsocial risk taking. These facts, passive linkages that were either arrived at or built on in Kosfeld et al.'s article, are mobilized as active linkages, or technical things, in a recent study by Thomas Baumgartner and colleagues (2008) in order to find out more about the mechanisms implicated in oxytocin's efficacy. The question of *how* oxytocin exerts its effect vis-à-vis trust was not answered by Kosfeld et al.; however, by setting up an experiment in which oxytocin administration and fMRI are combined, Baumgartner et al. attempted to unravel this process, thereby giving us a firmer grip on the neurobiology of trust.

The main conclusion from Baumgartner et al.'s study is that oxytocin exerts its influence on trust through its impact on several subcortical areas, specifically, the amygdala, the midbrain and the striatum. In order to interpret this finding, we must first know more about the experiment that gave rise to it.

The setup: Trust Game + breaches of trust and feedback + oxytocin + fMRI + controls

Baumgartner et al. follow up on all neuroimaging work concerning trust published so far, though most explicitly on the research by Kosfeld et al.. This is not surprising, given that three of Baumgartner's coauthors were also coauthors with Kosfeld. However, whereas Kosfeld et al. merely stated that "trust pervades human societies" (2005, p.673), according to Baumgartner et al.'s (2008, p.639) "[t]rust *and betrayal of trust* are ubiquitous in human societies" (my italics). This addition—that next to trust betrayal of trust is also inescapable in human societies—does not follow from new empirical data that Kosfeld et al. did not have at their disposal, but is rather justified on a priori conceptual grounds; the authors state that "whenever we trust there is also the possibility of trust betrayal" (Baumgartner et al. 2008, p.646). Because of this conceptual truth concerning the possibility of breaches of trust, Baumgartner et al. added not only fMRI examinations of experimental subjects to the experimental paradigm used by Kosfeld et al., but also breaches of trust as well as explicit feedback concerning those breaches.

Just as in Kosfeld et al.'s study, Baumgartner et al.'s used the Risk Game as a control, and had the Trust Game played as a one-shot game, in which investors played with a different trustee in every round. As a reminder, the latter is significant because it ensures that selfish trustees have no incentive to honor the trust that might be put in them (Baumgartner

et al. 2008, p.640). However, after they played twelve game periods, investors received feedback. Moreover, the two games (Trust and Risk) were pseudo-randomized such that, by the time they received feedback, subjects had played six Trust Game periods and six Risk Game periods. The feedback information investors received, told them that their trust was repaid in about 50 percent of all cases.

Altogether, three variables can be identified in the experimental paradigm of Baumgartner et al.: group (oxytocin or placebo), phase (prefeedback or postfeedback) and game (Trust Game or Risk Game). Thus, the impact of oxytocin, relative to placebo, on behavior and on brain activity is investigated in both prefeedback periods and postfeedback periods and in both the Risk Game and the Trust Game.

The observations

The effect of oxytocin on behavior in either of the games was as follows. Both in the Trust Game and in the Risk Game, oxytocin administration had no significant effect relative to placebo in the prefeedback period. Moreover, in the Risk Game, subjects in the oxytocin group and in the placebo group responded similarly to feedback (i.e., they made few adjustments, if any, to their investment pattern during the postfeedback period). Investors who received data concerning the meager trustworthiness of the trustees they played with in the Trust Game, however, were expected to adjust their behavior. And indeed, in the postfeedback period, investors in the placebo group showed decreased trust relative to the prefeedback period in the Trust Game. Subjects in the oxytocin group, on the other hand, did not show a significant behavioral adaptation after feedback in the Trust Game. In other words, the oxytocin and placebo groups had different trust adaptation responses to feedback regarding breaches of trust—the oxytocin group did not significantly alter their trust behavior, whereas the placebo group became less trusting.

Baumgartner et al.'s data on trust in the prefeedback period do not replicate Kosfeld et al.'s finding that oxytocin increases trust. However, Baumgartner et al. dismiss this incongruous finding and apparent anomaly by pointing out the stronger incentive to show trust behavior in this experimental setup relative to the setup used by Kosfeld et al., which is due to the addition of feedback. In the words of Baumgartner et al.:

> subjects faced additional incentives to transfer money to the trustees in the prefeedback periods because they knew they would receive feedback. The only way to learn about the degree of trustworthiness in the population of trustees is to transfer money to them. If OT [i.e., oxytocin] indeed reduces betrayal aversion, the incentive to explore the trustees' trustworthiness must obviously be weaker in the OT group than in the placebo group because subjects with OT are less afraid of betrayal. Thus, the placebo group has a

> stronger reason for extracting information from the feedback, implying that they also transfer more money relative to their natural inclination to trust. If placebo subjects experience this conflict between their natural inclination to trust and the incentive to explore their partners' trustworthiness, the brain should then represent this conflict. In this context it is therefore interesting to observe that the placebo subjects in the Trust Game exhibit higher activation in the dorsal anterior cingulate cortex (ACC), a brain region frequently implicated in conflict monitoring and cognitive control in social [. . .] and nonsocial paradigms. (Baumgartner et al. 2008, pp.644-645)

In other words, the data from the prefeedback period in Baumgartner et al.'s study may not have been expected in light of Kosfeld et al.'s results, but they can be explained relatively easily and do not constitute a veritable anomaly.

The lack of behavioral differences that subjects in both the placebo and oxytocin groups show between prefeedback period and postfeedback period in the Risk Game has its clear counterpart on the level of brain activations: "There are no differences in brain activation between the OT and the placebo group in the Risk Game, where we observe no behavioral differences" (Baumgartner et al. 2008, p.645). Things are different where the Trust Game is concerned. Here differences *are* found between the oxytocin and placebo group, both in terms of behavior and, in those subjects who exhibited behavioral differences, also in terms of brain activity. More precisely, during postfeedback periods, activity in the amygdala (bilaterally) and connected brainstem effector sites, as well as the midbrain and striatum of subjects in the placebo group is significantly higher than that of those in the oxytocin group. Thus, subjects who have received intranasal oxytocin (1) show no significant behavioral adaptation to feedback on meager "returns" on their trust and (2) show decreased activity in several subcortical brain structures (see also figure 5.3).

Trust as decreased fear: the amygdala The thing to do in response to such findings is to check the available literature concerning the brain areas at issue. Naturally, Baumgartner et al. did so, and it turns out that much light was shed on the findings concerning the activation of the amygdala (bilaterally) and the brainstem effector sites connected to the amygdala by a number of studies showing that these are brain areas critical to fear signaling and modulating responses to fear. These studies included human lesion research and animal research, as well as various neuroimaging studies, and all pointed in the same direction (Adolphs et al. 2005, Amaral 2003, Domes et al. 2007, Kirsch et al. 2005, Huber et al. 2005, Winston et al. 2002, Adolphs et al. 1998).[5]

The idea, therefore, is that, during the Trust Game, oxytocin reduces fear responses by decreasing activation in the amygdala and the brainstem effector sites connected to it. Trust,

then, is construed as the realization of the capacity to counteract the risk of betrayal. This capacity is enhanced by diminished amygdala activation, which in turn can be achieved by oxytocin administration.

This finding is consistent with several studies from behavioral economics concerning "betrayal aversion." These studies were done to further the modeling of social preferences, and they indicated that many people strongly dislike it when their trust is not honored. In the language of economists, this can be expressed by saying that subjects derive a disutility or negative utility from what they experience as the betrayal of trust they have put in others, given the economic loss they associate with it (cf. Bohnet & Zeckhauser 2004). With trust stylized as the decrease in disutility inferred from the risk of betrayal, we see that neuroeconomics helps better define utility and disutility, vital concepts in economics, but concepts which are often treated tautologically in conventional economic circles, where their material instantiation is abstracted away from and where utility is defined simply as that which subjects aim at maximizing.

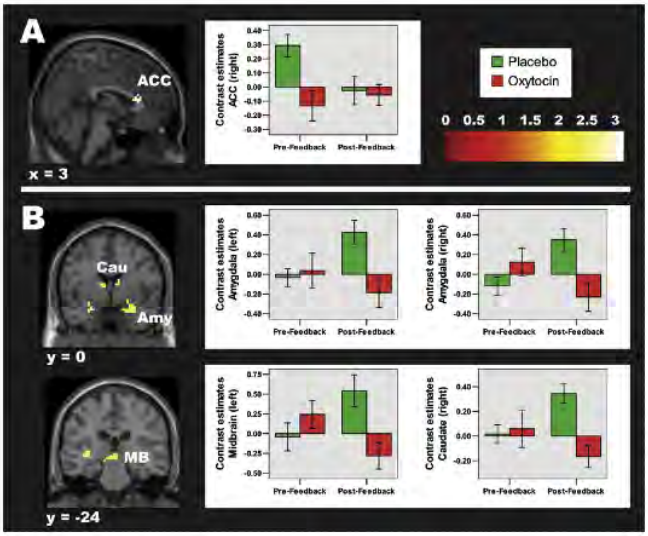

Figure 4. Brain Regions Showing Stronger Activation in the Placebo Compared to the Oxytocin Group
Depicted on sagittal or coronal slices is the increased activation in the placebo compared to the oxytocin group in trust periods played (A) prefeedback (including ACC) and (B) postfeedback (including bilateral amygdala, bilateral caudate nucleus and midbrain brain regions). All regions are significant at $p < 0.005$ with a cluster extent of ten voxels. However, for display purposes all regions are depicted at $p < 0.01$. Bar plots represent differences in contrast estimates (trust > risk) of functional ROI's (see Experimental Procedures section for details) for bilateral amygdala, right caudatus, midbrain regions and ACC, broken down for the oxytocin (in red color) and placebo group (in green color) as well as time phase (prefeedback/postfeedback). Univariate and repeated-measures ANOVAs calculated with and without control for potentially confounding variables (including general trust, sensation seeking, first feedback) confirmed for each depicted brain regions the interaction effect of *group* × *phase*, qualified by stronger activation in these brain regions in the placebo compared to the oxytocin group either *only* in prefeedback or *only* in postfeedback periods.

Figure 5.3: Oxytocin-induced trust: decreased activation in the amygdala and caudate nucleus

This figure, adapted from Baumgartner (2008, p.643), shows fMRI images displaying that, associated with trust and with intranasal oxytocin administration, is reduced amygdala activation and reduced caudate nucleus activation in the postfeedback period. Also visible in this figure is that in the postfeedback period the anterior cingulate cortex (ACC) shows more activation in the placebo group than in the oxytocin group. This is a result Baumgartner et al. do not extensively discuss in the main text of the article but is only dealt with briefly in a supplement.

Trust as compromised learning: the striatum The normal adaptation to feedback on breaches of trust, shown by subjects in the placebo group, is affected by oxytocin. The effect of oxytocin involves a decrease in amygdala activation, so Baumgartner et al. found and explained, but it is also associated with a decrease in striatal activity. More precisely, Baumgartner et al. state that oxytocin reduces activation in the part of the striatum called the caudate nucleus during Trust Games played in the postfeedback period (2008, p.645).[6]

Whereas the amygdala is associated with fear processing, the striatum is a structure well-known to be implicated in reward processing and learning (O'Doherty et al. 2004, Tricomi et al. 2004). Given that subjects can actually gain from "doing well" in the Trust Game, the game can help us understand reward processing. As is conventional in behavioral economics and its deployment in neuroeconomics, by participating in Baumgartner et al.'s experiment, subjects not only earned a lump sum payment of 80 Swiss francs (about €65), they could also exchange the monetary units they obtained during the games for real money, with an exchange rate of 5 units for 1 Swiss francs (≈ €0,81). Thus, subjects earned around 140 Swiss Francs, or €113, on average. Since the game was played in two periods, with feedback on the "success rate" of one's strategy in the first period, testing for learning was explicitly built into the experimental design of Baumgartner et al.'s (2008).

Baumgartner et al. appropriated the outcome of research by Mauricio Delgado and colleagues (2005) as a guide to their data on caudate nucleus activation. The main finding of Delgado et al.'s (2005) was that the way in which subjects perceive the moral character of their partners in the Trust Game modulates those neural systems implicated in reward. More precisely, Delgado et al. found that caudate nucleus activation diminished during the initial stages of the Trust Game if subjects played the game against others whom they perceived as morally good. The same effect, though smaller, was also found in subjects playing against people whom they perceived as being morally bad. It was not found when subjects played against people they perceived as being morally neutral.

Thus, Delgado et al. investigated whether caudate nucleus activation, which had been found to be implicated in adjusting behavior on the basis of reward feedback, is affected by social and moral information about one's exchange partners in the Trust Game. They did so by giving subjects vivid descriptions of their exchange partners, suggesting that they were morally bad, neutral or good. While playing an iterated version of the Trust Game, subjects received equivalent returns from all three types of characters, but were nonetheless more prone to put trust in the alleged "good" partner throughout the game. While caudate nucleus activation did differentiate between negative and positive feedback, it did so only with regard to the *neutral* exchange partner. In other words: "prior social and moral perceptions can diminish reliance on feedback mechanisms in the neural circuitry of trial-and-error reward learning" (Delgado et al. 2005, p.1611).

Experiment as theory test

The results Delgado et al. found constitute a problem for two relevant theories in this domain: rational choice theory in economics, and Bayesianism in cognitive neuroscience (2005, p.1615). The fact that subjects continued to put trust in a partner who was described to them as morally good, irrespective of that partner's actual behavior vis-à-vis those subjects, conflicted with predictions derived from the perspective of rational choice. In rational choice theory subjects are presumed to be alert to nonreciprocating behavior at all times and are therefore always expected to adjust their behavior in response to opportunistic behavior on the part of their exchange partners.

This same behavior is also predicted by Bayesianism, an increasingly influential theory of global brain functioning. According to this view, developed by the eminent neuroscientist Karl Friston, the brain is what is sometimes called a "Bayesian machine," meaning that the brain is assumed to form probabilistic models of the world around it and to continually update these models in light of whatever new stimuli it is confronted with, all in accordance with Bayesian statistics. "Bayesian brain", the brain as a "Bayesian machine" or the brain as a "Helmholtz machine" (see, e.g., Dayan et al. 1995) are three different ways of naming what is a key ingredient in an attempt at providing a unifying theory of brain function under the heading of "the free energy principle." In short,

> the Bayesian brain is a corollary of the free energy principle, which says that any self organizing system (like a brain or neuroimaging community) must maximize the evidence for its own existence, which means it must minimize its free energy using a model of its world (Friston 2012, p.1230)

Such models are always in the process of being updated in relation to the stimuli the system is provided, irrespective of whether the system concerned is an animal, a person, a brain or a scientific community.[7]

It is consistent with Bayesianism that subjects would create prior beliefs from the character profiles they were given, which, in the initial stages of the game, would affect their choices. However, feedback concerning the exchange partners' actual behavior that conflicts with those prior beliefs should in turn alter those beliefs. In other words, the new evidence provided should be reflected in subjects' behavior in later stages of the experiment, as the probabilistic models concerning the actions of subjects' exchange partners should adapt to the new evidence.[8]

When subjects play against a morally neutral partner, trial-and-error feedback processing and associated caudate nucleus activation was indeed observed. But this, as we saw, was not the case with regard to partners described as morally good—especially in the latter case, where it turns out that this profile "not only creates a prior belief, but also disrupts the

regular encoding of evidence, or learning, from surprising outcomes" (Delgado et al. 2005, p.1615). This is also reflected in caudate nucleus activation, as in those cases where learning seems absent, caudate nucleus activation is diminished.

This, then, is the major finding on which Baumgartner et al. model their interpretation of their own data concerning the caudate nucleus. According to Baumgartner et al. subjects playing against "good" partners in the experiment by Delgado et al. play as if they simply *know* their partners can be trusted, and for that reason show less behavioral adaptation to the feedback information they receive of those allegedly good partners—if you *know* something, so knowledge appears to be conceived of here, this means you stop learning. Furthermore, the fact that subjects playing "good" partners are not affected by feedback in their behavior, is associated with decreased caudate nucleus activation. Baumgartner et al. see something analogous happening in their experiment: "subjects with OT behave as if they implicitly 'know' that they can trust their partners, and this may be the reason why the brain structure that is critical for learning the contingency between actions and outcomes—the caudate nucleus—shows diminished activation" (2008, p.646). Intranasally administered oxytocin, in other words, not only diminishes fear of betrayal by acting on the amygdala. By acting on the caudate nucleus it also compromises reward learning.

The research by Baumgartner et al. and Delgado et al. are not the only sources suggesting that reward learning as mediated by the caudate nucleus plays a role in trust in situations in which subjects have prior knowledge of their exchange partners. For example, Brooks King-Casas and colleagues (2005) add to the modality of this statement.[9] In their experiment King-Casas et al. used a Trust Game in which two anonymous players played together for ten rounds and in which both partners in the exchange can be said to express trust, because both run the risk that their move is not positively reciprocated. Moreover, players can change their moves in the game so as to reflect their experience of their counterpart. In their view, this experimental design improved on the ecological validity of the task relative to most other experiments concerning trust (King-Casas et al. 2005, p.18). It can also be argued, however, that they have altered the conditions of play too much. For the move by the second player can be understood both as a response to the first player's move, and as anticipating the first player's future move, creating ambiguity in regard to trusting and reciprocating behavior. As they evaluate it themselves, the main finding by King-Casas et al. is that reputations of investing partners can be identified in caudate nucleus activation patterns (i.e., trustees form models of investors "in their brains"). Or as Baumgartner et al. put it, King-Casas et al. show "that caudate activation is reduced as learning progresses and rewards can be more reliably predicted. Thus, once subjects in these studies have learned the contingency between their actions and the associated outcomes, the caudate is less active" (2008, pp.645-646).[10]

In their study King-Casas et al. used hyperscan fMRI—that is, simultaneous fMRI scanning of two experimental subjects who were playing the Trust Game together—to investigate trust and reputation building. They found that previous reciprocation of a gaming partner's move is the best predictor of whether or not subjects will place trust in their partners. Moreover, they report that this behavioral finding is "mirrored by neural responses in the dorsal striatum" (King-Casas et al. 2005, p.78). The tit-for-tat strategy, in which a lack of trust by one player, exhibited through untrustworthy behavior, is repaid by the other player ceasing to trust, appeared to be almost the standard, even if it was deviated from in some cases. Especially when subjects deviate strongly from this "neutral reciprocity" strategy, changes in trust can be very well predicted.

The behavioral results, however, form only one part of the story. As King-Casas et al. state:

> Social decision making critically depends on internally represented models of social partners. In principle, such covert knowledge might be inferred from behavior observations. However, behavior signals are intrinsically lower dimensional than their underlying neural responses, and so behavior alone is an insufficient signal source for inferring neural representations. Put another way, an inference based only on the observable behavior of a social partner ignores many observable neural processes that give rise to that behavior. The measurement of both interacting brains directly sidesteps this problem and allows us to probe the cross correlation of internal models—replacing inference with a measurement. (King-Casas et al. 2005, p.78)

In other words: King-Casas et al. believe that from the activity of the brains of interacting subjects we can learn things about the internally or mentally represented models these subjects make of their counterparts, things which behavior alone does not tell us all too much about. In this formulation we find a meaningful turn of phrase: looking only at behavior 'ignores many *observable neural processes*' (my italics), and as subsequently becomes clear, when it comes to those neural processes, which are normally invisible but can be observed with fMRI, observing them equals *measuring* them.

Now, what is it that we can learn from the neuroimages about the models that subjects build of their counterparts, and how do we learn this? Most noteworthy is the fact that there is a time-shift in caudate nucleus activity. The caudate nucleus is thought to compute information concerning how fair or unfair the decision of someone's partner is and concerning "the intention to repay that decision with trust." Moreover:

> In early rounds of the game, the 'intention to trust' is evident only after an investment is revealed. With experience, this signal shifts to a time preceding the revelation of the investment. (King-Casas et al. 2005, p.82)

It is as if even without a bell ringing, a well-conditioned dog would start drooling in antic-

ipation of getting food. And indeed, in order to understand this, it is important to first get acquainted with the concept of "reward prediction error," developed in the context of research on reinforcement learning—neobehaviorist research on learning, that is. King-Casas et al. lucidly explain this as follows:

> This finding is reminiscent of analogous shifts of reward prediction error signals from reinforcement learning that have recently been identified by fMRI in human caudate and putamen and are thought to involve outputs of midbrain dopaminergic systems. These prediction error signals were identified using simple conditioning experiments in which lights predict the future delivery of rewards (e.g., squirt of juice or delivery of monetary return). The scheme is simple: An initially neutral light is flashed; it causes no change in dopaminergic activity, but the later (surprising) arrival of juice causes a burst of activity in the dopamine neurons. Repeated pairing of light followed at a consistent time later by juice causes two dramatic changes: (i) The response to juice delivery drops back to baseline and (ii) a burst response occurs just after the light is flashed. This temporal transfer of the burst response to the light is thought to represent the future value predicted by the light. (King-Casas et al. 2005, p.82)

Apparently, there is an analogy between this case of simple conditioning and learning about and responding to people's behavior in economic exchange games such as the repeated Trust Game. The neuroimaging results suggest that subjects build models of their counterparts; or put differently, that subjects develop reputations in the brains of their counterparts. And this happens in the caudate nucleus.[11]

The biological reality of social trust caught in an image, part II: *or*, Trust in the cortical brain

In the next chapter I will elaborate how similarities between humans and other species are mobilized to get neuroeconomic research going, to develop hypotheses, to explain experimental setups and to interpret data. In the area in which we are now headed, however, what is emphasized is precisely what distinguishes man from other species. For example, Frank Krueger and colleagues state that "[u]nlike other species, humans are trustful and cooperate with genetically unrelated strangers, with individuals they will never meet again, or even when reputation and gains are absent" (Krueger et al. 2007, p.20084). And although

> [r]eciprocal exchange is ubiquitous to the behavior of many species [...] [i]ncreased specialization by humans in productive activities, together with the advantages this has produced, likely has been built on improved adaptations for social exchange. [...] Such an adaptation would support more

> sophisticated reciprocity strategies (McCabe et al. 2001, p.11832)

Evolutionary thinking is also the engine behind these considerations, but now it is emphasized that humans show more complex social behaviors than other animals. McCabe et al. and Krueger et al. state this in the context of further investigations into the neurobiology of trust. But now the problem area is framed as one that includes not only behaviors that are distinctive of humans, there is also an associated move to "regions of interest" (as brain areas focused on during imaging experiments are conventionally designated) that are cortical rather than subcortical. More precisely, what these articles articulate is the role various parts of the prefrontal cortex (PFC) play in trust. The PFC is a part of the brain that in humans is much larger relative to body size than it is in other mammals.[12] Moreover, with regard to the density of neural pathways connecting the PFC to other cortical areas, for example those involved in processing sensory data, the human PFC is much more developed than that of our fellow mammals. This stands in contrast to what we have so far encountered in the neuroeconomics of trust, which has focused much more on what connects humans to other animals, both in terms of behavior and with regard to the neuroscientific findings concerning such behaviors. As for the latter, we have so far mostly found phylogenetically old structures or neurochemicals—both the neuropeptide oxytocin and the subcortical structures (amygdala, brain stem and caudate nucleus) fit this description. Research that focuses more on the distinctive mark of humans, so it will presently become clear, tells us that in addition to fear and learning, also mentalizing, empathy and delaying reward gratification are implicated in trust.

Theory of mind, empathy and overcoming immediate reward gratification In their 2001 study, McCabe et al. tested the hypothesis that the prefrontal cortex (PFC) is involved in the integration of so-called theory-of-mind processing and cooperative behavior. To this end they had subjects play the Trust Game while their brains were scanned with fMRI. The results can in fact be seen as the first explanation ever on the neural correlates of trust. However, in their research McCabe et al. did not attempt to decouple the neural correlates of trust and of trustworthiness. This makes sense, given that their goal was to investigate the difference between cooperative and noncooperative behavior and that both trust and repayment of trust are cooperative. Just as in Kosfeld et al.'s study, McCabe et al. used the Risk Game as a control condition. (As a reminder, this game is identical to the Trust Game except that it is played against a computer instead of against a human.) In this way McCabe et al. meant to assess the hypothesis that there would be differential activation patterns for subjects playing a computer and subjects playing humans, as only in the latter case would there be an occasion to learn about the neural activity associated with the integration of reasoning about others' intentions (i.e., "theory-of-mind processing) and economic decisions.

The idea entertained by McCabe et al. is that, if one investigates trust using the economic Trust Game, coupled with the Risk Game as control condition, this will allow for the situation in which it follows passively that mentalizing is implicated in trust. Put differently, the experimental set up is designed to probe the difference between strategic interaction and nonsocial risk taking. Thus, the attention of the scientists is immediately directed to investors' predictions concerning how their exchange partners will behave in the near future. Insofar as there is a difference between how investors act in the Trust Game and how they act in the Risk Game, this indicates what, from the point of view of the investors, distinguishes human counterparts from computer counterparts. Whereas for example Kosfeld et al. take this difference to be a difference between social trust, on the one hand, and nonsocial risk taking, on the other, McCabe et al. give a more complex description of what Kosfeld et al. label "social trust": According to McCabe et al. this implies that investors make inferences concerning the intentions of their counterparts.[13]

Not only did McCabe et al. find that indeed several subjects show a difference between how they play the Risk Game and how they play the Trust Game, they also found that in those individuals who consistently opt for cooperation in the Trust Game, there is more activation in the prefrontal region of the brain than when they play against a computer, i.e. play the Risk Game. More precisely, associated with cooperation is activation of the paracingulate cortex, especially the anterior part. In independent studies this brain region has been found to be critical to theory-of-mind processing or "mentalizing" (see e.g. Gallagher & Frith 2003, Frith & Frith 2006) and later research has confirmed the findings by McCabe et al. concerning the involvement of the PFC in trust (Krueger et al. 2007, Krueger et al. 2008).[14] What is social as opposed to nonsocial is unpacked as essentially involving mentalizing. On both counts, the social nature of an action is presumed identifiable in the action and/or neural activation patterns and/or impact of oxytocin on any one isolated half of the playing dyad—thus, what is social is assumed to exist and to be visible inside the skull of individual subjects.

Rather than simply replicating the work of McCabe et al., the investigations by Krueger et al. again tackled some slightly different questions, even though these too involved fMRI-scanning of subjects playing the Trust Game. This is typical of the field[15]: Experimental protocols might involve playing Risk Games as controls and Trust Games as primary data sources, but even though the games played are always recognizable as such in different protocols, they nonetheless tend to differ on multiple dimensions. This makes it hard to say something incontrovertible about the relationship between trust and risk (cf. Houser et al. 2010a) and to compare what each of these studies teach us about trust and its neural correlates. The difficulty is due to the fact that the epistemic things at issue, trust and its neurobiology, are by necessity always partly unknown. The epistemic things are

co-constituted by the various protocols we come across in this experimental system, and insofar as these protocols meander, so do the identities of the epistemic things under consideration. As long as trust and its neurobiology are co-constituted by a certain degree of vagueness, it is to be expected that the approaches to them may well differ (or even conflict) in some respects.

In Krueger et al.'s study (2007), the neuroeconomic investigation of trust involved a non-anynymous version of the repeated Trust Game, with investors and trustees alternating their roles. Accordingly, using this protocol enabled Krueger et al. to study the neural correlates of trust simultaneously with partnership building and maintenance. The latter, again, has behavioral, psychological and neural materializations. It comes with specific ways of interacting, specific mental models each player develop of their gaming partners, and with specific neural activation patterns underlying these models.

First, the paracingulate cortex (PcC) and the septal area (SA), when contrasted with the control of risk taking in the nonsocial Risk Game, are identified as underlying decisions to trust. The PcC is said to be a key player in the building up of trust relationships—this brain region is centrally involved in inferring the intentions of one's gaming partner and in this way it allows for the prediction of the partner's behavior. As Krueger et al. emphasize in connection to PcC involvement in trust, theory-of-mind reasoning "is a unique human characteristic and can be observed only in a most rudimentary form in great apes and has never been observed in monkeys" (2007, p.20087). The SA, on the other hand, is a part of the (phylogenetically old) limbic system, and is hypothesized to be recruited for its oxytocin receptors: "Because synthetic oxytocin increases trust, we surmised that partners recruited the SA to encode goodwill to maintain their trust partnership" (Krueger et al. 2007, p.20087).

All three levels of analysis key to neuroeconomics can be easily distinguished here, but I wish to emphasize the special significance of the psychological level for interpreting the data on SA activity. This is clear when we read that

> '[r]esults from pre- and postquestionnaire rating support our view demonstrating that partners felt significantly closer to each other and ranked themselves as being more of a partner to the other person after the experiment' (Krueger et al. 2007, p.20087)

In other words, subjects' experiences of partnership are actively engaged to account for activity in the SA, which only succeeds because of passive linkages connecting the activity found in the SA with the findings from Kosfeld et al.'s earlier study concerning oxytocin's capacity to increase trust (2005), and the fact that there are oxytocin receptors in the SA.[16]

Moreover, Krueger et al. distinguished between two different trust strategies, viz. "conditional trust" and "unconditional trust." In the strategy they labeled conditional trust,

investors assume their interacting partners to be self-interested. The expected value of this strategy is estimated by investors, with regard to the advantage associated with cooperating, the risk that one's partner defects, and the value one's past decisions might have in the future (p.20084). It is a cognitively more demanding strategy, which leads to greater variation in decisions than the alternative, unconditional strategy does. In the unconditional strategy investors assume their partners to be trustworthy. Depending on which of these strategies investors follow, the PcC's involvement in trust figures earlier or later in the development of the game and is associated with the SA differently, so Krueger et al. argue.

This is explained as follows. First the experiment is divided up with respect to two aspects; that is, two stages are distinguished and the group of experimental subjects is divided in two equal-sized groups based on whether or not either partner "ever defected on their partner's decision to trust" (Krueger et al. 2007, p.20087). The latter allows for the identification of a defector and a nondefector group, the former for the identification of a partnership-building ("building") and a partnership-maintaining ("maintenance") stage. The hypothesis, then, is that defectors and nondefectors implement different trust strategies in these different stages of the experiment. Behind this hypothesis is the idea that investors' trust only involves PcC activation when the trustworthiness of partners is evaluated using theory-of-mind reasoning, inasmuch as it is this that the PcC is involved in.

Krueger et al. found that subjects in the nondefector group had higher PcC activation in the building stage and higher SA activation in the maintenance stage than those in the defector group. Furthermore, those in the defector group showed higher activation in the ventral tegmental area (VTA) during the maintenance stage. In order to explain this, Krueger et al. reason that in the nondefector group mentalizing figured centrally in the building stage and that this led to better models of their exchange partners and, accompanying this, "sufficient mutual goodwill to become socially attached to each other" (Krueger et al. 2007, p.20087). This in turn explains why those in the nondefector group had higher SA activation in the maintenance stage, and lower PcC activations in that stage:

> Through early mentalizing, partners in the nondefector group must have balanced goodwill more quickly, allowing them to become synchronized in their decision patterns. Brain-to-brain correlations only increased in the SA region for the nondefector group across stages, and only partners in the nondefector group became synchronized in their SA BOLD amplitudes as first movers in adjacent trials of Trust Games. Synchronization in the SA led to social attachment associated with a significant decrease in activation in the PcC during the maintenance stage. (Krueger et al. 2007, p.20088)

Furthermore, whereas the unconditional strategy might be cognitively costly to start with, decreased decision times in the maintenance stage prove it is much less costly then.

The conditional trust strategy involves PcC activation or theory-of-mind reasoning only in the maintenance stage. Moreover, in the maintenance stage subjects in the defector group showed higher VTA activations, which are associated with encoding expected and realized reward. More than that, this region was most active in those pairs, across all groups, who shared the least trust reciprocity in their choices. For partners in the defector group, taking on a strategy that was more cognitively costly meant that they exhibited "significant increase in activation in the PcC over the experiment" (Krueger et al. 2007, p.20088). And not only is the conditional trust strategy more costly cognitively, it also pays less than does the unconditional strategy.

This is not all that one can infer from these experiments, however. In 2008, Krueger et al. reported new findings concerning the neuroeconomics of trust, which were based on novel analyses of their 2007 results (see Krueger et al. 2008, p.3867). These further analyses showed that, common to both trust and reciprocity, is the activation of the rostral medial prefrontal cortex (arMFC), known to be implicated in theory-of-mind reasoning, and the temporoparietal junction (TPJ), which is known to subserve empathy. According to Krueger et al., both cognitive and affective sharing with another person figure in trust and in trustworthiness.

Additionally, in this new report Krueger et al. emphasized that trust recruits a so-called evaluation system for likely outcomes, namely the bilateral frontopolar cortex (FPC). This is a relatively recently evolved system that facilitates subjects to weigh long-term rewards over immediate returns. And this, of course, is requisite to many of the complex forms of cooperation discernible in human societies (cf. Krueger et al. 2008, p.3870).

In other words, the phylogenetically young, and hence characteristically human, FPC enables us humans to overcome our immediate desires so that we can choose to have higher but later payoffs—on condition that one's exchange partner reciprocates, of course. In addition to the role of mentalizing in trust, delayed gratification is also something McCabe et al. capitalized on in their study (2001).[17] Since trustees might not cooperate, the cooperative or trusting strategy by investors involves a risk, the overcoming of which McCabe et al. then described as an "inhibition of immediate reward gratification to allow cooperative decisions" (2001, p.11834). Due to the experimenter's role in the Trust Game, such cooperative decisions are, in the long run, more rewarding than the immediately satisfied guaranteed reward that noncooperative behavior entails.

5.4 Conclusion

As we traced the neuroeconomics of trust we saw how science works its way through and to epistemic things by a process describable as "differential reproduction." Every new

investigation I described involved a slightly different experimental protocol. It is important that any new investigation, to some recognizable extent, replicates or reproduces previous results—be they results from ethological or neuroendocrinological animal studies, human neuroendocrinology, behavioral economics, social neuroscience, or neuroeconomics. But it is no less important that these repetitious moves are sufficiently distinct from each other for any of them to be of interest at all. It would be hard, if not impossible, to have exact replications of previous research published, despite the "official" rhetoric in methodology departments about the epistemological significance not only of possible replicability of research but also of factual replication. Likewise, it would be difficult find a publisher for research that did not (obviously) relate to or form part of "normal scientific" practices, while at the same time presenting something novel or original with respect to that practice. All in all, the evolutionary metaphor of differential reproduction seems appropriate in regard to the description of the neuroeconomics of trust.

Understanding the development of science on the model of evolution is not new of course. Such otherwise almost antipodal philosophers as Karl Popper, Thomas Kuhn and Hans-Jörg Rheinberger, to name but three, have all appropriated evolutionary thinking in their portrayal of the pathways of science—and accordingly have portrayed these pathways as less or more complex, winding, straightforward and rationally structured. The idea that "differential reproduction" describes one aspect of the development of science is clearly consistent with various ways of studying science philosophically and, relatedly, with various philosophical views on the nature of science and its products. These views can diverge widely on such issues as (1) the role or importance of timeless rational standards in evaluating scientific results vis-à-vis the historicity of claims to truth and modes of representation deemed suitable media for such claims, (2) the relative importance of experiments, logic, theory or technology for the process of science or (3) the nature or structure of the collective at work and its relation(s) to other collectives in the development of science.

This is not the right place for an in-depth analysis of the different ways in which positions concerning these issues cohere with different views on the aptness of using the notion of evolution in one's conception of science. What matters, though, is that the philosophical position defended and used here entails that scientific practices are at the center of attention. The delineation of epistemic objects and the production of scientific facts about these objects is what I am interested in, as are such issues as how facts travel from the confines of esoteric centers to the world outside of them and what the consequences of this happening are. Moreover, the approach I have taken implies a rather confined and specialized type of interest in the production of facts. What I do not aim at, then, is "competing" with the scientists in this chapter. The point is, as Andreas Roepstorff aptly put it in the context of discussing the neuroscientific study of consciousness, that

> the explicit criteria of truth are mainly probed, tested, and developed within a particular group, within a particular style, [and that therefore] they are in practice relatively inaccessible to outsiders. Outsiders have no way of validating whether, for instance, in a given article, the right hemodynamic response function has been used to fit the data or whether the right pulse sequence and gradient manipulations were used in the MRI scanner. To a large extent, an outsider to a scientific field has to trust, first, that things are done properly and, second, that there are internal mechanisms for checking and validating whether this was indeed the case. (Roepstorff 2004, pp.1112)

Having said that, and thus indeed trusting neuroeconomists and their peers to have done their jobs properly, let me review what facts the neuroeconomics of trust has delivered so far.

By linking behavioral economics with neuroscientific research technologies, neuroeconomists have succeeded in revealing the fact that oxytocin increases trust and that several brain structures are involved in either trust or in the efficacy of oxytocin with regard to trust. *Oxytocin increases trust in humans*, so this story goes.

As concerns the localization of trust in the brain, however, the results have not had the same (relative) simplicity. I have shown that the brain structures implicated in trust range from cortical structures to subcortical ones—from phylogenetically young to phylogenetically old structures. In addition, these structures are never simply or immediately referred to as *the* locus at which "trust happens." Rather, using their localizationist technologies, their behavioral economic operationalization of trust and their psychological interpretations, neuroeconomists have reduced this complex epistemic thing called trust to a number of more simple constituents, and it is these that, reportedly, each have their own *locale*. The investigation of trust has thus resulted in a picture in which trust most centrally involves (1) a decrease in social fear which is identifiable in distinctive activation patterns in the amygdala; (2) a compromised learning capacity which relates to activation patterns in the striatum (caudate nucleus); (3) mentalizing, as managed by the paracingulate cortex; (4) empathy, instantiated in the temporoparietal junction and (5) an increased capacity to overcome immediate desires facilitated by the frontopolar cortex.

Anyone acquainted with today's neuroscience will directly perceive, by this preliminary enumeration of structures involved in trust that the examination of trust in the brain is a complex endeavor entailing a myriad of connections that may take us in many possible different directions. There is, however, a simple explanation for this unfolding complexity, which brings us back to a point made earlier. Neuroimaging experiments concerning trust address widely divergent questions with trust fulfilling several different roles. In the motley array of investigations featuring trust, sometimes trust is not so much an epistemic

thing but rather a technical thing: it plays more of an instrumental role for learning about something else (for instance, reputation building), than that it is an end in itself to better understand trust. Moreover, trust sometimes is an epistemic thing insofar as its neural correlates are concerned, while being a technical thing insofar as it is understood in terms of particular behaviors (i.e., specific moves in the Trust Game). We also see that in these investigations a conventional reductionist research strategy is followed. Trust is not conceived of as an indivisible "primitive," but rather as something that can be broken down into smaller component parts. And it is these (psychological) component parts that are, in turn, correlated with specific neuronal activation patterns and with the effect of oxytocin.

In the next chapter I will analyze in more detail the style at work in neuroeconomic investigations of trust.