Chapter 9. Trust and Emotions

Federica Farolfi 1, Li-Ang Chang 1 and Jan B. Engelmann 1,2,3

1 Center for Research in Experimental Economics (CREED), Amsterdam School of Economics, University of Amsterdam, The Netherlands

2 Amsterdam Brain and Cognition (ABC), University of Amsterdam, The Netherlands

3 The Tinbergen Institute, University of Amsterdam, The Netherlands

Abstract

In the current chapter, we review research investigating the influences of affective states on trust. To delineate the behavioral and neural effects of emotions on trust decisions we consider research from multiple fields, including Behavioral Economics, Psychology and Neuroeconomics. We focus on behavioral and neural research that examined the impact of moods and emotions experienced at the moment of choice. In particular, we critically examine evidence concerning the role played by both positive and negative incidental and integral emotions in the decision to trust. Overall, a pattern emerges from previous findings that strongly suggests that both incidental and integral emotions can influence decisions to trust. Specifically, positive incidental emotions, such as happiness, can enhance trust while negative incidental emotions, such as anxiety, typically reduce trust. At the same time, evidence from neuroimaging studies suggests that this behavioral effect is paralleled by emotions having specific effects on decision-relevant neural circuitry. Brain regions affected by emotions during trust decisions include the temporoparietal junction and medial PFC, both of which have been implicated in theory of mind, as well as the anterior insula, which is commonly implicated in anticipatory negative affect. We conclude by pointing at important open avenues of research regarding the role of emotions in trust in the context of learning from past experiences with interaction partners, and chronic distortions of affect and social behavior, due to underlying psychopathological conditions.

Keywords: Trust, Affect, Emotion, Incidental Emotions, Integral Emotions, Betrayal Aversion, Anxiety, Happiness, Temporoparietal Junction, Anterior Insula
9.1 Introduction

Decisions whether to trust someone are complex and require the support of a multitude of cognitive, affective and motivational processes. A common conceptualization of the neurocognitive processes involved in social decisions, including trust decisions, posits the following steps that the decision-maker is required to undergo at minimum (see for instance Engelmann & Hare, 2018; Rangel, Camerer, & Montague, 2008): 1. forming a perceptual representation of the choice space, including what the choice options are and who is affected by the decision outcomes; 2. evaluating the different outcomes in terms of their costs and benefits to self and others; 3. planning and executing an action that reflects the choice; 4. consuming the outcome of the choice, which involves comparing the observed with the expected value of the outcome (which gives rise to the prediction error) to update the subjective value of the chosen option for future decisions. Multiple cognitive processes support the above decision steps, including attention (e.g., Lim, O’Doherty, & Rangel, 2011; Rangel & Hare, 2010 Hare, Malmaud, & Rangel, 2011), memory (e.g., Bechara & Martin, 2004; Hinson, Jameson, & Whitney, 2003), theory of mind (Cutler & Campbell-Meiklejohn, 2019; Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004; Schurz, Radua, Aichhorn, Richlan, & Perner, 2014), and learning (e.g., Niv, Edlund, Dayan, & O’Doherty, 2012; Schönberg, Daw, Joel, & O’Doherty, 2007; Schultz, 2002). A plethora of research in cognitive and affective neuroscience on emotion-cognition interactions has shown that many of these cognitive processes are readily influenced by affective processes (e.g., Pessoa, 2008). Attention, for instance, is commonly captured by emotional stimuli (Mulckhuyse, Crombez, & Van der Stigchel, 2013; Mulckhuyse, Engelmann, Schutter, & Roelofs, 2017; Yiend, 2010), leading to preferential processing of emotional stimuli over others. Given the tight integration between emotional and choice-relevant cognitive processes (Pessoa, 2008), as well as significant overlap in the neural circuitry of cognitive and affective processes (Engelmann & Hare, 2018), emotions can be expected to influence cognitive mechanisms that support trust decisions at all stages outlined above.

In the current chapter, we review research from multiple fields, including Behavioral Economics and Neuroeconomics, that investigates the influences of affect on trust. We focus on the relationship between emotional states and trusting behavior by looking at behavioral and neural evidence which examined moods and emotions

---

1 From a theoretical standpoint, there is a generally accepted distinction between emotions and moods. Emotions are characterized as affect which arises from a clear specific event, they occur at high intensity and they are short-lived. Moods are described as states without a clear point of formation, which present low intensity but a long duration such as hours or days (Capra, 2004; Forgas, 1995; Lench, Flores, & Bench, 2011; Schwarz, 1990). While it would be meaningful to investigate the different impact of emotions and moods on trusting behavior, in this chapter the term emotion refers to both emotions and moods (Pham, 2007).
experienced at the moment of the choice (Loewenstein, Weber, Hsee, & Welch, 2001). This type of emotions, known as immediate emotions (Dunning, Fetchenhauer, & Schlösser, 2017; Lerner, Li, Valdesolo, & Kassam, 2015; Loewenstein & Lerner, 2003; Rick & Loewenstein, 2008; Schlösser, Fetchenhauer, & Dunning, 2016), falls into two categories: incidental emotions, as those whose source is unrelated to the target decision; and integral emotions, which arise from directly contemplating features of the decision, such as thinking about the consequences of a choice (Lerner & Keltner, 2000; Pham, 2007).

The majority of the experiments we review here employ the trust game (TG), a game-theoretic approach developed in experimental economics to model the decision to trust (for a detailed review of the TG see Alós-Ferrer & Farolfi, 2019; Engelmann, 2010) (see also Chapter 3). In the TG, two agents are facing a social dilemma in a strategic incentivized scenario. In its continuous version, as proposed by Berg, Dickhaut, and McCabe (1995), two agents, a trustor and a trustee, engage in a sequential one-shot interaction. The trustor acts as a first mover and decides how much of her endowment to allocate to the second agent, the trustee, where zero is an option (effectively ending the game). The amount sent is then tripled by the experimenter and given to the trustee, who decides how much of this magnified amount to return to the trustor, where again, zero is an option. The portion of endowment sent to the trustee is considered to be a measure of trust, while the amount returned to the trustor represents a measure of trustworthiness or reciprocity. The rules of the game (i.e., monetary payoffs) are common knowledge to both agents, and the theoretical prediction for a rational self-interested trustor would be not to make any transfer to the trustee by anticipating that a selfish opponent would return nothing (sub-game perfect equilibrium). As might be expected, actual human behavior significantly differs from the economically rational choice to act selfishly as indicated by the positive levels of both trust and reciprocity commonly observed in laboratory settings (Berg et al., 1995; Camerer, 2003; Johnson & Mislin, 2011).

Initially seen as an emotionless cognitive task, the literature on the TG did not immediately include affective states as factors that could potentially affect or explain the observed “irrational” behavior in laboratory. Subsequently, social preferences such as altruism and other-regarding preferences (e.g. inequity aversion) and possible cognitive limits entered into the equation (Camerer, 2003; Fehr, Fischbacher, & Kosfeld, 2005). Nowadays, the tendency is to perceive emotions as an integral part of the decision-making process (e.g., as factors producing biases) and a valuable albeit not always behaviorally transparent, source of information (Engelmann & Fehr, 2017; Engelmann & Hare, 2018; Lerner et al., 2015; Loewenstein et al., 2001).
9.2 Effects of Incidental Emotions on Trust Behavior

Because real world decisions are rarely made within an emotional vacuum, incidental emotions are ubiquitous in real life and understanding their influence on decision processes is therefore important. From the perspective of economic rationality, incidental emotions should not affect the decision at hand, since the origin of these emotional states is unrelated to the targets of the decision. Psychological intuition, on the other hand, supports the view that our current affective states may influence our decisions. A large body of empirical evidence agrees with our psychological intuition and has already established that these emotional states can affect how people interpret the possible outcomes of their decisions (Andrade & Ho, 2007; Capra, 2004; Dunn & Schweitzer, 2005; Schwarz, 1990). Incidental emotions are therefore increasingly accepted as a core part of judgment and decision making in different domains such as risk (Cohn, Engemann, Fehr, & Maréchal, 2015; Engemann & Fehr, 2017; Engemann & Hare, 2018; Engemann, Meyer, Fehr, & Ruff, 2015; Lerner et al., 2015; Loewenstein et al., 2001; Mellers, Schwartz, Ho, & Ritov, 1997; Phelps, Lempert, & Sokol-Hessner, 2014) and interpersonal trust (Dunn & Schweitzer, 2005; Engemann, Meyer, Ruff, & Fehr, 2019).

To observe how emotions and moods affect behavior, experimentalists rely on a gamut of emotion-induction techniques in the laboratory (also known as Mood Induction Procedures, MIPs) that temporarily place the participants in a particular mood state (e.g. anger, sadness, happiness) immediately before performing an unrelated task. Examples are the Autobiographical Emotional Memory Task (AEMT) which asks participants to reactivate past memories; the Velten method (Velten, 1968), where subjects read emotionally loaded sentences to endogenously generate the suggested mood; hypnosis (Maccallum, McConkey, Bryant, & Barnier, 2000); or by using emotion-inducing material such as movie clips (Kirchsteiger, Rigotti, & Rustichini, 2006) and music (Kenealy, 1988). The effectiveness of such emotion manipulations is typically assessed by measuring neurophysiological responses that reflect autonomic nervous system arousal and include heart rate (Roelofs, 2017), Galvanic Skin Conductance Levels (Eimontaitė, Nicolle, Schindler, & Goel, 2013), and salivary cortisol levels (Hellhammer, Wüst, & Kudielka, 2009), self-report measures such as the Positive and Negative Affect Schedule (PANAS, Saadaoui, El Harbi, & Ibanez, 2019), correlation with the observed behavior, or by using face reading technology that assess moment-to-moment changes in facial expressions of emotion (Kugler, Ye, Motro, & Noussair, 2020). While all these procedures have been corroborated along the years, but less effectively by using an internet-based setting (Ferrer, Grenen, & Taber, 2015; Göritz & Moser, 2006), they present limitations in terms of timing of the effect, contamination of other nontarget affects (Myers & Tingley, 2016) and ethical admissibility in some settings (Ferrer et al., 2015; Gerrards-Hesse, Spies, & Hesse, 1994). More recently, studies have addressed some shortcomings of MIPs and induced negative affect via the Threat of Shock procedure.
(ToS; Schmitz & Grillon, 2012) in which participants receive unpredictable and mildly-painful electrical shocks in a threatening condition, and no or nonpainful shocks in a safe condition (Cohn et al., 2015; Engelmann, Meyer, et al., 2019). Importantly, converging evidence from different methods provides relatively strong support for the role of particular emotions in trust decisions.

**Positive Incidental Emotions**

Capra (2004) has conducted one of the first studies that tested the impact of positive (and negative) induced-mood on behavior in a one-shot TG. This initial study found no evidence that emotions influence trust behavior. As a mood-inducer, the author used both an AEMT, where subjects had to write an autobiographical essay, and experienced success or failure during the experiment (by receiving hard or easy sets of MENSA questions). In the same study, she found that subjects in a positive mood were more altruistic in the Dictator Game\(^2\) (suggesting that positive affect makes people more aware of the other player’s payoffs). In line with the latter result, Mellers, Haselhuhn, Tetlock, Silva, and Isen (2010) found that a positive mood (induced by giving subjects a candy bag and a movie clip to watch) generally increased the level of cooperation in both the Dictator Game and Ultimatum Game\(^3\). Saadaoui et al. (2019) explores the effects of positive (and negative) emotions on trust and risk attitude. In their study, emotion induction was performed by showing subjects a slideshow of pictures (International Affective Picture System) followed by the assessment of participants’ emotional state using the Self-Assessment Manikin (SAM) scale. Subjects performed a TG, a Dictator Game and ultimately, a risk and ambiguity game. The authors implemented two treatments, a *safe* treatment (TG/DG) and a *risky* treatment (RTG/RDG) with the main difference that the amount sent by the investor was multiplied by 3 with a probability of 50% and multiplied by 1 with the remaining probability in the risky conditions (RTG/RDG). Their results indicate that in the risky treatment, both positive and negative emotions decrease pro-social behavior in the Dictator Game, whereas emotions do not affect trusting behavior in the TG. In the safe

\(^2\) In the Dictator Game (Kahneman, Knetsch, & Thaler, 1986) the investors are facing a similar choice as in the trust game, but the recipients are passive (they have no chance to counteract) thus putting the investors in complete control and not exposed to financial risk (loss of the investment) and social risk (possibility of being betrayed by the opponent). See Alós-Ferrer and Farolfi (2019) for a critical survey of the use of the game in comparison with the trust game.

\(^3\) In the Ultimatum Game (Güth, Schmittberger, & Schwarze, 1982) the first mover (the proposer) makes an offer on how to split her endowment between her and the second mover (the receiver). The latter observes the proposed offer and decides between rejecting the offer (where both players get a null payoff) or accepting the offer (where the suggested split of the amount occurs). Experimental evidence indicates that proposers usually offer between 40% and 50% of their endowment (which is usually accepted), while rejections are observed when proposers offer below their 20% of the endowment (Camerer, 2003). Although inefficient from an economic point of view (i.e., the receivers would be better off by accepting any positive amount) the rejection of the offer signals to the proposer the unfairness of the implied division.
treatment, however, trustors showed higher levels of trust after negative compared to positive mood induction. Finally, individuals in the positive mood condition were more likely to be risk takers compared to the negative mood condition.

A more specific and discrete positive emotion that has been addressed in the literature is happiness. Mislin, Williams, and Shaughnessy (2015) asked subjects to recall and write about a happy and joyful memory and watch commercial film clips before engaging in a TG. Results indicate that happy people are more likely to trust, particularly in the condition that offered a relatively lower overall gain from the initial trust (Figure 9.1A). These results are in line with a study of emotions on interpersonal trust by Dunn and Schweitzer (2005). Myers and Tingley (2016) used a similar procedure to induce happiness (together with anger, anxiety, guilt and self-assurance) by means of the AEMT and subsequently asked subjects to complete the expanded form PANAS questionnaire (PANAS-X) to measure general positive and negative affect. The autobiographical memory task failed to manipulate positive emotions4. An interesting aspect that has been explored within the realm of incidental emotions and cooperative behavior is whether emotion manipulation has an impact on the cognitive processing style. In this strand of research, Hertel, Neuhof, Theuer, and Kerr (2000) used a modified version of a public good game and induced happiness and sadness by asking subjects to watch two film excerpts. The authors found that happy people were more likely to use a heuristics processing strategy, whereas sad people favoured a more careful, gathering-information approach.

---

4 Note that one reason for the relative sparsity of studies investing the causal role of positive affect is that it remains relatively difficult to induce positive emotional states, especially for longer periods of time.
Figure 9.1 Effects of the induction of specific incidental emotions on trust. A. shows that the positive emotion of happiness increases trust decisions under specific conditions in the trust game. B. shows the effect of the specific emotion of anxiety, induced via threat of shock, on trust. In particular, less subjects transfer low amounts (below 15 MU) in the no threat condition (ca. 1/2 in the threat condition vs. ca. 1/3 in the no threat condition), while more subjects transferred relatively large amounts in the no threat condition. Figure 9.1A is reproduced with permission from Mislin, Williams and Shaughnessy, 2015. Figure 9.1B was created using publicly available data from Engelmann et al., 2019.

In sum, the results reviewed above suggest that positive mood states have relatively little influence on trust decisions, with 2 studies showing a null effect of induced positive mood on trust, while positive effects were observed on cooperation and altruism. On the other hand, the specific emotion of happiness has been shown to lead to an increase in trust across multiple studies, likely by increasing participants’ optimism and therefore their expectations of a positive back-transfer (Fredrickson, Tugade, Waugh, & Larkin, 2003). This is in agreement with previous findings demonstrating a consistent association between happiness and enhanced risk seeking across a number of studies (Engelmann & Hare, 2018). The results reviewed above suggest that this effect of positive emotions on decision-making extends also to the domain of social decision-making, including decisions to trust. A likely mechanism for the specific effects of happiness on trust is by enhancing cognitive mechanisms that support approach behaviors, such as reward seeking, as well as inducing optimistic beliefs.

Negative Incidental Emotions
The study done by Capra (2004) found no impact of negative induced-mood on trusting behavior. An emotion-specific analysis conducted by Myers and Tingley (2016) revealed that anxiety (an emotion with low certainty about the negative consequences of the event) has a negative influence on trust, contrary to guilt or anger. Similar behavioral results were obtained in a recent neuroimaging study that induced incidental anxiety via Threat of Shock and found significant decreases in trust as a consequence (Figure 9.1B, Engelmann, Meyer, et al., 2019). Two other important negative emotions that have been tested within a TG environment are incidental regret and disappointment. Martinez and Zeelenberg (2015) induced these emotions via both a scrambled sentence task and a standard recall memory procedure. The authors showed that the feeling of regret decreased trusting behavior whereas disappointment increased it. The authors suggest that the positive effect of disappointment on trust levels might reflect the willingness to avoid future disappointments and the attempt to gain a better final outcome. The decreasing levels of trust observed by participants feeling regret might be seen as a correcting manoeuvre with the intent of avoiding re-experiencing the same feeling in the future.

Nelissen, Dijker, and deVries (2007) induced guilt and fear by using an autobiographical recall procedure and asked participants to play a Social Dilemma
Game (Give-some dilemma) that is similar to the TG in the respect that highly cooperative subjects can be taken advantage of by the interaction partner. Additionally, participants’ Social Value Orientation was assessed to account for possible individual differences in terms of pro-social or pro-self preferences. Their findings reveal that incidental fear reduced cooperation, but only for those subjects with a pro-social value orientation, while incidental guilt increased cooperation, but only for those individuals showing a pro-self orientation. In line with this result, De Hooge, Nelissen, Breugelmans, and Zeelenberg (2011) showed that incidental guilt promotes prosocial behavior toward the “victim” of a multi-player Dictator Game, even if it is done at the expenses of the others. Regarding interpersonal trust, Dunn and Schweitzer (2005) showed that incidental anger decreased levels of trust, but with emotional cognitive appraisal the effects on trust disappeared.

Affective dispositions in the form of character traits can be considered another form of incidental affect, as they are unrelated to the decision outcomes. A recent study assessing such affective character traits and investigating individual differences in trust behavior showed that trait anger significantly reduced trust (Engelmann et al., 2019b). Moreover, people with high levels of dispositional anger also had more pessimistic beliefs about the back-transfers of trustees, and changed their behavior significantly in the context of a punishment opportunity. These results generally agree with a recent meta-analysis of personality traits and behavior in social dilemmas (see Figure 4, Thielmann, Spadaro, & Balliet, 2020), showing that two related dispositional negative emotions (aggression and anger) reduce pro-social decisions in the context of multiple games. Interestingly, positive dispositional affect showed no influence on pro-social decisions, while the personality traits of honesty-humility and its negatively correlated construct of antisociality (including Machiavellianism and Psychopathy) significantly enhanced (honesty-humility) and reduced (antisocial traits) prosocial behaviors, such as trusting and reciprocating (Engelmann, Schmid, De Dreu, Chumbley, & Fehr, 2019; Thielmann et al., 2020).

In sum, induced negative mood seems to have little to no effect on trust decisions. Specific negative emotions, on the other hand, have varying effects. Reduced trust was observed in response to induced anxiety and fear, which is consistent with the action tendencies of these emotions that enhance avoidant and withdrawal-related behavior (Chen & Bargh, 1999; Roelofs et al., 2010). Anger, both as incidental and dispositional affect, has been shown to reduce trust, which at first glance is inconsistent with the approach motivation associated with this emotion. However, in the setting of the TG, it is likely that anger reduces the participants’ willingness to be placed in a situation that render participants vulnerable to the betrayal of another. In agreement with this hypothesis, Engelmann, Schmid, et al., 2019 find that under conditions in which the trustor is able to punish the trustee, and therefore is no longer powerless and can respond to potential betrayal, trust in dispositionally angry participants significantly increased. In addition to these basic
emotions, more complex social emotions, such as guilt, shame, disappointment and regret have also been successfully induced in previous work. Disappointment and guilt both consistently increased cooperation, while regret decreased trust.

While incidental emotions have been shown to have an important effect on trust behavior, further research will be necessary to identify whether and how emotions influence not just trust, but also belief formation and whether their effects on behavior and beliefs occur consciously or not. In tackling these questions, future research will need to circumvent methodological limitations related the MIPs, including subjects’ different susceptibility to the techniques and avoiding potential demand effects (for instance, by possibly avoiding self-report questions that might cause subjects to reflect on their emotional state). Recent approaches relying on the Threat of Shock procedure tackle many of the limitations of mood inductions (e.g., Engelmann et al., 2015; Engelmann, Meyer, et al., 2019), but are currently limited to negative affect only.

9.3 Behavioral Effects of Integral Emotions

Unlike incidental emotions, integral emotions are directly related to the decision outcomes. Within this category, a meaningful distinction concerns those emotions that are experienced before making the actual decision, known as anticipatory emotions (Engelmann & Hare, 2018; Loewenstein & Lerner, 2003); and those emerging as a consequence of the revealed outcome. Examples of the latter are disappointment in case of losses (Loomes & Sugden, 1986), regret about a non-chosen and better option (Loomes & Sugden, 1982), pleasure and gratitude if decisions are rewarded (Rabin, 1993), negative emotional reactions due to unfair offers (Sanfey, 2009; Tabibnia, Satpute, & Lieberman, 2008), or forgiveness after a transgression (Desmet, De Cremer, & van Dijk, 2011). These types of emotions are experienced once the outcome is known (post-decisional phase) and therefore are integral to the decision process. They are particularly important for the computation of prediction errors, as well as updating future expectations and decisions based on the experienced outcomes. Because here we are interested mainly in how emotions influence trust at the time the decision is made, we focus on the more impactful role of anticipatory emotions, as the affective responses triggered by thinking about the future consequences of one’s behavior (pre-decisional phase) (Loewenstein & Lerner, 2003)\(^5\).

---

\(^5\) Dunning, Fetchenhauer, and Schlösser (2012) and Schlösser et al. (2016) advocate the importance of emotions which are not attached to the outcome of the decision, but rather to the specific action a person is contemplating to take. In this less consequentialist vision, the authors argue that a person may feel anxiety just to the thought of engaging in a trust decision, and feel guilty to the thought or distrust the human counterpart (without cogitate the possible consequences).
Although the decision to trust could be rewarding in itself (Engelmann & Fehr, 2017), from a strictly monetary point of view, trust decisions are favorable only if the trustee is in fact trustworthy and reciprocates the initial investment of the trustor. There are two important sources that can influence the decision to trust: 1. The perceived trustworthiness of the interaction partner during the decision, which is commonly measured in experiments as an expectation or belief about how much the interaction partner will return and 2. The participant’s willingness to take a social risk, which, in the context of the trust game, is related to the perceived likelihood of betrayal by the trustee. Both sources can influence decision-relevant cognitive processes, such as the perceived social risk involved in trusting another person, and affective processes, such as the emotion involved in anticipating betrayal (betrayal aversion). In this respect, the trustor forms a belief about her opponent’s level of trustworthiness and subsequently relies on her expectations when forming a decision (Thielmann & Hilbig, 2015).

To form a belief the trustor could integrate information (if available) from cues which are directly related to her opponent’s reciprocity such as history and reputational mechanism (Charness, Du, & Yang, 2011; Delgado, Frank, & Phelps, 2005); or unrelated and normatively uninformative cues such as (fictional) opponent’s handwritten essays (Eimontaite et al., 2013), resemblance (DeBruine, 2002) and pupil mimicry (Prochazkova et al., 2018; Wehebrink, Koelkebeck, Piest, de Dreu, & Kret, 2018) (see also Chapter 7). Research in social neuroscience suggests that in the absence of such cues, trustors likely form reciprocity expectations based on how they would behave if placed in the shoes of the trustee, an approach based on self-projection (e.g., Engelmann, Meyer, et al., 2019; Waytz & Mitchell, 2011). A downside of self-projection is that it can lead to decision biases in overly pro- and anti-social individuals (Engelmann, Schmid, et al., 2019). Opponents’ facial expressions are also very influential for decisions to trust or distrust (Alguacil, Tudela, & Ruz, 2015; Ekman & Friesen, 2003), whether the visual clue is provided by means of photographs (Franzen et al., 2011; Jaeger, Evans, Stel, & van Beest, 2019; Scharlemann, Eckel, Kacelnik, & Wilson, 2001), line drawings (Eckel & Wilson, 2003), video clips of facial dynamics (Krumhuber et al., 2007) or virtual reality (Kugler et al., 2020).

Numerous studies have used facial stimuli to inspect the role of trustworthiness judgments in decisions to trust, because trustworthiness judgments are made spontaneously (Klapper, Dotsch, van Rooij, & Wigboldus, 2016) and quickly (Willis & Todorov, 2006) from other’s faces and have a clear neural representation in the amygdala (Engell, Haxby, & Todorov, 2007; Winston, Strange, O’Doherty, & Dolan, 2002). Campellone and Kring (2013) found that facial expressions of anger reduced trust, while facial expression of happiness, contrary to the authors’ expectations, had no influence on decisions to trust. Similarly, Tortosa, Strizhko, Capizzi, and Ruz (2013) found that trustors were more willing to trust happy partners compared to angry partners in a repeated TG environment. On a more subtle level,
Krumhuber et al. (2007) investigated the dual nature of a smile (fake versus genuine) showing video clips (less than 6s) of opponent players where only the mouth region was animated. Subjects chose more often a counterpart with an authentic smile and engaged more often in trusting behavior in the TG. Interestingly, participants perceived opponents showing a neutral expression (used as a control) as least trustworthy than either a genuine or fake smile. Informing trustors about their opponents’ incidental emotional state could also affect behavior. Kausel and Connolly (2014) found that trustors associated their opponent’s anger with untrustworthiness and acted accordingly.

In the TG, the trustor needs to overcome two different sources of risk: a financial one, which represents the risk of not receiving a return on the investment; and a social one, which is the risk to be betrayed (Aimone, Ball, & King-Casas, 2015; Aimone, Houser, & Weber, 2014; Bohnet, Greig, Herrmann, & Zeckhauser, 2008; Bohnet & Zeckhauser, 2004; Fehr, 2009; Fetchenhauger & Dunning, 2012; Quercia, 2016). The argument has been made that betrayal aversion has important emotional consequences that can affect trust decisions (Engelmann & Fehr, 2017) (see Chapter 6). One way to measure betrayal aversion has been developed by Bohnet and Zeckhauser (2004). In their paradigm, participants complete an equivalently framed binary trust, risk and risky dictator game, but unlike the standard TG, participants do not indicate how much they would invest in each scenario. Instead, participants are asked to indicate their minimum acceptable probabilities (MAP), i.e. the probability of positive and fair repayment of the investment that they require to make the (socially) risky investment (Bohnet & Zeckhauser, 2004; for a global replication, see Bohnet et al., 2008). The risky dictator game provides the most interesting comparison to trust behavior, as the risky dictator game includes a passive recipient who has no opportunity to betray the trustor. Both the trustor and this recipient receive the equivalent payment to the one in the TG, but how much is determined by a random mechanism (instead of the trustee who decides in the standard TG). Social preferences therefore should be (almost) equal across the two games. Results from multiple experiments employing this method (Bohnet et al., 2008; Bohnet & Zeckhauser, 2004) indicate that there is an additional cost when taking social compared to nonsocial risk, as reflected by a higher MAP in the TG, compared to all other games. Additional behavioral paradigms have been developed to measure betrayal aversion (Aimone & Houser, 2012), which have shown that the decision to trust can be affected by the possibility of finding out whether betrayal occurred. Interestingly, Aimone & Houser (2012) report that levels of trust are higher when investors are forced to stay in the dark about the trustee’s betrayal. There is an ongoing debate regarding how different methods of elicitation of betrayal aversion (i.e., the tendency to avoid social risk) could lead to mixed results in literature (e.g., use of Minimum Accepted Probability vs. Choice Lists; see Engelmann & Fehr, 2017 for a discussion; see Alós-Ferrer & Farolfi, 2019 for the analysis of the two sources of risk).
In sum, integral emotions have been shown to significantly affect trust decisions. Specifically, facial emotional expressions communicated by the interaction partner during the trust decision can have significant influences on trust, with facial expressions of negative affect, such as anger, reducing and positive facial expressions, such as happiness, enhancing trust. It is noteworthy that while participants readily use facial expressions of emotions and the facial appearance of counterparts to infer their trustworthiness, only conflicting evidence exists about whether such inferences can in fact be made accurately. While initial reports showed that trustworthiness can be detected at levels slightly above chance from neutral photographs (Bonnefon, Hopfensitz, & De Neys, 2017; De Neys, Hopfensitz, & Bonnefon, 2017; Tognetti, Berticat, Raymond, & Faurie, 2013; Verplaetse, Vanneste, & Braeckman, 2007), more recent work suggests that participants are in fact unable to do so (Efferson & Vogt, 2013; Jaeger et al., 2020). One important reason for this inaccuracy is that, when making trustworthiness inferences, participants seem to be easily swayed by the emotional features of faces that are in fact unpredictable of a counterparts' trait prosociality (Jaeger et al., 2020; Scharlemann et al., 2001; Jaeger et al., 2020; Scharlemann et al., 2001). Moreover, betrayal aversion (not an emotional state in itself, but a dispositional attitude favoring specific emotions) and its affective consequences during the time of choice, can significantly reduce trust decisions. Future research is needed to further understand the specific emotional consequences of betrayal aversion and their influence on trust decisions. One promising route to take may be to investigate individual differences in betrayal aversion and its relationship with dispositional character traits, emotions and theory of mind abilities.

9.4 The Neural Basis of Emotional Influences on Trust Decisions

The neural basis of trust has received significant attention in the field of neuroeconomics. It is generally acknowledged that two processes interact during trust decisions (Figure 9.2), which are 1. a social cognitive component that assesses the trustworthiness and intentions of interaction partners and is processed in a social cognition network consisting of temporoparietal junction, superior temporal sulcus and dorsomedial prefrontal cortex; and 2. an evaluative component that assesses the subjective value of the decision and is processed in ventral striatum and ventromedial prefrontal cortex (vmPFC, e.g., Engelmann & Fehr, 2017; Ruff & Fehr, 2014). Multiple neuroimaging studies employing the TG have shown activation in social cognition areas when participants were actively making trust decisions (e.g. Engelmann, Meyer, et al., 2019; Gromann et al., 2013; Krueger, Grafman, & McCabe, 2008; McCabe, Houser, Ryan, Smith, & Trouard, 2001; Sripada et al., 2009; Stanley et al., 2012). Such activations of social cognition regions during trust decisions likely reflect mentalizing, as the agent needs to take the perspective of the trustee and simulate how she will react to given transfers. This is because of the presence of a strong incentive to avoid being betrayed, which is unique to decisions to trust as opposed to matched risky
decisions. Moreover, a recent meta-analysis of the neuroimaging literature around the TG also implicated the anterior insula in one-shot trust decisions (Bellucci, Feng, Camilleri, Eickhoff, & Krueger, 2018), likely reflecting this region’s involvement in assessing social norm deviations (Krueger, Bellucci, Xu, & Feng, 2020).

Figure 9.2 Brain networks involved in trust decisions. The mentalizing, or “theory of mind” network that includes the temporoparietal junction (TPJ) and dorsomedial prefrontal cortex (dmPFC) is shown in the left column. The regions depicted are thought to be involved in social cognitive functions that include taking the perspective of another person to understand their intentions and beliefs. The right column shows the valuation network that includes ventral striatum (VS) and ventromedial prefrontal cortex (vmPFC). This network has been implicated in subjective value computations and reward processing. The figure shows meta-analyses (association tests) obtained from neurosynth.org using the search terms “mentalizing” and “value”.

Incidental Affect: Anxiety suppresses social cognitive neural circuitry
While converging evidence points to the involvement of social cognitive and valuation processes during trust decisions, relatively little is known about the effects of emotions on the neural networks supporting trust decisions. The effects of anxiety on the neural circuitry involved in trust decisions was recently investigated in an fMRI study by Engelmann, Meyer, et al. (2019). The authors employed a one-shot TG in which participants made decisions in either a trust or matched non-social control game. Importantly, participants made these decisions in the context of incidental anxiety, which was induced using a Threat of Shock (ToS) paradigm. At the behavioral level, the presence of ToS-induced anxiety led to a reduction in trust (Figure 9.1B), an observation that agrees with the behavioral results reviewed above (Myers & Tingley, 2016). Moreover, neuroimaging results showed that the left temporoparietal junction (TPJ) is specifically engaged during trust decisions, and this activation is suppressed by the presence of incidental anxiety. In addition, the researchers identified a “trust network” in which functional connectivity with the left TPJ correlated with transfers in the TG, but not in an equivalent non-social control game. This brain-behavior relationship reflects increased functional connectivity between left TPJ and dorsomedial prefrontal cortex (dmPFC), as well as ventrolateral prefrontal cortex.
(vlPFC, extending into anterior insula) with increasing trust levels. Importantly, the aversive emotion induced by the threat of shock significantly reduced the connectivity between left TPJ and posterior superior temporal sulcus (pSTS), and this neural breakdown was associated with reduced trust at the behavioral level. The neural effects of anxiety therefore reflect a suppression of the activity and connectivity of social cognition regions, particularly the left TPJ, while participants were making trust decisions. Similar suppressions of choice-relevant brain regions have also been observed in the context of risky decision making, specifically in ventral striatum and ventromedial Prefrontal Cortex (vmPFC, Engelmann et al., 2015). Moreover, social cognition regions have also been shown to be suppressed under conditions of stress, which is a natural consequence of anxiety (Nolte et al., 2013). Jointly, these results support the notion that the effects of anxiety play out at the neural level by suppressing specific neural circuitry that importantly supports decision-relevant cognitive processes (Engelmann & Hare, 2018).

A study by Kang, Williams, Clark, Gray, and Bargh (2011) used temperature priming, i.e. exposure to cold and warm packs, to induce feelings of (un)pleasantness immediately before participants played a TG. Behavioral results show that cold priming reduced investments in the TG relative to warm primes. A follow-up fMRI study by the same authors showed increased activity in the anterior insula during trust decisions in the cold (compared to warm) condition. The results suggest that enhanced insular activity after an unpleasant prime may be associated with reduced behavioral trust. The anterior insula is an interesting region that is relevant for trust decisions, as it has been implicated in multiple processes that support trust decisions, including negative affect and withdrawal motivation (Caria, Sitaram, Veit, Begliomini, & Birbaumer, 2010; Craig, 2009; Harlé, Chang, van’t Wout, & Sanfey, 2012; Wager & Barrett, 2017). It commonly activates during trust decisions as indicated by recent meta-analysis (Bellucci, Chernyak, Goodyear, Eickhoff, & Krueger, 2017), potentially reflecting negative affect in anticipation and response to unfair offers (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003; Tabibnia et al., 2008), which are likely triggered by deviations from expectations or social norms (Krueger et al., 2020). An additional possibility of anterior insula activation at the time of choice is that it might be involved in the anticipatory affective component of betrayal aversion (as opposed to the social cognitive component of mentalizing).

**Integral Affect: Betrayal Aversion is computed in Anterior Insula**

As discussed above, betrayal aversion is associated with negative integral emotions at the time of choice (Engelmann & Fehr, 2017). Betrayal aversion has been directly investigated in an imaging context. Aimone et al. (2014) adopted their approach used in behavioral studies of betrayal aversion (Aimone & Houser, 2011; Aimone & Houser, 2012; Bohnet et al., 2008; Houser, Schunk, & Winter, 2010), for use in the context of an fMRI experiment to investigate the neural correlates of betrayal aversion. Specifically, subjects made decisions about whether to trust or not to trust in the context of two
binary TGs: 1. In the standard version, the trustor makes a binary decision between a fair option, in which each agent receives an equal payout, and an option that leads to clear advantages for the trustee, but losses for the trustor; 2. In the computer-mediated condition is similar to the risky dictator game of Bohnet and Zeckhauser (2004). Specifically, a random mechanism (not the trustee) decides how much the trustor and trustee receive. Therefore, while payouts are matched across both conditions, the trustee can actively betray the trust of the trustor in the standard TG, but not in the computer-mediated condition. Behavioral results showed increased trust in the computer-mediated condition, mirroring the results of Bohnet and Zeckhauser (2004) and supporting the presence of betrayal aversion. Importantly, the anterior insula showed increased activation when participants made the decision to trust during the standard TG, but not when they made the same decision during the computer mediated game, nor when they decided to play it safe by not trusting. Since betrayal is only possible when subjects decided to trust in the TG, but none of the other conditions, the increased insula signal during this condition supports the view that this region is involved in betrayal aversion.

In sum, while neuroimaging results on the affective processes of trust decisions are relatively sparse at the moment, there is a clear picture emerging from these studies. Both incidental and anticipatory negative affect are associated with trust-suppressive consequences at the time of choice, and this behavioral effect is paralleled by the suppression of social cognition regions at the neural level. Specifically, incidental anxiety was shown to suppress activity and connectivity within the social cognition network, consisting of TPJ, STS and dmPFC (Figure 9.2). Additionally, experiencing an unpleasant (relative to a pleasant) stimulation immediately before making trust decisions suppresses trust and enhances activity in the anterior insula during trust decisions. Given the involvement of the anterior insula in aversive affective processes related to social decisions (Lamm & Singer, 2010), it is likely that contextual and anticipatory affective signals from anterior insula, and the amygdala (see Engelmann, Meyer, et al., 2019), are communicated to social cognition regions such as TPJ, STS and dmPFC. That the effect of negative affect on these regions is one of suppression makes much sense at the neural level: as the need to focus on immediate biological needs and safety increases under conditions of threat, the need to consider others in one’s decision decreases and thereby the need to recruit social cognitive regions involved in mentalizing. In fact, participants should be more selfish when threatened (by external circumstances or the likelihood of betrayal) and perform safer actions to protect their interests. In the context of the TG this should be reflected by decreased social risk taking and reduced considerations of the benefit of one’s actions for others. This conjecture, however, needs to be confirmed by future studies investigating how affective and cognitive components are integrated via the activity and connectivity patterns of affective and social cognitive neural circuits.
9.5 Conclusions, Future Directions and Open Questions

The evidence we reviewed strongly suggests that both incidental and integral emotions can influence decisions to trust and, at the same time, decision-relevant neural circuitry. Importantly, this largely occurs in a manner that is consistent with the action tendencies associated with specific emotions (Engelmann & Hare, 2018). On the one hand, anxiety (and the related emotion of fear) serves the function of initiating behaviors that protect the organism and in many cases are related to defensive behaviors such as withdrawal from threatening situations (Mulckhuyse et al., 2017). Situations that trigger withdrawal, such as Threat of Shock, therefore lead to reductions in trust. On the other hand, happiness is an approach-related emotion that can promote optimism (e.g., Lyubomirsky, King, & Diener, 2005) and generally leads to increased trust in game-theoretical settings. Interestingly, the effects of negative incidental and integral affect have recently been investigated at the neural level, with anxiety suppressing social cognitive neural circuitry that supports trust decisions, such as TPJ, STS and dmPFC, while unpleasant experiences preceding decisions, as well as the possibility of betrayal, enhanced activity in a region associated with negative affect, the anterior insula.

An open avenue of research on trust and emotions concerns the role played by past negative and positive experiences with other people. How we learn in social contexts, how our experiences are generalized to novel contexts, and what role emotions play in social learning are important aspects of trust behavior that need to be investigated by future studies. Real-life interactions with other people are typically repeated and differ from the one-shot, anonymous interactions commonly used in laboratory experiments (Engelmann, 2010). Despite its importance for the generalizability of the research on trust and emotions, this aspect of trust behavior remains to be assessed in depth. Promising first steps in this direction have been taken in a study conducted by FeldmanHall et al. (2018), in which subjects played an iterative TG against three partners exhibiting a high, neutral or low level of trustworthiness. Once they learned who can and can’t be trusted in part 1 of the study, subjects faced new players that had a certain resemblance with the previous players in part 2. In this second part of the game, participants picked their new interaction partners from a list of new players that had different levels of similarity with the players from part 1 that subjects had interacted with. The goal of this study was to investigate how perceptual similarity between known individuals and unfamiliar strangers shapes social learning. In part 2 of the study, subjects preferred to play with strangers that resembled the highly trustworthy players from the first part and refused to play with strangers that resembled the untrustworthy known players. Additionally, these behavioral tuning profiles were asymmetrically applied, in the sense that individuals were distrusted more when they presented a minimal resemblance with the untrustworthy players from the first game. At the neural level, during the second part of the game, amygdala activation was found to track the levels
of resemblance with the untrustworthy types. Given that the amygdala is commonly associated with negative affect (Lindquist, Wager, Kober, Bliss-Moreau, & Barrett, 2012), and, at the same time, trustworthiness judgments (Winston et al., 2002), such results suggest that affective processes might play a role in learning about the trustworthiness of interaction partners.

While negative experiences with interaction partners and their generalization are reflected in the amygdala, positive experiences have been shown to be tracked by the vmPFC (Bellucci, Molter, & Park, 2019). This recent experiment commenced in two stages: first, while in the scanner, participants played the Take Advice Game (TAG), in which they chose one of two cards that faced upside down. Importantly, participants had no information about the cards, but were paid for selecting the card with a higher numeric value, providing a 50% chance of winning. Advisers were able to improve those odds by providing information about one of the two cards that could either be correct (honest advice), or incorrect (dishonest advice). Importantly, participants could identify the intentions of the advisors when given feedback about their choices. After scanning, participants played a TG with the honest and dishonest advisors. Results indicate that advisors that proved to be trustworthy during the TAG game, were trusted more in the subsequent TG. Multivariate analyses of the fMRI responses during the TAG game identified that honesty was represented in a network consisting of vmPFC, dorsolateral PFC, intraparietal sulcus, and posterior cingulate cortex. Moreover, functional connectivity strength between vmPFC and TPJ was positively and significantly associated with trust in the subsequent TG. Jointly, results from FeldmanHall et al. (2018) and Bellucci et al. (2019) implicate both affective (learning from betrayal) and cognitive (e.g., learning and generalization to novel contexts) processes in learning about the trustworthiness of others. Future research is needed to more clearly identify the interactions between affective and cognitive processes during trust decisions and learning whom to trust.

Finally, recent research in the field of Psychiatry has begun investigating social interactions in chronic mood disorders such as depression, anxiety and Borderline Personality Disorder (BPD) (see Chapters 17, 18). Such investigations afford observing the effects of chronic changes of affective states on social decision-making and related neural circuitry. Indeed, perturbed social decision-making and difficulties maintaining functioning social relationships is frequently reported in mood disorders (Hirschfeld et al., 2000), and can be particularly distressing to patients. Investigating how subjects affected by mood disorders that experience chronic and relatively intense distortions of specific emotions and how these disturbances affect trusting behavior could provide a better understanding of the neural basis of the role of emotions in complex social interactions. BPD patients, for instance, have been shown to report lower trustworthiness in faces of interaction partners than normal controls (Fertuck, Fischer, & Beeney, 2018). Interestingly, rejection sensitivity mediates the relationship between mistrust and increased BPD symptoms, which agrees with core
symptoms of BPD, including mistrustful, negative representations of the self and others (Hallquist & Pilkonis, 2012).

At the neural level, BPD patients have been shown to exhibit abnormal anterior insula activity during repeated TGs in which they were not able to maintain trust (King-Casas et al., 2008). Abnormal responses in the anterior insula were observed in BPD patients: while the AI in healthy control subjects clearly tracked level of cooperation (with greater responses seen when norms were violated), this was not the case in BPD patients. Mood disorders have also been associated with abnormal neural responses during social interactions including the trust and ultimatum game. In an earlier study (Sripada et al., 2009), patients with social anxiety disorder showed decreased activation in medial prefrontal cortex during trust decisions compared to healthy controls. A study by Gradin et al. (2015) showed that, while healthy control participants showed increased responses in ventral striatum as the fairness of proposers’ offers increased, such fairness tracking was absent in depressed patients. Jointly, these results indicate that important choice-relevant and social cognitive brain regions show abnormal activation patterns in patients experiencing chronic distortions of emotions. This lends further support to the above conclusions that core neural circuitry involved in trust decisions, including valuation regions in ventral striatum and vmPFC, as well as social cognition regions such as dorsomedial PFC, and affective regions such as anterior insula, are intimately involved in integrating emotion and cognition during trust decisions.
9.6 References


Efferson, C., & Vogt, S. (2013). Viewing men's faces does not lead to accurate predictions of trustworthiness. *Scientific Reports, 3*, 1047. DOI: [https://doi.org/10.1038/srep01047](https://doi.org/10.1038/srep01047)


Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature reviews neuroscience, 9*(2), 148-158. DOI: [https://doi.org/10.1038/nrn2317](https://doi.org/10.1038/nrn2317)


Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature reviews neuroscience, 9*(7), 545-556. DOI: [https://doi.org/10.1038/nrn2357](https://doi.org/10.1038/nrn2357)

Rangel, A., & Hare, T. (2010). Neural computations associated with goal-directed choice. *Current opinion in neurobiology, 20*(2), 262-270. DOI: [https://doi.org/10.1016/j.conb.2010.03.001](https://doi.org/10.1016/j.conb.2010.03.001)


Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature reviews neuroscience, 15*(8), 549-562. DOI: [https://doi.org/10.1038/nrn3776](https://doi.org/10.1038/nrn3776)

Saadaoui, H., El Harbi, S., & Ibanez, L. (2019). Do people trust more when they are happy or when they are sad? Evidence from an experiment. *Managerial and Decision Economics, 40*(4), 374-383. DOI: [https://doi.org/10.1002/mde.3008](https://doi.org/10.1002/mde.3008)


Behavior, 28(4), 260-271. DOI: https://doi.org/10.1016/j.evolhumbehav.2007.04.006


Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. Psychological Science, 17(7), 592-598. DOI: https://doi.org/10.1111/j.1467-9280.2006.01750.x
