



universität
wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„How we learn to trust others -
the biopsychology of repeated trust interactions“

verfasst von / submitted by

Lisa Anna Rosenberger, BSc MSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doktorin der Naturwissenschaften (Dr.rer.nat.)

Wien, 2020 / Vienna 2020

Studienkennzahl lt. Studienblatt /
degree programme code as it appears on the student
record sheet:

UA 796 605 298

Dissertationsgebiet lt. Studienblatt /
field of study as it appears on the student record sheet:

Psychologie

Betreut von / Supervisor:

Univ.-Prof. Mag. Dr. Claus Lamm

Betreut von / Supervisor:

Dr. Michael Naef

CONTENT THESIS

Acknowledgement.....	7
Chapter 1 - General Introduction.....	9
Chapter 2 - The human basolateral amygdala is indispensable for social experiential learning.....	27
Chapter 3 - Fairness norm violations in anti-social psychopathic offenders in a repeated trust game	53
Chapter 4 - Interpersonal distance adjustments after interactions with a generous and selfish trustee during a repeated trust game	81
Chapter 5 - Slow touch targeting CT-fibres does not increase prosocial behaviour in economic laboratory tasks.....	121
Chapter 6 - General Discussion	161
Chapter 7 - Abstract	169
Chapter 8 - CV Lisa Anna Rosenberger.....	173

ACKNOWLEDGEMENT

Throughout the last six years, there were many people inside and outside of academia that have supported, helped, and guided me through my PhD journey. There are too many to name, therefore I would like to highlight only a few. To all of you who are not mentioned, please know that I am immensely grateful that you have travelled (parts of) this journey with me!

Claus, Michael, Jack

Uta, Dani

Jasminka, Hana, Nace

Sigrid

Marijn

- without you all, this thesis would not have been completed. Thank you for sticking with me!

To Chris.

I can only hope I would have made you proud.

CHAPTER 1 - GENERAL INTRODUCTION

OVERVIEW

Every day we manoeuvre a social world full of uncertainty. During an interaction people can behave in many different ways, which poses a potential problem for our own decision making. We strive to behave in a way that will lead for us to the best possible outcome. However, if other people can potentially react to our action in many different ways, how are we to assess what the best possible outcome is? This uncertainty makes decision making difficult and has negative consequences for our mental wellbeing (FeldmanHall & Shenhav, 2019). Therefore, we aim to reduce uncertainty with different strategies. One of them is to trust people (Lewis & Weigert, 1985). By trusting someone, we assume that the other person will behave in a trustworthy way, which reduces all the possible actions that the other person can take to one concrete (set of) trustworthy action(s). This helps us estimate which of our possible actions will have the best outcome and we thus act as if the other behaves in a trustworthy manner (Lewis & Weigert, 1985). Distrust also reduces uncertainty of our social world, through expectations that the other behaves untrustworthy. However, trust and distrust have opposing consequences for a relationship: while trust increases the development of a positive relationship, with distrust we close ourselves off from the possibility of a flourishing relationship (Lewis & Weigert, 1985).

The focus of my thesis is trust, specifically how we learn to trust others. I will often frame my research in the context of trust, not distrust, even though most of the studies described in the coming chapters measure both. In my research, I assume that trust and distrust are extremes of a continuous phenomenon. Even though a trust mindset might engage different cognitive processes than a distrust mindset (Mayo, 2015), a detailed reflection on the similarities and differences of trust and distrust is beyond the scope of my thesis.

All the studies I conducted for my PhD are based in a neuropsychoeconomic perspective of trust. In that framework, expressions of trust develop from behavioural, psychological, neurofunctional, neurochemical, as well as neurogenetical mechanisms that are forged by a nature-nurture-culture interplay (Krueger & Meyer-Lindenberg, 2019). Within this framework, trust can manifest itself in different ways on the cognitive, affective, behavioural, and situational level (Lewis & Weigert, 1985). The research for my thesis focussed on these different manifestations and addressed trust in understudied populations, such as those with mental health problems, and non-WEIRD (i.e. Western, educated, industrialized, rich, and democratic) backgrounds (Krueger & Meyer-Lindenberg, 2019). This introduction chapter will lay the theoretical basis and provide a guidance to the upcoming

chapters, where I will describe my research in greater detail. I will introduce the different decision neuroscientific, clinical, and social psychological concepts and methodologies used in my research and will place my research questions into a theoretical context.

TRUST, TRUSTWORTHINESS, AND TRUST PROPENSITY

My thesis builds on the neuropsychoeconomic framework of trust (Krueger & Meyer-Lindenberg, 2019), which means that I used both behavioural economic as well as experimental psychological paradigms to study the behavioural expression of trust within a dyad. Within this framework, trust is closely linked to two related concepts: trust propensity and trustworthiness. I will start with a definition of all three concepts from a decision neuroscience perspective, where trust is implemented as monetary transactions during a trust game (described in the next section). Even though social transactions are more ambiguous than economic transactions (Colquitt, Scott, & LePine, 2007), in decision neuroscience it is presumed that the underlying basic mechanisms are comparable in both cases. Thus while I frame many of my explanations in a monetary context, I assume that they can be generalized to a social context.

Trust is defined as a social exchange wherein a person (in a trust game called an investor) shares an exclusive source (anything from information to money) with an interaction partner (in a trust game called a trustee) without knowing whether the partner reciprocates that trust (Fehr & Camerer, 2007). Therefore, trusting a trustee is a risky decision, as the investor is dependent on the goodwill of the trustee (Fehr, 2009). However, trust and risk are not the same concepts. While trust signals a willingness to be vulnerable, risk taking is a behaviour where a person becomes vulnerable (Mayer, Davis, & Schoorman, 1995). Investors are believed to trust the trustee only when they expect a higher profit than without trust, thus investors have a positive expectation about the trustee's reciprocation. In chapters 2 and 4 I examine in detail neurofunctional aspects of trust, as well as what consequences trust has on interpersonal functioning.

Trustworthiness is defined as the amount of reciprocation by the trustee (Fehr, 2009). Thus the more the trustee reciprocates the investor's trust, the more trustworthy they are perceived to be. Reciprocating trust is a costly behaviour for the trustee, as their profits would be higher if they would not reciprocate the investor's trust (Berg, Dickhaut, & McCabe, 1995). In chapters 3 and 5 I examine trustworthy behaviour in a clinical sample (psychopathic offenders), as well as its malleability (through affective touch).

Lastly, trust propensity is defined as the investor's trust disposition (Colquitt et al., 2007). A high trust propensity is associated with an increased willingness to rely on others. While trust develops during repeated interactions with the same partner, trust propensity is independent of the

personal history with specific others. More specifically, an investor's trust propensity and their belief about the trustee's trustworthiness inform how much they will trust the trustee during their first interaction. However, this might not be an evenly balanced relationship, as people with a high trust propensity in general have more positive beliefs about their interaction partners than people with a low trust propensity (Colquitt et al., 2007). Even though trust propensity is an important factor in building a trusting relationship, my thesis only focuses on different aspects of trust and trustworthiness.

THE TRUST GAME: FROM ONE-SHOT TO REPEATED INTERACTIONS

In all upcoming chapters I report on behavioural aspects of trust that I studied with a repeated trust game. This task is a highly validated paradigm used in decision neuroscience (Tzieropoulos, 2013) and the only currently available behavioural measure of dynamic trust. A repeated trust game is an extension of the one-shot trust game (Berg et al., 1995). The major task difference between the two games is that in a one-shot trust game players interact once with each other, whereas in a repeated trust game players interact multiple times. A trust game is always conducted with two players, an investor and a trustee. A single interaction comprises one trust game round, which consists of 3 steps as illustrated in Figure 1. In step 1 both players get endowed with 10 monetary units (= MU) that they can use during that round. In step 2 the investor decides how much of these 10 MU they want to entrust the trustee with. This investment is a measure of trust in the trustee (Berg et al., 1995). The investment gets tripled by the experimenter and is subsequently send to the trustee. In step 3 the trustee decides on the size of their back-transfer based on their currently available balance ($10\text{MU endowment} + 3 * \text{investment}$). This back-transfer is a measure of reciprocity, and thus trustworthiness (Berg et al., 1995). After the trustee made their back-transfer a round is concluded. At the end of a round the investor's earnings are composed of $\text{endowment} - \text{investment} + \text{backtransfer}$. The trustee's earnings are composed of $\text{endowment} + 3 * \text{investment} - \text{backtransfer}$. In a repeated trust game the same two players interact with each other over multiple rounds through the same three steps as in a one-shot trust game. The history of actions of the interaction partner is called a prior (Civai & Hawes, 2016) and is part of their reputation. Earnings in every round are added up and usually paid out at the end of the game or the experiment.

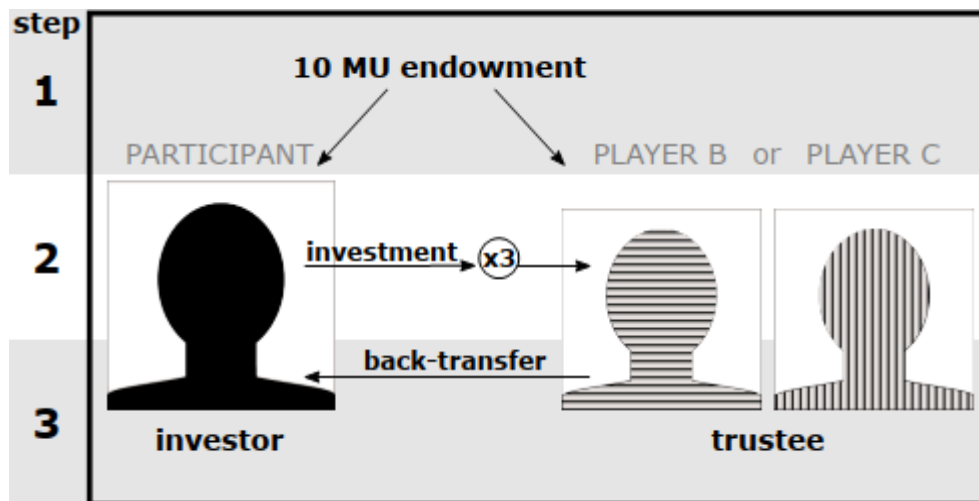


Figure 1. Illustration of one round of a repeated trust game with two potential interaction partners (player B or player C). Before a round starts, one of the two trustees is chosen as this round's interaction partner. A round consists of three steps: an endowment step, an investment step, and a back-transfer step. During the whole task, all players are represented with different avatars and neutral names for anonymization and easy distinction. The figure is created by Marijn Moerbeek and originally published in Rosenberger et al. (2019), see chapter 2.

With an investment, the investor signals two things to the trustee; first, they believe that the trustee will not betray their trust (i.e. make no back-transfer). Second, they intend to enter into a trusting relationship with the trustee (McCabe, Rigdon, & Smith, 2003). The trustee can reward the investor's signal with positive reciprocity, i.e. a back-transfer that is larger than the investment (Falk, Fehr, & Fischbacher, 2003). With this positive reciprocity, the trustee honours the investors trust by incurring a loss (as they give up part of their earnings), whilst at the same time ensuring that both earn money from the interaction. It is assumed that both players use this strategy to maximize their joint payoffs at the end of the game (McCabe et al., 2003). The trustee can also respond to the investor's investment with negative reciprocity, i.e. a back-transfer that is lower than the investment (Falk et al., 2003). Here, the trustee incurs a loss to decrease the investor's maximum possible payoffs. A reason for negative reciprocity might be that the trustee perceives the investment or the intention behind it as unfavourable and uses negative reciprocity to punish the investor's decision (Falk et al., 2003).

The trust game is rooted in game theory, which assumes that players are rational decision makers that aim to maximize their own payoffs (for an introduction see Civai & Hawes, 2016). It is assumed that players can predict the best trajectory of actions that maximizes their payoffs, based on complete information of the rules of the game, available decisions to both players, the history of

decisions taken by both players, as well as the value of these actions to the player. The strategy that maximizes payoffs based on available information is also called the Nash equilibrium. In a one - shot trust game, the Nash equilibrium prescribes that the trustee should not make a back-transfer and thus keep all their MU to themselves. Because the investor is believed to anticipate this decision, they ought to not make an investment in the first place (Civai & Hawes, 2016). However, as the vast research literature on trust games shows (Johnson & Mislin, 2011), both players in general deviate from this best response. Reasons for this deviation are manifold, ranging from theories why players do not compute the Nash equilibrium (e.g. due to cognitive constraints Gigerenzer, 2001) to theories about motivational concerns besides payoff maximization (e.g. altruistic concerns Fehr & Fischbacher, 2003).

LEARNING TO TRUST AN INTERACTION PARTNER

When two people meet for the first time, both experience a high level of uncertainty about how the other person might behave (FeldmanHall & Shenhav, 2019). The two people have very limited information about each other (and sometimes none at all) on which they can base decisions on how to act towards the other. Trusting or distrusting each other helps reduce this uncertainty (Lewis & Weigert, 1985). We decide whether we can trust the other based on our first impression of them. A first impression is a crude classification of the interaction partner into a social category (Fiske & Neuberg, 1990). This categorization helps to predict the other person's behaviour (Ames, Fiske, & Todorov, 2012) and to make a quick, initial decision on whether to trust the other. For example, facial information can help us decide whether a person looks trustworthy within 100 milliseconds (Willis & Todorov, 2006), and is positively related to initial trust decisions in a trust game (van 't Wout & Sanfey, 2008).

When more interactions follow, the two people have the opportunity to learn about each other's trustworthiness. The coarse categorization that was made during the first impression on which the initial (dis-)trust decision was based, gets updated with direct behavioural feedback about the interaction partner's trustworthiness during the repeated interactions (FeldmanHall & Shenhav, 2019). Therefore, the decision to (dis-)trust an interaction partner is not anymore based on the coarse first impression alone, but is supplemented with their actual behaviour. The belief about the interaction partner's trustworthiness changes especially when the actual behaviour of the interaction partner differs from their expected behaviour (Chang, Doll, van 't Wout, Frank, & Sanfey, 2010).

How people update their belief about the interaction partner's trustworthiness and integrate this into their (dis-)trust decisions is very akin to reinforcement learning processes (Sutton & Barto, 1998). Reinforcement learning accounts hold that such learning is driven by a so-called prediction

error, which signals the difference between expected outcomes (such as the belief about the interaction partner's trustworthiness) and obtained outcomes (such as the back-transfer from the interaction partner). This prediction error is then used to update predictions about an interaction partner's future behaviour (e.g. how trustworthy the partner will react during the next interaction) (Chang et al., 2010; Sutton & Barto, 1998). By learning about the interaction partner's trustworthiness, predictions about their behaviour match their actual behaviour more closely. This decreases the uncertainty about the interaction partner's character and behaviour, and consequently the uncertainty about whether to trust or distrust the other (FeldmanHall & Shenhav, 2019).

However, learning to (dis-)trust an interaction partner is not a uniform process and there are many factors (e.g. the informational value, the motivation, a situation) influencing how quickly we learn about others (Mende-Siedlecki, 2018). Since cognitive resources are limited and have to be used economically (FeldmanHall & Shenhav, 2019), our initial impression biases this learning process greatly (e.g. Delgado, Frank, & Phelps, 2005; Fareri, Chang, & Delgado, 2012; Fouragnan et al., 2013). If a trustee appears trustworthy and behaves accordingly, then the investor adjusts their investment behaviour towards the trustee much more readily than when the trustee's back-transfer is inconsistent with their first impression. This further highlights the significance of the first impression in the development of trusting relationships.

BRAIN NETWORKS INVOLVED IN LEARNING TO TRUST AN INTERACTION PARTNER

On a neural level, learning to trust others engages a domain-general reward network, which is based in dopaminergic pathways (Krueger & Meyer-Lindenberg, 2019). These pathways originate in midbrain regions and connect to ventral striatal regions (the so-called mesolimbic pathway), the prefrontal cortex (the so-called mesocortical pathway), and dorsal striatal regions (the so-called nigrostriatal pathway). The mesolimbic and mesocortical pathways are involved in the anticipation of the (un-)trustworthy reaction of the partner and how the partner's reaction is valued (Krueger & Meyer-Lindenberg, 2019). The in the last section described prediction error consistently activates the ventral striatum (part of the mesolimbic pathway) of the investor during their investment, thus during their decision to (dis-)trust their interaction partner (the second step of a repeated trust game round (see Figure 1); Bellucci, Chernyak, Goodyear, Eickhoff, & Krueger, 2016). The nigrostriatal pathway is involved in the implementation of a (dis-)trust decision, as well as the integration of the consequences of this decision (thus the (un-)trustworthy reaction of the partner) with the (dis-)trust decision (Krueger & Meyer-Lindenberg, 2019). The dorsal striatum (part of the nigrostriatal pathway) of the investor is activated when they get to see the trustee's back-transfer, thus their (un-)trustworthy reaction in step 3 of a repeated trust game round (see Figure 1; Bellucci et al., 2016).

Since learning about an interaction partner's trustworthiness is comparable to domain-general reinforcement learning processes, research on reinforcement learning in animals can also inform our understanding of the neural processes underlying trust building. Interestingly, in the rodent learning literature the focus of the neural processes does not only lie on striatal regions, but also on amygdalar regions (for a comprehensive review see e.g. Chesworth & Corbit, 2017). The human amygdala consists of several sub-nuclei. In a learning context the basolateral amygdala and the central-medial amygdala are of special importance, through their connections to both the ventral striatum and medial prefrontal cortex (see Janak and Tye (2015) for an anatomical cross-species comparison). While the rodent basolateral amygdala seems to be involved in updating the value of outcomes and integrating this with both stimulus and response associations (Wassum & Izquierdo, 2015), the rodent central-medial amygdala is heavily involved in adaptive responding (Chesworth & Corbit, 2017; Fadok, Markovic, Tovote, & Luthi, 2018). In the context of trust, this suggests that both subregions might have separate roles in trust learning: the basolateral amygdala in updating the belief about the interaction partner's trustworthiness, and the central-medial amygdala in the decision to (dis-)trust.

However, trust game meta-analyses suggest that the amygdalar subregions do not play a special role in trust learning; Bellucci et al. (2016) did not report any involvement of the amygdala (nor its subregions) in decisions made during a repeated trust game. Cutler and Campbell-Meiklejohn (2019) report whole amygdala involvement only in selfish and strategic decisions during decision-making paradigms (including the trust game). And a recent review by Krueger and Meyer-Lindenberg (2019) link amygdala functioning as a whole only to the emotional valuation of trust-related decisions, but not the learning process in itself. This lack of convergence of human and rodent learning literature points to the methodological difficulty of capturing neural activation in a dynamic and multifaceted process such as trust building. However, conceptualizing the amygdala as a homogeneous structure when studying its functionality might have also contributed to these translational gaps between the rodent and human learning literature. In chapter 2 I address the lacking focus on the amygdalar subregions in trust research, by studying the role of the human basolateral amygdala when building (dis-)trusting relationships with a unique sample of human basolateral amygdala-lesioned subjects. The lesion in these subjects is restricted to the bilateral basolateral amygdala and is due to a genetic mutation caused by the Urbach-Wiethe disease (see box 1 for a detailed description of the disease).

Box 1. Description of Urbach-Wiethe disease

Urbach-Wiethe disease is a genetic disorder, where the extracellular matrix protein 1 gene (ECM1) is mutated. Even though men and women are equally affected, the infrequent and recessive nature of the mutation contributes to its very limited spread, with about 100 known cases worldwide (Van Honk, Terburg, Thornton, Stein, & Morgan, 2016). Clinically, the disease manifests itself in differing degrees of dermatological aberrances (such as acne-like scars), a hoarse voice, and can also include neuropsychiatric abnormalities and epilepsy. Brain calcifications, with increasing severity over the lifespan, are common and mostly include the amygdala. Other affected brain areas are also reported, including the striatum, hippocampus, and the parahippocampal gyrus (Koen et al., 2016). Calcified brain areas lose their functionality.

The individuals involved in the study described in chapter 2 are part of a South-African group with Urbach-Wiethe disease. Van Honk et al. (2016) estimated that almost half of the Urbach-Wiethe population resides in South Africa, with a large part living in the rural Namaqualand area on the border between South Africa and Namibia. Since about a decade van Honk and his research group conduct research with a small group of women affected with Urbach-Wiethe disease who live in the Namaqualand area. These women do not display any neuropsychiatric abnormalities, have an IQ comparable to healthy individuals from this rural area, and their brain calcifications are limited to the bilateral basolateral subregion of the amygdala. A detailed case description of the women is included in Van Honk et al. (2016) and a review of the studies published with them is included in Koen et al. (2016).

NORM COMPLIANCE DURING REPEATED INTERACTIONS AND ITS CLINICAL APPLICATION

While learning about an interaction partner's trustworthiness helps decrease uncertainty in estimating their behaviour, a certain level of uncertainty remains about how to signal specific (e.g. benevolent) intentions. Here, normative behaviour can be helpful (FeldmanHall & Shenhav, 2019). Norms are accepted rules of behaviour that coordinate expectations between interacting people (Young, 2008). They aid decision-making in situations where there is a conflict between following self-interest or joint group gains (Bicchieri, 2005). The decisions of both players in a repeated trust game can be related to norms: the investor can adhere to the norm of trust, and the trustee can adhere to the norm of fairness (FeldmanHall, Son, & Heffner, 2018). Here, positive and negative reciprocity are linked to fairness considerations. A fair decision is a specific form of positive reciprocity, as it equalizes pay-offs between the players (Fehr & Schmidt, 1999). Negative reciprocity, on the other hand, is used to punish unfair behaviour (Fehr & Gächter, 2000). By adhering to the fairness norm, the trustee not only signals that they are adhering to a commonly accepted rule, but

at the same time they reward the investor's trust with a (for the investor) profitable transaction. By doing so, the trustee builds a reputation of being a trustworthy interaction partner, which increases the likelihood that the investor will trust the trustee again in the next interaction (Anderhub, Engelmann, & Guth, 2002). Therefore people are believed to adhere to norms, such as the fairness norm, for positive future outcomes.

Norm compliance is a well-validated mechanism to establish a trusting relationship and is therefore a powerful means to help us understand in which way specific mental health issues are associated with difficulties in developing a trusting relationship. For example, psychopathic individuals (for a description of the personality disorder see box 2) are well known for their norm violating (see e.g. Gacono, 2000) and manipulative nature (Hare, 2003). This combination suggests a poor ability to develop a trusting relationship. However, surprisingly little is known about their trusting nature (but see Ibanez et al., 2016; Richell et al., 2005), and little is published on their ability to maintain a trusting relationship using a decision neuroscience trust game approach. Therefore, in chapter 3 I investigated whether psychopathic offenders violate fairness norms during a repeated trust game and whether this is associated with their ability to establish a trusting relationship.

Box 2. Brief description of psychopathy

Psychopathy is a clinical construct where scholars have diverging views on its definition and neurobiological basis (for reviews see e.g. Blair, 2013; Fowles, 2011; Moul, Killcross, & Dadds, 2012). It is widely accepted as a continuum of personality traits that relate to specific interpersonal, affective, and behavioural characteristics that are classified in the clinical and juridical context with the Psychopathy Checklist-Revised (PCL-R; Hare, 2003). Within the framework of the PCL-R, interpersonal characteristics include a manipulative and lying nature, a superficial charm, as well as a grandiose sense of self-worth. On the affective side, individuals with high psychopathic traits display a lack of remorse, guilt and empathy, a shallow affect, as well as a failure to accept responsibility for their own actions. Behaviourally, psychopathic traits include proneness to boredom, impulsivity, irresponsibility, as well as antisocial behaviour. Within the juridical system, the PCL-R can be used to categorize offenders into psychopathic or non-psychopathic individuals, however the cut-off scores cannot be universally applied across countries (Brazil, van Dongen, Maes, Mars, & Baskin-Sommers, 2016). Therefore, a dimensional approach, clustering the personality traits into two factors (interpersonal/ affective, and lifestyle/ antisocial; Hare, 2003) is often used in research contexts. Since this characterization is based on personality traits, there is also a host of research on non-incarcerated individuals with high psychopathic traits. These individuals are often called the "successful psychopaths" as they are able to successfully function amidst our society (for a

description see Gao & Raine, 2010). Interestingly, economic decision-making is rather different in incarcerated and non-incarcerated individuals with high psychopathic traits (for a more detailed description see the discussion section of chapter 3). This might be attributed to its highly heterogeneous neurobiological basis (Brazil et al., 2016).

TRUST AND PSYCHOLOGICAL DISTANCE

After a period of learning and impression updating, we have formed stable trust. Uncertainty about the partner is reduced (FeldmanHall & Shenhav, 2019), and the (dis-)trust in the other person becomes resistant to change (Kim, Park, & Young, 2020). Our beliefs about the interaction partner's trustworthiness are very strong. Therefore, new information that does not fit the (un-)trustworthy impression does not change this belief easily. Rather, this non-fitting information gets attributed to external causes that are unrelated to the trustworthiness of the interaction partner (Kim et al., 2020). Once this stable (dis-)trust is established, the quality of our behaviour towards the interaction partner changes. This behaviour is an expression of our (dis-)trust in the partner, and is not anymore rooted in decreasing uncertainty about the other's actions. Thus, while our behaviour still signals (dis-)trust in our partner and therefore still reinforces the (dis-)trust dynamics, the root of the behaviour has changed.

Stable trust is amongst other things reflected in the partners closeness, both on a physical and on a psychological level. Psychological distance is defined as the divergence in space, time, hypotheticality, or social distance from the direct experience of the self, here, and now (Liberman & Trope, 2014). Especially social distance, which is operationalized as self vs. other, familiar vs. unfamiliar other, ingroup vs. outgroup, etc. is linked to trust; Close relationships, such as with a romantic partner, exhibit high levels of stable trust (Wieselquist, Rusbult, Foster, & Agnew, 1999). And the influential attachment theory, which describes different relationship styles in light of the affectional bond with an attachment figure, asserts that closeness is necessary to develop and maintain trust in close relationships (Collins & Feeney, 2004). There are many examples from diverse contexts where social distance is associated with trust: e.g. in ingroup and outgroup member settings (Henderson & Lount, 2011; Maddux & Brewer, 2005), in buyer and seller contacts in industry settings (Nielson, 1998), or in same-sex friendships (Salazar, 2015).

Less is known about the association between the spatial dimension (that is the physical distance) and stable trust. In proxemics research, famous for Hall's (1966) four interpersonal distance zones, the physical distance between two interacting people is characteristic of their relationship type. Close spatial distances are usually reserved for intimate and close relationships, while far spatial distances are for acquaintances and the public (Hall, 1966). Given the association between

social distance and stable trust, it seems intuitive that physical distance can be mapped onto trust in a similar vein: we stand closer to a person we trust than a person we distrust. However, in the field of decision neuroscience it has not been established how the distance to the interaction partner changes after forming stable (dis-)trust. In fact, in general little decision neuroscience research has addressed how the quality of the behaviour changes when have established stable (dis-)trusting relationships. As far as I am aware, only Hale, Payne, Taylor, Paoletti, and Hamilton (2018) have studied to what extend individuals follow advice from interaction partners they have learned to (dis-)trust during a repeated trust game. In chapter 4, I continue their line of research and study whether stable (dis-)trust after a repeated trust game affects the physical distance to a trusted and a distrusted interaction partner.

PSYCHOLOGICAL DISTANCE AND TRUST

Interestingly, Bryan, Perona, and Adolphs (2012) reported a reversed association between physical distance and trust: in their study, physical distance affected how trustworthy a face was judged to be. It is thus conceivable that not only trust might be related to physical closeness, but that this closeness might also signal trust. In proxemics research, the physical distance to an interaction partner is seen as a form of nonverbal communication, signalling information about the relationship (Andersen, Gannon, & Kalchik, 2013). In Hall's (1966) closest interpersonal distance zone, the so-called intimate distance, interaction partners stand so close to each other that they can embrace and touch each other. Even though touch is often put in the context of close relationships, it can also have professional or polite functions. Thus the context influences what kind of nonverbal signal a touch transfers (Andersen et al., 2013). However, independent of context, touch might affect trust in general, as a massage can influence how trustworthy people behave to unrelated interaction partners (Morhenn, Park, Piper, & Zak, 2008). The touch's intensity might be of special importance in eliciting trust. When touch is applied as a caress at a specific stroke speed (3 cm/s), it is activating C-tactile (CT) fibres in the touched area of the skin. These CT-fibres are implicated in so-called affective touch (McGlone, Wessberg, & Olausson, 2014). They transfer information about the pleasantness of touch and are therefore believed to play an important role in interpersonal bonding and communication, and influence our social perceptions (Morrison, Loken, & Olausson, 2010). Thus the above mentioned massage might have activated the receiver's CT fibres. This changed physiological state might have in turn elicited trustworthy behaviour towards interaction partners other than the masseur (Morhenn et al., 2008). In chapter 5 I address this reversed association between physical distance and trust. Specifically, I test whether touch, as a conceptualization of a close physical distance, affects trustworthy behaviour during a trust game, and what role CT-fibre activation play in the effect.

OUTLOOK

The next chapters describe the studies I conducted during my PhD in great detail. Chapter 2 focusses on the neurofunctional aspects of trust learning. Specifically, it highlights the role of the human basolateral amygdala in the expression of trust during repeated interactions with trusted and distrusted interaction partners. Chapter 3 concentrates on the clinical aspects of trust learning. The chapter addresses whether psychopathic offenders adhere to norms when expressing (un-)trustworthy behaviour during repeated interactions with a single partner. Chapter 4 examines the behavioural consequences of an established (dis-)trusted relationship. It describes how (dis-)trusting an interaction partner impacts the comfortable interpersonal distance to them during an imagined conversation. Chapter 5 further elaborates on the association between trust and (psychological) distance, but focusses on the reversed direction. Here, the focus lies on CT-fibre activating touch and whether it impacts trustworthy behaviour towards an interaction partner.

REFERENCES

- Ames, D. L., Fiske, S. T., & Todorov, A. T. (2012). Impression Formation: A Focus on Others' Intents. In J. Decety & J. T. Cacioppo (Eds.), *The Oxford Handbook of Social Neuroscience*: Oxford University Press.
- Anderhub, V., Engelmann, D., & Guth, W. (2002). An experimental study of the repeated trust game with incomplete information. *Journal of Economic Behavior & Organization*, 48(2), 197-216. doi:10.1016/S0167-2681(01)00216-5
- Andersen, P., Gannon, J., & Kalchik, J. (2013). Proxemic and haptic interaction: the closeness continuum. In P. Schulz & P. Cogley (Eds.), *Handbooks of Communication Science* (Vol. 2). Berlin: Walter de Gruyter GmbH.
- Bellucci, G., Chernyak, S. V., Goodyear, K., Eickhoff, S. B., & Krueger, F. (2016). Neural signatures of trust in reciprocity: A coordinate-based meta-analysis. *Human Brain Mapping*. doi:10.1002/hbm.23451
- Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *GEB*, 10, 122 - 142.
- Bicchieri, C. (2005). *The Grammar of Society: The Nature and Dynamics of Social Norms*. New York: Cambridge University Press.
- Blair, R. J. (2013). The neurobiology of psychopathic traits in youths. *Nat Rev Neurosci*, 14(11), 786-799. doi:10.1038/nrn3577
- Brazil, I. A., van Dongen, J. D., Maes, J. H., Mars, R. B., & Baskin-Sommers, A. R. (2016). Classification and treatment of antisocial individuals: From behavior to biocognition. *Neuroscience and Biobehavioral Reviews*. doi:10.1016/j.neubiorev.2016.10.010
- Bryan, R., Perona, P., & Adolphs, R. (2012). Perspective distortion from interpersonal distance is an implicit visual cue for social judgments of faces. *PLoS One*, 7(9), e45301. doi:10.1371/journal.pone.0045301
- Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: trustworthiness as a dynamic belief. *Cogn Psychol*, 61(2), 87-105. doi:10.1016/j.cogpsych.2010.03.001
- Chesworth, R., & Corbit, L. (2017). The Contribution of the Amygdala to Reward-Related Learning and Extinction. In B. Ferry (Ed.), *The Amygdala-Where Emotions Shape Perception, Learning and Memories*.
- Civai, C., & Hawes, D. R. (2016). Game Theory in Neuroeconomics. In M. Reuter & C. Montag (Eds.), *Neuroeconomics* (pp. 13-37). Berlin: Springer.
- Collins, N. L., & Feeney, B. C. (2004). An Attachment Theory Perspective on Closeness and Intimacy. In D. J. Mashek & A. Aron (Eds.), *Handbook of closeness and intimacy* (pp. 163-187): Psychology Press.
- Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology*, 92(4), 909-927. doi:10.1037/0021-9010.92.4.909
- Cutler, J., & Campbell-Meiklejohn, D. (2019). A comparative fMRI meta-analysis of altruistic and strategic decisions to give. *NeuroImage*, 184, 227-241. doi:10.1016/j.neuroimage.2018.09.009
- Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, 8(11), 1611-1618. doi:10.1038/nn1575

- Fadok, J. P., Markovic, M., Tovote, P., & Luthi, A. (2018). New perspectives on central amygdala function. *Current Opinion in Neurobiology*, 49, 141-147. doi:10.1016/j.conb.2018.02.009
- Falk, A., Fehr, E., & Fischbacher, U. (2003). On the nature of fair behaviour. *Economic Inquiry*, 41(1), 20-26.
- Fareri, D. S., Chang, L. J., & Delgado, M. R. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Front Neurosci*, 6, 148. doi:10.3389/fnins.2012.00148
- Fehr, E. (2009). On the economics and biology of trust. *Journal of the european economic association*, 7(2-3), 235-266.
- Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Sciences*, 11(10), 419-427. doi:10.1016/j.tics.2007.09.002
- Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960), 785-791. doi:10.1038/nature02043
- Fehr, E., & Gaechter, S. (2000). Fairness and Retaliation: The Economics of Reciprocity. *Journal of Economic Perspectives*, 14(3), 159-181.
- Fehr, E., & Schmidt, K. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, 114(3), 817 - 868.
- FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nat Hum Behav*, 3(5), 426-435. doi:10.1038/s41562-019-0590-x
- FeldmanHall, O., Son, J.-Y., & Heffner, J. (2018). Norms and the Flexibility of Moral Action. *Personality Neuroscience*, 1. doi:10.1017/pen.2018.13
- Fiske, S. T., & Neuberg, S. L. (1990). A Continuum of Impression Formation, from Category-Based to Individuating Processes: Influences of Information and Motivation on Attention and Interpretation. *Advances in experimental social psychology*, 23, 1-74. doi:10.1016/s0065-2601(08)60317-2
- Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational priors magnify striatal responses to violations of trust. *Journal of Neuroscience*, 33(8), 3602-3611. doi:10.1523/JNEUROSCI.3086-12.2013
- Fowles, D. C. (2011). Current Scientific Views of Psychopathy. *Psychol Sci Public Interest*, 12(3), 93-94. doi:10.1177/1529100611429679
- Gacono, C. B. (2000). Assessing Psychopathy: Interpersonal Aspects and Clinical Interviewing *The Clinical and Forensic Assessment of Psychopathy* (pp. 225-252): Routledge.
- Gao, Y., & Raine, A. (2010). Successful and unsuccessful psychopaths: a neurobiological model. *Behavioral Sciences and the Law*, 28(2), 194-210. doi:10.1002/bsl.924
- Gigerenzer, G. (2001). The adaptive toolbox. In G. Gigerenzer & R. Selten (Eds.), *Bounded Rationality: the adaptive toolbox*. Cambridge, Massachusetts: MIT Press.
- Hale, J., Payne, M. E., Taylor, K. M., Paoletti, D., & Hamilton, A. F. (2018). The virtual maze: A behavioural tool for measuring trust. *Q J Exp Psychol (Hove)*, 71(4), 989-1008. doi:10.1080/17470218.2017.1307865
- Hall, E. T. (1966). *The hidden dimension* (Vol. 609): Garden City, NY: Doubleday.
- Hare, R. (2003). *Manual for the Revised Psychopathy Manual for the Revised Psychopathy Checklist (2nd edn)*. Retrieved from Multi-Health Systems:
- Henderson, M. D., & Lount, R. B. (2011). Physical Distance in Intragroup and Intergroup Negotiations: Implications for Negotiator Judgment and Behavior. In E. A. Mannix, M. A. Neale, & J. R. Overbeck (Eds.), *Negotiation and Groups* (Vol. 14, pp. 35-56): Emerald Group Publishing Limited.

- Ibanez, M. I., Sabater-Grande, G., Barreda-Tarrazona, I., Mezquita, L., Lopez-Ovejero, S., Villa, H., . . . Georgantzis, N. (2016). Take the Money and Run: Psychopathic Behavior in the Trust Game. *Front Psychol*, 7, 1866. doi:10.3389/fpsyg.2016.01866
- Janak, P. H., & Tye, K. M. (2015). From circuits to behaviour in the amygdala. *Nature*, 517(7534), 284-292. doi:10.1038/nature14188
- Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, 32(5), 865-889. doi:10.1016/j.joep.2011.05.007
- Kim, M., Park, B., & Young, L. (2020). The Psychology of Motivated versus Rational Impression Updating. *Trends in Cognitive Sciences*, 24(2), 101-111. doi:10.1016/j.tics.2019.12.001
- Koen, N., Fourie, J., Terburg, D., Stoop, R., Morgan, B., Stein, D. J., & Van Honk, J. (2016). Translational neuroscience of basolateral amygdala lesions: studies of Urbach-Wiethe Disease. *Journal of Neuroscience Research*, 94, 504 - 512.
- Krueger, F., & Meyer-Lindenberg, A. (2019). Toward a Model of Interpersonal Trust Drawn from Neuroscience, Psychology, and Economics. *Trends in Neurosciences*, 42(2), 92-101. doi:10.1016/j.tins.2018.10.004
- Lewis, J. D., & Weigert, A. (1985). Trust as a Social Reality. *Social Forces*, 63(4).
- Liberman, N., & Trope, Y. (2014). Traversing psychological distance. *Trends in Cognitive Sciences*, 18(7), 364-369. doi:10.1016/j.tics.2014.03.001
- Maddux, W. W., & Brewer, M. B. (2005). Gender Differences in the Relational and Collective Bases for Trust. *Group Processes & Intergroup Relations*, 8(2), 159-171. doi:10.1177/1368430205051065
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, 20(3), 709 - 734.
- Mayo, R. (2015). Cognition is a matter of trust: Distrust tunes cognitive processes. *European Review of Social Psychology*, 26(1), 283-327. doi:10.1080/10463283.2015.1117249
- McCabe, K. A., Rigdon, M. L., & Smith, V. L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior & Organization*, 52(2), 267-275. doi:10.1016/s0167-2681(03)00003-9
- McGlone, F., Wessberg, J., & Olausson, H. (2014). Discriminative and affective touch: sensing and feeling. *Neuron*, 82(4), 737-755.
- Mende-Siedlecki, P. (2018). Changing our minds: the neural bases of dynamic impression updating. *Curr Opin Psychol*, 24, 72-76. doi:10.1016/j.copsyc.2018.08.007
- Morhenn, V., Park, J., Piper, E., & Zak, P. (2008). Monetary sacrifice among strangers is mediated by endogenous oxytocin release after physical contact. *Evolution and Human Behavior*, 29(6), 375-383. doi:10.1016/j.evolhumbehav.2008.04.004
- Morrison, I., Loken, L. S., & Olausson, H. (2010). The skin as a social organ. *Experimental Brain Research*, 204(3), 305-314. doi:10.1007/s00221-009-2007-y
- Moul, C., Killcross, S., & Dadds, M. R. (2012). A model of differential amygdala activation in psychopathy. *Psychological Review*, 119(4), 789-806. doi:10.1037/a0029342
- Nielson, C. C. (1998). An empirical examination of the role of "closeness" in industrial buyer-seller relationships. *European Journal of Marketing*, 32(5/6), 441-463. doi:10.1108/03090569810215812
- Richell, R. A., Mitchell, D. G. V., Peschardt, K. S., Winston, J. S., Leonard, A., Dolan, R. J., & Blair, R. J. R. (2005). Trust and distrust: the perception of trustworthiness of faces in psychopathic and non-psychopathic offenders. *Personality and Individual Differences*, 38(8), 1735-1744. doi:10.1016/j.paid.2004.11.017

- Rosenberger, L. A., Eisenegger, C., Naef, M., Terburg, D., Fourie, J., Stein, D. J., & Van Honk, J. (2019). The human basolateral amygdala is indispensable for social experiential learning. *Current Biology*.
- Salazar, L. R. (2015). Exploring the relationship between compassion, closeness, trust, and social support in same-sex friendships. *The Journal of Happiness & Well-Being*, 3(1), 15-29.
- Sutton, R., & Barto, A. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.
- Tzieropoulos, H. (2013). The Trust Game in neuroscience: a short review. *Soc Neurosci*, 8(5), 407-416. doi:10.1080/17470919.2013.832375
- van 't Wout, M., & Sanfey, A. G. (2008). Friend or foe: the effect of implicit trustworthiness judgments in social decision-making. *Cognition*, 108(3), 796-803. doi:10.1016/j.cognition.2008.07.002
- Van Honk, J., Terburg, D., Thornton, H., Stein, D. J., & Morgan, B. (2016). Consequences of selective bilateral lesions to the basolateral amygdala in humans. In D. G. Amaral & R. Adolphs (Eds.), *Living without an amygdala*.
- Wassum, K. M., & Izquierdo, A. (2015). The basolateral amygdala in reward learning and addiction. *Neuroscience and Biobehavioral Reviews*, 57, 271-283. doi:10.1016/j.neubiorev.2015.08.017
- Wieselquist, J., Rusbult, C. E., Foster, C. A., & Agnew, C. R. (1999). Commitment, Pro-Relationship Behavior, and Trust in Close Relationships. *Journal of Personality and Social Psychology*, 77(5).
- Willis, J., & Todorov, A. (2006). First impressions: making up your mind after a 100-ms exposure to a face. *Psychol Sci*, 17(7), 592-598. doi:10.1111/j.1467-9280.2006.01750.x
- Young, H. P. (2008). Social Norms. In S. N. Durlauf & L. E. Blume (Eds.), *New Palgrave Dictionary of Economics* (2 ed., pp. 6065-6069). London: Macmillan.

CHAPTER 2 - THE HUMAN BASOLATERAL AMYGDALA IS INDISPENSABLE FOR SOCIAL EXPERIENTIAL LEARNING

Original manuscript as published in Current Biology

DOI: doi.org/10.1016/j.cub.2019.08.078

Lisa A. Rosenberger^{1*}, Christoph Eisenegger¹, Michael Naef², David Terburg^{3,4}, Jorique Fourie⁴, Dan J. Stein⁴, and Jack van Honk^{3,4,5*}

Affiliations:

¹Neuropsychopharmacology & Biopsychology Unit, Department of Basic Psychological Research and Research Methods, Faculty of Psychology, University of Vienna, 1010 Vienna, Austria.

²Experimental Economics Laboratory, Department of Economics, Royal Holloway, University of London, Egham, TW20 0EX, United Kingdom.

³Department of Psychology, Utrecht University, 3584 CS Utrecht, the Netherlands.

⁴Department of Psychiatry and Mental Health, MRC Unit on Risk & Resilience in Mental Disorders, University of Cape Town, Observatory, 7925 Cape Town, South Africa.

⁵Institute of Infectious Disease and Molecular Medicine (IDM), University of Cape Town, Observatory, 7925 Cape Town, South Africa.

*Correspondence to: Lisa A. Rosenberger (lisa.anna.rosenberger@univie.ac.at) or Jack van Honk (j.vanhonk@uu.nl, lead contact).

© 2019. This manuscript version is made available under the CC-BY-NC-ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

SUMMARY

Trust and betrayal are central to our social world, and adaptive responses are crucial to our economic and social well-being (Axelrod & Hamilton, 1981). We learn about others' trustworthiness through trial-and-error during repeated interactions (Chang, Doll, van 't Wout, Frank, & Sanfey, 2010). Rodent research has established a crucial role for the basolateral amygdala (BLA) in social experiential learning (Hernandez-Lallement, van Wingerden, Schable, & Kalenscher, 2017), by respectively reinforcing and suppressing behavior during positive and negative interactions with conspecifics (Hernandez-Lallement, van Wingerden, Schable, & Kalenscher, 2016). The human BLA has undergone a reorganization with massive expansion relative to other amygdala nuclei (Janak & Tye, 2015), and there is no translational research on its role in experiential learning. The human amygdala is traditionally researched as a single structure (Gupta, Kosciak, Bechara, & Tranel, 2011), neglecting the sub-nuclei's structural and functional differences (B. W. Balleine & Killcross, 2006), which might explain inconsistent findings in research on social interactions (Baumgartner, Heinrichs, Vonlanthen, Fischbacher, & Fehr, 2008; FeldmanHall et al., 2018). Here we study whether the human BLA is necessary for social and non-social experiential learning, by testing a group of five humans with selective bilateral damage to the BLA. We compared their learning behavior in a repeated trust game, and a non-social control task, to healthy, matched controls. Crucially, BLA-damaged subjects, unlike control subjects, completely failed to adapt their investments when interacting with a trustworthy and an untrustworthy partner. In the non-social task, BLA-damaged subjects learned from positive outcomes, but differed from the controls by not learning from negative outcomes. Our data extend findings in rodent research by showing that the human BLA is essential for social experiential learning and provide confirmatory evidence of divergent mechanisms for differentially-valenced outcomes in non-social learning.

Keywords

Basolateral amygdala, Trust, Trust Game, social learning, Decision making, Urbach-Wiethe disease, Brain lesion, Neuroimaging, Neuroeconomics.

RESULTS AND DISCUSSION

To directly examine the role of the human basolateral amygdala (BLA) in social and non-social experiential learning, we investigated behavioral adaptation in 5 humans with selective bilateral calcifications of the BLA, and an intact and functional central-medial amygdala (CMA), and 17 healthy matched control subjects during a repeated trust game and a non-social control task (the Raffle task). In the repeated trust game two players make sequential decisions over multiple rounds of monetary transfers on whether to trust each other (Figure 1) (Berg, Dickhaut, & McCabe, 1995). Subjects interacted in turn with a generous trustee, who acted in a trustworthy way, and a selfish trustee, who acted in an untrustworthy way, for 52 rounds. A change in the subject's investments reflects the extent to which she learned about the trustee's trustworthiness (Chang et al., 2010). In a cross-species study with BLA-damaged humans and BLA-silenced rats, we recently demonstrated that rodent amygdala models of defensive behavior directly translate to humans (Terburg et al., 2018). Correspondingly, if the human BLA, like the rodent BLA (Hernandez-Lallement et al., 2016, 2017), is vital to social learning from experience, then BLA-damaged subjects should not be able to learn about the trustee's trustworthiness, and therefore should show no adaptive behavioral updating in the repeated trust game. In contrast, in non-social experiential learning, on the basis of rodent research (LeDoux & Daw, 2018), the involvement of the BLA might diverge for positive and negative outcomes. We used the Raffle task to explore whether human BLA involvement in experiential learning is specific to social context (Fareri, Chang, & Delgado, 2012). In this task subjects bought tickets for two different raffle boxes during 52 rounds. Outcomes and outcome probabilities exactly matched those used in the repeated trust game. Subjects had to learn from which raffle box they could win (the positive box) and from which they could lose money (the negative box), and adjust their ticket buying behavior accordingly.

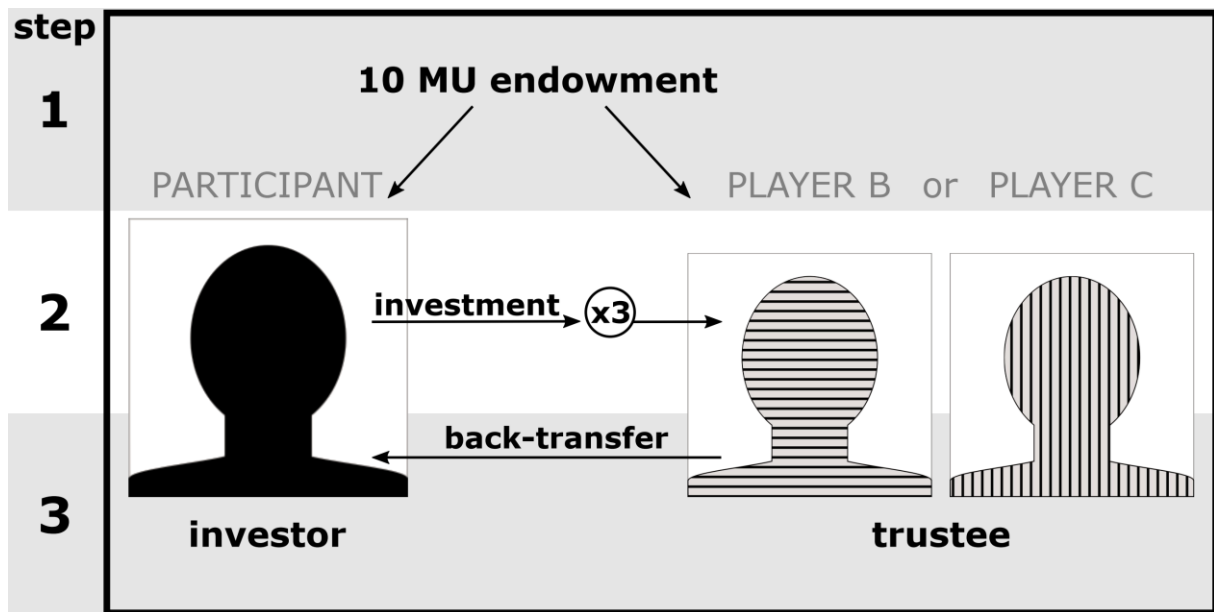


Figure 1. Repeated trust game design with original avatar pictures. Step 1: both players receive an endowment of 10 monetary units (MU) at the beginning of each round. Step 2: the subject makes an investment (between 1 – 10 MU), which gets tripled and send to the trustee (player b or c). Step 3: the trustee then makes a variable back-transfer. One trustee displayed generous behavior (back-transfers equal or higher than investments), the other displayed selfish behavior (back-transfers equal or lower than investments) (Fett, Gromann, Giampietro, Shergill, & Krabbendam, 2012).

In the repeated trust game, healthy control subjects learned that the two trustees were different and adapted their behavior accordingly (Figure 2A). They transferred significantly more money to the generous trustee than to the selfish trustee (Wilcoxon signed rank test $z = 3.29$, $p = 0.001$, $n = 17$), with average transfers of 7.51 MU ($SE = 0.14$) to the generous trustee and 4.14 MU ($SE = 0.14$) to the selfish trustee. The BLA-damaged subjects, on the other hand, did not learn to distinguish between the two trustees (Wilcoxon signed rank test $z = 0.13$, $p = 0.89$, $n = 5$), with average transfers of 5.33 MU ($SE = 0.28$) to the generous trustee and 5.15 MU ($SE = 0.27$) to the selfish trustee. The crucial difference between the transfers to the generous and selfish trustee was also significantly higher (3.37 vs 0.18 MU) in control subjects compared to BLA-damaged subjects (Mann-Whitney U-test, $p = 0.019$, $n = 22$). These comparisons are also significant when using a linear mixed model with random intercept per subject (Model 1 (M1) in Table S1).

Figure 2A (and Figure S1 with added individual data points) shows that the healthy subjects gradually learned that the generous trustee was more trustworthy than the selfish trustee, while the BLA-damaged subjects fluctuated around 5 MU without noticeable change in the two trustees. To

test for differences in learning, we estimated a linear mixed model with random intercept per subject (Model 2 (M2) in Table S1). The matched controls clearly learned about the trustee's trustworthiness, as they adjusted their investment behavior in an adaptive manner (trustee type x round number: $b = -0.201$, $p < 0.001$). They steadily increased investments in the generous trustee ($b = 0.074$, $p < 0.001$), and steadily decreased investments in the selfish trustee ($b = -0.126$, $p < 0.001$). These slopes differed from those of the BLA-damaged group (group x trustee type x round number: $b = 0.16$, $p = 0.001$). BLA-damaged subjects did not increase or decrease their investment behavior in the two trustees ($b_{generous} = 0.044$, $p = 0.155$; $b_{selfish} = 0.002$, $p = 0.940$; trustee type x round number: $b = -0.042$, $p = 0.341$), pointing to absent social experiential learning. This resulted in the control subjects earning more money than BLA-damaged subjects (Mann-Whitney U-test, $p = 0.034$, $n = 22$, Figure 2B).

We measured the subjects' self-reported trust in the trustees, reflecting their perception of the trustees' trustworthiness. During the last two rounds of the task we prompted the subjects after their investments, and before they saw any back-transfer feedback, to indicate on a VAS scale (ranging from 0 to 100) how much they trusted the trustee in this round. Generous trustees ($m_{controls} = 77$ ($SE = 7$), $m_{BLA-damage} = 63$ ($SE = 16$)) were rated as significantly more trustworthy than selfish trustees ($m_{controls} = 24$ ($SE = 7$), $m_{BLA-damage} = 35$ ($SE = 18$)), Wilcoxon signed rank test: $z = 2.841$, $p = 0.005$, $n = 22$. The groups did not differ in their ratings (Mann-Whitney U-test: $p = 0.967$, $n = 22$), and the difference in the ratings of the trustees did not differ between the two groups (Mann-Whitney U-test: $p = 0.319$, $n = 22$). This suggests that the BLA-damaged subjects have potentially formed an explicit memory of the trustworthiness of the trustees. However, the variability of their ratings was high, and they were also not associated with their investments in that round ($b = 0.00$, $p = 0.837$). In the healthy subjects, the association between their ratings and investments differed significantly from the BLA-damaged subjects (trustworthiness rating x group interaction: $b = -6.38$, $p = 0.011$, Table S2), such that their ratings were positively associated with the investments in that round ($b = 0.07$, $p < 0.001$). Thus, while the ratings point towards explicit memory formation in the BLA-damaged subjects, they were unable to use this memory to adjust their investment behavior. While this interpretation requires caution, it fits with the function of the rodent BLA, which is not part of the declarative memory system (Bechara et al., 1995; Eichenbaum, 2000), but rather is involved in non-declarative, emotionally significant memory formation (R. J. McDonald & White, 1993; Roozendaal, McEwen, & Chattarji, 2009).

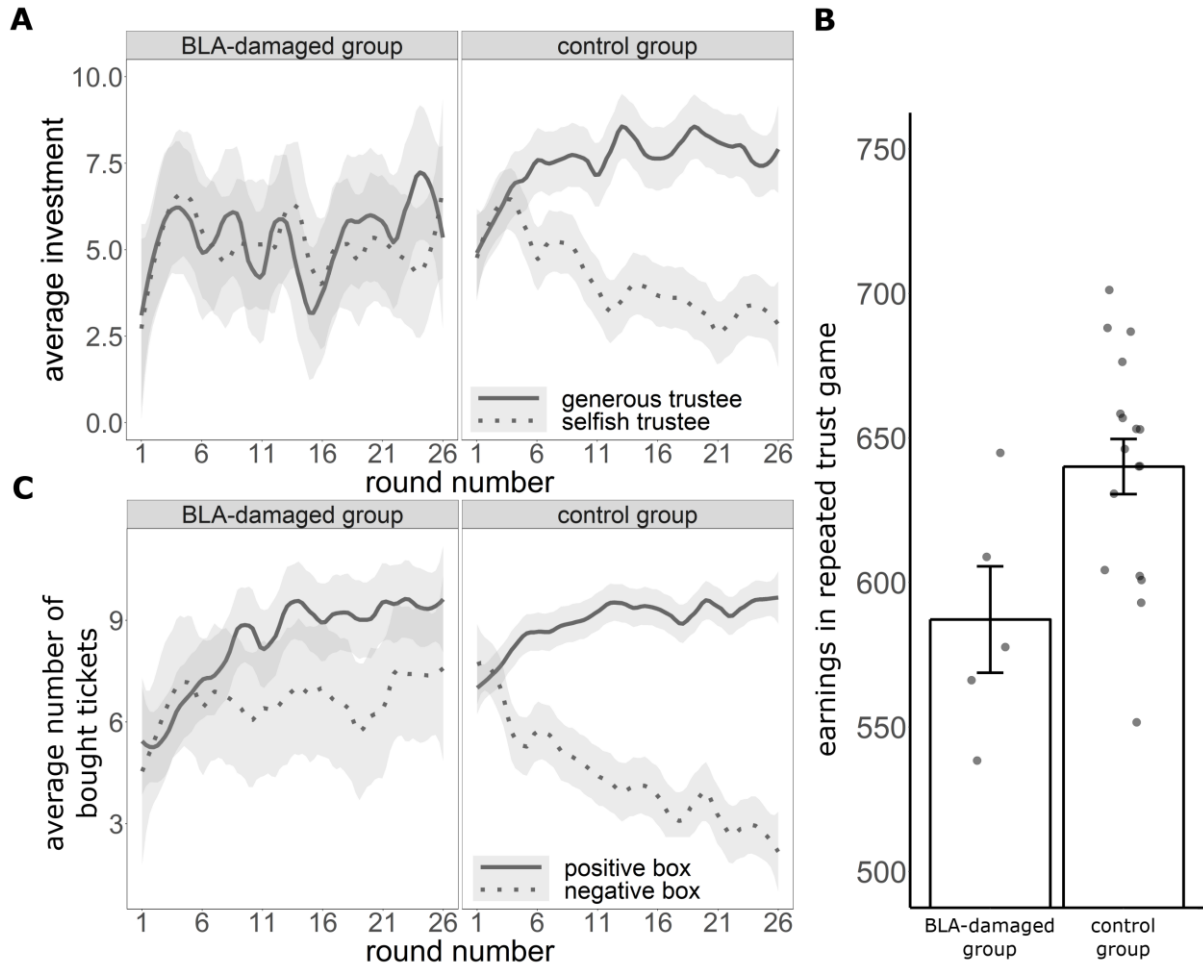


Figure 2. Learning in the repeated trust game and the raffle task. (A) Investments in the repeated trust game. Rounds are split into consecutive trustee interactions ($N = 26$ per trustee), plot is divided by group. Group data is smoothed with local regression (LOESS), shading represents 95% confidence intervals. (B) Total earnings (in MU) in repeated trust game, split by group. Dots represent individual subjects ($N_{\text{control group}} = 17$, $N_{\text{BLA-damaged group}} = 5$). (C) Bought tickets in the Raffle task. Rounds are split into consecutive box raffles ($N = 26$ per box), plot is divided by group. Group data are smoothed with local regression (LOESS), shading represents 95% confidence intervals. Individual data relating to panel A can be found in Figure S1. Analyses details can be found in Table S1 (for repeated trust game) and S3 (for Raffle task).

In the Raffle task, BLA-damaged subjects learned about the outcomes of the positive box and increased the amount of bought tickets, whereas healthy control subjects learned about the outcomes of both boxes and adapted their ticket buying behavior accordingly (see Figure 2C). To test the groups' learning behavior for the two boxes, we estimated a linear mixed model with random intercept per subject (Table S3). BLA-damaged subjects ($N = 5$) increased the amount of bought

tickets for the positive box from an average of 5.6 ($SE = 1.36$) to 9.8 ($SE = 0.2$), $b = 0.08$, $p < .001$, but did not decrease the amount of bought tickets for the negative box ($m_{round\ 1} = 4.6$ ($SE = 1.4$), $m_{round\ 52} = 7.6$ ($SE = 1.69$)), $b = 0.02$, $p = .066$ over the rounds (box x round number interaction: $b = 0.06$, $p = .004$). This differed from the control subjects' learning behavior for the two boxes (group x box x round number interaction: $b = -0.07$, $p < .001$). Control subjects ($N = 18$) adjusted their buying behavior both for the positive and the negative box over the rounds (box x round number interaction: $b = 0.13$, $p < .001$): they increased the amount of bought tickets for the positive box from an average of 6.94 ($SE = 0.61$) to 9.72 ($SE = 0.14$), $b = 0.04$, $p < .001$, and decreased the amount of bought tickets for the negative box from an average of 7.72 ($SE = 0.53$) to 2.17 ($SE = 0.2$), $b = -0.09$, $p < .001$ over the rounds. Thus, in the Raffle, BLA-damaged subjects were able to learn from positive outcomes but had impaired learning from negative outcomes where they could lose money.

During forced-choice control questions after the Raffle task, BLA-damaged subjects were able to discriminate between different reward magnitudes. They were as accurate as control subjects in choosing the box with which they earned more money during the task (Fisher's exact test: $p = .217$, $N = 23$), as well as the box which would earn them a higher reward when the reward options in the two boxes partly overlapped (Fisher's exact test: $p > .99$, $N = 23$), and when they did not overlap (Fisher's exact test: $p = .539$, $N = 23$).

Using the Raffle task, combined with control questions, we show that the experiential learning impairments in BLA-damaged subjects are specific to the social context. Furthermore, these learning impairments evidently are not due to problems in discrimination of reward magnitude, which is in line with spared reward discrimination abilities in BLA-lesioned rats (Hernandez-Lallement et al., 2016). The BLA-damaged subjects' inability to learn from negative outcomes in the Raffle task suggests different positive and negative non-social experiential learning mechanisms and related underlying neurocircuitry. In this respect our data are consistent with the theory of LeDoux and Daw (2018), which holds that rodent instrumental responses in a positive non-social context can be learned directly through stimulus-response associations (Lingawi & Balleine, 2012), rather than relying on BLA-dependent response–outcome associations (B. Balleine, Killcross, & Dickinson, 2003). Negative non-social experiential learning, in contrast, relies on BLA-dependent response–outcome associations in rodents (Killcross, Robbins, & Everitt, 1997). In sum, we show that in humans experiential learning in a social context is fully dependent on the BLA, while in a non-social context only negative experiential learning is dependent on the BLA. The latter finding supports the model of LeDoux and Daw (LeDoux & Daw, 2018), and is consistent with impairments in Pavlovian threat conditioning in BLA-lesioned subjects (Klumpers, Morgan, Terburg, Stein, & van Honk, 2015).

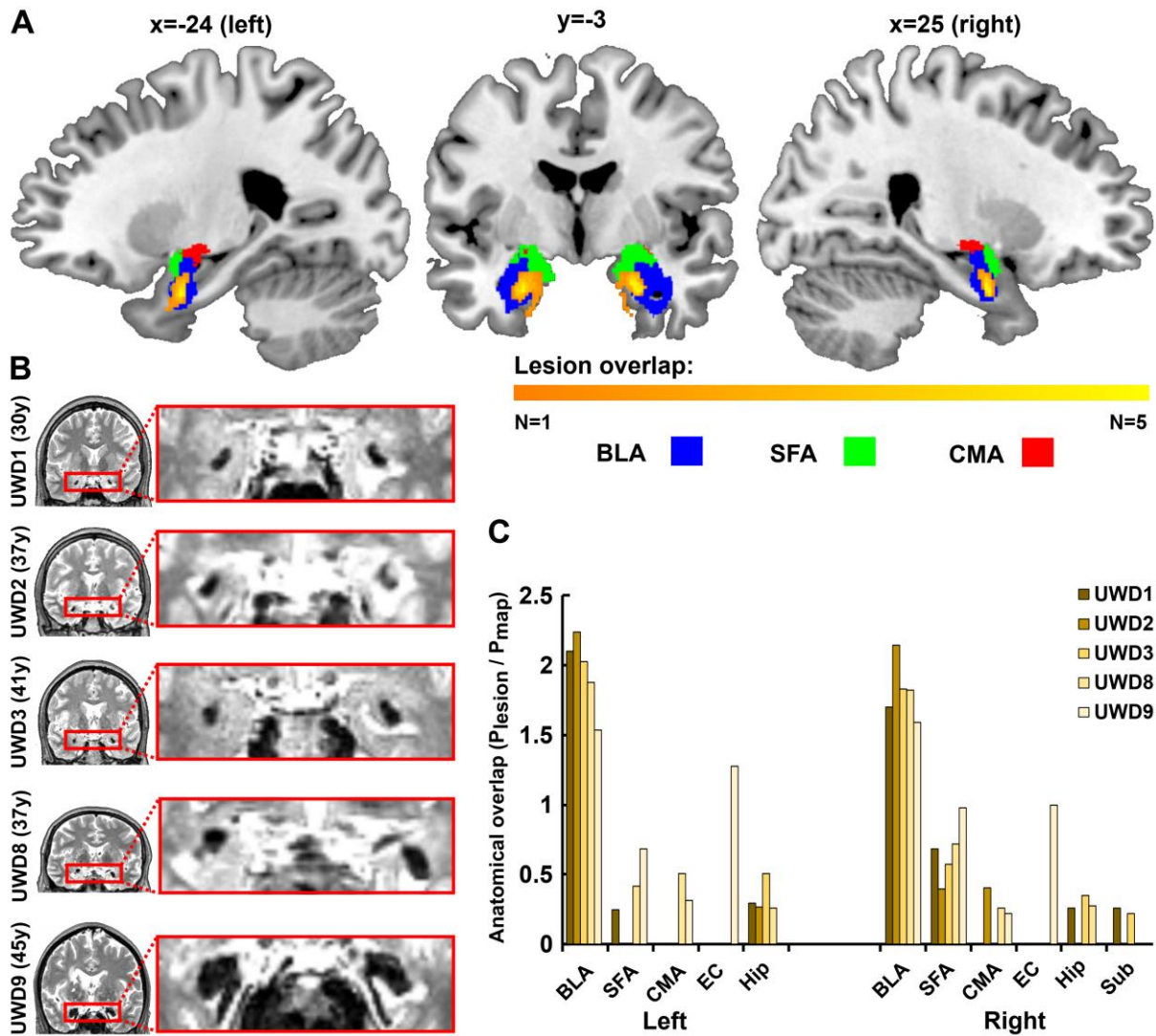


Figure 3. Bilateral and focal calcifications to the BLA. (A) Lesion-overlap image in MNI space plotted within the amygdala sub-regions defined as voxels with sub-region probability >50% (see Methods). (B) Coronal slices from each individual's raw T2-weighted MRI scan and age at time of scanning. (C) Bar-graph representing bilateral anatomical overlap with excess probability ($P(\text{lesion})/P(\text{map})$) values of the lesion volumes, whereby values >1 indicate a reliable match of volume and anatomical location of: BLA = Basolateral Amygdala, SFA = Superficial Amygdala, CMA = Central-Medial Amygdala complex, EC = Entorhinal Cortex, Hip = Hippocampus, Sub = Subiculum. UWD <X> (Urbach-Wiethe disease) represent the BLA-damaged subjects' identifiers.

Our results cannot be attributed to impairments in IQ (see Table 1), working memory (Morgan, Terburg, Thornton, Stein, & van Honk, 2012), or emotion perception (Terburg et al., 2012), as BLA-damaged subjects perform similar to and in some instances even better than healthy control subjects in these areas. Additionally, all subjects were instructed by a local native speaker who

ensured that everyone fully understood the task. Lastly, previously we showed generous investments in BLA-damaged subjects during a one-shot trust game (van Honk, Eisenegger, Terburg, Stein, & Morgan, 2013), where three BLA-damaged subjects invested 100% more than matched healthy controls. This generosity was not driven by increased trust in others, but arguably by altruistic motives, because the BLA-damaged subjects had similarly low expectations about their partners' back-transfers as the control subjects. In the present task we did not observe such generosity, defensibly because different behavioral mechanisms are present in the repeated trust game. That is, the repeated trust game we use not only requests continuous adjustments in response to feedback, but even before feedback there are reputational concerns that strongly influence decisions about investments.

In line with rodent models (B. W. Balleine & Killcross, 2006; Hernandez-Lallement et al., 2017), we demonstrate that the human BLA is necessary for experiential learning in a social context, that is, continuously adjusting behavior to distinct feedback from selfish and generous partners. In rodents the BLA–OMPFC (orbital medial prefrontal cortex) network seems to underlie learning that stimuli are associated with distinct outcomes, and subsequent behavioral updating (Wassum & Izquierdo, 2015). Our data suggest that the function and neural connectivity of the human BLA in this learning depends on an interaction between context (social or non-social) and valence (positive or negative) of the outcomes. These findings translate to rodent models of BLA function in social reinforcement learning (Hernandez-Lallement et al., 2017) and in learning from non-social positive and negative outcomes (LeDoux & Daw, 2018).

Our study furthermore highlights the importance of studying the functionality of human amygdala sub-nuclei separately, which is the typical approach in rodent research. Since amygdala sub-nuclei have different functionality and connectivity within the brain (A. J. McDonald, 1998), behavioral outcomes from human subjects with damage to the whole amygdala (Bhatt, Lohrenz, Camerer, & Montague, 2012) may be less informative for cross-species translation. Given the antagonistic properties of the BLA and CMA in decision-making in rodents, effects of damage on one sub-region might even be abolished by lesions to both sub-regions (B. W. Balleine & Killcross, 2006; Stalnaker, Franz, Singh, & Schoenbaum, 2007). We cannot exclude the possibility that selective CMA lesions in humans would result in similar behavioral patterns. There is no possibility to test this, as there are no humans with selective CMA lesions, as far as we know. Crucially however, our findings in the BLA-damaged subjects reported here and earlier (Klumpers et al., 2015; Terburg et al., 2018; van Honk et al., 2013) continuously translate to rodent models of the BLA. On that basis, we expect that the trust learning deficit in our BLA-damaged subjects depends on their selective BLA damage (see

Figure 3), and, also given our findings in the non-social Raffle task, that the different role for the rodent CMA in learning (B. W. Balleine & Killcross, 2006) might also apply to humans.

The BLA-damaged subjects' impairments in Pavlovian fear learning (Klumpers et al., 2015) are consistent with their impairments in experiential learning from negative outcomes in the Raffle task. In contrast, and most essentially, our findings with the repeated trust game show that the human BLA is necessary for all social experiential learning, that is, for both learning from negative and positive social outcomes. The extent to which the human BLA is involved in other aspects of reward processing (Wassum & Izquierdo, 2015) that are important for social decision-making remains to be established. In the interim, our study is the first that addresses the gap in cross-species translational research on the role of the human BLA in social experiential learning. We have established that the human BLA plays a crucial role in social learning, and specifically in deciding who can and cannot be trusted after repeated social-economic interactions, and in adjusting investment behavior accordingly. This parallels results and predictions from rodent research, which emphasize the role of the BLA in social learning (B. W. Balleine & Killcross, 2006; Hernandez-Lallement et al., 2017), but the relevance of these findings for humans was hitherto unresolved.

ACKNOWLEDGMENTS

In loving memory of Christoph Eisenegger, who was the mind and soul behind this project. Chris died in Namibia, 27 February 2017. We thank Marijn Moerbeek for creating Figure 1 and Claus Lamm for valuable feedback on the manuscript. The work in this paper was supported by the National Research Foundation (NRF) South Africa, the South African Medical Research Council, the University of Cape Town (UCT), Utrecht University (UU), Royal Holloway, the Vienna Science and Technology Fund (WWTF VRG13-007), and the Marietta-Blau grant (OeAD-GmbH and the Austrian Ministry for Science, Research, and Economics (BMWFW)).

AUTHOR CONTRIBUTIONS

Conceptualization and Methodology, L.A.R., C.E., M.N., and J.v.H.; Investigation, C.E., M.N., J.C., L.A.R., and J.v.H.; Formal Analysis, L.A.R., M.N. and D.T.; Writing – original draft, L.A.R., D.T., M.N., and J.v.H.; Writing – review & editing, L.A.R., D.T., M.N., D.J.S. and J.v.H.; Funding Acquisition, C.E., L.A.R., and J.v.H.; Supervision, C.E., D.J.S. and J.v.H.

DECLARATION OF INTERESTS

D.J.S. has received research grants and/or consultancy honoraria from Lundbeck, and Sun. All other authors declare no competing interests.

STAR METHODS

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Jack van Honk (j.vanhonk@uu.nl). This study did not generate new unique reagents.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

We tested 5 female subjects with focal, bilateral damage to the BLA, due to a genetic disease (Urbach-Wiethe disease) where parts of the brain calcify over time (Koen et al., 2016). Previous research in this population showed that the calcifications are bilateral and focal to the BLA, while the CMA remains functional (Terburg et al., 2012). The control group, consisting of 17 healthy subjects, were matched for age, gender, socio-economic status, cultural-ethnic status, religion, and IQ (measured with the Wechsler Abbreviated Scale of Intelligence (WASI) (Wechsler, 1999)), see Table 1. All subjects were recruited from the same rural Namaqualand area in South Africa, and all had similar socio-economic backgrounds. None of the subjects had any history of secondary psychopathology. Testing took place in the rural Namaqualand area with a local experimenter who spoke the same Afrikaans language as the subjects. Subjects provided informed consent before the beginning of the test session. The Health Sciences Faculty Human Research Ethics Committee of the University of Cape Town, South Africa, approved the study (HREC # 639/2016).

Table 1.

Age and IQ from the Wechsler Abbreviated Scale of Intelligence (WASI) for the BLA-lesioned subjects and healthy controls. VIQ, verbal IQ; PIQ, performance IQ; FSIQ, full-scale IQ; SD, standard deviation; UWD, Urbach-Wiethe disease.

	BLA-damaged subjects							Controls	
	UWD 1	UWD 2	UWD 3	UWD 8	UWD 9	Mean	SD	Mean	SD
Age	30	37	41	37	45	38.00	5.57	36.41	6.86
VIQ	97	84	93	89	87	90.00	5.10	89.71	3.77
PIQ	99	87	85	95	89	91.00	5.83	88.65	4.73
FSIQ	98	84	87	91	86	89.20	5.54	87.76	3.03

METHOD DETAILS

REPEATED TRUST GAME

The Repeated Trust Game was adapted from Fett and colleagues (Fett et al., 2012). Subjects played 52 rounds in the role of the investor, with two randomly alternating trustees. Subjects did not receive any priors about the trustees and had to learn through trial-and-error how to best adjust their investments (i.e. high investments with generous trustee, low investments with selfish trustee). All players were depicted with a neutral avatar picture (see Figure 1) and a neutral name (player B/C), which were randomly assigned to the trustees. At the beginning of every round both players received an endowment of 10 monetary units (=MU). The investor had to invest at least 1 MU. Investments were tripled and transferred to the trustee. Trustees could make a variable back-transfer. In the first four rounds of the task back-transfers of the trustees were 100% of the subject's investment, so that there were no immediate ceiling or floor effects on the investment and a learning curve could be measured easily. After the fourth-round back-transfers for the generous trustee were either 100%, 150%, or 200% of the investment (with an equal probability of every option occurring). When the subject increased their investment relative to the previous round with the generous trustee, the probability of receiving a back-transfer of 200% of the investment increased with 10% (and thus the other two options decreased with 5% each). Back-transfer options for the selfish trustee were either 100%, 75%, or 50% of the investment. When the subject decreased their investment relative to the previous round with the selfish trustee, the probability of receiving a back-transfer of 50% of the investment increased with 10%. Because of the semi-random nature of the back-transfer options, it is possible that some subjects received ambivalent back-transfers of 100% of their investments in consecutive rounds, making it difficult to discern the trustees and thus to learn about their trustworthiness. Analyses of the back-transfer options in rounds 5 - 16 (Rosenberger et al., 2019) did not reveal a group difference in the amount of ambivalent back-transfer options. In the last two rounds of the task, subjects were prompted to indicate the trustworthiness (with a slider on a scale from 0 – 100) of the trustee they were playing with in that round. This prompt appeared after the investment and before the back-transfer decision. Analyses of the association between the investments and the ratings in these two rounds are described in Table S2. Points earned throughout the task were exchanged to South African Rand. The task was played in z-tree (version 3.42 (Fischbacher, 2007)). The experimenter explained the task verbally in Afrikaans until subjects understood the task. In addition, subjects played 10 practice rounds in z-tree, while the experimenter was present to familiarize themselves with the software and the task. During the practice rounds the experimenter gave additional explanations of the task.

RAFFLE TASK

This task was designed to match exactly the outcomes, the outcome probabilities, as well as the adjustment of the outcome probabilities of the repeated trust game. The task was programmed in z-tree (version 3.42). In this task subjects played a raffle for 52 rounds. In the raffle, subjects could buy raffle tickets for two different boxes, delineated by different colors (blue and green) and patterns (horizontal stripes, no stripes), in consecutive rounds (order of box presentation was random and was restricted to two consecutive presentations per box). Subjects did not have any priors about the boxes. At the beginning of a round, one of the two closed raffle boxes was presented, indicating for which box subjects would buy raffle tickets in that round. Then in step 1, subjects indicated how many raffle tickets (between 1 and 10) they want to buy for the box. In step 2, the box was opened and the amount of money the subject won was presented. Earnings were paid out at the end of the task. Subjects played 5 practice rounds with different raffle boxes. After the task, subjects were presented with three forced-choice control questions. In the first question the two closed boxes from the Raffle task were shown side by side and subjects had to choose the one with which they earned more money during the task. In the other questions two new boxes were shown with each containing three possible outcomes. In question two these outcomes represented the outcome probabilities of the Raffle task (box 1: 5, 7, 10 MU; box 2: 10, 15, 20 MU), and in question three the possible outcomes were non-overlapping (box 1: 12, 15, 18 MU; box 2: 8, 6, 4 MU). Subjects were instructed to choose one of the two presented boxes, and informed that one of the possible outcomes from the chosen box would be paid out at the end of the task. All instructions and in-game text were displayed in the native language of the subjects (Afrikaans).

LESION METHODS

MRI-scans were acquired with a Siemens Magnetom Skyra 3-Tesla scanner at the Cape Universities Brain Imaging Centre (CUBIC) in Cape Town, South Africa. For lesion analysis, we obtained whole brain T2-weighted images using an iPAT acceleration-factor of 3, voxel-size 0.9x0.9x1mm, TR = 3200 ms, and TE = 410 ms.

QUANTIFICATION AND STATISTICAL ANALYSIS

BEHAVIORAL ANALYSIS

We analyzed investment behavior both with parametric and with non-parametric two-tailed tests with alpha set to 0.05. For the parametric tests we computed linear mixed models with the lme4 package (version 1.1-15 (Bates, Mächler, Bolker, & Walker, 2015)) in R (version 3.4.3 (R Core Team, 2017)), which included a random intercept per subject. Effects were calculated with a treatment contrast and p-values were computed via Wald-statistics approximation. The results table in the supplements was produced with the SjPlot package (version 2.4.1 (Ludecke & Schwemmer, 2017)),

and the graphs with the ggplot2 package (version 2.2.1 (Wickham, 2009)). For the non-parametric tests, we used the Mann-Whitney U test for between-subject comparisons, the Wilcoxon-signed rank test for within-subject comparisons, and Kendall's tau for the correlation between final rounds investment and trustworthiness rating.

LESION ANALYSIS

To estimate extent and anatomical location of the lesions, T2-weighted scans were normalized to MNI-space using unified segmentation, which is optimized for normalization of lesioned brains (Crinion et al., 2007). Lesion volumes were defined using the 3D volume-of-interest featured implemented in MRICron (Rorden & Brett, 2000). Based on MR-images the precise borders between amygdalae and neighboring structures, or between the sub-regions of the amygdala, cannot be established (Amunts et al., 2005; Solano-Castiella et al., 2010). To determine the precise location of the lesions in our BLA-damaged subjects, we therefore assigned the lesion volumes to cytoarchitectonic probability maps according to the method described by Eickhoff and colleagues (Eickhoff et al., 2007). In this method, which is implemented in the SPM8 anatomy toolbox (Eickhoff et al., 2005), a volume of interest is superimposed onto a cytoarchitectonic probability map of the medial-temporal lobe (Amunts et al., 2005). This map is based on microscopic analyses of ten post-mortem human brains and follows a generally accepted division of the human amygdala in three sub-regions. The first is the central-medial amygdala (CMA), which consists of the central and medial nuclei. The second is the basolateral amygdala (BLA), which includes the lateral, basolateral, basomedial, and paralaminar nuclei, and the third is the superficial (or corticoid) amygdala (SFA), which includes the anterior amygdaloid area, amygdalopyriform transition area, amygdaloid-hippocampal area, and the cortical nucleus (Amunts et al., 2005). This method assigns to any given voxel a value representing the probability that it belongs to an underlying structure. These structures are derived from an overlap analysis of ten post-mortem brains and are therefore divided in ten separate probability classes ranging from 10% to 100% probability.

To estimate how well the lesion volumes fit to the underlying structure, $P(\text{excess})$ values are computed using the following equation: $P(\text{excess}) = \frac{P(\text{lesion})}{P(\text{map})}$ whereby $P(\text{lesion})$ represents the average cytoarchitectonic probability of the voxels that are shared by the lesion and the cytoarchitectonic probability map, and $P(\text{map})$ represents the average probability of the whole structure's cytoarchitectonic map. These values thus represent how much the average probability of the overlapping voxels exceed the overall probability distribution of that particular structure, and thus indicate whether the lesion overlaps with relatively high or low probability classes of that structure. In other words, $P(\text{excess})$ represents how 'central' the location of the lesion is relative to

that structure's cytoarchitectonic map, whereby $P(\text{excess}) > 1$ indicates a more central, and $P(\text{excess}) < 1$ a more peripheral location (Eickhoff et al., 2007)

As depicted in Figure 3A and C, calcified brain-tissue of the BLA-damaged subjects is localized in the BLA and the CMA seems unaffected. In a quantitative analysis these results are confirmed. Figure 3C shows $P(\text{excess})$ values for the individual lesions and these lesions are, bilaterally, most central to the BLA as $P(\text{excess})$ values exceed 1 for each individual and hemisphere. Since this method is purely based on probability distributions, it is impossible to fully exclude the possibility that structures other than the BLA are affected by the calcifications. The fact that the lesion-volumes largely overlap with high probability classes in the bilateral BLA, and that $P(\text{excess})$ values greatly exceed the value of 1, can however be seen as strong support for our claim that these subjects have bilateral damage limited to the BLA. In subject UWD9 the calcifications might extend into neighboring structures, namely the right superficial amygdala (SFA) and bilateral entorhinal cortex (all main results reported above remain significant when excluding subject UWD9). We can however safely conclude that the CMA is unaffected by the bilateral calcifications found in all of these BLA-damaged subjects.

DATA AND CODE AVAILABILITY

Behavioral data and analyses scripts are deposited here: osf.io/hwc5q (Rosenberger et al., 2019). Data relating to lesion analyses will only be made available upon request from the lead contact due to the de-anonymized nature of the data.

REFERENCES

- Amunts, K., Kedo, O., Kindler, M., Pieperhoff, P., Mohlberg, H., Shah, N. J., . . . Zilles, K. (2005). Cytoarchitectonic mapping of the human amygdala, hippocampal region and entorhinal cortex: Intersubject variability and probability maps. *Anatomy and Embryology*, 210(5-6), 343-352. doi:10.1007/s00429-005-0025-5
- Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211(4489), 1390-1396.
- Balleine, B., Killcross, S., & Dickinson, A. (2003). The effect of lesions of the basolateral amygdala on instrumental conditioning. *Journal of Neuroscience*, 23(2), 666 - 675.
- Balleine, B. W., & Killcross, S. (2006). Parallel incentive processing: an integrated view of amygdala function. *Trends in Neurosciences*, 29(5), 272-279. doi:10.1016/j.tins.2006.03.002
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1 - 48.
- Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., & Fehr, E. (2008). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron*, 58(4), 639-650. doi:10.1016/j.neuron.2008.04.009
- Bechara, A., Tranel, D., Damasio, H., Adolphs, R., Rockland, C., & Damasio, A. (1995). Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. *Science*, 269(5227), 1115-1118. doi:10.1126/science.7652558
- Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *GEB*, 10, 122 - 142.
- Bhatt, M. A., Lohrenz, T., Camerer, C. F., & Montague, P. R. (2012). Distinct contributions of the amygdala and parahippocampal gyrus to suspicion in a repeated bargaining game. *Proceedings of the National Academy of Sciences of the United States of America*, 109(22), 8728-8733. doi:10.1073/pnas.1200738109
- Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: trustworthiness as a dynamic belief. *Cogn Psychol*, 61(2), 87-105. doi:10.1016/j.cogpsych.2010.03.001
- Crinion, J., Ashburner, J., Leff, A., Brett, M., Price, C., & Friston, K. (2007). Spatial normalization of lesioned brains: Performance evaluation and impact on fMRI analyses. *NeuroImage*, 37(3), 866-875. doi:10.1016/j.neuroimage.2007.04.065
- Eichenbaum, H. (2000). A cortical-hippocampal system for declarative memory. *Nat Rev Neurosci*, 1(1), 41-50. doi:10.1038/35036213
- Eickhoff, S. B., Paus, T., Caspers, S., Grosbras, M. H., Evans, A. C., Zilles, K., & Amunts, K. (2007). Assignment of functional activations to probabilistic cytoarchitectonic areas revisited. *NeuroImage*, 36(3), 511-521. doi:10.1016/j.neuroimage.2007.03.060

- Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, 25(4), 1325-1335. doi:10.1016/j.neuroimage.2004.12.034
- Fareri, D. S., Chang, L. J., & Delgado, M. R. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Front Neurosci*, 6, 148. doi:10.3389/fnins.2012.00148
- FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences of the United States of America*, 115(7), E1690-E1697. doi:10.1073/pnas.1715227115
- Fett, A. K., Gromann, P. M., Giampietro, V., Shergill, S. S., & Krabbendam, L. (2012). Default distrust? An fMRI investigation of the neural development of trust and cooperation. *Soc Cogn Affect Neurosci*, 9(4), 395-402. doi:10.1093/scan/nss144
- Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2), 171-178.
- Gupta, R., Kosciak, T. R., Bechara, A., & Tranel, D. (2011). The amygdala and decision-making. *Neuropsychologia*, 49(4), 760-766. doi:10.1016/j.neuropsychologia.2010.09.029
- Hernandez-Lallement, J., van Wingerden, M., Schable, S., & Kalenscher, T. (2016). Basolateral amygdala lesions abolish mutual reward preferences in rats. *Neurobiology of Learning and Memory*, 127, 1-9. doi:10.1016/j.nlm.2015.11.004
- Hernandez-Lallement, J., van Wingerden, M., Schable, S., & Kalenscher, T. (2017). A Social Reinforcement Learning Hypothesis of Mutual Reward Preferences in Rats. *Curr Top Behav Neurosci*, 30, 159-176. doi:10.1007/7854_2016_436
- Janak, P. H., & Tye, K. M. (2015). From circuits to behaviour in the amygdala. *Nature*, 517(7534), 284-292. doi:10.1038/nature14188
- Killcross, S., Robbins, T. W., & Everitt, B. J. (1997). Different types of fear-conditioned behaviour mediated by separate nuclei within amygdala. *Nature*, 388.
- Klumpers, F., Morgan, B., Terburg, D., Stein, D. J., & van Honk, J. (2015). Impaired acquisition of classically conditioned fear-potentiated startle reflexes in humans with focal bilateral basolateral amygdala damage. *Soc Cogn Affect Neurosci*, 10(9), 1161-1168. doi:10.1093/scan/nsu164
- Koen, N., Fourie, J., Terburg, D., Stoop, R., Morgan, B., Stein, D. J., & Van Honk, J. (2016). Translational neuroscience of basolateral amygdala lesions: studies of Urbach-Wiethe Disease. *Journal of Neuroscience Research*, 94, 504 - 512.
- LeDoux, J., & Daw, N. D. (2018). Surviving threats: neural circuit and computational implications of a new taxonomy of defensive behaviour. *Nature Reviews Neuroscience*, 19(5), 269-282. doi:10.1038/nrn.2018.22

- Lingawi, N. W., & Balleine, B. W. (2012). Amygdala central nucleus interacts with dorsolateral striatum to regulate the acquisition of habits. *Journal of Neuroscience*, 32(3), 1073-1081. doi:10.1523/JNEUROSCI.4806-11.2012
- Ludecke, D., & Schwemmer, C. (2017). sjPlot: Data visualization for statistics in social science (version 2.3. 1)(software).
- McDonald, A. J. (1998). Cortical pathways to the mammalian amygdala. *Progress in Neurobiology*, 55(3), 257-332.
- McDonald, R. J., & White, N. M. (1993). A triple dissociation of memory systems: hippocampus, amygdala, and dorsal striatum. *Behavioral Neuroscience*, 107(1), 3-22.
- Morgan, B., Terburg, D., Thornton, H. B., Stein, D. J., & van Honk, J. (2012). Paradoxical facilitation of working memory after basolateral amygdala damage. *PLoS One*, 7(6), e38116. doi:10.1371/journal.pone.0038116
- R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
- Roosendaal, B., McEwen, B. S., & Chattarji, S. (2009). Stress, memory and the amygdala. *Nat Rev Neurosci*, 10(6), 423-433. doi:10.1038/nrn2651
- Rorden, C., & Brett, M. (2000). Stereotaxic display of brain lesions. *Behavioural Neurology*, 12, 191 - 200.
- Rosenberger, L. A., Eisenegger, C., Naef, M., Terburg, D., Fourie, J., Stein, D. J., & Van Honk, J. (2019). *Data and analyses scripts accompanying manuscript: The human basolateral amygdala is indispensable for social experiential learning*. Retrieved from: osf.io/hwc5q
- Solano-Castiella, E., Anwender, A., Lohmann, G., Weiss, M., Docherty, C., Geyer, S., . . . Turner, R. (2010). Diffusion tensor imaging segments the human amygdala in vivo. *NeuroImage*, 49(4), 2958-2965. doi:10.1016/j.neuroimage.2009.11.027
- Stalnaker, T. A., Franz, T. M., Singh, T., & Schoenbaum, G. (2007). Basolateral amygdala lesions abolish orbitofrontal-dependent reversal impairments. *Neuron*, 54(1), 51-58. doi:10.1016/j.neuron.2007.02.014
- Terburg, D., Morgan, B. E., Montoya, E. R., Hooge, I. T., Thornton, H. B., Hariri, A. R., . . . van Honk, J. (2012). Hypervigilance for fear after basolateral amygdala damage in humans. *Transl Psychiatry*, 2, e115. doi:10.1038/tp.2012.46
- Terburg, D., Scheggia, D., Triana Del Rio, R., Klumpers, F., Ciobanu, A. C., Morgan, B., . . . van Honk, J. (2018). The Basolateral Amygdala Is Essential for Rapid Escape: A Human and Rodent Study. *Cell*, 175(3), 723-735 e716. doi:10.1016/j.cell.2018.09.028
- van Honk, J., Eisenegger, C., Terburg, D., Stein, D. J., & Morgan, B. (2013). Generous economic investments after basolateral amygdala damage. *Proceedings of the National Academy of Sciences of the United States of America*, 110(7), 2506-2510. doi:10.1073/pnas.1217316110

- Wassum, K. M., & Izquierdo, A. (2015). The basolateral amygdala in reward learning and addiction. *Neuroscience and Biobehavioral Reviews*, 57, 271-283. doi:10.1016/j.neubiorev.2015.08.017
- Wechsler, D. (1999). *Wechsler abbreviated scale of intelligence*. San Antonio, Texas: Psychological Corporation.
- Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

SUPPLEMENTS

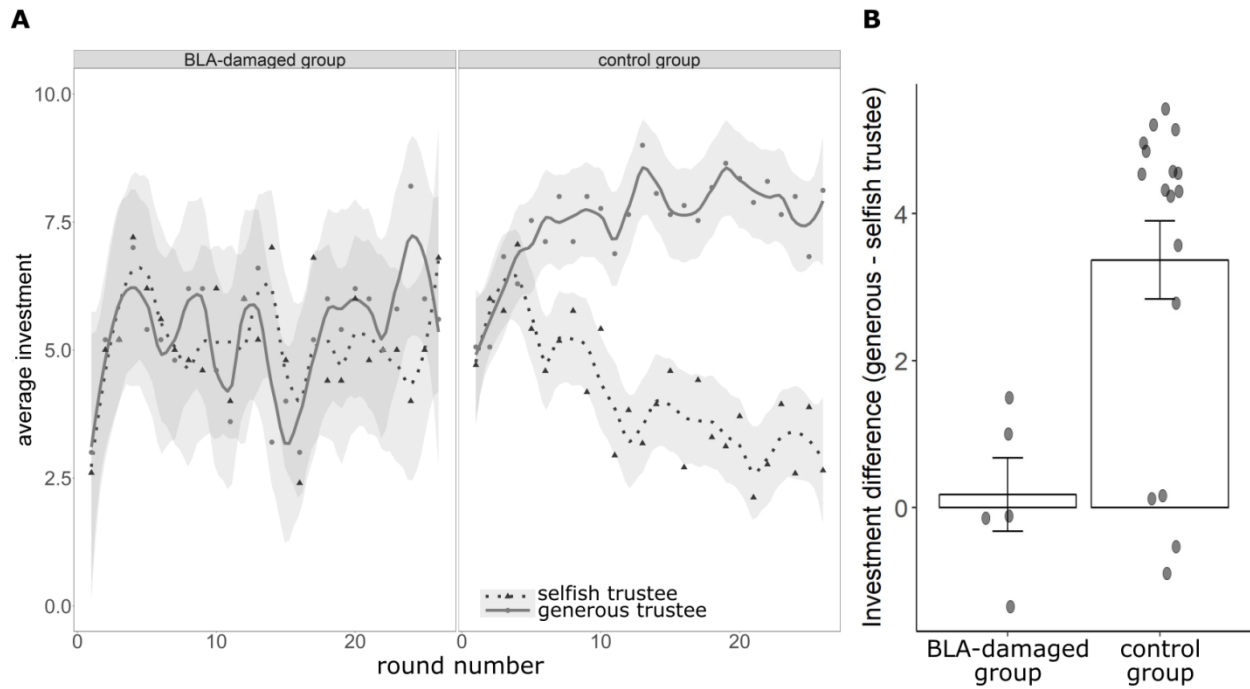


Figure S1. Learning in the repeated trust game. Related to Figure 2A. A) Investments over the course of the repeated trust game. Rounds are split into consecutive trustee interactions ($N = 26$ per trustee), plot is divided by group. Group data is smoothed with a local regression (LOESS), and shaded areas represent 95% confidence intervals. Individual data points ($N = 52$ per panel) represent averaged investment per trustee interaction (triangle for selfish trustee, circle for generous trustee). B) Investment difference scores (average investment generous trustee – average investment selfish trustee) per subject ($N_{BLA-damaged\ group} = 5$, $N_{control\ group} = 17$) with bar graph representing mean group value and error bars representing standard errors.

Table S1.

Beta statistics of linear mixed models for investment decisions during the repeated trust game.
Related to Figure 2A.

	M1: Investment			M2: Investment		
	<i>B</i>	<i>SE</i>	<i>p</i>	<i>B</i>	<i>SE</i>	<i>p</i>
Fixed Parts						
Dummy for selfish trustee	-3.37	0.18	<.001	-0.66	0.37	.070
Dummy for BLA-damaged group	-2.18	0.65	.001	-1.77	0.80	.027
Selfish trustee * BLA-damaged group	3.19	0.38	<.001	1.05	0.77	.173
Round Number				0.07	0.02	<.001
Selfish trustee * Round Number				-0.20	0.02	<.001
BLA-damaged group * Round Number				-0.03	0.04	.387
Selfish trustee * BLA-damage * Round No.				0.16	0.05	.001
Intercept	7.51	0.31	<.001	6.51	0.38	<.001
Random Parts						
σ^2		7.420			6.959	
$\tau_{00, ID}$		1.330			1.339	
N_{ID}		22			22	
ICC_{ID}		0.152			0.161	
Observations		1144			1144	
R^2		.330			.374	

Note. ICC, intra-class-correlation; N, number of random intercepts; σ^2 , random within group variance; τ_{00} , random between-group variance; R^2 , difference of the total variance of the null and full model divided by the total variance of the null model; SE, standard error; B, Beta. Results from two linear mixed models regressing investment against trustee type, group, and their interaction, including random intercept per subject. The second model on the right side has round number added as a predictor. Predictors are coded as follows: Dummy for selfish trustee (0 = generous, 1 = selfish), Dummy for BLA-damaged group (0 = control group, 1 = BLA-damaged group), round number (numeric ranging from 1 – 26).

Table S2.

Beta statistics of linear mixed models for investment decisions in the last two rounds of the repeated trust game. Related to STAR methods section METHOD DETAILS.

	Investment		
	<i>B</i>	<i>SE</i>	<i>p</i>
Fixed Parts			
Trustworthiness ratings	0.07	0.01	<.001
Dummy for BLA-damaged group	4.07	1.55	.009
Trustworthiness Ratings*BLA-damaged group	-0.06	0.02	.011
Intercept	1.90	0.74	.010
Random Parts			
σ^2		6.88	
$\tau_{00, ID}$		0.00	
ICC_{ID}		0.00	
Observations		44	

Note. ICC, intra-class-correlation; σ^2 , random within group variance; τ_{00} , random between-group variance; SE, standard error; B, Beta. Results from linear mixed models regressing investment in last two rounds of the task against trustworthiness rating, group, and their interaction, including random intercept per subject. Predictors are coded as follows: trustworthiness rating (numeric rating from 0 – 1), Dummy for BLA-damaged group (0 = control group, 1 = BLA-damaged group).

Table S3.

Beta statistics of linear mixed model for bought tickets in Raffle task. Related to Figure 2C.

	Bought raffle tickets		
	<i>B</i>	<i>SE</i>	<i>p</i>
Fixed Parts			
Dummy for positive box	1.12	0.28	<0.001
Dummy for BLA-damaged group	-0.88	0.67	0.185
Round No.	-0.09	0.01	<0.001
Positive box * BLA-damaged group	-0.86	0.61	0.155
Positive box * Round No.	0.13	0.01	<0.001
BLA-damaged group * Round No.	0.12	0.01	<0.001
Positive box * BLA-damaged group * Round No.	-0.07	0.02	<0.001
Intercept	6.85	0.31	<0.001
Random Parts			
σ^2		4.55	
τ_{00} ID		1.02	
N ID		23	
ICC _{ID}		0.18	
Observations		1196	
R ²		0.585	

Note. ICC, intra-class-correlation; N, number of random intercepts; σ^2 , random within group variance; τ_{00} , random between-group variance; R^2 , difference of the total variance of the null and full model divided by the total variance of the null model; SE, standard error; B, Beta. Results from a linear mixed model regressing the number raffle tickets bought against box type, group, round number and their interactions, including random intercept per subject. Predictors are coded as follows: dummy for positive box (0 = negative, 1 = positive), Dummy for BLA-damaged group (0 = control group, 1 = BLA-damaged group), round number (numeric ranging from 1 – 26).

CHAPTER 3 - FAIRNESS NORM VIOLATIONS IN ANTI-SOCIAL PSYCHOPATHIC OFFENDERS IN A REPEATED TRUST GAME

Original manuscript as published in Translational Psychiatry

DOI: doi.org/10.1038/s41398-019-0606-3

Running title: anti-social psychopathic offenders norm violations

Lisa A. Rosenberger^{1*}, Daniela M. Pfabigan^{2,3}, Benjamin Lehner², Katinka Keckeis⁴, Eva-Maria Seidel², Christoph Eisenegger¹ and Claus Lamm^{1,2*}

1 Neuropsychopharmacology and Biopsychology Unit, Department of Basic Psychological Research and Research Methods, Faculty of Psychology, University of Vienna, Austria

2 Social, Cognitive and Affective Neuroscience Unit, Department of Basic Psychological Research and Research Methods, Faculty of Psychology, University of Vienna, Austria

3 Faculty of Medicine, Institute of Basic Medical Sciences, Dept. of Behavioural Sciences in Medicine, University of Oslo, Oslo, Norway

4 Prison for Mentally Disordered Offenders (Justizanstalt Wien Mittersteig), Vienna, Austria

* To whom correspondence may be addressed. Email: lisa.anna.rosenberger@univie.ac.at or claus.lamm@univie.ac.at, address: Liebiggasse 5, 1010 Wien, telephone: +43-1427747130, fax: +43-1427747139

ABSTRACT

Psychopathic offenders have a high propensity to violate social norms, as indicated for instance by their widespread lying and cheating behaviour. The reasons for their norm violations are not well understood, though, as they are able to recognize norms in a given situation and also punish norm violators. In the present study we investigated whether psychopathic offenders would violate fairness norms during a repeated trust game because of increased profit-maximizing concerns. We measured back-transfer decisions in the repeated trust game, and affective arousal by means of skin conductance responses (SCR) in violent offenders with varying degrees of psychopathy, and non-offenders with low trait psychopathy. Psychopathy in offenders was measured with the Psychopathy Checklist-Revised (PCL-R). In the task, a participant and an interaction partner entrusted each other money for multiple rounds with the goal to earn as much money as possible. Fairness norm violations were positively associated with Factor 2 scores (the lifestyle/ antisocial psychopathy subscale) of the PCL-R, but this was not accompanied by clear profit-maximizing behaviour. Additionally, anticipatory arousal to self-advantageous decisions was higher in all offenders, independent of their degree of psychopathy, compared to non-offenders. The results of our study widen our understanding of social decision making in psychopathy. They also suggest treatment possibilities in offenders scoring high on Factor 2, targeting empathic concern and related prosocial intentions to overcome norm violating behaviour.

INTRODUCTION

Adhering to social norms is an integral part of successful functioning in everyday life. They help coordinate behaviour between people in situations that carry a conflict between self-interest and cooperative outcomes (Young, 2008). Norm adherence usually is prosocial and increases outcomes of the group, even though this often comes at the expense of one's own profits (Bicchieri, 2005). However, not all individuals are willing to adhere to social norms to the same extent. For example, psychopathic individuals have a reputation of violating social norms by lying, cheating, and deceiving on a regular basis (Hare, 2003). They are also highly reward sensitive (Geurts et al., 2016; Glenn, Raine, Yaralian, & Yang, 2010; Pujara, Motzkin, Newman, Kiehl, & Koenigs, 2014), which is why they might favour their own over group profits, and hence neglect social norms. This implies, though, that they should build a negative reputation and function very poorly in our society, but many studies on psychopathic personality traits in corporate environments suggest the opposite, (e.g. Gao & Raine, 2010). We aimed to shed light on this contradiction by studying social norm adherence and underlying affective arousal in offenders with varying degrees of psychopathy. Affective arousal was assessed by means of skin conductance responses (SCR), and fairness norm violations were assessed in a behavioural economics approach, using a repeated trust game. This approach is unique, as most previous decision-making research on psychopathy had used single-shot games, whose limitations could be overcome by the repeated nature of our task design. Most importantly, this enabled us to disentangle behaviour driven by profit-maximizing self-interest from group-benefiting norm adherence.

To adhere to norms it is important to recognize a shared norm in a given situation, as well as a deviation from it (Montague & Lohrenz, 2007). The large majority of individuals not only expects others to adhere to these norms, but also punishes others if they do not do so (Young, 2008), even if this comes at costs to the self (Fehr & Gächter, 2002). Psychopathic offenders are able to understand social norms in two player decision-making paradigms: They punish norm violators in an ultimatum game as often (Radke, Brazil, Scheper, Bulten, & de Bruijn, 2013), or even more often than healthy control participants (Koenigs, Kruepke, & Newman, 2010). However, much less is known about psychopathic offenders' norm adherence in these contexts. Evidence from single-shot decision-making studies – in which two people interact with each other exactly once and reputation effects do not play a role – suggest that even though psychopathic offenders behave highly profit-maximizing, they seem to adhere to fairness norms when there is the threat of retaliation from the interaction partner. In a dictator game, where the interaction partner has no opportunity to respond with retaliation or other ways of punishment of unfair offers, psychopathic offenders behave in their own

self-interest and keep most of an allocated amount of money to themselves (Koenigs et al., 2010; Mayer, Rauss, Pourtois, Jusyte, & Schonenberg, 2018; Scheeff, Jusyte, & Schonenberg, 2018). In contrast, in the ultimatum game an interaction partner gets the opportunity to reject an unfair offer. Here, psychopathic offenders favour joint outcomes and thus adhere to the norm of the game by splitting their allocated money in the same fair way as non-psychopathic participants (Koenigs et al., 2010; Mayer, Rauss, et al., 2018). To our knowledge only one study investigated psychopathic offenders' behaviour during repeated decision-making interactions (Mokros et al., 2008). Paradigms where participants interact with each other multiple times resemble real-world situations more closely than single-shot versions, as decisions are not taken in isolation but are informed by the possible consequences of the interaction partner's extended behaviour. These repeated interactions can either be simultaneous, or sequential in nature. In a repeated prisoner's dilemma, where two players decide simultaneously whether they want to cooperate with each other, psychopathic offenders manipulated their partners into believing they would adhere to a cooperative norm. In reality they betrayed them, maximized their own outcomes, and at the end earned more money than their partners (Mokros et al., 2008). Thus, decision-making studies have shown so far that, even though psychopathic offenders expect others to adhere to norms that maximize group profits and punish them if they do not do so, they do not seem to adhere to them themselves in simultaneous decision-making paradigms when there is no clear threat of retaliation.

Oftentimes, we do not make simultaneous decisions (as in Mokros and colleagues' paradigm(2008)), but rather react to our interaction partner in a sequential manner. In these situations decisions are more easily interpreted as direct responses to the interaction partner's behaviour, rather than an independent strategy. Negative behaviour gets punished immediately and can lead to a break-down of cooperation or even of interactions altogether (King-Casas et al., 2008). The repeated trust game is such a paradigm with which repeated sequential decision-making is studied (Eisenegger et al., 2013). In this task two players (an investor and a trustee) make sequential decisions to entrust each other money during multiple interactions, which are rooted in different motivations. A decision can be informed by inequity aversion, which is a dislike for situations where either of the players is worse off (Fehr & Schmidt, 1999). In that case a participant would adhere to a fairness norm that favours equal profits, and would therefore equalise the earnings of the investor and the trustee in a given round. From previous research it is not clear whether psychopathic offenders' behaviour would be driven by inequity aversion to the same extent as healthy participants during repeated sequential interactions. We address this by studying the psychopathic offender's fairness norm adherence during a repeated trust game.

Decisions in a repeated trust game can also be rooted in monetary self-interest and the valuation of monetary rewards. Psychopathic offenders are known to process and react to rewards differently than non-psychopathic offenders. They have difficulties inhibiting their impulses when they can earn rewards (Newman, Patterson, Howland, & Nichols, 1990), and prefer immediate over delayed rewards (Newman, Kosson, & Patterson, 1992). Additionally, there is an increased activation (Pujara et al., 2014) and a greater connectivity with frontal brain areas (Geurts et al., 2016) during reward processing in the ventral striatum of psychopathic offenders compared to healthy control subjects. Arnett(1997) described the motivational processes in psychopathic offenders as an imbalance between the punishment and reward system. Even though many studies have supported Arnett's hypothesis of psychopathic offenders' hypo-responsiveness to negative stimuli, Arnett's prediction of hyper-responsive arousal to rewarding stimuli has only attracted little research with mixed findings, (e.g. Arnett & Newman, 2000). Even though economic decision-making paradigms lend themselves perfectly to test hyper-responsive arousal to rewards, we are not aware of any decision-making study that included arousal measures in psychopathic offenders. We addressed this in our study and hypothesized that psychopathic offenders' aberrant reward processing would manifest itself both on a behavioural level, i.e. in terms of their decisions during the repeated trust game, and on the affective arousal level, as measured by skin conductance responses. On a behavioural level we expected a preference for own rather than joint profits in psychopathic offenders, and therefore a neglect of the fairness norm. On the psychophysiological level, we expected the psychopathic offenders to display hyper-responsive arousal to both the anticipation and the receipt of rewards.

We tested these hypotheses by comparing fairness norm adherence behaviour during a repeated trust game in two groups: Violent offenders with varying degrees of psychopathy, and a matched non-delinquent group scoring low on psychopathy. Within the offender group we hypothesized that the effects would be mainly driven by Factor 1, which is the interpersonal and affective subscale of our psychopathy measure (Psychopathy Checklist – Revised (PCL-R; Hare, 2003)). Factor 1 captures the deceptive and manipulative nature that is linked to norm violations. Participants played the task as trustees, i.e. as the person who could decide through their back-transfer whether or not to fully or partially reciprocate the trust invested in them.

METHODS

PARTICIPANTS

We recruited 30 male incarcerated offenders from an Austrian prison (Wien Mittersteig), a facility for prisoners with a psychiatric disorder who are deemed criminally responsible according to the Austrian criminal code §21, para. 2. We excluded three participants with intelligence scores one standard deviation below the average of the offender sample, and one due to non-violent offenses. One offender aborted the study after the first task (the task battery consisted of four tasks in total, findings for those outside the scope of the present paper will be described elsewhere). Thus the final offender sample consists of 25 male violent offenders. On average the offenders were imprisoned for 4.58 years ($SD = 3.8$). Education levels of all offenders were limited to compulsory education level (i.e., 9 years of schooling in Austria, $N = 7$), vocational training level ($N = 16$), and secondary education level ($N = 2$). Eleven offenders were taking psychiatric medication at the date of testing, which included antipsychotics ($N = 7$), antidepressants ($N = 2$), a mix of the two ($N = 1$), or tranquilizers ($N = 1$). Importantly, the analyses reported below did not change depending on medication or duration of imprisonment (Rosenberger, 2019). Note that while the PCL-R total scores of our offender sample lies below the average of American cohorts (see e.g. Koenigs et al., 2010), they are comparable to other European samples (see e.g. Mokros et al., 2008; Pfabigan et al., 2015; Seidel et al., 2013).

In addition, we recruited 28 male non-delinquent participants from a community cohort that were matched with the offender sample on age, education level, and intelligence. Participants were recruited via public advertisements in local hospitals, supermarkets, job centres, and through online postings on local job search sites. Participants were excluded because of missing skin conductance data ($N = 2$), psychopathy scores three times above the sample average ($N = 1$), and back-transfers of 0 MU ($N = 2$), which made it impossible to analyse behavioural variation. Thus the final non-delinquent group consisted of 23 male participants. Education levels of the non-offenders included compulsory education ($N = 7$), vocational training ($N = 12$), and secondary education ($N = 4$). None of the non-offenders had any history of psychiatric or neurological disorders or drug abuse (screened with the SCIDPIT light (Demal, 1999) and the SCID II (Fydrich, Renneberg, Schmitz, & Wittchen, 1997)). The study was approved by the local university ethics committee, as well as the institutional review board of the correctional treatment facility, and was conducted according to the declaration of Helsinki (2001). All participants provided informed consent before the start of the test session. Descriptives of age and questionnaire scores of all groups can be found in Table 1.

QUESTIONNAIRES

To assess psychopathic personality traits, we used the Psychopathy Checklist – Revised (PCL-R; Hare, 2003) in the offenders, and the Psychopathic Personality Inventory – revised (PPI-R; Alpers & Eisenbarth, 2008) in the non-offenders. High scores reflect a high degree of psychopathy. We used the continuous PCL-R scores to analyse the effect of psychopathy on norm violations in the offender sample. Impulsivity was measured with the Barratt Impulsiveness Scale (BIS-11; Preuss et al., 2008). Intelligence was assessed with Raven’s Standard Progressive Matrices (SPM; Raven, 1969), and self-reported trait interpersonal trust with the Interpersonal Trust Scale (ITS; Amelang, Gold, & Kuelbel, 1984). Details about the questionnaires can be found in the supplemental materials. Since these questionnaire scores were not the main focus of the study, we only included them as covariates of no-interest in all group comparison analyses to control for between-group variations.

Table 1.

Questionnaire scores and age of the two groups.

	offenders		non-offenders			
	mean (range)	SD	mean (range)	SD	t-value	p-value
age	35.41	10.19	34.65	10.12	-0.26	0.79
PCL-R						
<i>total</i>	20.36 (8 - 32)	6.97				
<i>Factor 1</i>	9.2 (2 - 14)	3.07				
<i>Factor 2</i>	9.0 (1 - 16)	4.56				
PPI-R						
<i>total</i>			294 (253 - 322)	17.83		
<i>Factor 1</i>			110.76	12.15		
<i>Factor 2</i>			151.32	17.75		
<i>coldheartedness</i>			31.4	5.82		
BIS 11	53.96	9.3	62.13	9.56	3.00	< 0.01
SPM	32.88	30.39	26.7	3.46	-1.01	0.32
ITS	71.92	13.81	74.26	11.56	0.64	0.53

TASK – REPEATED TRUST GAME

Participants performed the task on a Dell Latitude D630 laptop (Intel Core Dual 800MHz, 14.100) and they were led to believe that they were connected through the internet with an interaction partner. Data collection was conducted by two male experimenters. One served as the experimenter, while the other one served as an interaction partner for the current task. This confederate was introduced as player B, without providing any additional information on his background. The two laptops were separated by a screen that prevented the participant and the confederate from seeing each other. Contact between the confederate and the participant was restricted to in-game interactions. Participants were explicitly instructed that they would not meet again after the test session and would not be able to interact with each other outside the task. The task was programmed in E-Prime 2.0 (Psychology Software Tools, 2015).

In the task two players, playing as investor and trustee, interacted with each other for 20 rounds. Each round the investor got endowed with 20 monetary units (MU) and, as the first-mover, had to send any amount from his endowment to the trustee. This investment got automatically tripled by the experimenter and was sent to the trustee. The trustee, as the second-mover, had to make a back-transfer decision that could be any amount between 0 and the tripled investment. At the end of the round both players were presented with the MU they earned in this round. The profits from each round were added up, exchanged to Euros (exchange rate: 1 MU = 2 Eurocent), and paid out at the end of the game. The exact timing of a trial is displayed in Figure 1. Participants always played the role of trustee and were paired throughout the task with the same investor. Investments were preprogrammed (see supplements for details) and resembled a tit-for-tat strategy based on investment decisions of healthy participants in King-Casas et al. (2008), and King-Casas et al. (2005). The repeated trust game task was always administered at the end of the task battery (consisting of a dictator game, an ultimatum game, and a promise-deception task), and results of the other three tasks will be reported elsewhere.

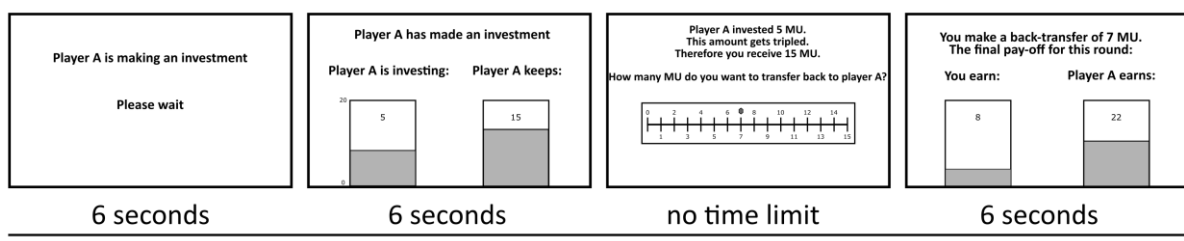


Figure 1. Time course of a single interaction in the repeated trust game including the translated screens (from German) presented to the participants.

SKIN CONDUCTANCE RESPONSES (SCR)

SCR was measured with an 8-channel bioamplifier (Mobi8-BP; TMSI B. V., Enschede, The Netherlands) with a 24 bit analogue-to-digital (A/D) conversion rate. A custom-specific skin conductance sensor, which consisted of two flat Ag/AgCl electrodes, was used to obtain unfiltered raw skin conductance data. Before applying the electrodes, participants had to wash their hands with curd soap to ensure similar skin condition at the beginning of the measurement across participants. Electrodes were applied to the distal phalanges of the index and middle finger of the non-dominant hand. Portilab 2.0 software was used to acquire time-logged skin conductance responses (SCR). The data was preprocessed in Matlab (version 7.9.0) via the Ledalab toolbox (V3.2.3). Raw skin conductance data was first filtered with a 1 Hz 4th order IIR Butterworth low-pass filter, and then down-sampled from 1024 Hz to 10.24 Hz. Phasic skin conductance responses, depicting the average magnitude within a predefined time window, were extracted with a continuous decomposition analysis (CDA; Benedek & Kaernbach, 2010) using a threshold of 0.1 μ S.

We measured SCRs at three time points during each trial, each with a duration of 6000 ms. *SCR investment* is measured from the onset of the investor's investment decision screen. Statistics are not reported, as they are out of scope of the present study. *SCR anticipation* is measured within six seconds before the participant made his back-transfer decision. In case latencies from *SCR investment* and *SCR anticipation* overlapped for more than 3000 ms (i.e. when the participant made a back-transfer decision quicker than 3000 ms) we excluded these trials from analyses. In the offender sample these were 9.40% (47 out of 500 trials) of the trials. In the non-delinquent group these were 6.30% (29 out of 460 trials) of the trials. These were 8.13% of the other-advantageous decision trials, 6.10% of the fair decision trials, and 8.81% of the self-advantageous decisions trials. Finally, we also measured *SCR outcome* from the onset of the outcome screen. We corrected the three extracted SCR measures for non-normal distribution with a log-transformation and performed an outlier correction based on boxplots, deleting datapoints that lie outside 1.5 times the interquartile range.

CODE AVAILABILITY

For research transparency all our analysis scripts are published online (Rosenberger, 2019). We are not able to publish our data due to confidentiality concerns related to the highly specific prisoner sample.

STATISTICAL ANALYSES

We analysed the raw data with (generalized) linear mixed models with orthogonal contrasts and always included the highest order within-participant predictor as random slope (Barr, Levy,

Scheepers, & Tily, 2013), and participant as random intercept. We used return-ratios ($= \frac{\text{back-transfer}}{\text{investment}}$) instead of the raw back-transfers in the models for an easier interpretation of the effects, since raw back-transfers cannot be easily interpreted as high or low, as they always depend on the investment size. Return-ratios < 1 reflect back-transfers that were smaller than the investments, and return-ratios > 1 reflect back-transfers that were higher than the investments. Since we were specifically interested in fairness violations, we categorized return-ratios into fair (back-transfer amount resulted in an equal share of overall profits for investor and participant in a given round, i.e. profit investor = participant), and unfair (i.e., profit investor $\neq$ participant) decisions. Unfair decisions were further split up into self-advantageous (profit participant $>$ investor) and other-advantageous (profit participant $<$ investor) decisions. For all behavioural and arousal analyses we first compared offenders and non-offenders, and in case of a group difference we repeated the same analysis with only the offender sample to test for effects depending on Factor 1 and Factor 2 scores of the PCL-R. In the group comparisons we included ITS, SPM, and BIS 11 scores as covariates of no-interest to control for variations in these scores between the groups. All tests were two-sided with an alpha of 0.05.

RESULTS

We hypothesized that offenders scoring high on Factor 1 would adhere less to the fairness norm than non-offenders. To test this we first compared odds-ratios of choosing a fair rather than an unfair decision between the two groups, and second analysed the association between these odds-ratios with PCL-R Factor 1 and 2 scores in the offender sample. The odds-ratio (= OR) for choosing a fair rather than an unfair decision was lower in the offender group (OR = 0.10, 95% CI = [0.06, 0.16]) than in the non-offenders (OR = 0.24, 95% CI = [0.15, 0.39]), $\chi^2(1) = 6.54$, $p = .01$, see. The lower the OR, the lower the likelihood is to make a fair back-transfer. Offenders adhered to the fairness norm (and thus behaved inequity averse) only in 12.6% of the trials, compared to 21.96% of the trials in the non-offenders. Within the offender group, Factor 2 was negatively associated with ORs for choosing a fair decision, $\chi^2(1) = 6.92$, $p = .01$, OR = 0.88 (95% CI = [0.80, 0.97]), but not Factor 1, $\chi^2(1) = 0.53$, $p = .47$, OR = 0.95 (95% CI = [0.82, 1.10]). Thus higher Factor 2 scores were associated with a lower likelihood to make a fair back-transfer. Offenders scoring high on Factor 2 (median split) adhered only in 7.5% of the trials to the fairness norm, while it was 19.17% of the trials in offenders scoring low on Factor 2. This is against our hypothesis, as we expected norm adherence to be negatively associated with Factor 1 in the offenders.

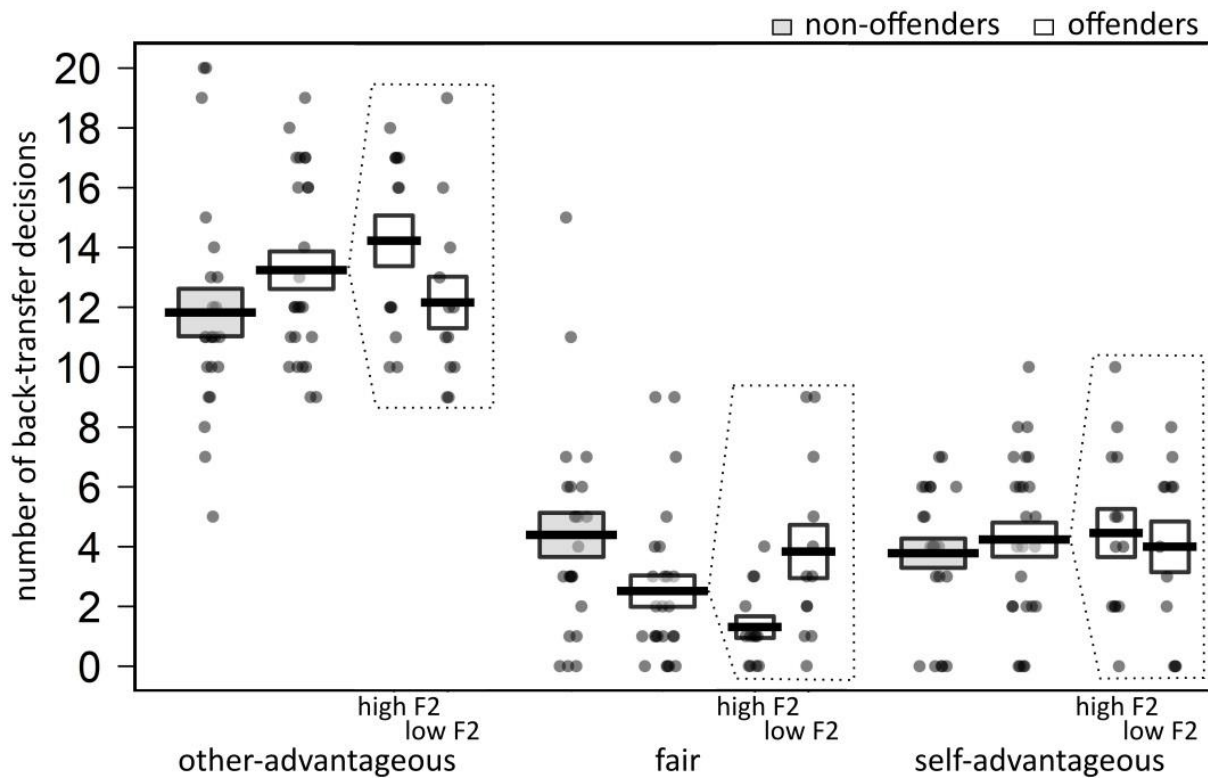


Figure 2. Offenders made less fair back-transfer decisions than non-offenders and this was negatively associated with Factor 2 scores. To visualize effects of Factor 2, we categorized offenders into high scoring (high F2) and low scoring (low F2) participants based on median split. Dots represent number of back-transfer decision for each decision type per each participant. Lines represent mean values per group, and boxes their standard errors.

We further hypothesized that the offenders scoring high on Factor 1 would adhere less to the fairness norm in order to maximize their own profits, thus using a self-advantageous decision more often than a fair decision compared to the non-offenders. To test this we first compared odds-ratios of choosing self-advantageous rather than fair decision between groups, and second analysed the association between these odds-ratios with PCL-R Factor 1 and 2 scores in the offender sample. The OR for choosing a self-advantageous rather than a fair decision was higher in the offenders (OR = 2.14, 95% CI = [1.16, 3.96]) than in the non-offenders (OR = 0.80, 95% CI = [0.44, 1.47]), $\chi^2(1) = 4.50$, $p = .03$. Within the offender group, Factor 2 was positively associated with the likelihood of using a self-advantageous rather than a fair decision, $\chi^2(1) = 4.47$, $p = .03$, OR = 1.19 (95% CI = [1.01, 1.40]), but not Factor 1, $\chi^2(1) = 0.29$, $p = .59$, OR = 1.07 (95% CI = [0.84, 1.34]). Thus again against our hypothesis, increased self-advantageous decisions were not associated with Factor 1, but rather with Factor 2, the lifestyle/ antisocial subscale.

To further characterize the offenders back-transfer behaviour, we also tested whether there were group differences when contrasting the ORs for choosing self-advantageous and other-advantageous decisions, as well as other-advantageous and fair decisions. The ORs for choosing a self-advantageous rather than an other-advantageous decision were not different between the groups, $\chi^2(1) = 0.04$, $p = .83$. However, the ORs for choosing an other-advantageous decision compared to a fair decision differed between the groups, $\chi^2(1) = 6.16$, $p = .01$. The offenders had a higher OR (7.59, 95% CI = [4.59, 12.56]) than the non-offenders (OR = 3.07, 95% CI = [1.90, 4.95]). We followed this up with an analysis of the association between the ORs for choosing an other-advantageous rather than a fair decision and PCL-R Factor 1 and 2 scores in the offender sample. Within the offender group the ORs were positively associated with Factor 2, $\chi^2(1) = 7.05$, $p = .01$, OR = 1.14 (95% CI = [1.03, 1.25]), but not with Factor 1, $\chi^2(1) = 0.37$, $p = .54$, OR = 1.05 (95% CI = [0.91, 1.21]). Thus offenders were not only more likely to make a self-advantageous, but also an other-advantageous rather than a fair decision, and both were positively associated with Factor 2. However, since the ORs of choosing a self-advantageous vs. other-advantageous decision did not differ between the groups, the OR differences in self-advantageous and other-advantageous decisions seem to be driven by lower fair decision likelihoods in the offenders, and not increased self-advantageous or other-advantageous decision likelihoods.

Part of being profit-maximizing is that the actual back-transfers during a self-advantageous decision are as low as possible. Therefore, we hypothesized that the self-advantageous return-ratios are negatively associated with Factor 1 in the offenders. We tested this by first comparing return-ratios between the two groups, the decision types, and their interaction, and second by analysing the association between the return-ratios and PCL-R Factor 1 and 2 scores, the back-transfer decision types, and their interactions in the offender sample. The return-ratios differed significantly for the three decision types (main effect back-transfer decision: $\chi^2(2) = 12.61$, $p < .01$), but not for the groups ($\chi^2(1) < 0.001$, $p = .98$). Importantly, however, there was a significant back-transfer decision x group interaction, $\chi^2(2) = 8.84$, $p = .01$, see Figure 3. Decomposing this interaction, we found that the return-ratios did not differ overall for back-transfer decisions between the groups (all $p > .1$), and also not in the non-offenders (all $t < 0.81$, $p > .90$). In the offenders, however, return-ratios were lower for self-advantageous decisions than for other-advantageous decisions ($t(43.32) = 4.60$, $p < .01$), and also compared to fair decisions ($t(47.24) = 2.82$, $p = .04$). Return-ratios did not differ between other-advantageous and fair decisions ($t(47.46) = 2.37$, $p = .11$). Within the offender group, there was a significant back-transfer decision x Factor 2 interaction, $\chi^2(2) = 8.35$, $p = .02$, but no back-transfer decision x Factor 1 interaction, $\chi^2(2) = 0.97$, $p = .61$. Decomposing the back-transfer

decision x Factor 2 interaction, we found that only self-advantageous return-ratios were negatively associated with Factor 2, $b = -0.04$ (95% CI = [-0.07, -0.02]), $p < .01$, but neither the other-advantageous ($b = 0.03$, 95% CI = [-0.03, 0.09]), nor the fair return-ratios ($b = -0.03$, 95% CI = [-0.08, 0.03]) were associated with Factor 2, both $p > .31$. Thus, the offenders' lower self-advantageous return-ratios were negatively associated with Factor 2 scores.

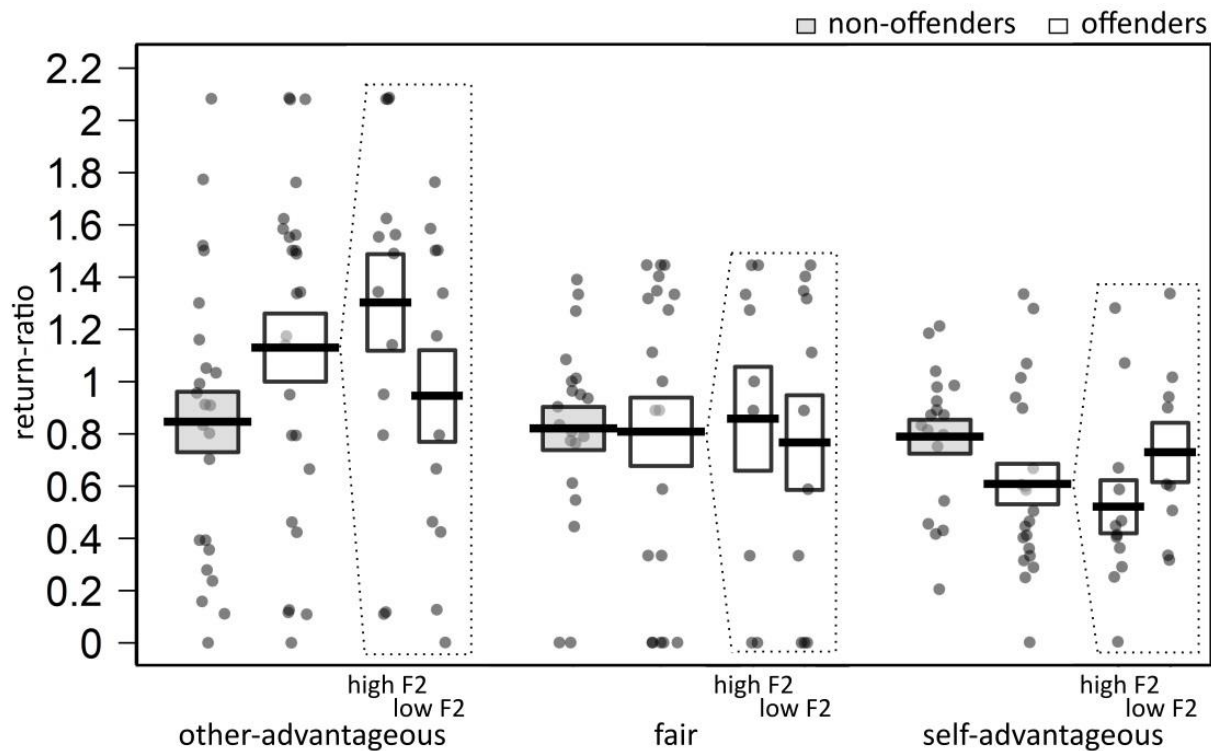


Figure 3. Self-advantageous return-ratios were lower than fair and other-advantageous return-ratios in the offenders, and this was negatively associated with Factor 2 scores. To visualize effects of Factor 2, we categorized offenders into high scoring (high F2) and low scoring (low F2) participants based on median split. Dots represent average return-ratio for every participant, per decision. Lines represent mean values per group, and boxes their standard errors.

Since the offenders displayed only mixed profit-maximizing behaviour, they did not earn more money than the non-offenders at the end of the game, $F(1, 46) = 0.42$, $p = .52$ ($m_{\text{offenders}} = 357.12$ MU (SD = 52.84) vs $m_{\text{non-offenders}} = 347.61$ MU (SD = 48.84)). Total earnings were also neither associated with Factor 1 ($F(1, 22) = 1.24$, $p = .28$), nor with Factor 2 ($F(1, 22) = 1.08$, $p = .31$).

Our final analysis addressed our hypothesis of increased SCRs in anticipation and during receipt of self-advantageous decisions in offenders scoring high on Factor 1. We tested this by first comparing SCR anticipation between the groups, the back-transfer decision types, and their

interaction, and second by analysing the association between SCR anticipation and PCL-R Factor 1 and 2 scores, the back-transfer decision types, as well as their interactions. The same analyses were performed with SCR outcome as dependent variable. First, for SCR anticipation there was a main effect of group ($\chi^2(1) = 4.13, p = .04$), with higher SCR in offenders than non-offenders, but no main effect for back-transfer decision ($\chi^2(2) = 0.59, p = .75$). Importantly, there was a significant back-transfer decision $\times$ group interaction, $\chi^2(2) = 10.79, p < .01$. Anticipatory SCR to self-advantageous decisions were higher in the offenders than in the non-offenders, $t(45.40) = 2.35, p = .05$ ($b = -0.94$, 95% CI = [-1.71, -0.17]). Surprisingly, anticipatory SCR for a fair decision were also marginally larger in the offenders than in the non-offenders, $t(44.91) = 2.26, p = .06$ ($b = -0.94$, 95% CI = [-1.75, -0.14]). Anticipatory SCR for an other-advantageous decision did not differ between the groups, $t(43.17) = 0.90, p = .58$ ($b = -0.33$, 95% CI = [-1.03, 0.38]). Within the offender group there were no anticipatory SCR differences depending on Factor 1 ($\chi^2(2) = 0.27, p = .87$), or Factor 2 ($\chi^2(2) = 1.51, p = .47$), but there was a main effect of back-transfer decision, $\chi^2(2) = 7.80, p = .02$, mirroring the decomposed interaction pattern described above. Thus higher anticipatory arousal to self-advantageous decisions in offenders were not associated with psychopathy scores.

SCR outcome neither differed for back-transfer decisions ($\chi^2(2) = 1.86, p = .39$), nor for group ($\chi^2(1) = 1.51, p = .22$), but there was a significant back-transfer decision $\times$ group interaction ($\chi^2(2) = 7.34, p = .03$). However, post-hoc t-tests did not reveal any significant group differences for the three back-transfer decisions. There was only a trend for higher SCR outcome to self-advantageous decisions in the offenders than in the non-offenders, $t(46.35) = 1.93, p = .11$ (all other $t < 0.89$). Within the offender group there were also no SCR outcome differences for the three back-transfer decisions associated with Factor 1 ($\chi^2(2) = 0.36, p = .84$), or Factor 2 ($\chi^2(2) = 1.05, p = .59$). However, there was a main effect of back-transfer decision, $\chi^2(2) = 6.35, p = .04$, mirroring the decomposed interaction pattern described above. Thus, SCR outcome did not decisively reveal increased arousal to self-advantageous decisions in offenders.

DISCUSSION

In the present study, we investigated whether the personal/affective dimension of psychopathy (Factor 1) is associated with fairness norm violations during a repeated trust game, and whether they prefer to maximize their own profit because of increased reward sensitivity (Geurts et al., 2016; Pujara et al., 2014). Offenders indeed violated fairness norms more frequently and were more likely to make self-advantageous decisions with lower return-ratios than non-offenders. However, these findings were not positively associated with Factor 1, but rather with Factor 2 scores in the offenders.

Factor 2 comprises the lifestyle (impulsive and irresponsible behaviour) and antisocial dimension of the PCL-R (Hare & Neumann, 2008). Our hypothesis of hyper-responsive arousal to reward in the psychopathic offenders was not supported, as there was only a general increase in anticipation of self-advantageous decisions for all offenders. Surprisingly, the norm-violating behaviour in the offenders scoring high on Factor 2 did not lead to different total earnings compared to the non-offenders.

Offenders scoring high on Factor 2, hence called anti-social offenders, of the PCL-R are characterized as being easily bored, impulsive, and having a long history of antisocial behavior even before any conviction (Hare, 2003). In prison, they break rules and are non-compliant during treatment (Hildebrand & de Ruiter, 2012), and are also more likely to commit violent crimes again in the future than offenders scoring high on Factor 1 of the PCL-R (Walters, 2003). Our results reinforce that extra attention should be paid to the treatment of these anti-social offenders. We not only provide empirical support to the norm violations that forensic clinicians encounter frequently in their treatment programs (Walters, 2003), but they can also be a starting-point for specialized treatment. It is possible that the norm violations in our sample are related to deficits in spontaneous empathizing abilities (Meffert, Gazzola, den Boer, Bartels, & Keyzers, 2013). Research by Mayer, Jusyte, Klimecki-Lenz, and Schonenberg (2018) suggests that these abilities might be trainable, since violent offenders increased their monetary distributions in a dictator game after an empathy enhancement intervention. Future research should establish whether anti-social offenders would decrease their norm violations with a treatment program that increases their empathic concerns.

We were unable to demonstrate decisively why anti-social offenders violate fairness norms, as our profit-maximization hypothesis was only partly supported. The differences in the decision likelihoods of the three back-transfer options were mainly driven by a decreased number of fair decisions, and were not driven by an increased number of self-advantageous decision in the anti-social offenders. Additionally, even though the offenders' anticipatory arousal was reflective of hyper-responsiveness to rewards, this was not related to any psychopathy measures. Only the low self-advantageous return-ratios in the anti-social offenders clearly supported the profit-maximizing hypothesis. Therefore we conclude that, while relevant, profit-maximization might not be the main driver of anti-social offenders' behavior in situations with repeated, anonymous interactions. The current experimental situation provided only limited external validity and subjects were only able to gain a comparably low amount of money. Additional research is needed to determine the translatability of our findings to criminal behavior in anti-social offenders.

It is conceivable that the anti-social offenders' norm-violations were motivated by a different behavioural strategy, which we were unable to detect with our paradigm. Indeed, their most frequent back-transfer decision was other-advantageous, which suggests that they were prosocial towards their interaction partner. However, according to current theorizing, psychopathic offenders lack many of the prerequisites that induce a concern for prosociality and would facilitate behaviour that benefits others (Barclay & Van Vugt, 2014). They show deficits in affective empathy (Blair, 2005; Decety, Skelly, & Kiehl, 2013; Keyesers & Gazzola, 2014; Meffert et al., 2013; Pfabigan et al., 2015), display little shame and guilt (Tangney, Stuewig, Mashek, & Hastings, 2011), and make deviant moral decisions (Koenigs, Kruepke, Zeier, & Newman, 2012). Therefore, we suggest that their frequent other-advantageous decisions were not based on other-related concern, but were rather part of a manipulative strategy that benefitted them in the long run. Indeed, in the literature a manipulative nature is widely regarded as a core feature of psychopathy (Mealey, 1995). Even though we were unable to measure a manipulative strategy in our task, the intention of the anti-social offenders might have still been to manipulate the investor. After all, their norm violations and low self-advantageous back-transfers did not have any repercussions, as they earned the same amount of money as the non-offenders. Notably, we are not making any assumptions nor do we have data to show whether these strategies operated intentionally and explicitly. Other decision-making paradigms that investigate non-prosocial cooperative strategies (Barclay & Van Vugt, 2014) will help understand psychopathic offenders' motivation and strategies during repeated interactions. The anti-social offenders' other-advantageous behaviour in our study might be a case of by-product reciprocity of otherwise selfish intentions. However, this remains to be tested in future studies.

Our arousal findings only partially support Arnett's notion of hyper-responsiveness to rewards in psychopathic offenders (Arnett, 1997). Even though we did find increased anticipatory arousal to self-advantageous decisions in the offenders, this was not associated with any psychopathy measure. Studies on empathy have found similar patterns of arousal discrepancies between offenders and non-offenders, which were not specific to any psychopathy measure (Carmen Pastor, Moltó, Vila, & Lang, 2003; Pfabigan et al., 2015; Seidel et al., 2013; Wahlund, Sorman, Gavazzeni, Fischer, & Kristiansson, 2010). However, the psychopathy scores of our offenders are (similar to other European offender samples) lower than those of American psychopathic offenders. We cannot exclude that with a larger variation between high- and low-scoring psychopathic individuals, arousal patterns might diverge between these groups. We believe that our rather unspecific results are reflective of a general effect of hyper-responsiveness to rewards in violent

offenders, but this needs replication in samples with more extreme psychopathy scores as the ones present in the sample to which we had access to.

We used two different psychopathy measures in our study, since we tested two different cohorts (a community and an offender sample). As a result we could not directly compare how the degree of psychopathy differently affected the normative behaviour in the offenders and the non-offenders. Our study was also not designed to test this, as we only recruited individuals scoring low on psychopathy and with little inter-subject variance in the community sample. Nonetheless future studies should directly compare decision-making behaviour in psychopathic individuals from a community sample with an offender sample, as the decision-making literature points to vastly different behaviour between them. Non-delinquent psychopathic individuals only rarely abide to norms (Berg, Lilienfeld, & Waldman, 2013; Gervais, Kline, Ludmer, George, & Manson, 2013; Gillespie, Mitchell, Johnson, Dawson, & Beech, 2013; Rilling et al., 2007) and punish those who deviate from social norms to a much lesser degree than incarcerated psychopaths (Berg et al., 2013; Masui, Iriguchi, & Ura, 2014; Osumi & Ohira, 2010). This might be due to a distorted perception of offer fairness, as participants from a community sample scoring high on psychopathy rated unfair offers as more fair than those scoring low on psychopathy (Vieira et al., 2014). In a recent study by Gong and colleagues (Gong, Brazil, Chang, & Sanfey, 2019), non-delinquent psychopathic individuals behaved much more profit-maximizing than the psychopathic offenders in our study. The non-delinquent psychopathic individuals ignored their interaction partners' expectations and felt less guilty when they violate them than non-delinquent subjects scoring low on psychopathy (Gong et al., 2019). This highlights the necessity that future work compares participants scoring high on psychopathy from a prison cohort with those from a community cohort to establish similarities and differences in norm adherence behaviour during repeated interaction decision-making paradigms.

With the present work, we aim to set a first step in characterizing anti-social offenders' behaviour during repeated sequential decision-making interactions, by establishing that they do not adhere to social norms and that this is only minorly driven by profit-maximizing back-transfer behaviour. Repeated decision-making paradigms enable psychopathic offenders to display complex, strategic behaviour which creates space for researchers to study how and why psychopathic offenders manipulate their interaction partners. We propose that future studies employ non-prosocial cooperative decision-making paradigms to specifically target manipulative strategies in psychopathic offenders. Additionally, we suggest treatment programs targeting empathic abilities particularly in anti-social offenders, since even in standardized experimental settings anti-social offenders violate norms which might extrapolate to criminal behaviour in real life. Overall, our study

outlines a promising new path for advancing our understanding of the norm violating nature of (anti-social) psychopathic offenders.

ACKNOWLEDGEMENT

We thank Daniel Prattinger for help with data collection and Jack van Honk for feedback on the manuscript. The work in this paper was supported by the Vienna Science and Technology Fund (WWTF VRG13-007). The University of Vienna funded participants' financial remuneration. In loving memory of Christoph Eisenegger, who tragically died in February 2017.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPLEMENTARY METHODS

REPEATED TRUST GAME PRE-PROGRAMMED INVESTMENT DECISIONS

In general, investments in the following round increased when the participant made a high back-transfer. The investments changed in the following ways: for extreme negative back-transfers, where the participant did not make a back-transfer, investments in the next round were low and ranged from 2 to 4 MU. In case that the participants did not make a back-transfer in the entire task, the investments in the last two rounds increased to 9 and 12 MU to coax the participant into making a back-transfer. However neither of the participants displayed such behaviour. For negative back-transfers, where the participant's back-transfer was lower than the investment, investments in the next round were low as well and ranged from 2 to 5 MU. To make the investment behaviour more realistic (and less static) we randomly interspersed these trials with noisy trials where investments ranged from 9 MU to 14 MU. For positive back-transfers, where the back-transfer was larger than the investment, investments in the next round were also positive and ranged from 6 and 10 MU. Noise trials were randomly interspersed with investments of either 4 MU, 5 MU, 15 MU, or 16 MU. Extremely positive back-transfers, which were more than twice the size of the investment, were followed by extremely positive investments in the next round, with a range between 14 and 18 MU. Noise trials comprised investments of either 5 MU, 9 MU, or 10 MU.

QUESTIONNAIRES

PSYCHOPATHY ASSESSMENT OFFENDERS

The offender sample's degree of psychopathy was assessed with the Psychopathy Checklist – Revised (Hare, 2003). The scale comprises a 20-item semi-structured interview consisting of a questionnaire (items are scored on a scale from 0 (*absent*) to 2 (*present*)) and a review of the offenders' records. Total scores can range from 0 to 40. Our samples average total score is comparable to other Austrian psychopathic offender samples ($N = 1046$, $m = 20.3$, $SD = 7.6$) (Neumann, Hare, & Pardini, 2015). Besides the total score, we also assessed the influence of Factor 1 and Factor 2 subscales on norm adherence and arousal in our study. Factor 1 comprises items from the interpersonal and affective subscale, whereas Factor 2 captures items from the lifestyle and antisocial subscales (Hare, 2003).

PSYCHOPATHY ASSESSMENT NON-OFFENDERS

In the non-offenders we assessed psychopathy with the German version of the Psychopathic Personality Inventory – revised (PPI-R, Alpers & Eisenbarth, 2008). The self-administered scale contains 154 items, which were scored on a 4-point scale ranging from *false* to *true*. Each item corresponds to one of nine scales, which are grouped into a three factor model (Lilienfeld & Widows,

2005). Our non-offender group's average total score is lower than in a German male student sample ($N = 148$, $m = 325.42$, $SD = 24.92$) and a German male offender sample ($N = 57$, $m = 338.71$, $SD = 25.68$) (Eisenbarth & Alpers, 2007). Therefore, we can conclude that our non-offenders did not display a high degree of psychopathy.

IMPULSIVENESS ASSESSMENT

Psychopathy is tightly interconnected with impulse-control problems (see Poythress and Hall (2011) for a discussion). Both the PCL-R and the PPI-R questionnaires have items that specifically assess impulsivity, and in both scales Factor 2 specifically captures the erratic and reactive side of psychopathy. To control for differences in normative behaviour as a result of impulsivity differences in the samples, we included the German version of the 11th revision of the Barratt Impulsiveness Scale (BIS-11; Preuss et al., 2008) in our analyses as a covariate of no-interest. It is a self-report instrument comprising 30 items with answers ranging from *rarely/never* to *almost always/always*. High scores signify high impulsivity. Each item corresponds to one of three factors: attentional, motor, and non-planning impulsivity.

INTELLIGENCE ASSESSMENT

We matched the non-offenders with the offender samples' intelligence scores. For this we used Raven's Standard Progressive Matrices (SPM; Raven, 1969). This is a Rasch homogenous, language-free intelligence test for the assessment of non-verbal reasoning. The long version of the SPM is divided into five sets, each including 12 items. For each item a participant has to determine which of several alternative parts fits into a pattern. The offenders at the correctional facility were all routinely assessed with a short version of the SPM, which only includes 32 items.

TRUST ASSESSMENT

The German version of the Interpersonal Trust Scale (ITS; Amelang et al., 1984) is a unipolar scale consisting of 17 items measuring interpersonal trust. Items are scored on a 5-point Likert scale ranging from *strongly disagree* to *strongly agree*. High scores signify high interpersonal trust.

STATISTICAL ANALYSES

Data were analysed in R (version 3.4.2 (R Core Team, 2017)) with the lme4 package version 1.1.14 (Bates, Mächler, Bolker, & Walker, 2015). The reported chi-square tests are calculated with type 3 sum of squares with the car package version 3.0 (Fox & Weisberg, 2011). Post-hoc tests for interactions only including factorial variables were performed with the emmeans package version 1.2.3 (Lenth, 2018) and p-values for multiple comparisons were Tukey-adjusted. The reported odds ratios serve as effect sizes. Post-hoc comparisons for interactions including continuous variables

were performed with the SjPlot package version 2.4.1 (Lüdecke, 2018) where p-value estimations were based on Wald z-statistics. Here, the reported betas serve as effect sizes. All the questionnaire scores were mean-centred. Including these predictors increased the fit of the models (as measured with the BIC). In the cases where the investor made investments lower than 6 MU the back-transfer decision options were limited, as no fair (with investments ≤ 4 MU) or self-advantageous (with investments ≤ 5 MU) decisions were possible. Our results do not change when excluding investments smaller than 6 (see (Rosenberger, 2019) for results).

Post-hoc power analyses with Westfall's online calculator (Westfall, 2016) for our primary group comparison of fair over unfair back-transfer decisions resulted in a power of 0.815. Analyses are based on a 2-group between participants design with random factor for participants, with a Cohen's d of 0.51 (transformed log odds-ratio of 0.931 of group effect), 20 observations per participant, 24 participants per group, and the pre-defined variances.

REFERENCES

- Alpers, G. W., & Eisenbarth, H. (2008). *Psychopathic Personality Inventory-Revised (PPI-R), German Version. Manual*. Göttingen, Germany: Hogrefe.
- Amelang, M., Gold, A., & Kuelbel, E. (1984). Ueber einige Erfahrungen mit einer deutschsprachigen Skala zur Erfassung zwischenmenschlichen Vertrauens (Interpersonal Trust). *Diagnostica*, 30, 198 - 215.
- Arnett, P. (1997). Autonomic responsivity in psychopaths: a critical review and theoretical proposal. *Clinical Psychology Review*, 17(8), 903 - 936.
- Arnett, P., & Newman, J. P. (2000). Gray's three-arousal model: an empirical investigation. *Personality and Individual Differences*, 28, 1171-1189.
- Association, W. M. (2001). World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Bulletin of the World Health Organization*, 79(4), 373.
- Barclay, P., & Van Vugt, M. (2014). The Evolutionary Psychology of Human Pro-sociality: Adaptations, Byproducts, and Mistakes. In D. A. Schroeder & W. G. Graziano (Eds.), *The Oxford Handbook of Prosocial Behavior* (pp. 37 - 60). Oxford: Oxford University Press.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J Mem Lang*, 68(3). doi:10.1016/j.jml.2012.11.001
- Bates, D. M., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1 - 48.
- Benedek, M., & Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, 190(1), 80-91. doi:10.1016/j.jneumeth.2010.04.028
- Berg, J. M., Lilienfeld, S. O., & Waldman, I. D. (2013). Bargaining with the devil: Using economic decision-making tasks to examine the heterogeneity of psychopathic traits. *Journal of Research in Personality*, 47(5), 472-482. doi:10.1016/j.jrp.2013.04.003
- Bicchieri, C. (2005). *The Grammar of Society: The Nature and Dynamics of Social Norms*. New York: Cambridge University Press.
- Blair, R. J. (2005). Responding to the emotions of others: dissociating forms of empathy through the study of typical and psychiatric populations. *Conscious Cogn*, 14(4), 698-718. doi:10.1016/j.concog.2005.06.004
- Carmen Pastor, M., Moltó, J., Vila, J., & Lang, P. J. (2003). Startle reflex modulation, affective ratings and autonomic reactivity in incarcerated Spanish psychopaths. *Psychophysiology*, 40(6), 934-938.
- Decety, J., Skelly, L. R., & Kiehl, K. A. (2013). Brain response to empathy-eliciting scenarios involving pain in incarcerated individuals with psychopathy. *JAMA Psychiatry*, 70(6), 638-645. doi:10.1001/jamapsychiatry.2013.27
- Demal, U. (1999). SKIDPIT-light Screeningbogen. Vienna, Austria: Universität Wien.
- Eisenbarth, H., & Alpers, G. W. (2007). Validierung der deutschen Übersetzung des Psychopathy Personality Inventory (PPI). *Zeitschrift für Klinische Psychologie und Psychotherapie*, 36(3), 216-224. doi:10.1026/1616-3443.36.3.216

- Eisenegger, C., Pedroni, A., Rieskamp, J., Zehnder, C., Ebstein, R., Fehr, E., & Knoch, D. (2013). DAT1 polymorphism determines L-DOPA effects on learning about others' prosociality. *PLoS One*, 8(7), e67820. doi:10.1371/journal.pone.0067820
- Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868), 137 - 140. doi:10.1038/415137a
- Fehr, E., & Schmidt, K. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, 114(3), 817 - 868.
- Fox, J., & Weisberg, S. (2011). *An {R} Companion to Applied Regression*. Thousand Oaks CA: Sage.
- Fydrich, T., Renneberg, B., Schmitz, B., & Wittchen, H. U. (1997). *SKID-II. Strukturiertes Klinisches Interview für DSM-IV. Achse II: Persönlichkeitsstörungen. Interviewheft*. Göttingen, Germany: Hogrefe.
- Gao, Y., & Raine, A. (2010). Successful and unsuccessful psychopaths: a neurobiological model. *Behavioral Sciences and the Law*, 28(2), 194-210. doi:10.1002/bsl.924
- Gervais, M. M., Kline, M., Ludmer, M., George, R., & Manson, J. H. (2013). The strategy of psychopathy: primary psychopathic traits predict defection on low-value relationships. *Proc Biol Sci*, 280(1757), 20122773. doi:10.1098/rspb.2012.2773
- Geurts, D. E., von Borries, K., Volman, I., Bulten, B. H., Cools, R., & Verkes, R. J. (2016). Neural connectivity during reward expectation dissociates psychopathic criminals from non-criminal individuals with high impulsive/antisocial psychopathic traits. *Soc Cogn Affect Neurosci*, 11(8), 1326-1334. doi:10.1093/scan/nsw040
- Gillespie, S. M., Mitchell, I. J., Johnson, I., Dawson, E., & Beech, A. R. (2013). Exaggerated intergroup bias in economical decision making games: differential effects of primary and secondary psychopathic traits. *PLoS One*, 8(8), e69565. doi:10.1371/journal.pone.0069565
- Glenn, A. L., Raine, A., Yaralian, P. S., & Yang, Y. (2010). Increased volume of the striatum in psychopathic individuals. *Biological Psychiatry*, 67(1), 52-58. doi:10.1016/j.biopsych.2009.06.018
- Gong, X., Brazil, I. A., Chang, L. J., & Sanfey, A. G. (2019). Psychopathic traits are related to diminished guilt aversion and reduced trustworthiness during social decision-making. *Sci Rep*, 9(1), 7307. doi:10.1038/s41598-019-43727-0
- Hare, R. (2003). *Manual for the Revised Psychopathy Manual for the Revised Psychopathy Checklist (2nd edn)*. Retrieved from Multi-Health Systems:
- Hare, R., & Neumann, C. S. (2008). Psychopathy as a clinical and empirical construct. *Annu Rev Clin Psychol*, 4, 217-246. doi:10.1146/annurev.clinpsy.3.022806.091452
- Hildebrand, M., & de Ruiter, C. (2012). Psychopathic traits and change on indicators of dynamic risk factors during inpatient forensic psychiatric treatment. *Int J Law Psychiatry*, 35(4), 276-288. doi:10.1016/j.ijlp.2012.04.001
- Keysers, C., & Gazzola, V. (2014). Dissociating the ability and propensity for empathy. *Trends Cogn Sci*, 18(4), 163-166. doi:10.1016/j.tics.2013.12.011
- King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science*, 321(5890), 806-810. doi:10.1126/science.1156902
- King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. *Science*, 308(5718), 78-83. doi:10.1126/science.1108062

- Koenigs, M., Kruepke, M., & Newman, J. P. (2010). Economic decision-making in psychopathy: a comparison with ventromedial prefrontal lesion patients. *Neuropsychologia*, 48(7), 2198-2204. doi:10.1016/j.neuropsychologia.2010.04.012
- Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2012). Utilitarian moral judgment in psychopathy. *Soc Cogn Affect Neurosci*, 7(6), 708-714. doi:10.1093/scan/nsr048
- Lenth, R. V. (2018). emmeans: Estimated Marginal Means, aka Least-Squares Means. (Version 1.2.3) [R package]. Retrieved from <https://CRAN.R-project.org/package=emmeans>
- Lilienfeld, S. O., & Widows, M. R. (2005). *Psychopathic Personality Inventory - Revised*. Retrieved from
- Lüdecke, D. (2018). sjPlot: Data Visualization for Statistics in Social Science. Retrieved from <https://CRAN.R-project.org/package=sjPlot>
- Masui, K., Iriguchi, S., & Ura, M. (2014). Assured Rewards Facilitate Non-Intervention in Unfair Situations by High Psychopathy Individuals. *International Journal of Psychological Studies*, 6(4). doi:10.5539/ijps.v6n4p56
- Mayer, S. V., Jusyte, A., Klimecki-Lenz, O. M., & Schonenberg, M. (2018). Empathy and altruistic behavior in antisocial violent offenders with psychopathic traits. *Psychiatry Res*, 269, 625-632. doi:10.1016/j.psychres.2018.08.035
- Mayer, S. V., Rauss, K., Pourtois, G., Jusyte, A., & Schonenberg, M. (2018). Behavioral and electrophysiological responses to fairness norm violations in antisocial offenders. *European Archives of Psychiatry and Clinical Neuroscience*. doi:10.1007/s00406-018-0878-2
- Mealey, L. (1995). The sociobiology of sociopathy: an integrated evolutionary model. *Behavioral and Brain Sciences*, 18, 523-599.
- Meffert, H., Gazzola, V., den Boer, J. A., Bartels, A. A., & Keysers, C. (2013). Reduced spontaneous but relatively normal deliberate vicarious representations in psychopathy. *Brain*, 136(Pt 8), 2550-2562. doi:10.1093/brain/awt190
- Mokros, A., Menner, B., Eisenbarth, H., Alpers, G. W., Lange, K. W., & Osterheider, M. (2008). Diminished cooperativeness of psychopaths in a prisoner's dilemma game yields higher rewards. *Journal of Abnormal Psychology*, 117(2), 406-413. doi:10.1037/0021-843X.117.2.406
- Montague, P. R., & Lohrenz, T. (2007). To detect and correct: norm violations and their enforcement. *Neuron*, 56(1), 14-18. doi:10.1016/j.neuron.2007.09.020
- Neumann, C. S., Hare, R. D., & Pardini, D. A. (2015). Antisociality and the Construct of Psychopathy: Data From Across the Globe. *Journal of Personality*, 83(6), 678-692. doi:10.1111/jopy.12127
- Newman, J. P., Kosson, D. S., & Patterson, C. M. (1992). Delay of gratification in psychopathic and nopsychopathic offenders. *Journal of Abnormal Psychology*, 101(4), 630 - 636.
- Newman, J. P., Patterson, C. M., Howland, E. W., & Nichols, S. L. (1990). Passive avoidance in psychopaths: the effects of reward. *Personality and Individual Differences*, 11(11), 1101 - 1114.
- Osumi, T., & Ohira, H. (2010). The positive side of psychopathy: Emotional detachment in psychopathy and rational decision-making in the ultimatum game. *Personality and Individual Differences*, 49(5), 451-456. doi:10.1016/j.paid.2010.04.016
- Pfabigan, D. M., Seidel, E. M., Wucherer, A. M., Keckeis, K., Derntl, B., & Lamm, C. (2015). Affective Empathy Differs in Male Violent Offenders with High- and Low- Trait

- Psychopathy. *Journal of Personality Disorders*, 29(1), 42 - 61. doi:10.1521/pedi_2014_28_145
- Poythress, N. G., & Hall, J. R. (2011). Psychopathy and impulsivity reconsidered. *Aggression and Violent Behavior*, 16(2), 120-134. doi:10.1016/j.avb.2011.02.003
- Preuss, U. W., Rujescu, D., Giegling, I., Watzke, S., Koller, G., Zetzsche, T., . . . Moller, H. J. (2008). Psychometric evaluation of the German version of the Barratt Impulsiveness Scale. *Nervenarzt*, 79(3), 305-319. doi:10.1007/s00115-007-2360-7
- Psychology Software Tools, I. (2015). E-Prime 2.0. Pittsburgh, PA. Retrieved from <http://www.pstnet.com>
- Pujara, M., Motzkin, J. C., Newman, J. P., Kiehl, K. A., & Koenigs, M. (2014). Neural correlates of reward and loss sensitivity in psychopathy. *Soc Cogn Affect Neurosci*, 9(6), 794-801. doi:10.1093/scan/nst054
- R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
- Radke, S., Brazil, I. A., Scheper, I., Bulten, B. H., & de Bruijn, E. R. (2013). Unfair offers, unfair offenders? Fairness considerations in incarcerated individuals with and without psychopathy. *Front Hum Neurosci*, 7, 406. doi:10.3389/fnhum.2013.00406
- Raven, J. C. (1969). Standard progressive matrices: sets A, B, C, D, and E. London, UK: Lewis.
- Rilling, J. K., Glenn, A. L., Jairam, M. R., Pagnoni, G., Goldsmith, D. R., Elfenbein, H. A., & Lilienfeld, S. O. (2007). Neural correlates of social cooperation and non-cooperation as a function of psychopathy. *Biological Psychiatry*, 61(11), 1260-1271. doi:10.1016/j.biopsych.2006.07.021
- Rosenberger, L. A. (2019). Analyses scripts to manuscript "Fairness norm violations in anti-social psychopathic offenders in a repeated trust game". Retrieved from <https://osf.io/x5sav>
- Scheeff, J., Jusyte, A., & Schonenberg, M. (2018). Investigating social-contextual determinants of cooperation in incarcerated violent offenders. *Sci Rep*, 8(1), 17204. doi:10.1038/s41598-018-35450-z
- Seidel, E. M., Pfabigan, D. M., Keckeis, K., Wucherer, A. M., Jahn, T., Lamm, C., & Derntl, B. (2013). Empathic competencies in violent offenders. *Psychiatry Res*, 210(3), 1168-1175. doi:10.1016/j.psychres.2013.08.027
- Tangney, J. P., Stuewig, J., Mashek, D., & Hastings, M. (2011). Assessing Jail Inmates' Proneness to Shame and Guilt: Feeling Bad About the Behavior or the Self? *Crim Justice Behav*, 38(7), 710-734. doi:10.1177/0093854811405762
- Vieira, J. B., Almeida, P. R., Ferreira-Santos, F., Barbosa, F., Marques-Teixeira, J., & Marsh, A. A. (2014). Distinct neural activation patterns underlie economic decisions in high and low psychopathy scorers. *Soc Cogn Affect Neurosci*, 9(8), 1099-1107. doi:10.1093/scan/nst093
- Wahlund, K., Sorman, K., Gavazzeni, J., Fischer, H., & Kristiansson, M. (2010). Attenuated subjective ratings and skin conductance responses to neutral and negative pictures in non-psychopathic mentally disordered offenders with various diagnoses. *Psychiatry Res*, 180(1), 30-34.
- Walters, G. D. (2003). Predicting institutional adjustment and recidivism with the psychopathy checklist factor scores: a meta-analysis. *Law Hum Behav*, 27(5), 541-558.
- Westfall, J. (2016). PANGEA: Power ANalysis for GEneral Anova designs. *Working paper*.

Young, H. P. (2008). Social Norms. In S. N. Durlauf & L. E. Blume (Eds.), *New Palgrave Dictionary of Economics* (2 ed., pp. 6065-6069). London: Macmillan.

CHAPTER 4 - INTERPERSONAL DISTANCE ADJUSTMENTS AFTER INTERACTIONS WITH A GENEROUS AND SELFISH TRUSTEE DURING A REPEATED TRUST GAME

Original manuscript as published in Journal of Experimental Social Psychology

DOI: doi.org/10.1016/j.jesp.2020.104001

Lisa A. Rosenberger^{a*}, Michael Naef^b, Christoph Eisenegger^a, and Claus Lamm^{a,c*}

a Neuropsychopharmacology and Biopsychology Unit, Department of Basic Psychological Research and Research Methods, Faculty of Psychology, University of Vienna, Austria.

b Experimental Economics Laboratory, Department of Economics, Royal Holloway, University of London, Egham, TW20 0EX, United Kingdom.

c Social, Cognitive and Affective Neuroscience Unit, Department of Basic Psychological Research and Research Methods, Faculty of Psychology, University of Vienna, Austria

* To whom correspondence may be addressed. Email: lisa.anna.rosenberger@univie.ac.at or claus.lamm@univie.ac.at, address: Liebiggasse 5, 1010 Wien, telephone: +43-1427747130, fax: +43-1427747139

© 2020. This manuscript version is made available under the CC-BY-NC-ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

ABSTRACT

How physically close to an individual we choose to stand—our interpersonal distance preference—is thought to closely reflect the trust placed in the individual. However, the exact relationship between interpersonal distance and trust is not well understood and the causal effects of trust on interpersonal distance has not yet been demonstrated and quantified experimentally. Here, we induced trust (vs distrust) in an interaction partner through generous (vs selfish) monetary transactions in a repeated trust game. Subjects' investments in, experienced trust in, and expected back-transfers from the two interaction partners indicated that they learned to distrust more the selfish (vs generous) interaction partner. Further, across two experiments, using a pre-post design in which interpersonal distance preferences were measured before and after the trust game, preferred reflective and reflexive interpersonal distance from the selfish interaction partner increased, whereas interactions with the generous partner did not change these preferences. Importantly, the post-game distance was associated with subjects' experienced trust in the selfish interaction partner, and not their monetary decisions made during the game. These findings are a starting point for our understanding of how interactions with untrusted others can be successful. We speculate that an increased interpersonal distance might be accompanied by an increased psychological distance (e.g. through impersonal communication), which might help prevent a break-down of an interaction with a distrusted partner.

Keywords

Interpersonal distance, trust game, social interaction, belief updating

Introduction

We can intuitively imagine standing further away from someone we trust less (vs more) during a conversation. This preferred distance from another person is a largely automatic behavior (Hall, 1966) that co-determines how successfully we navigate everyday social interactions. Too close a distance feels intrusive and threatening, yet too far a distance feels standoffish and distrusting (Lloyd, 2009). Inappropriate interpersonal distances may also hamper our interactions, because they implicitly signal the relationship type and quality we perceive ourselves having with the person (Hall, 1966), and are closely linked to how much we trust them (Jourard & Friedman, 1970; Skotko & Langmeyer, 1977; Wheelless & Grotz, 1977). Surprisingly, there is little research on the exact nature of the relationship between these two phenomena; To our knowledge it has only been described that interpersonal distance causally affects how much we trust others (Bryan, Perona, & Adolphs, 2012). However, the perhaps more intuitive causal effects of trust on interpersonal distance have not yet been studied. Understanding the directionality of the association between these two concepts is integral for determining how both work alone and in combination to affect social interactions and decision-making. Hence, in two experiments, we tested whether (dis-)trusting another person can cause healthy subjects to adjust their preferred interpersonal distance from them.

The distance we stand from interaction partner influences our trust in them, with uncomfortably close or wide distances detrimental to trust. For example, even though a close distance is associated with enhanced self-disclosure (Jourard & Friedman, 1970; Skotko & Langmeyer, 1977) (an indirect measure of trust (Wheelless & Grotz, 1977)), too close a distance instead decreases trust. When an interaction partner stands within one's personal space (i.e., the area around a person which causes discomfort when intruded; (Hayduk, 1983)), the amount of money entrusted to them and the extent to which they are perceived as trustworthy decreases (Bryan et al., 2012). Running in the opposite causal direction, there is also evidence—albeit somewhat indirect—that trust impacts interpersonal distance. For instance, when observing an interaction where a cartoon agent's actions have a larger (vs smaller) effect on another cartoon agent's outcome, the physical distance between the two is perceived to be smaller (Shao, Yin, Ji, Yang, & Song, 2020). Thus, the quality of the interaction affects the perceived distance between the cartoon agents. Direct positive social experiences, such as joint book reading or Lego play, have been shown to decrease the distance to the interaction partners (Candini et al., 2017; Gessaroli, Santelli, di Pellegrino, & Frassinetti, 2013). Negative experiences, such as an aggressive conversation tone or angry facial expression, on the other hand, increase the distance one prefers between themselves and the interaction partner (Cartaud, Ruggiero, Ott, Iachini, & Coello, 2018; Ruggiero et al., 2017; Vagnoni, Lewis, Tajadura-Jimenez, & Cardini, 2018). Further, manipulating the moral impression of

an interaction partner affects the interpersonal distance one prefers to have from them (Iachini, Pagliaro, & Ruggiero, 2015; Pellencin, Paladino, Herbelin, & Serino, 2018).

To provide a direct test of whether trust impacts interpersonal distance, we made use of the Trust Game (Berg, Dickhaut, & McCabe, 1995). In a typical setup of this game, two people, an investor and a trustee, aim to maximize their profits through monetary interactions. The amount of money the investor sends to the trustee signals the degree of trust in the trustee, who in turn can reciprocate that trust by making a back-transfer that is profitable for both of them. After multiple interactions, the investor learns to trust or distrust the trustee based on whether they reciprocate or exploit the trust placed in them (Chang, Doll, van 't Wout, Frank, & Sanfey, 2010). In this way, trust and distrust can be experimentally induced through interactions with a generous or a selfish trustee (Fett, Gromann, Giampietro, Shergill, & Krabbendam, 2012; Rosenberger et al., 2019). We predicted that, after interacting with them in the repeated trust game, subjects would prefer a smaller interpersonal distance from the generous trustee and larger interpersonal distance from the selfish trustee. In experiment 1, we measured static and reflective interpersonal distance with a paper and pencil task in which subjects indicate explicitly their preferred interpersonal distance from imagined interaction partners (Duke & Nowicki, 1972). This task measures a deliberative expression of comfortable interpersonal distance on a continuum, while holding stimulus appearance and context constant (Duke & Nowicki, 1972). In experiment 2, we measured more dynamic, reflexive interpersonal distance preferences with a digital stop-motion task (Perry, Rubinsten, Peled, & Shamay-Tsoory, 2013) in which subjects view an approaching interaction partner and indicate on the keyboard when they should stop in their approach. This task is an extension of the paper and pencil measure, where the expression of comfortable interpersonal distance is substituted with a reflexive halt response to approaching figures. Even though both measures are similarly sensitive to distance requirements with interaction partners of different gender and age (Iachini et al., 2016), the digital version taps more into automatic approach – avoidance motivational preferences (Cohen, Perry, Mayseless, Kleinmintz, & Shamay-Tsoory, 2018; Perry, Mankuta, & Shamay-Tsoory, 2014) than the paper and pencil version.

METHODS

SUBJECTS

Two experiments were conducted using two independent student samples but involving similar task protocols. In both experiments, subjects first performed the interpersonal distance task, then played the repeated trust game, and then—immediately after the trust game—completed the interpersonal distance task again (see Figure 1). The experiments differed in the implementation of

the interpersonal distance task (paper and pencil version used in experiment 1 and a digital task used in experiment 2). In experiment 1, 54 students participated (34 female; $m_{\text{age}} = 25.02$, $SD = 7.55$, range = 18 - 70) who were recruited through the university study recruitment website. In experiment 2, 55 students participated (35 female; $m_{\text{age}} = 27.43$, $SD = 8.83$, range = 19 - 68) who were all recruited through the same university study recruitment website. The large age range is due to outliers in both samples. Removing the outliers did not change the results (data not shown). Therefore, no subjects were excluded from analyses. Subjects received a flat-fee reimbursement plus the money earned during the repeated trust game (exchange rate points to euro: 1:0.007). We collected data in a group laboratory where two to nine students participated simultaneously. A-priori sample size requirements were based on practical considerations of a minimum of 50 subjects during 7 group test sessions. Data analyses were performed after finishing data collection and thus did not influence the number of subjects. In the lab, dividers between the computers prevented subjects from seeing each other during testing. Verbal or non-verbal communication during testing was not allowed and—subjects were warned—would lead to exclusion from the study. All subjects adhered to this rule and none had to be excluded during the test sessions. Before the start of the experiment, all subjects gave written informed consent. The study was conducted according to the declaration of Helsinki and was approved by the local university ethics committee. The tasks described in the present study were part of a larger test battery, aiming to improve the characterization of trust development between interacting people. The test battery consisted of questionnaires (experiment 1 and 2: general trust, social value orientation; experiment 2 only: mood, behavioural motivation), implicit association tests (experiment 1: sequential priming task; experiment 2: affect misattribution procedure), and a measure of cue-controlled behaviour (only in experiment 2: pavlovian to instrumental transfer task). Results of these tasks are not reported, as these measures address different aspects of trust development than the one described in this paper. All measures, manipulations, and exclusions in the study are disclosed.

TASKS

REPEATED TRUST GAME

The repeated trust game was programmed in z-Tree (version 3.3.7 (Fischbacher, 2007)) that allows for an easy implementation of multiplayer behavioural economic tasks. The task design and implementation was exactly the same in experiment 1 and experiment 2. The script is deposited online (Rosenberger, 2019). Subjects were instructed on paper and completed profit calculation questions as a comprehension check. In addition to the paper instructions, subjects played five practice rounds to familiarize themselves with the layout of the task. After the practice rounds, the experimenter ostensibly made a phone call to a laboratory in another building where the same study

was thought to be taking place. During the phone call they audibly told the “other experimenter” that the subjects were ready to start the task. This belief induction was preceded by an explanation that the interaction partners for the task were seated in the other laboratory. The initial study invitation also stated that the experiments would take place in two separate laboratories simultaneously.

In the repeated trust game two players entrust each other money over several rounds. Subjects always played in the role of investor with two pre-programmed trustees who ostensibly were fellow students participating in the same study in another laboratory. We chose ostensible partners in another room, so that subjects had no specific prior beliefs about them except that they were fellow students in a similar situation. All players were referred to with letters (subjects were player A, trustees were player B and C) and were represented with neutral avatars (see Figure 1A), with either horizontal or vertical stripes allowing to visually distinguish between them. Subjects played 52 rounds of the task, randomly alternating between the two trustees (but never more than two consecutive rounds with the same trustee). Subjects were unaware of the exact length of the task to prevent end-game effects. At the beginning of a round, both players were endowed with 10 points. The subject started the round by making a variable investment (between 1 and 10 points of their endowment) to the trustee, see Figure 1A. This investment got tripled by the experimenter and was then sent to the trustee. Next, the trustee made a variable back-transfer (between 1 and (endowment + 3x investment) points) to the investor. This concluded a round and both players saw how many points they earned in this round.

The back-transfers of the trustees were pre-programmed (Rosenberger et al., 2019). The generous trustee behaved trustworthy by making back-transfers that were either 100%, 150%, or 200% of the investment. At the start of the task, there was an equal probability of these three back-transfers, but the probability of a back-transfer of 200% of the investment increased with 10% every time the subject increased her investment to the generous trustee in consecutive interactions. Thus, generous investments were rewarded with a higher probability of generous back-transfers. The selfish trustee behaved untrustworthy by making back-transfers that were either 100%, 75%, or 50% of the investment. At the start of the task there was an equal probability of these three back-transfers. The probability of a back-transfer of 50% of the investment increased with 10% every time the subject increased their investment in consecutive interactions. Thus the subject’s generous investments got punished with a higher probability of selfish back-transfers. We fixed the back-transfers in the first four rounds to 100% of the investments (irrespective of trustee type) to induce the same prior beliefs about the back-transfers across subjects. The conceptualization of trustworthy and untrustworthy behaviour as respectively generous and selfish back-transfers is based on neuro-

economic definitions of trust and reciprocity. Here, trust is defined as sharing an exclusive source with a second person with the expectation that this will lead to a higher profit than not sharing this source. A trustee behaves trustworthy when they reciprocate this trust and thus increases the investor's profit (*Fehr & Camerer, 2007*). In the context of a trust game this means that the investor is placing trust in the trustee by making an investment, with the expectation that the back-transfer from the trustee is profitable, thus higher than the amount invested (Berg et al., 1995).

Every ten rounds subjects got prompted to rate the trustees. This prompt happened after the subjects' investments, and before they saw the back-transfers. In these prompts, subjects had to rate how much they trust the trustee in this round on a visual analogue scale (VAS) ranging from 0 to 100, how high they expect the back-transfer to be (ranging from 1 to $(10 + 3 \times \text{the investment})$ in this round), and how certain they are about their back-transfer expectations (on a VAS from 0 to 100). Subjects were instructed beforehand that their ratings were not shared with the trustees and would therefore have no influence on the actual back-transfers.

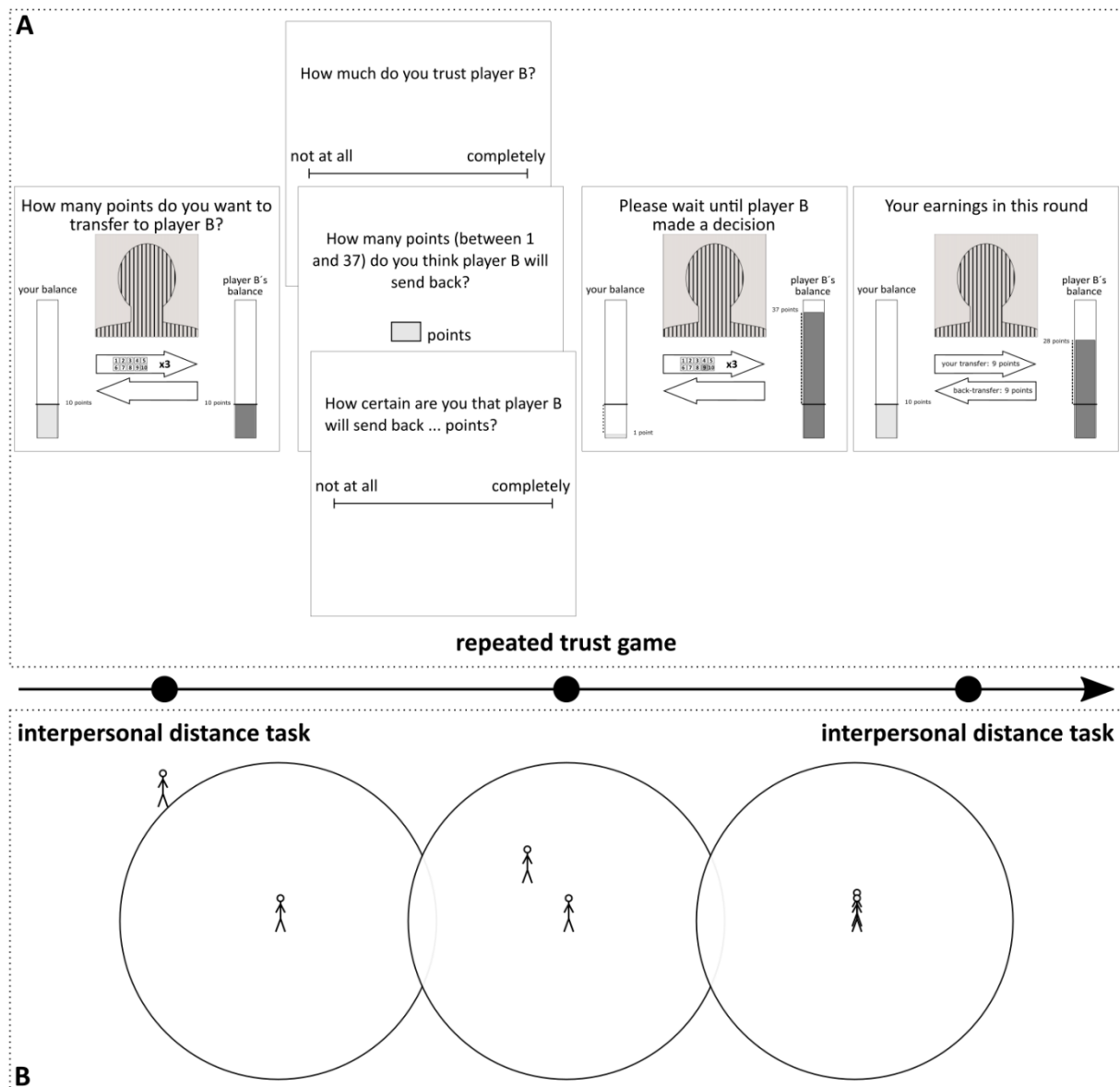


Figure 1. Task progression in experiment 1 and 2: interpersonal distance task, followed by a repeated trust game, which was followed by a repetition of the interpersonal distance task. A) Trial progression of repeated trust game. Subjects played for 52 rounds alternating between player B and player C (depicted with horizontally- and vertically-striped avatars). Every 10 rounds, subjects were prompted about their trust in the trustee (screen 2), the back-transfer expectations (screen 3), and the certainty about the back-transfer expectations (screen 4). B) Trial progression of digital interpersonal distance task. The figure in the middle represented the subject, and the other figure (player B, player C, mother, father, experimenter, best friend, unknown person) approached from the edge of the room. Distance between the figures is expressed in percentages ranging from 100% (approaching figure stands at the edge – left frame) to 0% (two figures colliding – right frame).

INTERPERSONAL DISTANCE TASK

Subjects had to imagine they were standing in the middle of a round room, and a person would approach them from the side of the room (see Figure 1B). Subjects indicated where the approaching person should stand in the room (experiment 1), or stop (experiment 2) for a comfortable conversation. Every trial, subjects indicated their comfortable interpersonal distance for a different approaching figure, with a total of 7 trials. Subjects were told that they were to imagine the approaching figures were—across the different trials—player B, player C, their mother, their father, the experimenter, their best friend, and an unknown person. The order of the approaching figures was randomized between subjects. The principle of the task was the same in experiment 1 and experiment 2, but the administration of the task differed. In experiment 1, a paper and pencil version of the task was administered (Duke & Nowicki, 1972). In this version, every trial was printed on a separate A4-sheet, where a round circle represented a round room (similar to left panel in Figure 1B). The figure in the middle of the room represented the subject, and the figure on the edge of the room the approaching person for whom the comfortable interpersonal distance had to be evaluated. A title identified the approaching figure. Subjects had to draw a cross where the approaching figure should stand for a comfortable distance. The sheets were stapled so that subjects could not lay them next to each other and use their distance from previous sheets as reference points. Distance from the figure in the middle of the room to the cross was measured in mm by two independent raters. Inter-rater correlation was high ($\rho = 0.98$, $p < .001$).

In experiment 2, a digital version of the task was administered (Perry et al., 2013) in E-Prime (version 2.0 (Psychology Software Tools, 2015)). The script is deposited online (Rosenberger, 2019). Here, subjects received three practice trials before the start of the task to familiarize themselves with the task procedure. The layout of the trials was similar to experiment 1, except that the figure at the edge of the room was animated and thus actually approached the figure in the middle. The subject could stop the approaching figure with a button click. If the subject failed to do so, the approaching figure would collide with the figure in the middle and a new trial would begin. The animation consisted of 49 frames, transitioning from the approaching figure standing at the edge of the room to colliding with the figure in the middle of the room. Each frame was displayed for 90 milliseconds. In this digital version, we measured the distance between the figure standing in the middle (i.e. the subject) and the approaching figure in number of frames.

STATISTICAL ANALYSES

We analysed the raw investment data and interpersonal distances with linear mixed models in R (version 3.5.0 (R Core Team, 2017)), with the *lme4* package (version 1.1-21 (Bates, Mächler, Bolker, & Walker, 2015)). Linear mixed models are an extension of linear regressions. For this analysis

individual subject data is not aggregated and individual variation can be specified in the model through the specification of subject specific (and possibly task item specific) intercepts and slopes. This improves the fit of the statistical model to the individual subject data, and reduces error variance in the data due to, for example, testing different individuals, or using different task items, and takes into account time effects (e.g. fatigue) due to repeated observations from the same individual. This increases the statistical test's power, as well as the accuracy of the test results (e.g. Baayen, Davidson, & Bates, 2008).

The test statistics (from Wald chi-square tests) for the main effects and interactions were calculated with orthogonal sum contrasts, and type 3 sum of squares from the *car* package (version 3.0-3 (Fox & Weisberg, 2011)). To decompose interactions with factorial variables, we performed *post-hoc* t-tests with the *emmeans* package with a multivariate correction for multiple comparisons (version 1.3.5.1 (Lenth, 2019)). Degrees of freedom for these post-hoc tests were approximated with the default Kenward-Roger method of the *emmeans* package. To decompose interactions with continuous variables, we interpreted the regression tables generated with the *sjPlot* package (version 2.6.3 (Lüdtke, 2018)). For this we switched to a treatment contrast of the factorial variables and changed their reference levels to obtain the appropriate slope statistics. The reported p-values for these beta statistics were obtained through the default setting in the *sjPlot* package which treats the t-statistics as Wald z-statistics. Standardized beta-statistics for all post-hoc results were also obtained with a treatment contrast from the same regression tables from the *sjPlot* package. They serve as effect size estimates. Alpha for all analyses was set to 0.05. To compare the distances from the paper and pencil version with the digital version, we normalized data of each experiment by computing relative distances, which was done by dividing the measured distance by the maximum possible distance. The transformed interpersonal distance is expressed in percentage of the maximum distance. All the analysis steps described below were performed separately for the datasets of experiment 1 and 2. We performed a sensitivity power analysis in G*Power (version 3.1.9.4) for our main hypothesis that the interpersonal distance before compared to after the trust game differs between the two trustees. Based on a within-subject repeated measures ANOVA with 2 groups, 2 measurements, two-tailed $\alpha = 0.05$, power = 0.80, N = 55, repeated-measure correlation = 0.5, and nonsphericity correction = 1, the minimum expected effect size is $f = 0.19$. The data sets and analysis scripts are deposited online (Rosenberger, 2019). Inter-variable correlations of the mixed models are reported in the supplementary results Tables 6 - 21.

First, we evaluated the change in interpersonal distance resulting from the interactions in the trust game with the following linear mixed model:

$$relative\ distance = \beta_0 + \beta_1 approaching\ people + \beta_2 time\ point + \beta_3 approaching\ people *$$

time point + ε with individual intercepts per subject. Approaching people was a factorial variable with the levels: generous trustee, selfish trustee, mother, father, experimenter, best friend, unknown person. Time point was also a factor with the levels: before trust game, after trust game. With *post hoc* t-tests we compared the relative distance change for each imagined approaching person (thus before vs. after the trust game), as well as the relative distance change between the generous and selfish trustee.

Second, we analysed how well the subjects learned during the trust game, through their change in investments, trust ratings, back-transfer expectations, and certainty ratings. For this we ran four mixed models with the following structure: *outcome variable* = $\beta_0 + \beta_1 \text{trustee type} + \beta_2 \text{round number} + \beta_3 \text{trustee type} * \text{round number} + \varepsilon$. The outcome variables for the four models were investments, trust ratings, back-transfer expectations, and certainty ratings. All models included a separate intercept per subject, as well as two slopes for the trustee types per subject. Trustee type was a factor with the levels generous and selfish, and round number was a continuous variable ranging from 1 to 52. We report the round number beta statistics of the two trustees for an interpretation of the interactions.

Third, to quantify the difference in learning for the two trustee types, we analysed absolute change in the trust game measures (investment, trust ratings, back-transfer expectations, certainty about back-transfer expectations) across the task. For this we calculated difference scores, subtracting the value of the measures in the first trust game round from those in the last round, separately for generous and selfish trustees. We compared these difference scores between trustees with paired t-tests. We also tested for too high initial back-transfer expectations by comparing the actual back-transfer and the back-transfer expectation in the first round with a one-sample t-test. Finally, we tested whether the total earnings at the end of the task differed between the trustees with a paired t-test.

Fourth, to analyse associations between individual performance during the trust game and the change in interpersonal distance, we tested whether total earnings were associated with the change in interpersonal distance (subtracting pre trust game distance from post trust game distance) with the following linear mixed model: *distance change* = $\beta_0 + \beta_1 \text{total earnings} + \beta_2 \text{trustee type} + \beta_3 \text{total earnings} * \text{trustee type} + \varepsilon$, which included a separate intercept for each subject. To increase interpretability of the beta coefficient we changed the scale of total earnings (original range: 173 to 468 points, new range 1.73 to 4.68 points) to be closer to the scale to the distance change variable (range: -0.36 to 0.80), by dividing total earning by 100. In addition, we explored whether the subject's investment and beliefs about the interaction partner during the final interaction was associated with the

distance after the trust game with the following linear mixed model: $distance_{post\ trust\ game} = \beta_0 + \beta_1 trustee\ type + \beta_2 investment_{final\ interaction} + \beta_3 back - transfer\ expectation_{final\ interaction} + \beta_4 trust\ rating_{final\ interaction} + \beta_5 trustee\ type * investment_{final\ interaction} + \beta_6 trustee\ type * back - transfer\ expectation_{final\ interaction} + \beta_7 trustee\ type * trust\ rating_{final\ interaction} + \varepsilon$, which included a separate intercept for each subject. Finally, to explore whether the generosity/selfishness of the trustees, or the trust ratings of the trustees were associated with the distance after the trust game, we ran the following linear mixed model: $distance_{post\ trust\ game} = \beta_0 + \beta_1 trustee\ type + \beta_2 back - transfer_{final\ interaction} + \beta_3 trust\ rating_{final\ interaction} + \beta_4 trustee\ type * back - transfer_{final\ interaction} + \beta_5 trustee\ type * trust\ rating_{final\ interaction} + \varepsilon$, which included a separate intercept for each subject.

RESULTS

EXPERIMENT 1 - REFLECTIVE INTERPERSONAL DISTANCE TASK

Subjects adjusted the distance to the approaching people after (compared to before) the repeated trust game (approaching people x time point interaction: $\chi^2(6) = 23.32, p < .001$), see Figure 2A. Decomposing this interaction with *post-hoc* t-tests, we found that before the trust game interpersonal distance was not different for the two trustees, $t(667.02) = 0.132, p > .99, std.b = 0.01$ (95% $CI = (-0.09, 0.11)$), but it was different afterwards, $t(667.02) = 6.405, p < .001, std.b = 0.31$ (95% $CI = (0.22, 0.41)$). As hypothesized, the distance to the selfish trustee increased from an average of 25% ($SE = 2\%$) before the trust game to an average of 37% ($SE = 4\%$) after the trust game, $t(667.93) = -4.376, p < .001, std.b = 0.31$ (95% $CI = (0.17, 0.45)$). The post-game distance was largest from the selfish trustee compared to the other approaching people (see supplementary results Table 1). The post-game distance from the selfish trustee differed significantly from the post-game distance of all the other approaching people, except the unknown person ($t(667) = 1.683, p = .860, std.b = -0.08$, see also supplementary Table 2).

The distance to the generous trustee did not decrease significantly, $t(667.93) = 1.792, p = .413, std.b = -0.13$ (95% $CI = (-0.27, 0.01)$), from an average of 24% ($SE = 2\%$) to an average of 19% ($SE = 1\%$) after the trust game, and thus did not confirm our hypothesis. The post-game distance from the generous trustee was not different from the post-game distance from the mother ($t(667) = 1.511, p = .931, std.b = -0.07$), father ($t(667) = -0.188, p > .999, std.b = 0.01$), and the experimenter ($t(667) = -0.365, p > .999, std.b = 0.02$). However, it was larger than the distance from the friend ($t(667) = -3.627, p < .011, std.b = -0.18$), and smaller than the distance from the unknown person

($t(667) = -4.722, p < .001, std.b = 0.23$), see also supplementary results Tables 1 and 2. Importantly, for all the other approaching people, the interpersonal distances did not change when comparing before and after trust game measurements (see supplementary results Table 1).

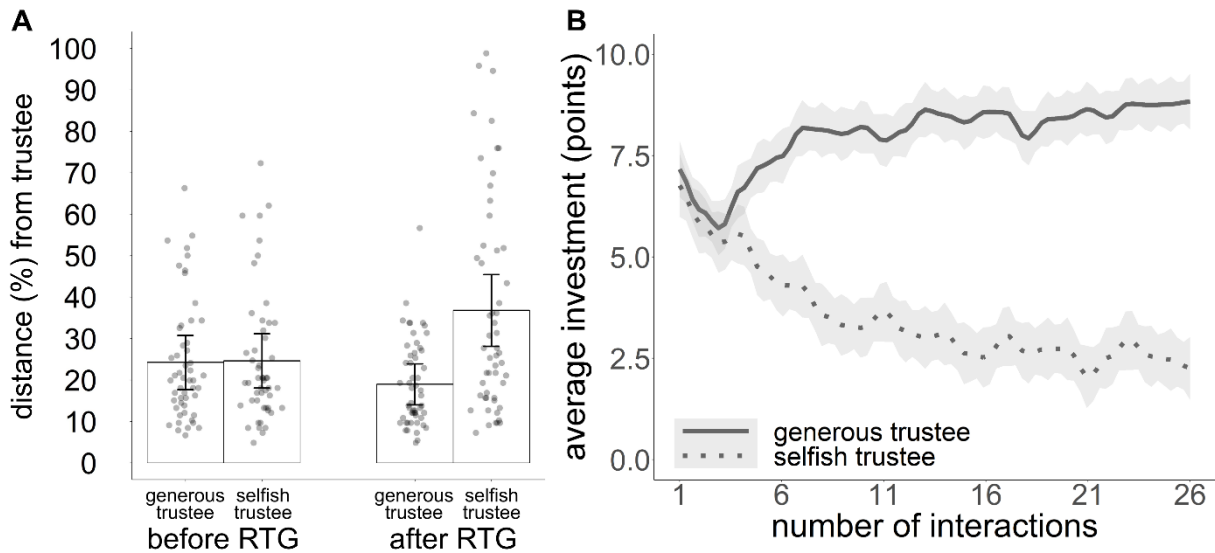


Figure 2. Trust and reflective interpersonal distance adaptation in experiment 1. A) Subjects increased their distance from the selfish trustee after the repeated trust game (RTG) in the interpersonal distance task. Dots represent individual data points, error bars are 95% confidence intervals (CI) adjusted for within-subject comparisons. B) Subjects learned to (dis-)trust the trustees. They increased their investments to the generous trustee (solid line) and decreased them to the selfish trustee (dashed line) during 26 interactions with each trustee in the repeated trust game. Group data are smoothed with local regression (LOESS). Shaded areas around lines represent 95% CI.

EXPERIMENT 2 – REFLEXIVE INTERPERSONAL DISTANCE TASK

In experiment 2, interpersonal distance also changed after the repeated trust game, as indicated by an approaching people $\times$ time point interaction: $\chi^2(6) = 51.43, p < .001$, see supplementary results Figure 1A. Decomposing this interaction, we found that before the trust game, the interpersonal distance did not differ between the selfish and generous trustee, $t(707) = -0.720, p > .999, std.b = 0.03$ (95% CI = (-0.06, 0.13)), but afterwards it did, $t(707) = 7.831, p < .001, std.b = 0.37$ (95% CI = (0.28, 0.47)). The distance from the selfish trustee increased significantly from 20% ($SE = 2\%$) to 38% ($SE = 4\%$), $t(708) = -7.663, p < .001, std.b = 0.52$ (95% CI = (0.39, 0.66)). As in experiment 1, the post-game distance was largest from the selfish trustee compared to the other approaching

people (see supplementary results Table 3). The post-game distance from the selfish trustee differed significantly with all of the other approaching figures (see supplementary results Table 4).

The distance decrease from the generous trustee was very small (from 22% ($SE = 2\%$) to 19% ($SE = 2\%$)) and not significant, $t(708) = 0.841$, $p = .972$, $std.b = -0.06$ (95% $CI = (-0.19, 0.08)$). Similar to experiment 1, the post-game distance from the generous did not differ from the distance from the mother ($t(707) = 2.244$, $p = .471$, $std.b = -0.11$), the father ($t(707) = 2.838$, $p = .134$, $std.b = -0.13$), and the experimenter ($t(707) = -0.30$, $p > .999$, $std.b = 0.00$). However it was larger than the distance from the friend ($t(707) = -4.651$, $p < .001$, $std.b = -0.22$), and smaller than the distance from the unknown person ($t(707) = -3.774$, $p = .007$, $std.b = 0.18$). As in experiment 1, the interpersonal distance did not change for the other approaching people when comparing before and after trust game interactions, see supplementary results Table 3.

Even though the increase in interpersonal distance from the selfish trustee was larger in the 2nd experiment than in the 1st experiment, this difference was not significant, $t(102.61) = -1.444$, $p = 0.152$, $std.b = 0.14$ (95% $CI = (-0.05, 0.33)$).

EXPERIMENT 1 - REPEATED TRUST GAME

At the beginning of the trust game, subjects overestimated the trustees back-transfers, see Table 1: their back-transfer expectations were significantly more positive than the actual back-transfers they received from the two trustees, $t_{generous}(59) = -7.57$, $p < .001$, $d = -0.98$, $t_{selfish}(59) = -6.87$, $p < .001$, $d = -0.89$. In line with that, the first investments to both trustees were higher than the median endowment (= 5 points), and subjects expected higher back-transfers than their investments from both trustees.

Table 1.

Descriptive statistics of measures in the repeated trust game in experiment 1

	first interaction		last interaction	
	generous trustee	selfish trustee	generous trustee	selfish trustee
	m (SD)	m (SD)	m (SD)	m (SD)
investment in points	7.01 (2.88)	6.87 (3.08)	8.63 (2.57)	2.33 (2.83)
trust rating in %	52.7 (25.43)	52.18 (25.77)	70.05 (26.99)	14.79 (27.34)
back-transfer expectation in %	12.09 (6.95)	11.84 (7.17)	15.38 (5.98)	2 (2.89)
difference (BT exp. – BT) in %	-5.08 (4.93)	-4.97 (4.89)	0.5 (4.47)	-0.3 (2.26)
certainty about BT exp. in %	32.19 (25.89)	32.92 (27.05)	63.67 (31.84)	69.79 (39.21)
total earnings in points			408.53 (37.19)	238.37 (18.09)

Note on abbreviations: m = mean, SD = standard deviation, BT = back-transfer, BT exp. = back-transfer expectations

Over the rounds, subjects changed their investments to the generous trustee in a different manner than to the selfish trustee, trustee x round number interaction: $\chi^2(1) = 444.48$, $p < .001$, see Figure 2B. Decomposing this interaction, we found that the investments to the generous trustee increased over the rounds, $b = 0.04$ (95% CI = (0.04, 0.05)), $p < .001$, $std.b = 0.17$ (95% CI = (0.14, 0.20)), and decreased to the selfish trustee, $b = -0.07$ (95% CI = (-0.08, -0.06)), $p < .001$, $std.b = -0.28$ (95% CI = (-0.31, -0.25)). The investment decrease with the selfish trustee was larger than the investment increase with the generous trustee, $t(59) = -11.675$, $p < .001$, $std.b = 0.64$ (95% CI = (0.53, 0.75)).

Subjects also changed their back-transfer expectations about the generous trustee differently than about the selfish trustee over the rounds, trustee x round number interaction: $\chi^2(1) = 158.06$, $p < .001$. *Post-hoc* tests revealed that they increased their back-transfer expectations about the generous trustee ($b = 0.10$ (95% CI = (0.07, 0.12)), $p < .001$, $std.b = 0.21$ (95% CI = (0.15, 0.27))) and decreased them about the selfish trustee ($b = -0.16$ (95% CI = (-0.19, -0.13)), $p < .001$, $std.b = -0.34$ (95% CI = (-0.40, -0.28))) over the rounds. The back-transfer expectations decreased more for the selfish trustee than they increased for the generous trustee, $t(59) = -12.875$, $p < .001$, $std.b = 0.66$ (95% CI = (0.56, 0.76)). This back-transfer expectation change was accompanied by an increase in certainty about the back-transfer expectations over the rounds, round number main effect: $\chi^2(1) = 118.62$, $p < .001$. The certainty did not change differently for the two trustees over the rounds, trustee x round number interaction: $\chi^2(1) = 0.02$, $p = .879$.

Trust ratings about the generous trustee changed differently than about the selfish trustee over the rounds, trustee x round number interaction: $\chi^2(1) = 147.80$, $p < .001$. *Post-hoc* tests revealed that they increased for the generous trustee ($b = 0.44$ (95% CI = (0.33, 0.56)), $p < .001$, $std.b = 0.23$ (95% CI = (0.17, 0.28))), and decreased for the selfish trustee ($b = -0.58$ (95% CI = (-0.70, -0.47)), $p < .001$, $std.b = -0.30$ (95% CI = (-0.35, -0.24))) over the rounds. The decrease in trust ratings was larger for the selfish trustee than the increase in trust ratings for the generous trustee, $t(59) = -10.998$, $p < .001$, $std.b = 0.64$ (95% CI = (0.53, 0.76)).

Total earnings at the end of the trust game were larger with the generous trustee than with the selfish trustee, $t(59) = 29.187$, $p < .001$, $std.b = 0.95$ (95% CI = (0.89, 1.00)), see Table 1. The behavior and beliefs in the repeated trust game confirmed our hypothesis that subjects learned to trust the generous trustee more than the selfish trustee. Interestingly, the change in investments and beliefs was larger for the selfish trustee than the generous trustee, which can be partly attributed to biased back-transfer expectations at the beginning of the repeated trust game.

EXPERIMENT 1 – REPEATED TRUST GAME AND INTERPERSONAL DISTANCE TASK ASSOCIATIONS

To quantify the differences in interpersonal distance changes between the generous and selfish trustee, we explored whether individual variation in learning during the trust game was related to distance changes. First we explored whether total earnings, as a crude measure of learning in the trust game, was associated with the change in interpersonal distance (= after trust game – before trust game): indeed the association between distance change and total earnings differed significantly between the trustees (trustee x total earnings interaction: $\chi^2(1) = 7.912$, $p = .005$). *Post-hoc* tests revealed that higher total earnings with the selfish trustee were significantly associated with larger distance changes from the selfish trustee ($b = 0.33$ (95% $CI = (0.08, 0.59)$), $p = 0.01$, $std.b = 1.53$ (95% $CI = (0.36, 2.69)$)). Total earnings with the generous trustee were not associated with distance change from the generous trustee ($b = -0.08$ (95% $CI = (-0.20, 0.05)$), $p = .233$, $std.b = -0.34$ (95% $CI = (-0.91, 0.22)$)).

Second, we explored whether the investment, back-transfer expectation, or trust rating in the final interaction with the two trustees were associated with their post-game distance. There was no main effect, other than a trust rating main effect ($\chi^2(1) = 6.103$, $p = .014$). There were no significant trustee type x investment interaction ($\chi^2(1) = 2.75$, $p = .097$) or trustee type x back-transfer expectation interaction ($\chi^2(1) = 0.872$, $p = .350$). However, the trustee type x trust rating interaction was significant, $\chi^2(1) = 5.495$, $p = .019$. *Post-hoc* tests revealed that higher trust ratings of the selfish trustee were associated with a smaller post-game distance from the selfish trustee, $b = -0.33$ (95% $CI = (-0.59, -0.07)$), $p = .013$, $std.b = -0.59$ (95% $CI = (-1.05, -0.12)$). The association between the trust rating of the generous trustee and the post-game distance from the generous trustee was not significant, $b = 0.05$ (95% $CI = (-0.13, 0.24)$), $p = .578$, $std.b = 0.10$ (95% $CI = (-0.24, 0.43)$).

As a final step, we explored whether the selfishness/ generosity of the trustees (and thus their back-transfers) or the trust rating of the two trustees in the final interaction were associated with their post-game distance. There was only a significant trustee type x trust rating interaction ($\chi^2(1) = 5.226$, $p = .022$), and no trustee type x back-transfer interaction ($\chi^2(1) = 1.289$, $p = .256$) or back-transfer main effect ($\chi^2(1) = 0.348$, $p = .555$). *Post-hoc* tests revealed, that higher trust ratings of the selfish trustee were associated with smaller post-game distances from the selfish trustee, $b = -0.32$ (95% $CI = (-0.57, -0.06)$), $p = .014$, $std.b = -0.57$ (95% $CI = (-1.02, -0.12)$). The trust rating of the generous trustee was not associated with the post-game distance from the generous trustee, $b = 0.04$ (95% $CI = (-0.14, 0.22)$), $p = .646$, $std.b = 0.08$ (95% $CI = (-0.25, 0.40)$).

EXPERIMENT 2 - REPEATED TRUST GAME

Learning in the repeated trust game was similar as in experiment 1 and is described in the supplementary results, the supplementary results Table 5, as well as the supplementary results

Figure 1B. Here we only highlight the different findings. First, the certainty about the back-transfer expectations changed differently for the two trustees over the rounds (trustee x round number interaction: $\chi^2(1) = 5.51, p = .019$). *Post-hoc* tests revealed that the increase in certainty was marginally steeper for the selfish trustee ($b = 0.73, 95\% CI = (0.55, 0.91), p < .001, std.b = 0.35 (95\% CI = (0.27, 0.44))$) than for the generous trustee ($b = 0.43, 95\% CI = (0.25, 0.61), p < .001, std.b = 0.21 (95\% CI = (0.12, 0.29))$), $t(56) = 1.908, p = .06, std.b = -0.17 (95\% CI = (-0.34, 0.00))$. Second, the association between total earnings and distance change did not significantly differ for the two trustees (earnings x trustee interaction: $\chi^2(1) = 0.10, p = .756$). As in experiment 1, there was also no main effect of earnings on distance change, $\chi^2(1) = 0.73, p = .393$. Third, when exploring whether the investment, back-transfer expectation, or trust rating in the final interaction with the two trustees were associated with their post-game distance, there was only a significant main effect of trust ratings ($\chi^2(1) = 6.795, p = .009$), but no significant interactions (all other $p > .37$). Here, post-game distances decreased with increasing trust ratings, independent of trustee type. Finally, when exploring whether the generosity/ selfishness of the trustees or the trust ratings during the final interaction were associated with the post-game distance, there was only a trust rating main effect ($\chi^2(1) = 8.924, p = .003$), but no interactions (all other $p > .53$). Here, higher trust ratings were associated with a smaller distance from the trustees.

DISCUSSION

In the present study, we demonstrate that trust in interaction partners—developed during a repeated trust game—modulates both reflective and reflexive interpersonal distance preferences. While interactions with a selfish trustee increased interpersonal distance, those with a generous trustee did not significantly decrease distance preferences. These preferences are not caused by learning differences in the repeated trust game, as subjects correctly adapted their behavior and beliefs by increasing their investments, back-transfer expectations, and experienced trust to the generous trustee, and decreasing them to the selfish trustee as the task progressed. Rather, the distances after the trust game were associated with how much the subjects trusted the interaction partners during their final interaction in the trust game. For the reflective interpersonal distance this was specific to the selfish trustee, whereas for the reflexive interpersonal distance there was a general association for both trustees. Interestingly, in both experiments, subjects were biased at the beginning of the repeated trust game as they made high investments and were expecting high back-transfers of both trustees. These overly optimistic expectations were violated by the selfish trustee which resulted in larger changes in investments and beliefs for the selfish than the generous trustee. These expectation violations might have put more attention on the selfish than the generous trustee and might have brought about the greater change in interpersonal distance to the selfish trustee.

Another reason the change in interpersonal distance might have been more pronounced for the selfish than for the generous trustee may be because the selfish (vs generous) trustee's behavior was more salient in the repeated trust game. Negative feedback, but specifically non-cooperative behavior (e.g., low back-transfers), draw more attention than positive feedback (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001) or cooperative behavior (e.g., high back-transfers) (Rankin & Eggimann, 2009; Vanneste, Verplaetse, Vanhiel, & Braeckman, 2007). While high expectations at the beginning of a decision-making task are a general phenomenon (Branas-Garza, Rodriguez-Lara, & Sanchez, 2017), the selfish trustee violated them continuously by making low back-transfers. The non-cooperative behavior and expectation violations might have made more salient the selfish (vs generous) trustee's behavior, resulting in a more pronounced adjustment in interpersonal distance preferences.

The salience of the selfish trustee mirrors saliency of negative decision-makers when inferring their moral character (Siegel, Mathys, Rutledge, & Crockett, 2018). Interestingly, belief formation about bad moral behavior is characterized by a higher level of uncertainty than good moral behavior. This finding contrasts with our results where certainty about back-transfer expectations were similar for the generous and selfish trustee in experiment 1, and higher for the selfish than the generous trustee in experiment 2. A reason for these different results might lie in differences in the decision outcome about which subjects had to form beliefs: while subjects in our study actually received the back-transfers about which they formed beliefs, subjects in Siegel et al.'s study formed beliefs about a behavior that did not affect them (i.e. the moral character of a decision-maker giving electric shocks to a third person). Thus, while we both show that negatively perceived people appear to be more salient than positively perceived people during belief formation, the certainty about these beliefs might depend on whether these people's behavior can directly impact the subject.

At the beginning of the trust game in both experiments, subjects displayed overly optimistic expectations about the interaction partners' back-transfers. These expectations might indicate that the subjects started out the task in a trust mindset. With a trust mindset, specific mental representations that are in that moment highly accessible can influence behaviour and decisions, such as priming effects (Kleiman, Sher, Elster, & Mayo, 2015; Mayo, 2015). With this trust mindset, subjects in the present study might have been receptive to the trust game manipulation and therefore adjusted their subsequent interpersonal distance. In contrast, with a distrust mindset, alternatives are considered to specific mental representations, which decreases, or even blocks the influence of the highly accessible mental representations on the current behaviour and decisions (Kleiman et al., 2015; Mayo, 2015). Curiously, interactions with the selfish trustee did not appear to

change their mindset to distrust. Specifically, we found that interactions with the selfish trustee increased distance preferences. However, with a distrust mindset, the influence of interacting with a selfish trustee on distance preferences should have been diminished. This discrepancy between our findings and the distrust mindset's prediction might indicate that the subjects had a trust mindset throughout the experiment.

We demonstrate that the trust learned during the repeated trust game generalizes beyond monetary exchanges and influences subsequent imagined interactions in a different context with the same partner. Specifically, the experienced trust—but not the monetary aspects of the trust game (i.e., the investments, the back-transfers from the trustee, nor the back-transfer expectations)—were associated with distance preferences after the trust game. Our study validates findings from Hale, Payne, Taylor, Paoletti, and Hamilton (2018) who have suggested, but could not unequivocally demonstrate, that interactions with a fair and unfair trustee during a repeated trust game impacted how much of their advice was followed in navigating a subsequent virtual maze task. In their virtual reality study, almost half of the subjects based their decision to follow advice on the character of the voice of the trustee. The other half of the subjects reported that they based their decision to follow advice on the perceived fairness of the trustee during the trust game. This second half of the subjects followed the advice of the fair (vs unfair) trustee more. However, this effect was not present in the group that was guided by the character of the trustee's voice. By limiting the source of information to one (i.e., the back-transfer behavior during the trust game), we validated Hale et al.'s finding in a larger sample. In addition, our study suggests that Hale et al.'s findings might be attributable to the subjectively experienced fairness of the trustees, and not solely the monetary interactions during the trust game.

The trust learned during the repeated trust game only changed interpersonal distance to the trustees, and did not affect distance preferences to imagined approaching people unrelated to the trust game. Negative social experiences with the selfish trustee did not generalize to affect interpersonal distance preferences with other approaching people. For example, it neither made subjects more cautious—by increasing interpersonal distance to all of the approaching people—nor did it increase affiliative behavior with close others (Loseth, Ellingsen, & Leknes, 2014)—by decreasing interpersonal distance after the trust game with an approaching friend or parents. Our reported effects are thus specific to the interaction partners of the trust game. These are important notions for the specificity of our manipulation as well as for the external validity of our study, as trusting a person in everyday life is not confined to a specific situation and typically does not extend to unrelated others.

We have demonstrated that both reflective and reflexive interpersonal distance preferences change when we distrust an interaction partner. This finding is in line with work on peripersonal, as well as reflective and reflexive interpersonal distance adaptations to interaction partners of different age and gender (Iachini et al., 2016). Interestingly, in our reflective interpersonal distance task, total earnings and experienced trust were specifically associated with the change in interpersonal distance to the selfish trustee. Opposite to that, in the reflexive interpersonal distance task, experienced trust was generally associated with the distance after the trust game in both trustees, and not at all with total earnings. These effects hint at possible differential mechanisms underlying the reflective and reflexive measures. Future studies with a well-matched non-social trust game control task are needed to exclude the possibility that our findings are driven by changes in risk-aversion instead of trust. For now, we can only conclude that the distrust in the selfish trustee was so strong that it not only affected the conscious deliberative aspect of interpersonal distance, but also the more automatic, reflexive aspect.

The increase in interpersonal distance to the selfish trustee might be due to a change in the subject's personal space caused by the interactions with the selfish trustee. Personal space is theorized to serve as a buffer-zone protecting against harms or threats (Dosey & Meisels, 1969), and extends when interacting with unfavourable others (Iachini et al., 2015; Pellencin et al., 2018; Teneggi, Canzoneri, di Pellegrino, & Serino, 2013). As the permeability of this personal space is thought to depend on the trustworthiness of the interaction partner (Scott, 1993), the selfish trustee might have intruded the subject's extending personal space. Such an intrusion is associated with physiological arousal (Sawada, 2003), which might be related to increased stress (Dosey & Meisels, 1969) and anxiety (Perry et al., 2013; Rogers, Rearden, & Hillner, 1981). Because an intrusion of personal space triggers a withdrawal reaction (Hayduk, 1983; Sundstrom & Altman, 1976), the increase in interpersonal distance from the selfish trustee might be a way to relieve this physiological and emotional discomfort. Future research has yet to establish whether a change in trust affects the size of a subject's personal space.

A consequence of an increased physical distance from negative interaction partners might be that the psychological distance increases as well, as communication becomes more impersonal (Jourard & Friedman, 1970; Skotko & Langmeyer, 1977), and empathising behaviour potentially decreases (Perry et al., 2014; Strayer & Roberts, 1997). Relatedly, increasing spatial distance might have a similar function to decreasing similarity, which has a negative impact on empathy and prosocial behaviour as well (e.g. Majdandzic, Amashafer, Hummer, Windischberger, & Lamm, 2016). An increased distance might be necessary for successful outcomes with a negative interaction partner (Burgoon, Stacks, & Woodall, 1979), potentially preventing a break-down of the interaction

(King-Casas et al., 2008). Future research has to establish whether the success of an interaction with a distrusted interaction partner in different contexts (such as collaboration, or competition) depend on the interpersonal distances from them.

In conclusion, we have established that developing distrust in a person behaving in a consistently selfish manner during a repeated trust game affects subsequent reflective, as well as reflexive interpersonal distance towards them. These findings extend our current understanding of trust and interpersonal distance, such that interpersonal distance does not only influence levels of trust (Bryan et al., 2012), but trust also impacts interpersonal distance. Furthermore, our results are a starting point to deepen our understanding of how we successfully navigate interactions with distrusted others without a breakdown or escalation of an interaction.

ACKNOWLEDGEMENT

In loving memory of Christoph Eisenegger. The work in this paper was supported by the Vienna Science and Technology Fund (WWTF VRG13-007). We thank Shawn Geniole for his invaluable help with English editing our manuscript. We thank Sebastijan Veselic for his help with data collection and distance measurements for experiment 1. We also thank Xenia Raufeisen, Eugen Secara, and Caspar Matzhold for their help with data collection for experiment 2.

The authors declare no conflict of interest.

REFERENCES

- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390-412. doi:10.1016/j.jml.2007.12.005
- Bates, D. M., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1 - 48.
- Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, 5(4), 323-370. doi:10.1037/1089-2680.5.4.323
- Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *GEB*, 10, 122 - 142.
- Branas-Garza, P., Rodriguez-Lara, I., & Sanchez, A. (2017). Humans expect generosity. *Sci Rep*, 7, 42446. doi:10.1038/srep42446
- Bryan, R., Perona, P., & Adolphs, R. (2012). Perspective distortion from interpersonal distance is an implicit visual cue for social judgments of faces. *PLoS One*, 7(9), e45301. doi:10.1371/journal.pone.0045301
- Burgoon, J., Stacks, D., & Woodall, W. G. (1979). A communicative model of violations of distancing expectations. *The Western Journal of Speech Communication*, 43, 153 - 167.
- Candini, M., Giuberti, V., Manattini, A., Grittani, S., di Pellegrino, G., & Frassinetti, F. (2017). Personal space regulation in childhood autism: Effects of social interaction and person's perspective. *Autism Res*, 10(1), 144-154. doi:10.1002/aur.1637
- Cartaud, A., Ruggiero, G., Ott, L., Iachini, T., & Coello, Y. (2018). Physiological Response to Facial Expressions in Peripersonal Space Determines Interpersonal Distance in a Social Interaction Context. *Front Psychol*, 9, 657. doi:10.3389/fpsyg.2018.00657
- Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: trustworthiness as a dynamic belief. *Cogn Psychol*, 61(2), 87-105. doi:10.1016/j.cogpsych.2010.03.001
- Cohen, D., Perry, A., Mayseless, N., Kleinmuntz, O., & Shamay-Tsoory, S. G. (2018). The role of oxytocin in implicit personal space regulation: An fMRI study. *Psychoneuroendocrinology*, 91, 206-215. doi:10.1016/j.psyneuen.2018.02.036
- Dosey, M., & Meisels, M. (1969). Personal space and self-protection. *Journal of Personality and Social Psychology*, 11(2), 93 - 97.

- Duke, M. P., & Nowicki, S. (1972). A new measure and social-learning model for interpersonal distance. *Journal of Experimental Research in Personality*.
- Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Sciences*, 11(10), 419-427. doi:10.1016/j.tics.2007.09.002
- Fett, A. K., Gromann, P. M., Giampietro, V., Shergill, S. S., & Krabbendam, L. (2012). Default distrust? An fMRI investigation of the neural development of trust and cooperation. *Soc Cogn Affect Neurosci*, 9(4), 395-402. doi:10.1093/scan/nss144
- Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2), 171-178.
- Fox, J., & Weisberg, S. (2011). *An {R} Companion to Applied Regression*. Thousand Oaks CA: Sage.
- Gessaroli, E., Santelli, E., di Pellegrino, G., & Frassinetti, F. (2013). Personal space regulation in childhood autism spectrum disorders. *PLoS One*, 8(9), e74959. doi:10.1371/journal.pone.0074959
- Hale, J., Payne, M. E., Taylor, K. M., Paoletti, D., & Hamilton, A. F. (2018). The virtual maze: A behavioural tool for measuring trust. *Q J Exp Psychol (Hove)*, 71(4), 989-1008. doi:10.1080/17470218.2017.1307865
- Hall, E. T. (1966). *The hidden dimension* (Vol. 609): Garden City, NY: Doubleday.
- Hayduk, L. A. (1983). Personal-Space - Where We Now Stand. *Psychological Bulletin*, 94(2), 293-335. doi:10.1037/0033-2909.94.2.293
- Iachini, T., Coello, Y., Frassinetti, F., Senese, V. P., Galante, F., & Ruggiero, G. (2016). Peripersonal and interpersonal space in virtual and real environments: Effects of gender and age. *Journal of Environmental Psychology*, 45, 154-164. doi:10.1016/j.jenvp.2016.01.004
- Iachini, T., Pagliaro, S., & Ruggiero, G. (2015). Near or far? It depends on my impression: moral information and spatial behavior in virtual interactions. *Acta Psychologica*, 161, 131-136. doi:10.1016/j.actpsy.2015.09.003
- Jourard, S. M., & Friedman, R. (1970). Experimenter-subject "distance" and self-disclosure. *Journal of Personality and Social Psychology*, 15(3), 278-282.

- King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science*, 321(5890), 806-810. doi:10.1126/science.1156902
- Kleiman, T., Sher, N., Elster, A., & Mayo, R. (2015). Accessibility is a matter of trust: Dispositional and contextual distrust blocks accessibility effects. *Cognition*, 142, 333-344. doi:10.1016/j.cognition.2015.06.001
- Lenth, R. V. (2019). emmeans: Estimated Marginal Means, aka Least-Squares Means. (Version R package version 1.3.5.1). Retrieved from <https://CRAN.R-project.org/package=emmeans>
- Lloyd, D. M. (2009). The space between us: a neurophilosophical framework for the investigation of human interpersonal space. *Neuroscience and Biobehavioral Reviews*, 33(3), 297-304. doi:10.1016/j.neubiorev.2008.09.007
- Loeth, G. E., Ellingsen, D. M., & Leknes, S. (2014). State-dependent mu-opioid modulation of social motivation. *Front Behav Neurosci*, 8, 430. doi:10.3389/fnbeh.2014.00430
- Lüdtke, D. (2018). sjPlot: Data Visualization for Statistics in Social Science. Retrieved from <https://CRAN.R-project.org/package=sjPlot>
- Majdandzic, J., Amashauer, S., Hummer, A., Windischberger, C., & Lamm, C. (2016). The selfless mind: How prefrontal involvement in mentalizing with similar and dissimilar others shapes empathy and prosocial behavior. *Cognition*, 157, 24-38. doi:10.1016/j.cognition.2016.08.003
- Mayo, R. (2015). Cognition is a matter of trust: Distrust tunes cognitive processes. *European Review of Social Psychology*, 26(1), 283-327. doi:10.1080/10463283.2015.1117249
- Pellencin, E., Paladino, M. P., Herbelin, B., & Serino, A. (2018). Social perception of others shapes one's own multisensory peripersonal space. *Cortex*, 104, 163-179. doi:10.1016/j.cortex.2017.08.033
- Perry, A., Mankuta, D., & Shamay-Tsoory, S. G. (2014). OT promotes closer interpersonal distance among highly empathic individuals. *Soc Cogn Affect Neurosci*. doi:10.1093/scan/nsu017
- Perry, A., Rubinsten, O., Peled, L., & Shamay-Tsoory, S. G. (2013). Don't stand so close to me: a behavioral and ERP study of preferred interpersonal distance. *NeuroImage*, 83, 761-769. doi:10.1016/j.neuroimage.2013.07.042

- Psychology Software Tools, I. (2015). E-Prime 2.0. Pittsburgh, PA. Retrieved from <http://www.pstnet.com>
- R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
- Rankin, D. J., & Eggimann, F. (2009). The evolution of judgement bias in indirect reciprocity. *Proc Biol Sci*, 276(1660), 1339-1345. doi:10.1098/rspb.2008.1715
- Rogers, P., Rearden, J., & Hillner, W. (1981). Effects of distance from interviewer and intimacy of topic on verbal productivity and anxiety. *Psychological Reports*, 49, 303 - 307.
- Rosenberger, L. A. (2019). *Scripts for: Interpersonal distance adjustments after interactions with a generous and selfish trustee during a repeated trust game*. Retrieved from: osf.io/4r8sd
- Rosenberger, L. A., Eisenegger, C., Naef, M., Terburg, D., Fourie, J., Stein, D. J., & Van Honk, J. (2019). The human basolateral amygdala is indispensable for social experiential learning. *Current Biology*.
- Ruggiero, G., Frassinetti, F., Coello, Y., Rapuano, M., di Cola, A. S., & Iachini, T. (2017). The effect of facial expressions on peripersonal and interpersonal spaces. *Psychol Res*, 81(6), 1232-1240. doi:10.1007/s00426-016-0806-x
- Sawada, Y. (2003). Blood pressure and heart rate responses to an intrusion on personal space. *Japanese Psychological Research*, 45(2), 115 - 121.
- Scott, A. (1993). A beginning theory of personal space boundaries. *Perspectives in Psychiatric Care*, 29(2).
- Shao, M., Yin, J., Ji, H., Yang, Y., & Song, F. (2020). Distance perception warped by social relations: Social interaction information compresses distance. *Acta Psychologica*, 202, 102948. doi:10.1016/j.actpsy.2019.102948
- Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*, 2(10), 750-756. doi:10.1038/s41562-018-0425-1
- Skotko, V. P., & Langmeyer, D. (1977). The effects of interaction distance and gender on self-disclosure in the dyad. *Sociometry*, 40(2), 178-182.
- Strayer, J., & Roberts, W. (1997). Children's personal distance and their empathy: indices of interpersonal closeness. *International Journal of behavioural Development*, 20(3), 385 - 403.

- Sundstrom, E., & Altman, I. (1976). Interpersonal relationships and personal space: research review and theoretical model. *Human Ecology*, 4(1).
- Teneggi, C., Canzoneri, E., di Pellegrino, G., & Serino, A. (2013). Social modulation of peripersonal space boundaries. *Current Biology*, 23(5), 406-411. doi:10.1016/j.cub.2013.01.043
- Vagnoni, E., Lewis, J., Tajadura-Jimenez, A., & Cardini, F. (2018). Listening to a conversation with aggressive content expands the interpersonal space. *PLoS One*, 13(3), e0192753. doi:10.1371/journal.pone.0192753
- Vanneste, S., Verplaetse, J., Vanhiel, A., & Braeckman, J. (2007). Attention bias toward noncooperative people. A dot probe classification study in cheating detection. *Evolution and Human Behavior*, 28(4), 272-276. doi:10.1016/j.evolhumbehav.2007.02.005
- Wheless, L. R., & Grotz, J. (1977). The measurement of trust and its relationship to self-disclosure. *Human Communication Research*, 3(3), 250-257.

SUPPLEMENTARY RESULTS

EXPERIMENT 1

Supplementary Table 1.

Descriptives of distance measured during paper & pencil interpersonal distance task and test statistics of distance change due to trust game

	Before trust game Mean (SD)	After trust game Mean (SD)	t-statistic	p-value (corrected)	std. b
Friend	0.09 (0.05)	0.09 (0.05)	0.012	> .999	-0.00
Mother	0.15 (0.20)	0.15 (0.20)	0.160	> .999	-0.01
Unknown person	0.29 (0.18)	0.32 (0.22)	-1.037	.917	0.07
Father	0.21 (0.27)	0.19 (0.26)	0.424	.999	-0.03
Experimenter	0.22 (0.15)	0.20 (0.14)	0.640	.994	-0.05
Selfish trustee	0.25 (0.16)	0.37 (0.26)	-4.376	<.001	0.31
Generous trustee	0.24 (0.14)	0.19 (0.10)	1.792	.413	-0.13

Supplementary Table 2.

Test statistics of distance measured during paper & pencil interpersonal distance task comparing distances between trustees and other approaching people before and after trust game

Comparison	Estimated marginal means (SE)	Degrees of freedom	t-statistic	p-value (corrected)	std. b
<u>Before trust game</u>					
Selfish trustee -					
Generous trustee	0.004 (0.029)	667	0.132	1.0000	-0.01
Friend	0.155 (0.029)	667	5.415	<.0001	-0.27
Mother	0.092 (0.029)	667	3.213	0.0458	-0.16
Father	0.037 (0.029)	667	1.291	0.9790	-0.07
Unknown person	-0.048 (0.029)	667	-1.662	0.8706	0.08
Experimenter	0.026 (0.029)	667	0.916	0.9991	-0.05
Generous trustee -					
Friend	-0.151 (0.029)	667	-5.283	<.0001	-0.27
Mother	0.088 (0.029)	667	3.081	0.0681	-0.16
Father	0.033 (0.029)	667	1.160	0.9916	-0.06
Unknown person	-0.051 (0.029)	667	-1.794	0.7987	0.09
Experimenter	0.022 (0.029)	667	0.784	0.9998	-0.04
<u>After trust game</u>					
Selfish trustee -					
Generous trustee	0.178 (0.028)	667	6.405	<.0001	-0.31
Friend	0.279 (0.028)	667	10.033	<.0001	-0.49
Mother	0.220 (0.028)	667	7.916	<.0001	-0.39
Father	0.173 (0.028)	667	6.217	<.0001	-0.30
Unknown person	0.047 (0.028)	667	1.683	0.8602	-0.08
Experimenter	0.168 (0.028)	667	6.040	<.0001	-0.30
Generous trustee -					
Friend	-0.101 (0.028)	667	-3.627	0.0114	-0.18
Mother	0.042 (0.028)	667	1.511	0.9308	-0.07
Father	-0.005 (0.028)	667	-0.188	1.0000	0.01
Unknown person	-0.131 (0.028)	667	-4.722	0.0001	0.23
Experimenter	-0.010 (0.028)	667	-0.365	1.0000	0.02

EXPERIMENT 2

Supplementary Table 3.

Descriptives of distance measured during digital interpersonal distance task and test statistics of distance change due to trust game

	Before trust game Mean (SD)	After trust game Mean (SD)	t-statistic	p-value (corrected)	Std. b
Friend	0.10 (0.12)	0.08 (0.06)	0.652	.994	-0.04
Mother	0.13 (0.17)	0.14 (0.18)	-0.417	.999	0.03
Unknown person	0.27 (0.19)	0.29 (0.20)	-0.774	.982	0.05
Father	0.12 (0.12)	0.12 (0.15)	-0.353	.999	0.02
Experimenter	0.19 (0.12)	0.19 (0.12)	-0.586	.997	0.04
Selfish trustee	0.20 (0.19)	0.38 (0.29)	-7.663	<.001	0.52
Generous trustee	0.22 (0.16)	0.19 (0.15)	0.841	.972	-0.06

Supplementary Table 4.

Test statistics of distance measured during digital interpersonal distance task comparing distances between trustees and other approaching people before and after trust game

Comparison	Estimated marginal means (SE)	Degrees of freedom	t-statistic	p-value (corrected)	std. b
<u>Before trust game</u>					
Selfish trustee -					
Generous trustee	-0.018 (0.025)	707	-0.720	0.9999	0.03
Friend	0.101 (0.025)	707	4.078	0.0021	-0.20
Mother	0.068 (0.025)	707	2.759	0.1641	-0.13
Father	0.081 (0.025)	707	3.284	0.0370	-0.16
Unknown person	-0.071 (0.025)	707	-2.849	0.1297	0.14
Experimenter	0.017 (0.025)	707	0.675	1.0000	-0.03
Generous trustee -					
Friend	-0.119 (0.025)	707	-4.798	0.0001	-0.23
Mother	0.086 (0.025)	707	3.479	0.0195	-0.17
Father	0.099 (0.025)	707	4.003	0.0027	-0.19
Unknown person	-0.053 (0.025)	707	-2.129	0.5586	0.10
Experimenter	0.035 (0.025)	707	1.394	0.9615	-0.07
<u>After trust game</u>					
Selfish trustee -					
Generous trustee	0.192 (0.025)	707	7.831	<.0001	-0.37
Friend	0.306 (0.025)	707	12.482	<.0001	-0.59
Mother	0.247 (0.025)	707	10.075	<.0001	-0.48
Father	0.262 (0.025)	707	10.669	<.0001	-0.51
Unknown person	0.099 (0.025)	707	4.057	0.0021	-0.19
Experimenter	0.191 (0.025)	707	7.801	<.0001	-0.37
Generous trustee -					
Friend	-0.114 (0.025)	707	-4.651	0.0001	-0.22
Mother	0.055 (0.025)	707	2.244	0.4713	-0.11
Father	0.070 (0.025)	707	2.838	0.1337	-0.13
Unknown person	-0.093 (0.025)	707	-3.774	0.0066	0.18
Experimenter	-0.001 (0.025)	707	-0.030	1.0000	0.00

REPEATED TRUST GAME

In experiment 2, subjects learned successfully to differentiate between the generous and selfish trustee during the repeated trust game. They changed their investments in a different manner for the two trustees over the rounds, interaction trustee x period: $\chi^2(1) = 665.272$, $p < .001$. Decomposing this interaction, the investments to the generous trustee increased over the rounds, b

= 0.05 (95% CI = (0.04, 0.06)), $p < .001$, $std.b = 0.19$ (95% CI = (0.17, 0.22)), and decreased to the selfish trustee, $b = -0.08$ (95% CI = (-0.09, -0.07)), $p < .001$, $std.b = -0.31$ (95% CI = (-0.34, -0.28)). In addition, subjects changed their trust ratings about the two trustees in a different manner of the rounds, interaction trustee x period: $\chi^2(1) = 107.948$, $p < .001$. Decomposing this interaction, trust ratings about the generous trustee increased over the rounds, $b = 0.36$ (95%CI = (0.24, 0.49)), $p < .001$, $std.b = 0.17$ (95% CI = (0.11, 0.23)), and decreased with the selfish trustee, $b = -0.57$ (95% CI = (-0.69, -0.44)), $p < .001$, $std.b = -0.27$ (95% CI = (-0.33, -0.21)). Back-transfer expectations about the two trustees also changed in a different manner over the rounds, trustee x period interaction: $\chi^2(1) = 135.665$, $p < 0.001$. *Post-hoc* tests revealed that the back-transfer expectations about the generous trustee increased over the rounds ($b = 0.08$, (95% CI = (0.05, 0.11)), $p < .001$, $std.b = 0.18$ (95% CI = (0.11, 0.24))), while they decreased for the selfish trustee over the rounds ($b = -0.17$ (95% CI = (-0.20, -0.14))), $p < .001$, $std.b = -0.38$ (95% CI = (-0.44, -0.31))).

Supplementary Table 5.

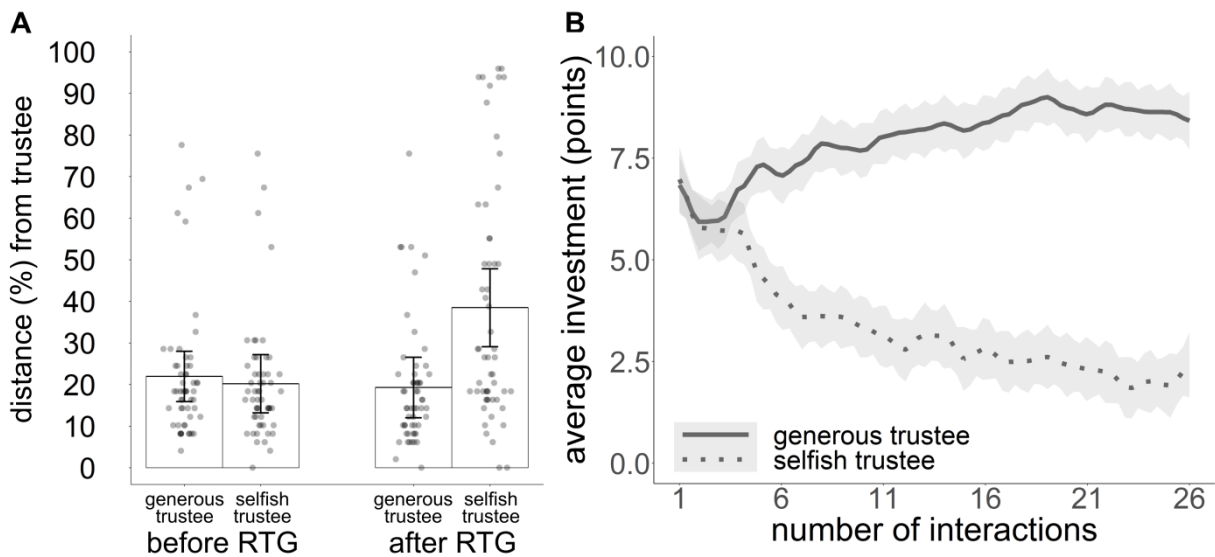
Descriptive statistics of measures in the repeated trust game in experiment 2

	first interaction		last interaction	
	generous trustee	selfish trustee	generous trustee	selfish trustee
	m (SD)	m (SD)	m (SD)	m (SD)
Investment in points	6.84 (3.02)	6.98 (3.16)	8.42 (2.59)	2.42 (2.90)
trust rating in %	54.46 (26.21)	53.11 (26.26)	70.07 (27.07)	15.82 (28.46)
back-transfer expectation in %	12.28 (7.42)	12.51 (7.37)	14.89 (5.61)	2.09 (3.07)
difference (BT exp. – BT) in %	5.44 (5.03)	5.53 (4.76)	-0.11 (4.83)	0.23 (2.56)
certainty about BT exp. in %	36.56 (26.35)	32.79 (26.35)	62.67 (30.77)	71.89 (38.34)
total earnings in points			405.28 (37.47)	242.93 (15.07)

Note on abbreviations: m = mean, SD = standard deviation, BT = back-transfer, BT exp. = back-transfer expectations

As in the first experiment, subjects displayed a bias in their trust game behavior. During the first interaction with the generous and selfish trustee, their back-transfer expectations did not match the actual back-transfers from the two trustees ($t_{selfish}(56) = -8.77$, $p < .001$, $d = -1.082$; $t_{generous}(56) = -8.16$, $p < .001$, $d = -1.162$). As can be seen in the supplementary Table 3, the decrease in investments with the selfish trustee was larger than the increase in investments with the generous trustee, $t(56) = -12.58$, $p < .001$, $std.b = 0.66$ (95% CI = (0.55, 0.76)). In addition, the decrease in back-transfer expectations was larger with the selfish trustee than the increase with the generous trustee, $t(56) = -$

13.94, $p < .001$, $std.b = 0.62$ (95% CI = (0.54, 0.71)). Finally, the decrease in trust ratings was also larger in the selfish trustee, than the increase with the generous trustee, $t(56) = -11.66$, $p < .001$, $std.b = 0.66$ (95% CI = (0.55, 0.77)). Total earnings were larger for the generous trustee than the selfish trustee, $t(56) = 29.85$, $p < .001$, $std.b = -0.94$ (95% CI = (-1.01, -0.88)).



Supplementary Figure 1. Trust and static interpersonal distance adaptation in experiment 2. A) Subjects increased their distance to the selfish trustee after the repeated trust game (RTG) in the interpersonal distance task. Dots represent individual data points, error bars are 95% confidence intervals (CI) adjusted for within-subject comparisons. B) Subjects learned to trust the trustees. They increased their investments to the generous trustee (solid line) and decreased them to the selfish trustee (dashed line) during 26 interactions with each trustee in the repeated trust game. Group data are smoothed with local regression (LOESS). Shaded areas around lines represent 95% CI.

Supplementary Table 6.

Inter-variable correlation matrix of mixed model regressing relative distance in experiment 1 on approaching people (trustee), time point (session), and their interaction.

	intercept	trustee 1	trustee 2	trustee 3	trustee 4	trustee 5	trustee 6	session 1	t1:s1	t2:s1	t3:s1	t4:s1	t5:s1
trustee 1	0												
trustee 2	0	-0.166											
trustee 3	0	-0.166	-0.166										
trustee 4	0	-0.166	-0.166	-0.166									
trustee 5	0	-0.166	-0.166	-0.166	-0.166								
trustee 6	0.001	-0.167	-0.167	-0.167	-0.167	-0.167							
session 1	-0.017	0.001	0.001	0.001	0.001	0.001	-0.004						
t1:s1	0	-0.029	0.004	0.004	0.004	0.004	0.006	-0.001					
t2:s1	0	0.004	-0.029	0.004	0.004	0.004	0.006	-0.001	-0.166				
t3:s1	0	0.004	0.004	-0.029	0.004	0.004	0.006	-0.001	-0.166	-0.166			
t4:s1	0	0.004	0.004	0.004	-0.029	0.004	0.006	-0.001	-0.166	-0.166	-0.166		
t5:s1	0	0.004	0.004	0.004	0.004	-0.029	0.006	-0.001	-0.166	-0.166	-0.166	-0.166	
t6:s1	-0.001	0.006	0.006	0.006	0.006	0.006	-0.038	0.004	-0.167	-0.167	-0.167	-0.167	-0.167

Supplementary Table 7.

Inter-variable correlation matrix of mixed model regressing relative distance in experiment 2 on approaching people (trustee), time point (session), and their interaction.

	intercept	trustee 1	trustee 2	trustee 3	trustee 4	trustee 5	trustee 6	session 1	t1:s1	t2:s1	t3:s1	t4:s1	t5:s1
trustee 1	0												
trustee 2	0	-0.167											
trustee 3	0	-0.167	-0.167										
trustee 4	0	-0.167	-0.167	-0.167									
trustee 5	0	-0.167	-0.167	-0.167	-0.167								
trustee 6	0	-0.167	-0.167	-0.167	-0.167	-0.167							
session 1	-0.005	0	0	0	0	0	0						
t1:s1	0	-0.009	0.002	0.002	0.002	0.002	0.002	0					
t2:s1	0	0.002	-0.009	0.002	0.002	0.002	0.002	0	-0.167				
t3:s1	0	0.002	0.002	-0.009	0.002	0.002	0.002	0	-0.167	-0.167			
t4:s1	0	0.002	0.002	0.002	-0.009	0.002	0.002	0	-0.167	-0.167	-0.167		
t5:s1	0	0.002	0.002	0.002	0.002	-0.009	0.002	0	-0.167	-0.167	-0.167	-0.167	
t6:s1	0	0.002	0.002	0.002	0.002	0.002	-0.009	0	-0.167	-0.167	-0.167	-0.167	-0.167

Supplementary Table 8.

Inter-variable correlation matrix of mixed model regressing investment in experiment 1 on trustee type, round number, and their interaction.

	intercept	trustee 1	round nr
trustee 1	0.057		
round nr	-0.332	0	
t1:round nr	0	-0.463	-0.001

Supplementary Table 9.

Inter-variable correlation matrix of mixed model regressing back-transfer expectation in experiment 1 on trustee type, round number, and their interaction.

	intercept	round nr	trustee 1
round nr	-0.497		
trustee 1	-0.259	0.001	
round nr:t1	0	-0.001	-0.687

Supplementary Table 10.

Inter-variable correlation matrix of mixed model regressing certainty about back-transfer expectation in experiment 1 on trustee type, round number, and their interaction.

	intercept	round nr	trustee 1
round nr	-0.538		
trustee 1	0.124	0.001	
round nr:t1	0	-0.001	-0.793

Supplementary Table 11.

Inter-variable correlation matrix of mixed model regressing trust ratings in experiment 1 on trustee type, round number, and their interaction.

	intercept	round nr	trustee 1
round nr	-0.502		
trustee 1	-0.05	0	
round nr:t1	0	-0.001	-0.63

Supplementary Table 12.

Inter-variable correlation matrix of mixed model regressing change in interpersonal distance in experiment 1 on trustee type, total earnings, and their interaction.

	intercept	total earnings	trustee 1
total earnings	-0.967		
trustee 1	0.186	-0.417	
te:t1	-0.41	0.619	-0.969

Supplementary Table 13.

Inter-variable correlation matrix of linear mixed model regressing post trust game interpersonal distance in experiment 1 on trustee type, investment, trust rating, back-transfer expectation, and their interactions of the final two rounds of the trust game.

	intercept	trustee	investment	back-transfer expectation	trust rating	trustee1: investment	trustee1: back-transfer expectation
trustee1	-0.764						
investment	-0.422	0.319					
back-transfer expectation	0.063	-0.113	-0.725				
trust rating	-0.162	0.192	-0.452	0.159			
trustee1: investment	0.223	-0.379	0.156	-0.265	-0.373		
trustee1: back-transfer expectation	-0.056	0	-0.337	0.663	0.268	-0.677	
trustee1: trust rating	0.225	-0.204	-0.348	0.201	0.312	-0.505	0.157

Supplementary Table 14.

Inter-variable correlation matrix of mixed model regressing post trust game interpersonal distance in experiment 1 on trustee type, back-transfer, trust rating, and their interaction of the final two rounds of the trust game.

	intercept	trustee	back-transfer	trust rating	trustee1: back-transfer
trustee1	-0.61				
back-transfer	-0.426	-0.179			
trust rating	-0.089	0.322	-0.611		
trustee1: back-transfer	-0.247	-0.381	0.942	-0.53	
trustee1: trust rating	0.358	-0.127	-0.582	0.319	-0.649

Supplementary Table 15.

Inter-variable correlation matrix of mixed model regressing investment in experiment 2 on trustee type, round number, and their interaction.

	intercept	trustee 1	round nr
trustee 1	0.162		
round nr	-0.259	-0.001	
t1:round nr	-0.001	-0.415	0.002

Supplementary Table 16.

Inter-variable correlation matrix of mixed model regressing back-transfer expectation in experiment 2 on trustee type, round number, and their interaction.

	intercept	round nr	trustee 1
round nr	-0.548		
trustee 1	-0.05	-0.004	
round nr:t1	-0.003	0.004	-0.697

Supplementary Table 17.

Inter-variable correlation matrix of mixed model regressing certainty about back-transfer expectation in experiment 2 on trustee type, round number, and their interaction.

	intercept	round nr	trustee 1
round nr	-0.573		
trustee 1	0.195	-0.004	
round nr:t1	-0.003	0.004	-0.764

Supplementary Table 18.

Inter-variable correlation matrix of mixed model regressing trust ratings in experiment 2 on trustee type, round number, and their interaction.

	intercept	round nr	trustee 1
round nr	-0.44		
trustee 1	0.003	-0.003	
round nr:t1	-0.002	0.004	-0.594

Supplementary Table 19.

Inter-variable correlation matrix of mixed model regressing change in interpersonal distance in experiment 2 on trustee type, total earnings, and their interaction.

	intercept	total earnings	trustee 1
total earnings	-0.973		
trustee 1	0.35	-0.543	
te:t1	-0.539	0.708	-0.976

Supplementary Table 20.

Inter-variable correlation matrix of linear mixed model regressing post trust game interpersonal distance in experiment 2 on trustee type, investment, trust rating, back-transfer expectation, and their interactions of the final two rounds of the trust game.

	intercept	trustee	investment	back-transfer expectation	trust rating	trustee1: investment	trustee1: back-transfer expectation
trustee1	-0.749						
investment	-0.473	0.353					
back-transfer expectation	0.103	-0.117	-0.679				
trust rating	-0.293	0.216	0.101	-0.456			
trustee1: investment	0.269	-0.444	-0.28	0.213	-0.165		
trustee1: back-transfer expectation	0.044	-0.138	-0.109	0.396	-0.026	-0.408	
trustee1: trust rating	0.182	-0.248	0.062	-0.222	0.03	-0.137	-0.377

Supplementary Table 21.

Inter-variable correlation matrix of mixed model regressing post trust game interpersonal distance in experiment 2 on trustee type, back-transfer, trust rating, and their interaction of the final two rounds of the trust game.

	intercept	trustee	back-transfer	trust rating	trustee1: back-transfer
trustee1	-0.666				
back-transfer	-0.3	-0.023			
trust rating	-0.347	0.251	-0.371		
trustee1: back-transfer	0.018	-0.419	0.735	-0.094	
trustee1: trust rating	0.252	-0.343	-0.188	-0.062	-0.376

CHAPTER 5 - SLOW TOUCH TARGETING CT-FIBRES DOES NOT INCREASE PROSOCIAL BEHAVIOUR IN ECONOMIC LABORATORY TASKS

Original manuscript as published in Scientific Reports

DOI: doi.org/10.1038/s41598-018-25601-7

Lisa Anna Rosenberger^{1*}, Anbjørn Ree², Christoph Eisenegger¹, Uta Sailer²

1 Neuropsychopharmacology and Biopsychology Unit, Department of Basic Psychological Research and Research Methods, Faculty of Psychology, University of Vienna, Vienna, Austria,

2 Faculty of Medicine, Institute of Basic Medical Sciences, Dept. of Behavioural Sciences in Medicine, University of Oslo, Oslo, Norway.

*corresponding author

ABSTRACT

Field studies have demonstrated that humans become more generous, helpful and compliant after having been touched by another person. Here, we explored whether these effects are larger for touch activating the C-tactile (CT) fibres, as it is ascribed a particular role in establishing and maintaining bonds and affiliative interactions. The role of CT-targeted and non-targeted touch on prosocial behaviour was investigated in three different experiments using a trust game and a task measuring individual differences in social value orientations (the SVO task). Whereas participants in general acted prosocially, there was no influence of CT-targeted touch on prosocial behaviour, both in comparison to non-CT-targeted control touch and visual (non-tactile) stimulation. The null findings were further corroborated by Bayesian statistics. Thus, under the controlled laboratory conditions employed, CT-targeted touch did not play a particular role in prosocial behaviour. This indicates that touch does not increase prosocial behaviour in the absence of meaningful social and psychological connotations. Any touch related effects on prosocial behaviour likely depends on the ecological validity of the situation.

INTRODUCTION

Touch plays an important role in initiating and maintaining social affiliation and attachment, as has been shown for primates (Dunbar, 1993, 1998; Silk, 2002) rats (Broad, Curley, & Keverne, 2006), and humans (Debrot, Schoebi, Perrez, & Horn, 2013; Gullledge, Gullledge, & Stahmann, 2003; Triscoli, Croy, Olausson, & Sailer, 2017). Besides its role for establishing and upholding bonding, touch has also been found to increase helping behaviour and compliance with requests. As a now classic example, restaurant customers who were briefly touched by a waitress gave larger tips (Crusco & Wetzel, 1984; Stephen & Zweigenhaft, 1986), and supermarket shoppers were more willing to take part in alleged research interviews following touch (Hornik & Ellis, 1988). Touch also increased compliance with the request to look after a “large and very excited” dog while the owner went to a shop where dogs were prohibited (Guéguen & Fischer-Lokou, 2002). Many more field studies (Brockner, Pressman, Cabitt, & Moran, 1982; Joule & Guéguen, 2007; Kleinke, 1977) and one study using an economic game (Morhenn, Park, Piper, & Zak, 2008) showed that touch recipients are more likely to be unselfish and generous (for reviews, see Gallace & Spence, 2010; Schirmer, Wijaya, & Liu, 2016).

These studies have in common that they only applied one type of touch and compared it to no use of touch. However, this is not sufficient to determine what kind of touch is most effective (for example, tapping versus stroking), and what the underlying mechanisms may be. Moreover, being touched may induce arousal and other psychological (e.g., assumed affection) and physiological (e.g., heart rate increase) effects which would not be present in a control condition of no touch. Therefore, in order to understand the specific effects of touch, experimental approaches are needed that disentangle these potential factors.

Since most of these existing studies were field studies, there was also a limit to how well the variable touch could be controlled. However, different characteristics of the touch applied may lead to different effects on prosocial behaviour. Humans have both a rapid touch system, where signals are transmitted via myelinated, fast conducting A β afferent nerves, and a slow touch system consisting of unmyelinated C fibres with a conduction velocity about 50 times lower. A particular type of C fibres, the C-tactile (CT) fibres, are present only in the hairy skin of humans (Nordin, 1990; Vallbo, Olausson, & Wessberg, 1999; Wessberg, Olausson, Fernström, & Vallbo, 2003). They are believed to filter information relevant to hedonic or social aspects of gentle touch, before transmitting it to the brain where it is processed within affective brain networks (for a recent review, see McGlone, Wessberg, & Olausson, 2014). CT-fibres are maximally activated to stimuli at skin temperature (Ackerley et al., 2014) and by gently stroking the participant at a velocity similar to that of caressing (Croy et al., 2016; Löken, Wessberg, McGlone, & Olausson, 2009). Since they appear to

respond to stimuli signalling interpersonal caress, they are thought to represent a peripheral mechanism to promote affiliative behaviour (Ackerley et al., 2014; Morrison, Loken, & Olausson, 2010). Indeed, the “social touch hypothesis” assumes a relationship between CT-targeted touch and bonding, attachment, and social perceptions (McGlone et al., 2014). Given these previous findings and theoretical assumptions, it seems plausible to hypothesise that the positive effect of touch on social behaviour is mediated via CT-fibres. The present study set out to test this hypothesis in an experimentally controlled laboratory setting, and to establish the specific role of CT-targeted tactile stimulation in prosocial behaviour. To this end, we conducted three experiments with two different economic tasks and different control conditions under controlled laboratory conditions. Based on the literature review above, we hypothesised that prosocial behaviour would be enhanced following CT-targeted touch.

To determine the influence of CT-targeted touch on prosocial behaviour, we employed a repeated trust game (Berg, Dickhaut, & McCabe, 1995) in two experiments, a classic two-player task of behavioural trust and reciprocity. The participants always played in the role of trustee, where they had to make decisions on how much of the entrusted money they wanted to send back to their interaction partner. Thus, the prosocial measure in experiment 1 and 2 was the back-transfer amount, or in other words the amount of reciprocity displayed to the interaction partner. The other player was either identical to the person providing touch (experiment 1), or not (experiment 2). In experiment 3 we employed a Social Value Orientation task (SVO (Murphy, Ackermann, & Handgraaf, 2011)) which measures prosocial behaviour in terms of concern for other people’s pay-off. Here, the participant makes decisions on how to allocate a monetary amount between themselves and an anonymous other. In all three experiments, the monetary decision of the participant was measured after having either received CT-targeted touch (3cm/s) or control touch (tapping in experiment 1 and 2, and very slow touch in experiment 3), both executed with a soft brush. A brush was preferred to the more ecologically valid hand, because previous studies have shown reliable CT-fibre activation using a brush (Ackerley et al., 2014; Löken et al., 2009). Moreover, a brush allows for more controlled tactile stimulation, since it is independent of skin temperature and humidity, and also because the force can be better monitored via the bending of the hairs.

In experiment 3, an additional control condition with a moving visual stimulus was employed in order to investigate whether a potential increase in prosociality was specific to the tactile stimuli, or whether it is just the pleasantness of any sensory stimulus that leads to an effect. In order to provide evidence to support the null hypothesis of no differences between conditions, Bayesian analysis was performed. Instead of computing a p-value, Bayesian analysis computes the odds favoring the null hypothesis or a plausible alternative in light of the data (Gallistel, 2009). This allows

us to determine the strength of the evidence for CT-targeted touch (not) having an effect on prosociality.

RESULTS

EXPERIMENT 1

Pleasantness for CT-targeted touch was rated higher than for control touch ($F(1, 98) = 6.907, p < 0.001$). Intensity was not rated as being different for the two touch types ($F(1,98) = 0.053; p = 0.818$). The return ratio was larger with a multiplication factor of 6, that is when the investment was multiplied with 6, than with a multiplication factor of 3 ($\chi^2(1) = 53.553, p < 0.001$), but it did not differ for touch types ($\chi^2(1) = 0.098, p = 0.754$), or pleasantness ($\chi^2(1) = 0.578, p = 0.447$). There was also no interaction between multiplication factor, touch type, or pleasantness (see supplementary material Results 1 for details on the results, and Table 1 for means and standard deviations of the ratings and return ratios).

EXPERIMENT 2

As in experiment 1, pleasantness for CT-targeted touch was rated higher than for control touch ($F(1, 156) = 29.65, p < 0.0001$). Intensity was not rated as being different for the two touch types ($F(1, 156) = 0.022, p = 0.882$). As in experiment 1, the return ratio was larger with a multiplication factor of 6, than with a multiplication factor of 3 ($\chi^2(1) = 131.020, p < 0.001$). Pleasantness was found to influence the return ratio ($\chi^2(1) = 4.602, p < 0.05$): higher pleasantness ratings were associated with higher return ratios. The return ratio was not different for the two touch types ($\chi^2(1) = 0.441; p = 0.506$), and there was also no interaction between multiplication factor, touch type, or pleasantness (see supplementary material Results 1 for details on the results, and Table 1 for means and standard deviations of the ratings and return ratios).

POOLED DATA FROM EXPERIMENT 1 AND 2

The data from experiment 1 and 2 were pooled in order to investigate potential differences between the experiments, and to additionally explore the effect of hypothesis guessing.

PLEASANTNESS AND INTENSITY

CT-targeted touch was rated as more pleasant than control touch (see Figure 1) ($F(1, 250) = 34.819, p < 0.0001$), with a Bayes factor (BF) of 7.714. Thus, there is almost 8 times more evidence for a difference in pleasantness ratings between CT-targeted touch and control touch, than there is for no difference between the two. At the same time, both touch types were rated as equally intense ($F(1, 250) = 0.066, df = 1, p = 0.798$), with a BF of 0.107. There were no differences in pleasantness or

intensity ratings between experiment 1 and 2, nor was there an effect of hypothesis guessing on the ratings (all $F < 2.139$).

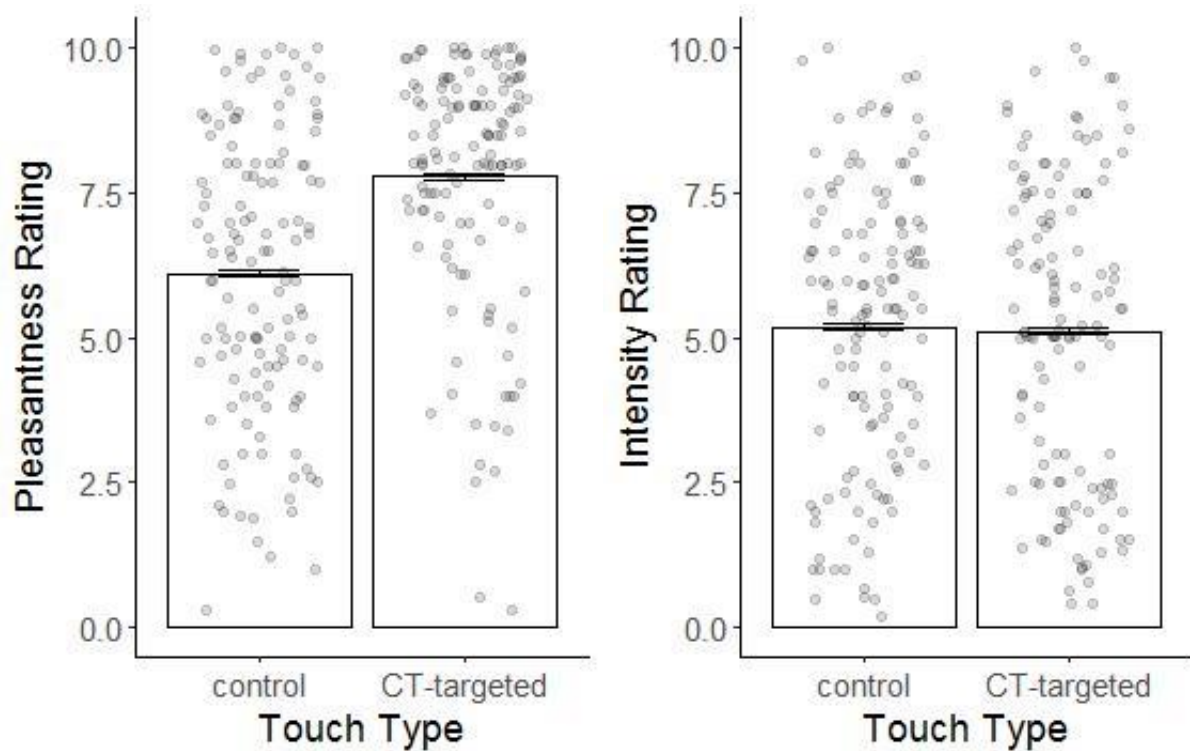


Figure 1. Mean pleasantness and intensity with within-subject standard error of the mean (SEM) and individual single-trial ratings of experiment 1 and 2 for CT-targeted and control touch.

RETURN RATIO

Return ratio following both types of touch was larger than 1 (see Figure 2), indicating that participants on average returned more than the investor had invested. However, there was no general difference in the return ratio depending on the type of touch ($\chi^2(1) = 0.173$, $p = 0.678$, see supplementary material Results 2 for a full model description). The Bayes factor analysis revealed a BF of 0.098 for touch type differences, thus, 10.25 times more evidence for the null hypothesis of no difference between the two touch types. The BF for touch type differences with multiplication factors of 3 and 6 were both 0.136. Again, this is much stronger evidence (7.34 times more) for an absent difference between the two touch types. Furthermore, the BF on the interaction between pleasantness and touch type (thus evidence that a return ratio difference between the touch types depends on pleasantness ratings) was 0.225. Thus, in contrast to our hypothesis, both the mixed model analysis and the Bayes factor analysis indicate that there was no effect of touch type on return ratios.

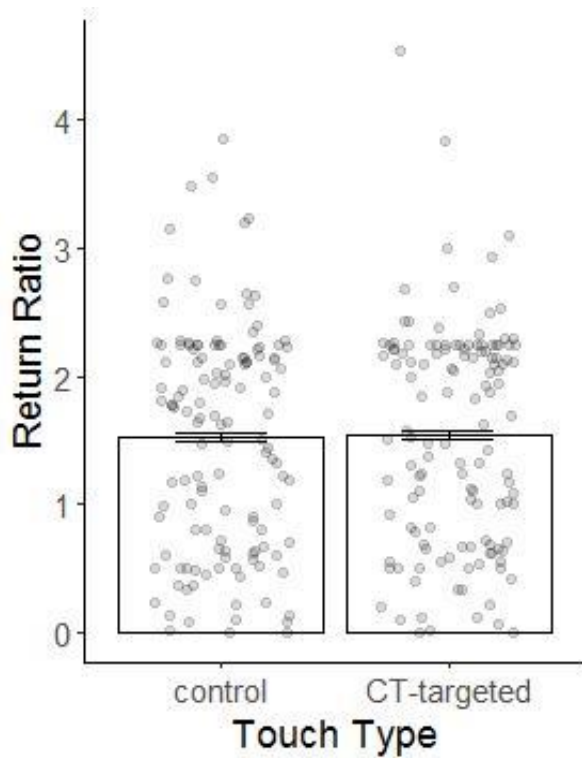


Figure 2. Mean return ratio with SEM and individual data points (averaged for each participant from experiment 1 and 2) for CT-targeted and control touch.

As to hypothesis guessing, 6 participants in experiment 1 (12%) and 14 participants (18%) in experiment 2 correctly guessed the hypothesis of our study despite the cover story. Hypothesis guessing affected the return ratio, as shown by a significant interaction between multiplication factor and hypothesis guessing ($\chi^2(1) = 15.789, p < 0.01$), as well as a significant interaction between multiplication factor, hypothesis guessing and pleasantness ($\chi^2(1) = 7.226, p < 0.01$, see supplementary material Results 2). To further explore this 3-way interaction, an additional analysis was performed separately for participants who guessed and did not guess the hypothesis. For participants who guessed the hypothesis, there was a significant interaction between multiplication factor and pleasantness ($\chi^2(1) = 3.999, p < 0.05$), see Figure 3. In trials with higher multiplication factor, the return ratio increased with increasing pleasantness ratings.

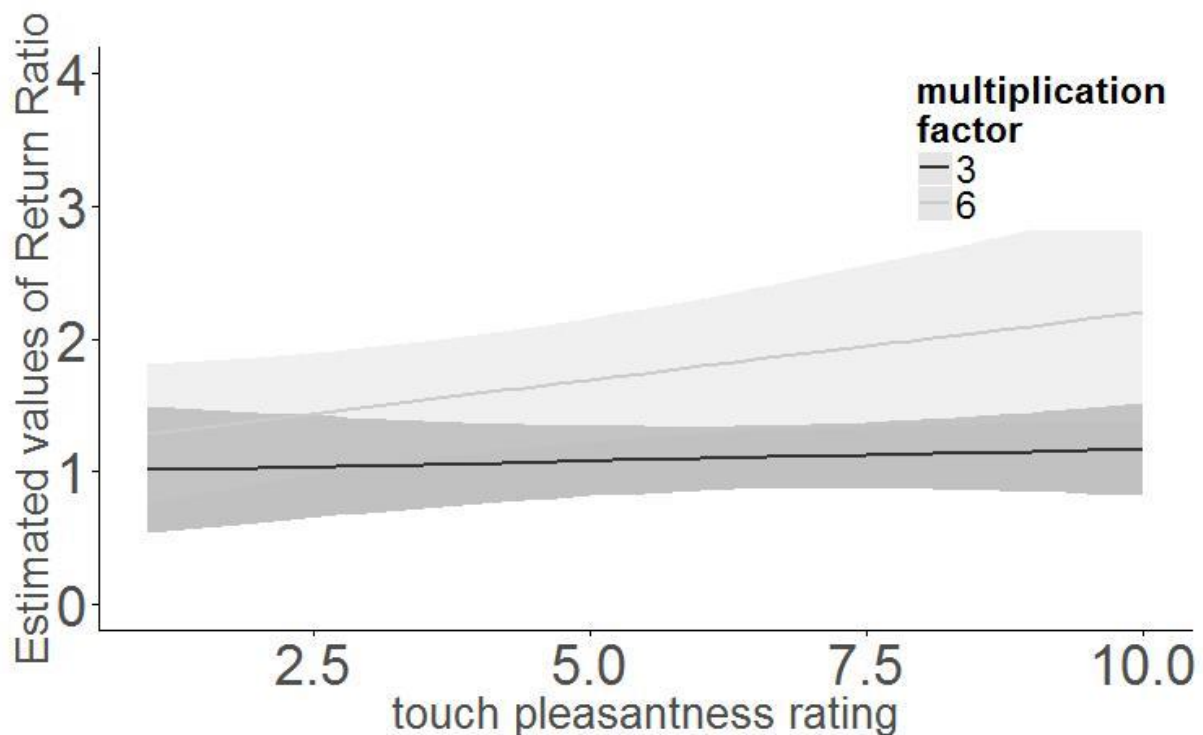


Figure 3. Regression line of estimated return ratios from linear mixed model analysis with 95% confidence intervals (shaded areas) for touch pleasantness ratings (both CT-targeted and control touch) of participants who guessed the study's hypothesis split by multiplication factor.

For participants who did not guess the hypothesis, there was a significant interaction between multiplication factor, touch type, and pleasantness ($\chi^2(1) = 4.912, p < 0.05$). Calculating separate analyses for both touch types (see supplementary material Results 3 for full model description) showed that this interaction was due to an effect of control touch, where higher pleasantness ratings in trials with a high multiplication factor were associated with lower return ratios (interaction between multiplication factor and pleasantness; $\chi^2(1) = 13.234, p < 0.001$), see Figure 4.

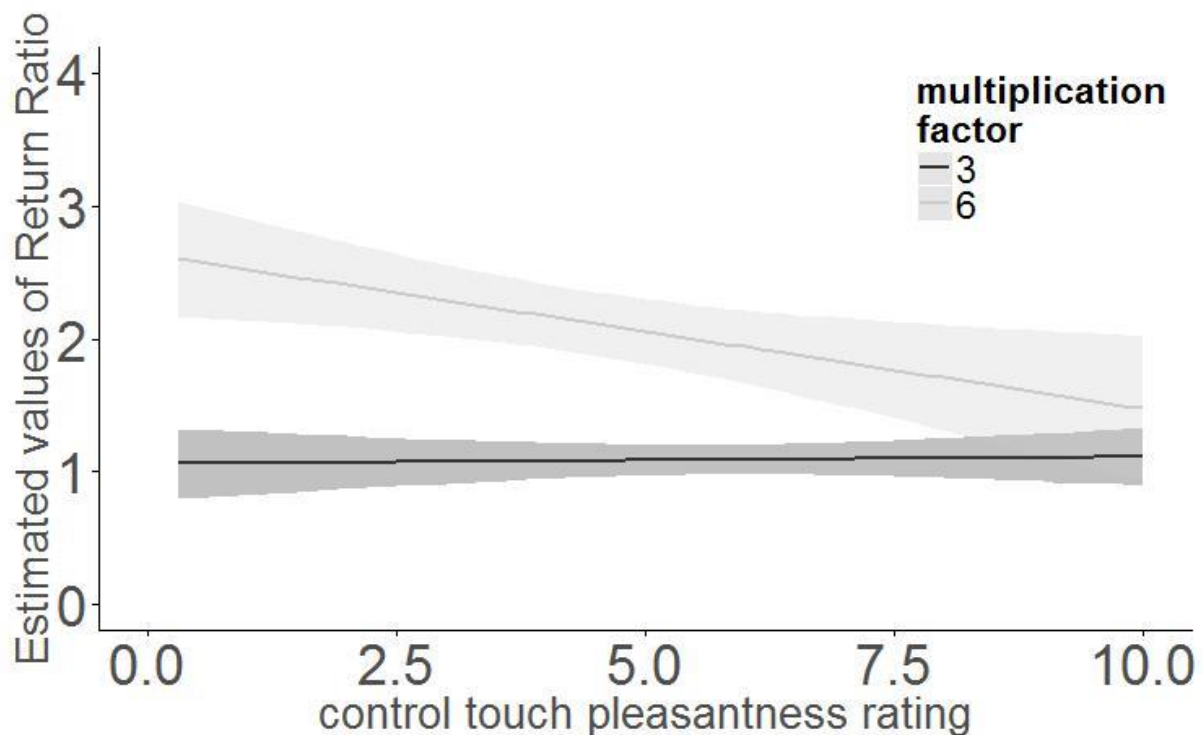


Figure 4. Regression line of estimated return ratios from linear mixed model analysis with 95% confidence intervals (shaded area) for different control touch pleasantness ratings of participants who did not guess the study's hypothesis split by multiplication factor.

Thus, whereas participants who correctly guessed the hypothesis returned more money the more pleasant they evaluated the touch, participants who did not guess the hypothesis returned less money the more pleasant they evaluated the touch. The latter did, however, only apply to control touch and with the higher multiplication factor.

EXPERIMENT 3

As in experiment 1 and 2, pleasantness was rated differently for the three types of stimulation (see Figure 5, $\chi^2(2) = 6.961$, $p < 0.05$), with CT-targeted touch being evaluated as more pleasant than control touch ($t(86)=2.542$, $p < 0.05$). Bayes factor analysis revealed that there is 2.135 times more evidence for a difference between the two touch types than no difference between the two ratings. However, CT-targeted touch was not rated as more pleasant than visual stimulation ($t(86) = 0.661$, $p = 0.787$), which is supported by a BF of 0.257. There was a trend for intensity to be rated differently for the three types of stimulation ($\chi^2(2) = 5.639$, $p = 0.060$), which was driven by tendentially higher intensity ratings for CT-targeted touch than for control touch ($t(86) = 2.371$, $p = 0.052$). However, Bayes factor analysis actually does not support the existence of a difference in intensity ratings between the three types of stimulation ($BF = 0.242$).

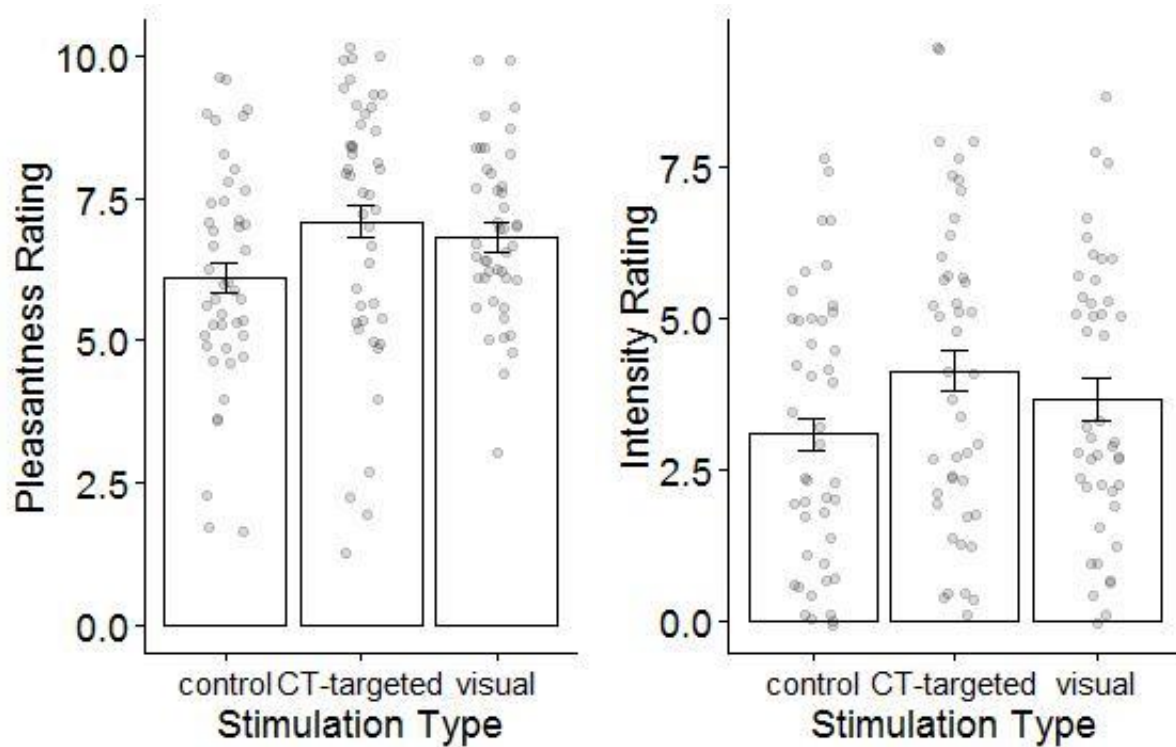


Figure 5. Mean pleasantness and intensity with within-subject SEM and individual ratings for different sensory stimulation.

Following all three types of stimulation, all of the participants had an SVO angle that classified them as either individualistic or – the majority – as prosocial (see Figure 6). None of them fell into the competitive or altruistic range. For three participants following control touch and for one participant following visual stimulation, an SVO angle could not be calculated because of inconsistent behaviour.

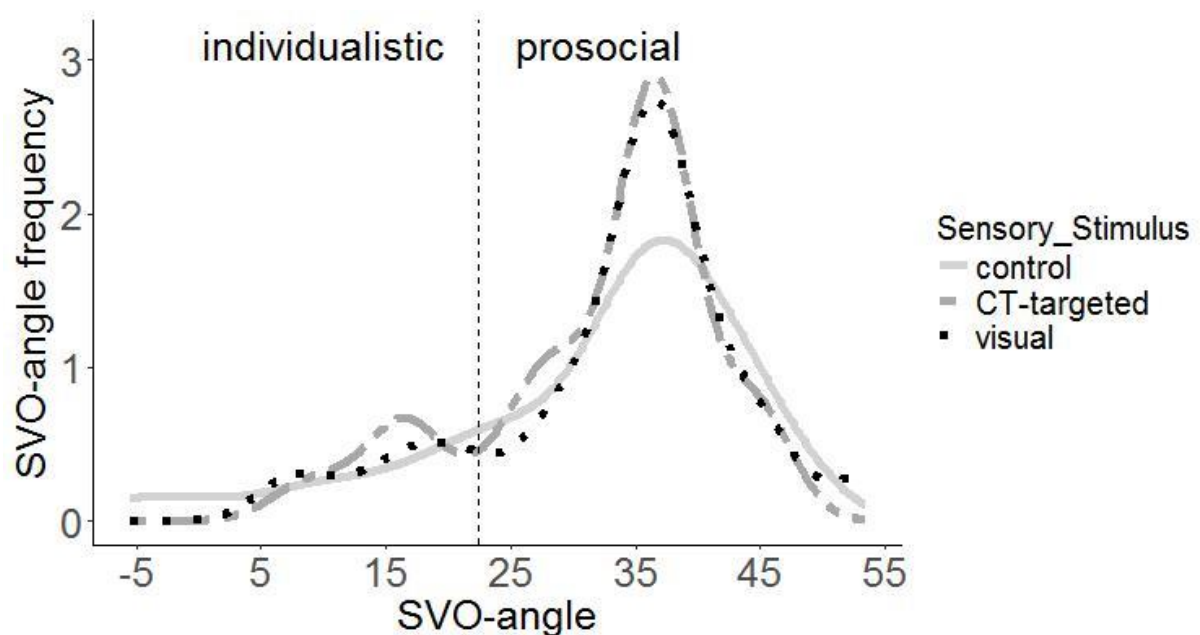


Figure 6. Distribution of SVO angles following different sensory stimulation. Behavioural categorization is based on the following ranges: “altruistic” (> 57.15 deg), “individualistic” (between -12.04 – 22.45 deg), “prosocial” (between 22.45-57.15 deg), or “competitive (< -12.04 deg)”.

SVO angles did not change depending on type of stimulation, or its pleasantness. There was only a tendency for the SVO angle to differ between the three types of stimulation ($\chi^2(2) = 5.756, p = 0.056$, see supplements Results 4). No other effects or interactions were significant. Post-hoc tests revealed that the SVO angle was higher following visual stimulation ($M = 33.519, SD = 10.470$) than control touch ($M = 31.463, SD = 12.491; t(78.52) = -2.662; p < 0.05$). The SVO angle for CT-targeted touch ($M=32.148; SD=9.491$) was in-between visual stimulation and control touch, and did not differ significantly from any of them (all $t_s < 1.53$). The BF for differences in the SVO angle between stimulation types was 0.104. Thus, the BF revealed much stronger evidence (9.593 times stronger) for an absent difference between the stimulation types. The BF for differences in SVO between stimulation types depending on pleasantness (thus the stimulation type x pleasantness interaction) was 0.146, and therefore, gave strong evidence for an absent effect. Thus, as in experiment 1 and 2, and against our hypothesis, CT-targeted touch did not increase prosocial behaviour.

DISCUSSION

The goal of the study was to test the hypothesis that slow brushing with an optimal velocity to activate CT-fibres increases prosocial behavior. To test the hypothesised prosociality-enhancing effect of touch, three experiments with two different behavioural economic laboratory tasks and different control conditions were performed. However, following CT-targeted touch (and compared to control touch or visual stimulation) neither reciprocity to the person who touched the participant, nor reciprocity to an unrelated person in the trust game were altered. Additionally, prosocial dispositions as measured with the SVO also did not shift.

In general, our data replicated some typical findings from both the Repeated Trust Game and the SVO. For example, it is well-established that participants are willing to return positive amounts (Camerer, 2011; Johnson & Mislin, 2011). It has also been shown previously that the same participants return a larger amount if the investment has been multiplied by a larger factor compared to a smaller factor (Ackert, Church, & Davis, 2011; Güth, Muegera, Musau, & Ploner, 2012; Mislin, Williams, & Shughnessy, 2015) (however, see Johnson and Mislin (2011) for opposite results in between-subjects designs). We were also able to replicate the effects from the Repeated Trust Game during a second experiment, where the change in some experimental details did not affect the participants' behaviour. Importantly, experiment 2 showed that the absent effect of CT-targeted

stimulation on trust game behaviour in experiment 1 was not due to playing with the experimenter, or to conducting the task in an Excel-sheet.

Based on their SVO-angle, the majority of participants were classified as prosocial, and the remainder as individualistic. That is, prosocial participants chose to allocate the money in a manner that evenly distributed the money between “self” and “other”, as opposed to individualistic participants who opted to allocate a larger portion of the allocated money to themselves. This is in accordance with findings from the original study (Murphy et al., 2011), where 59% of participants were prosocial, 35% were individualist and the remainder competitive. Thus, our participants seem to behave similarly to other healthy participants in studies with the same tasks.

Although we were able to replicate basic findings of both tasks reported in literature, and from the Repeated Trust Game in two experiments with different samples, our data did not support our main hypothesis regarding CT-targeted touch. Many field studies in contexts as different as stores, restaurants, the street and the classroom demonstrated that casual touch leads to more favourable attitudes and behaviour towards the toucher, as well as to larger compliance and increased helping behaviour (for reviews, see Gallace & Spence, 2010; Schirmer et al., 2016). Our data did not provide evidence for increased prosocial behaviour following CT-targeted touch, or following touch compared to visual stimulation. This was despite the fact that CT-targeted touch was consistently experienced as more pleasant than control touch. Such an absent influence of different touch types on behaviour, despite differences in their experienced pleasantness, is in line with findings that gentle touch in general, irrespective of CT-optimality, elicits affiliative emotions (Kirsch et al., 2017).

CT-fibres are known to show depressed responses with repeated stimulation, a phenomenon called fatigue. However, human CT's continue to respond to repeated touch (Nordin, 1990; Vallbo et al., 1999) as opposed to non-human animals, whose responses quickly depress with repeated mechanical stimulation (Iggo & Kornhuber, 1977). When stimulated every 3 seconds, CT responses in humans became shorter, but maintained the initial peak firing frequency (Vallbo et al., 1999). This may suggest that the extended brushing period on the same area on the arm in experiment 1 and 2 led to decreased CT responding. However, in experiment 3, stroking was alternated between 2 different areas on the arm, which did not lead to different pleasantness responses. Moreover, each brush stroke can be expected to activate a large number of CTs. Therefore, it appears unlikely that the absent effect of CT-targeted touch on prosocial behaviour is due to fatigue of CT-fibres. Nevertheless, as little is known about fatigue in CTs and its effect on subjective experience, further studies are warranted.

Regarding the comparison of tactile versus visual stimulation, prosocial behaviour was not different for visual stimulation either. Interestingly, participants showed opposing effects of touch on prosocial behaviour during the Repeated Trust Game, as measured with the return ratio, depending on whether they guessed the hypothesis correctly or not. This highlights the general importance of assessing the presence of hypothesis guessing when using the trust game. It needs to be taken into consideration, however, that the specific sample in experiment 1 and 3 may have increased the tendency and particularly the accuracy to guess the hypothesis, as it contained a number of psychology and medical students. Both can be expected to know about the use of deception in experiments and/or have some knowledge about the tactile system. However, in experiment 2 only a minority of the participants were students, but still the results were similar to the ones from experiment 1.

Besides showing no advantage of CT-targeted touch compared to control touch in eliciting prosocial behaviour, the findings of experiment 3 also documented a lack of enhancement of prosocial behaviour following touch compared to a visual control condition. This was a rather surprising finding. One reason for this absent difference in prosocial behaviour might lie in the experienced pleasantness of the two stimulations which were rated as similarly pleasant. However, comprehensive analyses showed that pleasantness did not influence the SVO angle. It therefore seems that it was not the similarity in pleasantness for CT-targeted touch and visual stimulation that underlay the absent difference in prosocial behaviour. Similarly, arousal cannot serve to interpret the results, since different arousal following the different types of stimulation would be expected to alter intensity ratings, which was not the case. What could then be the reason for the failure to replicate the field findings of increased prosocial behaviour following touch in our study? It could be speculated that an individual receiving touch may, consciously or subconsciously, interpret the meaning of this behavior (Jakubiak & Feeney, 2017). Accordingly, being touched by a waitress or a person asking for a favour in a natural context may be interpreted as a sign of affection (Floyd, 2006) or need (Rose, 1990). Receiving touch may also indicate physical and emotional closeness to the toucher, social inclusion, and/or felt security (Jakubiak & Feeney, 2017). This may in turn lead to positive feelings and beneficial physiological changes. It appears likely that the effect of touch on prosocial behaviour is dependent on one or several of these mediators that we were unable to replicate in a laboratory setting. Moreover, it is well-known that CT-fibres are subject to modulation from higher-order brain regions via descending influences (e.g. Melzack & Wall, 1965). Therefore, affective processing of CT-targeted touch is also influenced by top-down cognitive modulation (as reviewed in Ellingsen, Leknes, Løseth, Wessberg, & Olausson, 2016). For example, the assumed sex of the touch provider (Gazzola et al., 2012; Scheele et al., 2014), concomitant visual (Ellingsen et al.,

2014) and olfactory (Croy, D'Angelo, & Olausson, 2014) stimuli have all been found to influence the response to soft touch. Thus, early sensory processing will be altered by contextual mechanisms related to aspects such as expectations or the interpretation of touch.

To our knowledge, previous studies investigating the effect of touch on prosocial behaviour in economic games are scarce (Koppel, Andersson, Morrison, Vastfjall, & Tinghog, 2017; Morhenn et al., 2008; Spapé, Hoggan, Jacucci, & Ravaja, 2015). One recent study compared the effect of vibratory touch, an auditory cue or silence on behaviour in the Ultimatum game (Spapé et al., 2015). In this game, one player offers a portion of an endowment to another player who can either accept the offer or reject it. If the second player accepts the offer, the money is divided according to the proposed division, otherwise, both players leave with nothing. The authors found higher acceptance rates of offers following vibratory touch than silence, but no difference in acceptance rates following touch compared to sound (Spapé et al., 2015). In a different study on emotional face processing (Spapé, Harjunen, & Ravaja, 2017), the authors concluded that auditory and tactile primes rather seemed to affect stimulus evaluation in general by facilitating cognitive processing instead of having differential effects. Such a general facilitating effect may also exist for visual and tactile stimuli in the present study.

In a different study, massage increased reciprocity in a one-shot trust game (Morhenn et al., 2008), with participants returning more than twice as much following massage than following rest. However, in that study the difference between the two conditions may have been maximized because of the length and high intensity of the massage on the one hand, and because of the low arousal and possibly boredom in the rest condition on the other hand. Nevertheless, in comparing touch to “nothing”, the design of this study is similar to the many field studies on touch effects on prosociality, which may explain why similar results were found. The current control condition with visual stimulation necessitated task engagement, because participants were required to rate the videos and therefore, to attentively watch them. This raises the possibility that task engagement, or the lack thereof, may have an influence on prosociality. It has been shown previously that boredom can affect decision-making in terms of choosing a simpler strategy (Slovic, Lichtenstein, & Edwards, 1965). A potential effect of boredom on prosocial behaviour could be an interesting question to investigate in further studies.

To conclude, it has been suggested that CT-fibres constitute a specific pathway for signaling touch with social-affective meaning (Morrison et al., 2010; Olausson, Wessberg, McGlone, & Vallbo, 2010). Our data suggest that touch that particularly activates this pathway does not influence social behaviour in the present laboratory context. We reason that in order to obtain such an effect, the

social context may be crucial. Potentially, touch needs to be understood as expressing affection, not as a rather neutral stimulation as part of an experiment. Therefore, future studies on similar research questions should increase the potential affiliative meaning even in the laboratory context.

METHODS

PARTICIPANTS

Only participants who had not previously participated in other touch or economics related experiments were eligible for the studies. This was assessed when appointments were booked. For experiment 1, 50 healthy participants, 29 women and 21 men, were recruited via announcements on boards at the University of Gothenburg, Sweden. They were aged between 19 and 72 years, with a mean age of 30 years ($SD = 12$), and the majority of them were students. For experiment 2, 79 participants, 56 women and 23 men, were recruited via announcements on university boards and on social media. They were aged between 19 and 76 years, with a mean age of 41 years ($SD = 19$). 8 of them were medical students and 2 were psychology students. Following experiment 1 and 2, participants were asked what they thought the aim of the study was. The response was registered and coded as hypothesis guessing yes/no to be used in later analyses. Subsequently, participants were fully debriefed about the real aim and all deceptive elements in the experiments. All participants received financial compensation for participation which included a show-up-fee of 200 SEK (~20 EUR) and an additional amount that depended on their actual choices during the Repeated Trust Game (see detailed description below).

For experiment 3, 45 healthy participants were recruited by announcements in various academic institutions in Oslo, Norway. The theme of the announcements focused on how various forms of sensory stimuli could potentially affect decision making, empathy and perceptions of various phenomena. Most participants (approximately 75%) were physiotherapy, osteopathy, psychology or medical students. One participant was subsequently excluded due to an acute bout of rheumatoid arthritic joint pain reported in the written consent form. Thus, a total of 27 women and 17 men, aged between 18-47 ($M=25$ years, $SD=6$) completed the study. Participants were informed that both their own and a future participant's compensation for participating in the study could be affected by how they chose to allocate the money. However, during the debriefing, the participants were informed that the deception of "self" and "other" was included to ensure full commitment to the procedure of distributing money and all participants received a voucher with a value of 200 NOK (~20 EUR).

All participants provided informed consent before participating in the study. Experiments 1 and 2 were approved by the regional ethical board at the University of Gothenburg, Sweden (Dnr288-15), and experiment 3 by the regional committee for medical and health research ethics South East, Norway (2011/1337), and all were carried out according to the Declaration of Helsinki.

PROCEDURE

EXPERIMENT 1

After having signed the informed consent, participants were asked to fill in a form about demographics, potential skin diseases and medication use. The task procedure was similar in experiment 1 and 2 and consisted of the following blocks: tactile stimulation 1, Repeated Trust Game 1, filler task, tactile stimulation 2, and Repeated Trust Game 2. The order of tactile stimulation type (CT-targeted touch and control touch) was counterbalanced between participants. Participants were told that the aim of the experiment was to investigate the effect of different sensory experiences, among which tactile stimulation and hunger, on gaming behaviour. To increase the credibility of the cover story, they were also told not to eat for at least 2 hours prior to the experiment, and later answered two items about their hunger feelings. During the experiment, participants were seated at a table with a PC where all the tasks were administered. The experimenter sat next to the participant at the same table during the tactile stimulations and the repeated trust games. At the end of the experiment, participants were asked what they thought the aim of the experiment was, and if they had noticed anything particular, and were subsequently debriefed and paid.

TACTILE STIMULATION

Participants were asked to put their left forearm on a vacuum pillow which lay on a table and was adjusted to their arm. A distance of 10 cm was marked with a pen in the middle of the dorsal surface on their forearm. They were asked to put on occluding glasses to block view of their arm, to look straight ahead and to concentrate on the upcoming tactile stimulation. Using a hand-held soft watercolour brush that was 75mm wide, the experimenter then delivered one of two different tactile stimulations. In the experimental condition “CT-targeted touch”, repeated brushing-strokes were performed in the proximal-distal direction for 3 minutes with a 0.5s break in-between strokes, i.e. 32 stroking trials altogether. Brushing was performed with a velocity of 3cm/sec and an approximate force of 0.4 N. Brushing with these parameters has previously been found to maximally activate CT-fibres (Ackerley et al., 2014; Löken et al., 2009). The experimenter was trained in delivering this type of stimulation by performing the brush strokes on a scale and under the guidance of a bar on a computer monitor moving at 3 cm/s to maintain constant pressure and velocity. During the

experiment, the experimenter was also guided by the same bar shown on a monitor not visible to the participant.

In the control condition “control touch”, the experimenter had the brush “jump” across the marked area in four equally large steps, i.e. giving the impression that the arm was being prodded at four different positions within a 10 cm distance. The four hopping steps took as much time as one brushing stroke and were performed with a similar force and in the same direction as the brushing, and on the same area of the arm. This stimulation was also performed for 3 minutes. This kind of stimulation was chosen because it was assumed to activate CT-fibres to a lesser extent than slow stroking, while at the same time providing an experience that qualitatively was considerably distinct from stroking in order to increase the probability of an effect. It has previously been shown that this kind of tactile stimulation is perceived as less pleasant than stroking with CT-targeted parameters (Kress, Minati, Ferraro, & Critchley, 2011).

Whereas the majority of field studies on prosocial touch effects used short stimulation, a longer stimulation period of 3 minutes was chosen in order to increase the reliability of ratings, and because previous studies of our group showed that pleasantness and physiological effects of CT-targeted touch change very slowly with repeated stimulation (Sailer et al., 2016; Triscoli, Ackerley, & Sailer, 2014; Triscoli, Croy, Steudte-Schmiedgen, Olausson, & Sailer, 2017). 3 minutes is also an interval in-between the short tactile interventions in field studies and the 15 minute massage in a different laboratory study (Morhenn et al., 2008).

Following each type of stimulation, participants were asked to grade the pleasantness and intensity of the stimulation on a visual analogue scale (VAS) on a paper. Above the VAS for pleasantness, the question “How did you experience the touch?” was written, and the endpoints were “unpleasant” (scored as -10) and “pleasant” (+10). Above the VAS for intensity, the question “How intensive did you experience the touch?” was written, and the endpoints were “weak” (-10) and “strong” (+10).

REPEATED TRUST GAME

The trust game (Berg et al., 1995) is a frequently used and replicated measure of behavioural trust and reciprocity involving the sequential exchange of a monetary endowment (Johnson & Mislin, 2011). It is played by two players, an investor and a trustee. The investor, who is the first mover, receives an endowment and can transfer a variable amount from his/ her endowment to the trustee. This investment gets multiplied by a specific factor, called a multiplication factor, by the experimenter. For example if the multiplication factor is 2, the investment gets doubled by the experimenter. The trustee receives this multiplied investment and can make a back-transfer to the

investor based on the multiplied investment (Berg et al., 1995). The amount passed by the investor is assumed to capture trust, and the amount returned by the trustee is assumed to reflect reciprocity (Berg et al., 1995).

Participants played this game with the experimenter, i.e. the person touching them. Participants always played in the role of trustee, but were led to believe that this was assigned randomly with a lottery drawing. The endowment size was set to 100 SEK (~11 EUR) and the multiplication factor was either 3 (round 1 – 5) or 6 (round 6 – 10). Participants played with the same investor for 10 rounds. Investments were fixed to 70, 90, 80, 70, and 80 SEK for rounds 1 - 5, and were repeated in rounds 6 – 10. Participants were instructed that the investor would not see the size of the back-transfer in each round, but would only see the final result of all trials at the end of the experiment. This was done so that participants did not believe their choice would influence the investor's allocation in the next trial.

Participants were also informed that 10% of the amount they had decided to keep in two randomly selected rounds would be added to their show-up-fee. In fact, these two rounds were not randomly selected, but were the fifth round from block 1 and the last round from block 2. They were also told that the investor would receive the amount transferred back by the trustee from these randomly selected rounds, so that the actual payoff for both players depended on the trustee's behaviour in the game.

The game was implemented in an Excel sheet (see Figure 7). Prior to the actual experiment, participants played 2 practice rounds (analogous to the “real rounds” during the experiment) in order to get familiar with the procedure.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Period	I received from player 1	Thus I have				I send back			My profit	Player 1's profit					
2																
3	Practice	85		255 kr			100 kr			155 kr	100 kr					
4	Practice	37		111 kr			15 kr			96 kr	15 kr					
5																
6																
7	1	70		210 kr			kr			210 kr	0 kr					
8	2			0 kr			kr			0 kr	0 kr					
9	3			0 kr			kr			0 kr	0 kr					
10	4			0 kr			kr			0 kr	0 kr					
11	5			0 kr			kr			0 kr	0 kr					
12																
13																
14	6			0 kr			kr			0 kr	0 kr					
15	7			0 kr			kr			0 kr	0 kr					
16	8			0 kr			kr			0 kr	0 kr					
17	9			0 kr			kr			0 kr	0 kr					
18	10			0 kr			kr			0 kr	0 kr					
19																
20														Total profit	210	
21																

Figure 7. Implementation of the Repeated Trust Game in Experiment 1 (originally in Swedish). Participants always saw the complete Excel sheet. In a given round, the participant received a paper with the investment decision from the experimenter and filled this amount in the column “I received from player 1”. The multiplied amount was automatically displayed in the “Thus I have” column. Then the participant filled in his back-transfer decision in the “I send back” column, which led to the automatic calculation and display of the players profits in the last two columns. The Excel sheet was only visible to the participant throughout the task.

FILLER TASK

Participants were asked to repeatedly estimate the length of a 1 second interval and were given a sweet or salty snack, depending on the participant’s preferences, during 3 breaks. Subsequently, participants again answered a question on paper about how hungry they felt. The snack as well as this question was given in order to support the cover story. Another aim of this filler task was to distract the participants in order to avoid carry-over effects from the first stimulation. The break in between the two conditions was on average 31 min ($SD = 6$).

EXPERIMENT 2

Experiment 2 served to replicate the results from experiment 1 with a larger sample and revised procedure. To increase differences between the two stimulation types, a watercolour brush was used to deliver CT-targeted touch as in experiment 1, but the control touch was delivered using a cooled-down steel brush. This was done in order to minimise CT-fibre responding, since they are reported to respond preferentially to tactile stimuli at skin temperature (Ackerley et al., 2014). In experiment 2, participants sat on a hospital bed with a keyboard on their lap and a screen at the height of their eyes in front of them, see Figure 8. The experimenter was sat to the left of the participants with a dividing curtain in-between, so that participants were unable to see the experimenter/ the other alleged player at all times. In addition, the participant wore occluding glasses that prevented peripheral vision.



Figure 8. Setup of experiment 2. The participant faced a monitor. The experimenter was seated beside the participant and stroked the participant's left dorsal forearm with a paintbrush or tapped it with a steel brush.

In order to reduce a potential influence of the experimenter's impression on the participants' behaviour in the Repeated Trust Game, the participants were instructed that they would play the Repeated Trust Game with a colleague of the experimenter, a person they would neither see nor talk to during the experiment. The participants' view of the experimenter and their own arm was obscured by a dividing curtain. Unbeknown to the participant, the experimenter played the investor. The experimenter opened and closed doors, walked and breathed distinctly in the role of the investor, to increase believability of the deception. Prior to post-experimental de-briefing, the participants were asked about their impression of the investor. None of them mentioned any suspicions about the investor being the same person as the experimenter.

The design of the repeated trust game was exactly the same as in experiment 1, except that the task was now implemented in the software z-tree (Fischbacher, 2007) instead of in an excel sheet (see Figure 9). Instead of entering back-transfer decisions in an excel sheet, participants now clicked through an interactive task. We opted for a different task interface for the participants to ensure that the results of experiment 1 were not due to the rather bare, mathematical layout of the excel sheet, and also because a standardised and commonly used way of presentation was preferred (see e.g. Baran, Sapienza, & Zingales, 2010; Fehr & Gächter, 2000; Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005). Again, participants always played in the role of trustee, but were led to believe that this was assigned randomly by ztree.

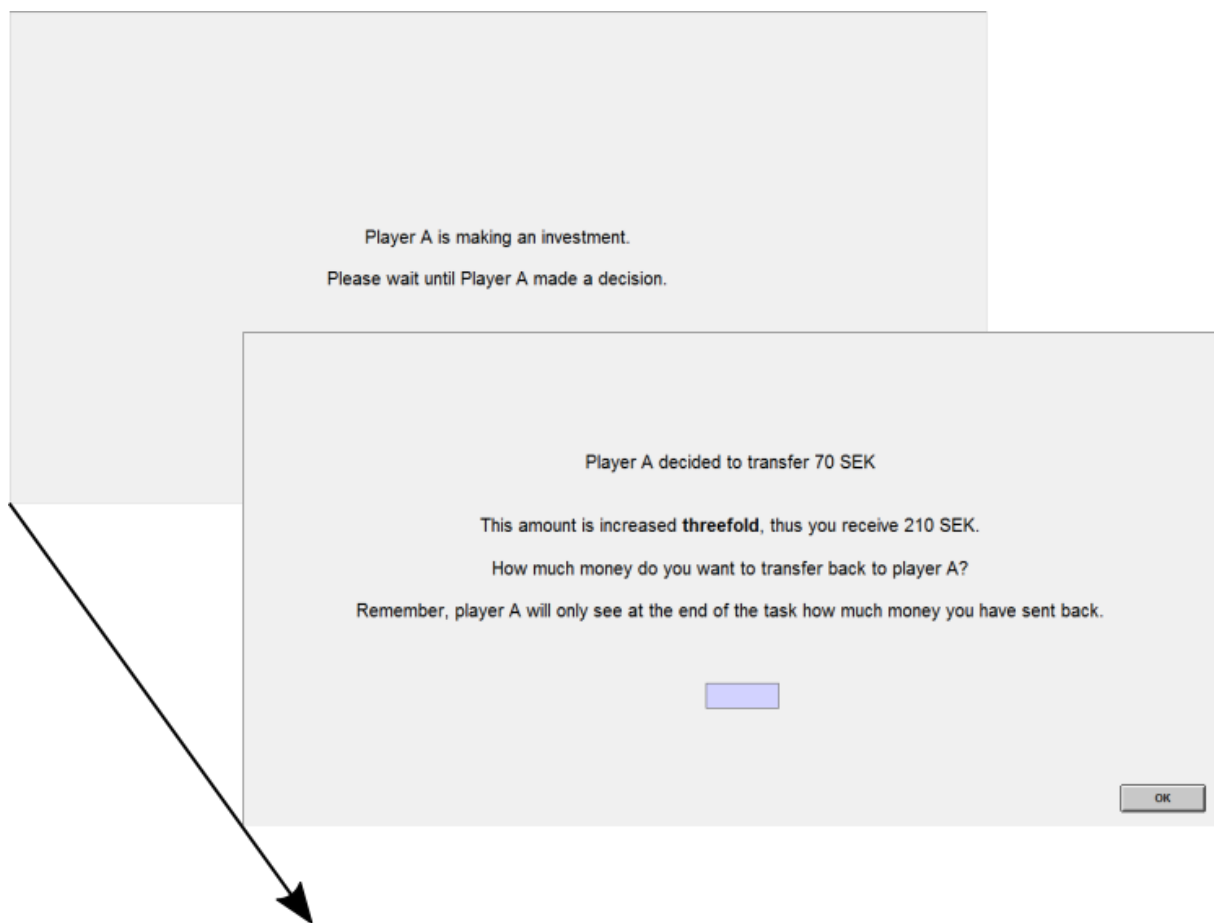


Figure 9. Layout of a trial of the trust game implemented in z-tree (originally in Swedish). The investment screen was displayed between 4 – 6 seconds. Participants had unlimited time for their back-transfer decision.

EXPERIMENT 3

Participants were informed that the main purpose of the experiment was to investigate if and how various forms of sensory stimuli influence our values and perceptions of different phenomena. They were further informed that they would be asked to rate the pleasantness and intensity of the sensory stimuli before performing a task that involved allocating money between themselves and a different player. Following signature of the written consent form, electrodes for facial EMG and skin conductance recordings were attached to the participant (results not reported here). Participants were then seated in front of a computer screen with the left arm placed on an inflatable cushion that was adjusted to their arm. A custom-made cap which limited lateral view was provided in order to ensure that the participants did not see their left arm or the experimenter during the experiment. The participants were also equipped with a set of headphones throughout the experiment (AKG, K 271 MkII). The task procedure for experiment 3 consisted of the following blocks: sensory stimulation 1, Social Value Orientation scale 1 (SVO; Murphy et al., 2011), sensory stimulation 2, SVO 2, sensory stimulation 3, SVO 3.

To further mask the main objective of the study, the participants were also exposed to auditory stimulation with 2 minutes of pink noise which was not followed by an economic task but by a questionnaire pertaining to ratings of facial attractiveness, questions of soft drink preferences and how one would spend an imaginary lottery win. The questionnaire served to further mask the main objective of the study and was therefore not analysed. The order of sensory stimulation types (CT-targeted, control touch, visual and auditory stimulation) was counterbalanced between participants. Visual stimulation was presented as an alternative to a passive “no touch” control condition. A moving, non-tactile control stimulus was chosen, that that was not expected to increase prosociality.

CT-targeted touch was performed as in experiments 1 and 2, however, to control for the possibility of CT-fibre fatigue, the stroking alternated between two equally large areas of the dorsal forearm. Additionally, control touch consisted here of slow brushing at 0.3 cm/s to complement the findings from the previous two experiments which used 30cm/sec as control touch. The duration of CT-targeted and control touch was 2 minutes each to ensure equal contact time between the conditions. For visual stimulation, participants watched a two minute long film which was edited from six films (Shutterstock Inc; New York, USA) displaying colourful geometrical objects moving at various speeds. Immediately after the tactile or visual stimulation, the participants rated the pleasantness and intensity of the sensory stimulus using two subsequently appearing visual analogue scales (VAS) as in experiments 1 and 2 on the screen in front of them.

SOCIAL VALUE ORIENTATION TASK

Participants performed the online version of the Social Value Orientation task (SVO; Murphy; Murphy et al., 2011). The SVO is designed to measure prosocial behaviour, i.e., the concern for other people's pay-off, with social value orientation corresponding "to the quantity of how much a decision maker is willing to sacrifice in order to make another decision maker better off (or perhaps worse off)." (Murphy et al., 2011; p.772). Participants are categorized into one of four different types, based on an angle score calculated from their decisions during the task. A larger SVO angle indicates that participants more often chose the option that maximised the allocation to the other person. On the other hand, a small SVO angle indicates that the participant opted for a larger portion of the allocated money themselves. The angle ranges for the four categories are as follows: "altruistic" (> 57.15 deg), "individualistic" (between $-12.04 - 22.45$ deg), "prosocial" (between $22.45-57.15$ deg), or "competitive (< -12.04 deg)".

In the present SVO-task, participants were instructed to allocate between NOK 100 and NOK 150 between themselves and a future random participant. They were told that there were no right or wrong answers, but that they had to carefully consider how they chose to allocate the money as the allocations could affect their own payment and that of the future randomized participant. The participants played the short version, whereby six slots of monetary allocations constitute one game (see Figure 10). Angle calculations from the website were verified by manual computations by the experimenter. Participants practiced 2-4 rounds prior to the actual experiment.

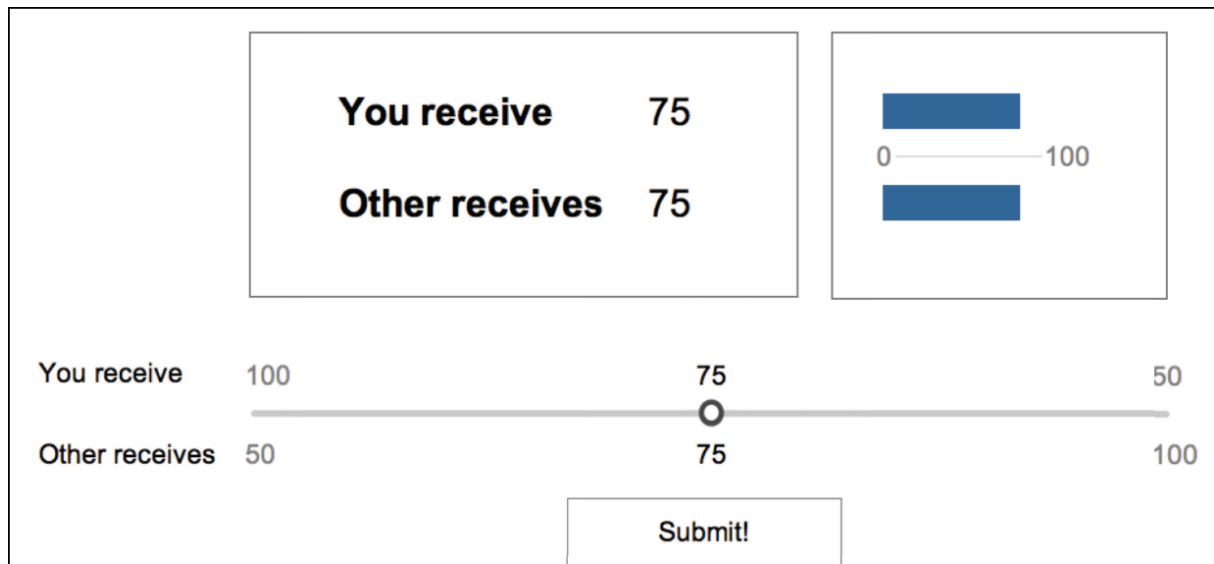


Figure 10. Example trial involving the distribution of NOK 150. The participants were instructed that "you receive" and "other receives" referred to their own and a future randomized participant's payment. Picture reproduced from ²⁵ under Creative Commons Attribution 3.0 License.

DATA ANALYSIS

All data were analysed with linear mixed models, which are ideally suited for repeated measures designs. There is no need for data aggregation, which increases power and makes it possible to model individual response styles, while at the same time controlling for within-subject dependences of observations (Barr, Levy, Scheepers, & Tily, 2013). We modelled participants as random intercept and the highest order within-subject interaction as random slope (as recommended by Barr et al., 2013) with the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) in R. For the reported interactions we calculated Type 3 Sum of Squares and used orthogonal contrasts. Reported main effects are based on Type 2 Sum of Squares. P-values are based on Wald-Chi-square tests from the car package (Fox & Weisberg, 2011). Post-hoc Tukey tests for factors were computed with the lsmeans package (Lenth, 2016). Results were supplemented with Bayes Factors, which were computed with the BayesFactor package (Morey, Rouder, & Jamil, 2015). Plots were created with the packages ggplot2 (Wickham, 2009) and sjPlot (Ludecke & Schwemmer, 2017) in R.

Dependent variables were pleasantness and intensity, and for the Repeated Trust Game the return ratio ($\frac{\text{back-transfer}}{\text{investment}}$). The return ratio was preferred to the absolute amount returned as a measure of prosocial behaviour because the size of the return ratio is not dependent on the amount invested. For experiment 1, pleasantness and intensity were compared for the two touch types (CT-targeted, control) in two separate analyses. The return ratio was analysed using a model with the predictors multiplication factor (3,6), touch type (CT-targeted, control), and pleasantness. An analogous analysis was done for experiment 2. Next, the data of both experiments were collapsed and analysed together. This was done, first, to evaluate whether there were any differences between the experiments, and second, to analyse the effect of hypothesis guessing in both experiments. Participants may correctly guess the intent of the study and consciously, or subconsciously, alter their behaviour. To investigate whether this was the case, hypothesis guessing was added as a predictor to this analysis. Pleasantness and intensity were separately analysed, and the return ratio was analysed with the predictors multiplication factor (3,6), touch type (CT-targeted, control), experiment number (1,2), hypothesis guessing (yes, no), and pleasantness.

The central question of experiment 1 and 2 was to determine whether the return ratio is larger after CT-targeted than control touch (hypothesis H1). In contrast, the null-hypothesis (H0) would assume no difference between the two types of touch. Conventional frequentist inference statistics only allows testing H1-hypotheses by rejecting the null-hypothesis. However, if the null-hypothesis cannot be rejected, it does not automatically mean that it is true. Bayesian statistics (Dienes, 2014; Rouder, Speckman, Sun, Morey, & Iverson, 2009) allow to determine whether the

data provide stronger evidence for H1 or the null-hypothesis. The Bayes factor B means that the data are B times more likely under the H1 than under the null-hypothesis. If the Bayes factor (BF) is larger than 1, support for the H1 is stronger, whereas a BF smaller than 1 means that support for the H0 is stronger, given the data. Conventionally, a $BF > 3$ can be interpreted as substantial evidence, whereas a $BF > 10$ is considered strong evidence for the model under consideration (Jeffreys, 1961). However, these thresholds should be used cautiously as rules of thumbs (see Lee & Wagenmakers, 2014; Rouder, Haaf, & Aust, 2017 for an overview of BF interpretation), which is why we refrain from referring to these thresholds in the results section. BF were calculated for pleasantness and return ratio for the factor touch type, and for return ratio also on the interaction between pleasantness and touch type. BF were calculated for the pooled data only for variables where the mixed model found no difference between the experiments. We used the default prior setting of the BayesFactor package, which is described in detail in (Rouder & Morey, 2012) for regression models and in (Rouder, Morey, Speckman, & Province, 2012) for ANOVA designs. The default settings specify priors for the variance, the intercept (in the regression models), and the mean of standardized effect sizes of the null and the alternative hypotheses. For the null hypothesis the prior for the mean is 0, and the variance is a non-informative Jeffreys prior. For the alternative hypothesis the priors for the variance and the intercept are also non-informative Jeffreys priors, for the mean it is a multivariate Cauchy distribution with a scale value of 0.5.

For experiment 3, the dependent variables were pleasantness, intensity, and the SVO angle. Pleasantness and intensity were separately analysed for the 3 stimulus types (CT-targeted, control, visual). The SVO angle was analysed using a model with the predictors stimulus type (CT-targeted, control, visual) and pleasantness. BF were calculated for SVO angle and pleasantness.

REFERENCES

- Ackerley, R., Wasling, H. B., Liljencrantz, J., Olausson, H., Johnson, R. D., & Wessberg, J. (2014). Human C-tactile afferents are tuned to the temperature of a skin-stroking caress. *Journal of Neuroscience*, 34(8), 2879-2883.
- Ackert, L. F., Church, B. K., & Davis, S. (2011). An experimental examination of the effect of potential revelation of identity on satisfying obligations. *New Zealand Economic Papers*, 45(1-2), 69-80.
- Baran, N. M., Sapienza, P., & Zingales, L. (2010). *Can we infer social preferences from the lab? Evidence from the Trust Game*. Retrieved from
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3), 255-278.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1 - 48.
- Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *GEB*, 10, 122 - 142.
- Broad, K. D., Curley, J. P., & Keverne, E. B. (2006). Mother–infant bonding and the evolution of mammalian social relationships. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 361(1476), 2199-2214.
- Brockner, J., Pressman, B., Cabitt, J., & Moran, P. (1982). Nonverbal intimacy, sex, and compliance: A field study. *Journal of Nonverbal Behavior*, 6(4), 253-258.
- Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction*: Princeton University Press.
- Croy, I., D'Angelo, S., & Olausson, H. (2014). Reduced pleasant touch appraisal in the presence of a disgusting odor. *PLoS One*, 9(3), e92975.
- Croy, I., Luong, A., Triscoli, C., Hofmann, E., Olausson, H., & Sailer, U. (2016). Interpersonal stroking touch is targeted to C tactile afferent activation. *Behavioural Brain Research*, 297, 37-40.
- Crusco, A. H., & Wetzel, C. G. (1984). The Midas touch: The effects of interpersonal touch on restaurant tipping. *Personality and Social Psychology Bulletin*, 10(4), 512-517.
- Debrot, A., Schoebi, D., Perrez, M., & Horn, A. B. (2013). Touch as an interpersonal emotion regulation process in couples' daily lives: the mediating role of psychological intimacy. *Personality and Social Psychology Bulletin*, 39(10), 1373-1385.
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Front Psychol*, 5, 781.
- Dunbar, R. (1993). Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, 16(4), 681-694.
- Dunbar, R. (1998). *Grooming, gossip, and the evolution of language*: Harvard University Press.
- Ellingsen, D.-M., Leknes, S., Løseth, G., Wessberg, J., & Olausson, H. (2016). The neurobiology shaping affective touch: expectation, motivation, and meaning in the multisensory context. *Front Psychol*, 6, 1986.
- Ellingsen, D.-M., Wessberg, J., Chelnokova, O., Olausson, H., Laeng, B., & Leknes, S. (2014). In touch with your emotions: oxytocin and touch change social impressions while others' facial expressions can alter touch. *Psychoneuroendocrinology*, 39, 11-20.
- Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4), 980-994.

- Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2), 171-178.
- Floyd, K. (2006). *Communicating affection: Interpersonal behavior and social context*: Cambridge University Press.
- Fox, J., & Weisberg, S. (2011). *An R companion to applied regression*: Sage Publications.
- Gallace, A., & Spence, C. (2010). The science of interpersonal touch: an overview. *Neuroscience & Biobehavioral Reviews*, 34(2), 246-259.
- Gallistel, C. (2009). The importance of proving the null. *Psychological Review*, 116(2), 439.
- Gazzola, V., Spezio, M. L., Etzel, J. A., Castelli, F., Adolphs, R., & Keysers, C. (2012). Primary somatosensory cortex discriminates affective significance in social touch. *Proceedings of the National Academy of Sciences*, 109(25), E1657-E1666.
- Guéguen, N., & Fischer-Lokou, J. (2002). An evaluation of touch on a large request: A field setting. *Psychological Reports*, 90(1), 267-269.
- Gulledge, A. K., Gulledge, M. H., & Stahmann, R. F. (2003). Romantic physical affection types and relationship satisfaction. *The American Journal of Family Therapy*, 31(4), 233-242.
- Güth, W., Muegera, H., Musau, A., & Ploner, M. (2012). *Intentions and consequences: An experimental investigation of trust and reciprocity determinants*. Retrieved from
- Hornik, J., & Ellis, S. (1988). Strategies to secure compliance for a mall intercept interview. *Public Opinion Quarterly*, 52(4), 539-551.
- Iggo, A., & Kornhuber, H. (1977). A quantitative study of C - mechanoreceptors in hairy skin of the cat. *The Journal of physiology*, 271(2), 549-565.
- Jakubiak, B. K., & Feeney, B. C. (2017). Affectionate touch to promote relational, psychological, and physical well-being in adulthood: A theoretical model and review of the research. *Personality and Social Psychology Review*, 21(3), 228-252.
- Jeffreys, H. (1961). *The theory of probability* (3 ed.). Oxford: Oxford University Press.
- Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, 32(5), 865-889. doi:10.1016/j.joep.2011.05.007
- Joule, R.-V., & Guéguen, N. (2007). Touch, compliance, and awareness of tactile contact. *Perceptual and Motor Skills*, 104(2), 581-588.
- Kirsch, L. P., Krahé, C., Blom, N., Crucianelli, L., Moro, V., Jenkinson, P. M., & Fotopoulou, A. (2017). Reading the mind in the touch: Neurophysiological specificity in the communication of emotions by touch. *Neuropsychologia*.
- Kleinke, C. L. (1977). Compliance to requests made by gazing and touching experimenters in field settings. *Journal of Experimental Social Psychology*, 13(3), 218-223.
- Koppel, L., Andersson, D., Morrison, I., Vastfjall, D., & Tinghog, G. (2017). The (Null) Effect of Affective Touch on Betrayal Aversion, Altruism, and Risk Taking. *Front Behav Neurosci*, 11, 251. doi:10.3389/fnbeh.2017.00251
- Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature*, 435(7042), 673-676. doi:10.1038/nature03701
- Kress, I. U., Minati, L., Ferraro, S., & Critchley, H. D. (2011). Direct skin-to-skin vs. indirect touch modulates neural responses to stroking vs. tapping. *Neuroreport*, 22(13), 646.
- Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*: Cambridge University Press.
- Lenth, R. V. (2016). Least-squares means: the R package lsmeans. *Journal of Statistical Software*, 69(1), 1-33.

- Löken, L. S., Wessberg, J., McGlone, F., & Olausson, H. (2009). Coding of pleasant touch by unmyelinated afferents in humans. *Nature Neuroscience*, 12(5), 547.
- Ludecke, D., & Schwemmer, C. (2017). sjPlot: Data visualization for statistics in social science (version 2.3. 1)(software).
- McGlone, F., Wessberg, J., & Olausson, H. (2014). Discriminative and affective touch: sensing and feeling. *Neuron*, 82(4), 737-755.
- Melzack, R., & Wall, P. D. (1965). Pain mechanisms: a new theory. *Science*, 150(3699), 971-979.
- Mislin, A., Williams, L. V., & Shaughnessy, B. A. (2015). Motivating trust: Can mood and incentives increase interpersonal trust? *Journal of Behavioral and Experimental Economics*, 58, 11-19. doi:10.1016/j.socec.2015.06.001
- Morey, R. D., Rouder, J. N., & Jamil, T. (2015). BayesFactor: Computation of Bayes factors for common designs. *R package version 0.9*, 9, 2014.
- Morhenn, V. B., Park, J. W., Piper, E., & Zak, P. J. (2008). Monetary sacrifice among strangers is mediated by endogenous oxytocin release after physical contact. *Evolution and Human Behavior*, 29(6), 375-383.
- Morrison, I., Loken, L. S., & Olausson, H. (2010). The skin as a social organ. *Experimental Brain Research*, 204(3), 305-314. doi:10.1007/s00221-009-2007-y
- Murphy, R. Retrieved from <http://ryanomurphy.com/styled-2/index.html>
- Murphy, R., Ackermann, K. A., & Handgraaf, M. J. (2011). Measuring Social Value Orientation. *Judgment and Decision Making*, 6(8), 771 - 781.
- Nordin, M. (1990). Low - threshold mechanoreceptive and nociceptive units with unmyelinated (C) fibres in the human supraorbital nerve. *The Journal of Physiology*, 426(1), 229-240.
- Olausson, H., Wessberg, J., McGlone, F., & Vallbo, Å. (2010). The neurophysiology of unmyelinated tactile afferents. *Neuroscience & Biobehavioral Reviews*, 34(2), 185-191.
- Rose, S. A. (1990). The sense of touch. *Touch: The Foundation of Experience*. International Universities Press, Madison, WI, 299-324.
- Rouder, J. N., Haaf, J. M., & Aust, F. (2017). From theories to models to predictions: A Bayesian model comparison approach. *Communication Monographs*, 85(1), 41-56. doi:10.1080/03637751.2017.1394581
- Rouder, J. N., & Morey, R. D. (2012). Default Bayes Factors for Model Selection in Regression. *Multivariate Behav Res*, 47(6), 877-903. doi:10.1080/00273171.2012.734737
- Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, 56(5), 356-374. doi:10.1016/j.jmp.2012.08.001
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon Bull Rev*, 16(2), 225-237. doi:10.3758/PBR.16.2.225
- Sailer, U., Tricoli, C., Häggblad, G., Hamilton, P., Olausson, H., & Croy, I. (2016). Temporal dynamics of brain activation during 40 minutes of pleasant touch. *NeuroImage*, 139, 360-367.
- Scheele, D., Kendrick, K. M., Khouri, C., Kretzer, E., Schläpfer, T. E., Stoffel-Wagner, B., . . . Hurlmann, R. (2014). An oxytocin-induced facilitation of neural and emotional responses to social touch correlates inversely with autism traits. *Neuropsychopharmacology*, 39(9), 2078.

- Schirmer, A., Wijaya, M. T., & Liu, S. (2016). The Midas effect: How somatosensory impressions shape affect and other-concern *Affective touch and the neurophysiology of CT afferents* (pp. 283-299): Springer.
- Silk, J. B. (2002). Using the 'F'-word in primatology. *Behaviour*, 139(2-3), 421.
- Slovic, P., Lichtenstein, S., & Edwards, W. (1965). Boredom-induced changes in preferences among bets. *The American journal of psychology*, 78(2), 208-217.
- Spapé, M. M., Harjunen, V., & Ravaja, N. (2017). Effects of touch on emotional face processing: a study of event-related potentials, facial EMG and cardiac activity. *Biological Psychology*, 124, 1-10.
- Spapé, M. M., Hoggan, E. E., Jacucci, G., & Ravaja, N. (2015). The meaning of the virtual Midas touch: An ERP study in economic decision making. *Psychophysiology*, 52(3), 378-387.
- Stephen, R., & Zweigenhaft, R. L. (1986). The effect on tipping of a waitress touching male and female customers. *The Journal of Social Psychology*, 126(1), 141-142.
- Triscoli, C., Ackerley, R., & Sailer, U. (2014). Touch satiety: differential effects of stroking velocity on liking and wanting touch over repetitions. *PLoS One*, 9(11), e113425.
- Triscoli, C., Croy, I., Olausson, H., & Sailer, U. (2017). Touch between romantic partners: being stroked is more pleasant than stroking and decelerates heart rate. *Physiology & Behavior*, 177, 169-175.
- Triscoli, C., Croy, I., Steudte-Schmiedgen, S., Olausson, H., & Sailer, U. (2017). Heart rate variability is enhanced by long-lasting pleasant touch at CT-optimized velocity. *Biological Psychology*, 128, 71-81.
- Vallbo, Å., Olausson, H., & Wessberg, J. (1999). Unmyelinated afferents constitute a second system coding tactile stimuli of the human hairy skin. *Journal of Neurophysiology*, 81(6), 2753-2763.
- Wessberg, J., Olausson, H., Fernström, K. W., & Vallbo, Å. B. (2003). Receptive field properties of unmyelinated tactile afferents in the human skin. *Journal of Neurophysiology*, 89(3), 1567-1575.
- Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

ACKNOWLEDGEMENTS

This study was supported by the Swedish Research Council (grant number 2011-1529), and the Vienna Science and Technology Fund (WWTF VRG13-007). We thank Patrik Andersson and in particular Petra Valej for refinement of the procedure and for collecting the data for experiment 1 and 2. Thanks also to Claus Lamm for helpful comments on a previous version of the manuscript, to Rochelle Ackerley for useful information regarding CT fatigue, and to Ralph Endemann for drawing Figure 8.

AUTHORS' CONTRIBUTIONS

The studies were designed collaboratively by US, LAR, CE, and AR (exp. 3). LAR analysed the data, made the graphs, wrote large parts of the results and prepared the supplementary materials. AR

collected data (exp. 3), and helped drafting the manuscript. CE contributed critically to data interpretation of exp. 1 and 2. US developed the study ideas and coordinated the experiments, wrote the main text with contributions from LAR and AR. The authors LAR, AR, US gave final approval for publication.

ADDITIONAL INFORMATION

We have no competing interests.

DATA AVAILABILITY STATEMENT

Data and analyses scripts are available at the open science foundation website ⁵⁵

SUPPLEMENTARY MATERIALS

In all tables the following information about the random effects of the linear mixed models are added: ICC = intra-class-correlation, N = amount of random intercepts, σ^2 = random within group variance, τ_{00} = random between- group variance, ρ_{01} = random-slope-intercept correlation, R^2 = difference of the total variance of the null and full model divided by the total variance of the null model. This information is derived from the “sjPlot” package in R (Ludecke & Schwemmer, 2017). All reported main effects are computed with type 2 sum of squares and all interactions with type 3 sum of squares.

RESULTS 1

Table 1.

Means (m) and standard deviations (SD) of the ratings and the return ratios of experiment 1 and 2.

	CT-targeted touch		control touch	
	m	SD	m	SD
pleasantness rating				
experiment 1	7.78	2.07	6.63	2.30
experiment 2	7.79	2.15	5,78	2.48
intensity rating				
experiment 1	5.36	2.52	5.47	2.30
experiment 2	4.94	2.64	5.00	2.50
return ratios				
experiment 1	1.37	1.23	1.29	1.24
experiment 2	1.65	1.01	1.67	1.12

Table 2.

Linear mixed model results of return ratio in experiment 1 predicted by multiplication factor, touch type, touch pleasantness rating and their interactions.

	Return Ratio - experiment 1				
	χ^2	df	p	std. Beta	95% CI
Fixed Part					
Multiplication Factor	53.553	1	< 0.001	-0.31	-0.50 - -0.12
Touch Type	0.098	1	0.754	-0.05	-0.28 - 0.19
Pleasantness Rating	0.578	1	0.447	0.04	-0.04 - 0.11
Multiplication Factor x Touch Type	0.378	1	0.539	0.05	-0.12 - 0.23
Multiplication Factor x Pleasantness Rating	0.001	1	0.973	-0.00	-0.18 - 0.18
Touch Type x Pleasantness Rating	0.360	1	0.549	0.07	-0.16 - 0.31
Multiplication Factor x Touch Type x Pleasantness Rating	1.046	1	0.306	-0.09	-0.27 - 0.09
Random Part					
σ^2	0.238				
$\tau_{00, ID}$	0.428				
ρ_{01}	-0.280				
N_{ID}	50				
ICC_{ID}	0.643				
Observations	1000				
R^2	0.864				

Note. touch type (CT-targeted, control), multiplication factor (3, 6). Random slopes: multiplication factor x touch type.

Table 3.

Linear mixed model results of return ratio in experiment 2 predicted by multiplication factor, touch type, touch pleasantness rating and their interactions.

	Return Ratio - experiment 2				
	χ^2	df	p	std. Beta	95% CI
Fixed Part					
Multiplication Factor	131.020	1	<0.001	-0.50	-0.66 - -0.34
Touch Type	0.442	1	0.506	0.13	-0.07 - 0.34
Pleasantness Rating	4.602	1	0.032	0.02	-0.05 - 0.10
Multiplication Factor x Touch Type	0.097	1	0.756	-0.02	-0.16 - 0.12
Multiplication Factor x Pleasantness Rating	0.734	1	0.392	0.06	-0.08 - 0.21
Touch Type x Pleasantness Rating	2.411	1	0.121	-0.16	-0.36 - 0.04
Multiplication Factor x Touch Type x Pleasantness Rating	0.200	1	0.655	0.03	-0.11 - 0.18
Random Part					
σ^2	0.211				
$\tau_{00, ID}$	0.517				
ρ_{01}	-0.959				
N_{ID}	79				
ICC_{ID}	0.710				
Observations	1580				
R^2	0.837				

Note. touch type (CT-targeted, control), multiplication factor (3, 6). Random slopes: multiplication factor x touch type.

RESULTS 2

Table 4.

Linear mixed model results of combined data from experiment 1 and 2, of return ratio predicted by multiplication factor, touch type, touch pleasantness rating, hypothesis guessing and their interactions.

	Return Ratio – experiment 1 and 2 pooled together				
	χ^2	df	p	std. Beta	95% CI
Fixed Part					
Multiplication Factor (MF)	347.913	1	<0.001	-0.34	-0.45 - -0.23
Touch Type (TT)	0.173	1	0.678	0.00	-0.16 - 0.17
Pleasantness rating (PR)	0.432	1	0.511	0.03	-0.04 - 0.10
Hypothesis guessed (HG)	0.368	1	0.544	-0.13	-0.29 - 0.03
MF x TT	1.329	1	0.249	0.06	-0.04 - 0.17
MF x PR	0.028	1	0.868	-0.01	-0.17 - 0.15
TT x PR	0.011	1	0.915	-0.01	-0.20 - 0.18
MF x HG	15.788	1	<0.001	0.22	0.11 - 0.33
TT x HG	0.013	1	0.910	-0.01	-0.18 - 0.16
PR x HG	2.038	1	0.153	0.12	-0.04 - 0.28
MF x TT x PR	1.314	1	0.252	-0.08	-0.22 - 0.06
MF x TT x HG	0.139	1	0.709	-0.02	-0.13 - 0.09
MF x PR x HG	7.226	1	0.007	-0.22	-0.38 - -0.06
TT x PR x HG	0.059	1	0.808	-0.02	-0.21 - 0.17
MF x TT x PR x HG	0.387	1	0.534	0.04	-0.10 - 0.18
Random Part					
σ^2	0.227				
$\tau_{00, \text{experiment:ID}}$	0.365				
ρ_{01}	-0.475				
$N_{\text{experiment:ID}}$	129				
$ICC_{\text{experiment:ID}}$	0.616				
Observations	2580				
R^2	0.850				

Note. touch type (CT-targeted, control), multiplication factor (3, 6), hypothesis guessed (yes, no). Random slopes: pleasantness rating x multiplication factor x touch type.

Table 5.

Results of post-hoc linear mixed model analysis of significant multiplication factor x pleasantness rating x hypothesis guessing interaction. Results of linear mixed model for participants who correctly guessed hypothesis.

		Return Ratio - Hypothesis guessed correctly				
		χ^2	df	p	std.Beta	95% CI
Fixed Part						
Multiplication (MF)	Factor	11.992	1	< 0.001	-0.08	-0.26 - 0.10
Touch Type (TT)		1.622	1	0.203	0.03	-0.21 - 0.27
Pleasantness rating (PR)		0.451	1	0.502	0.12	-0.03 - 0.28
MF x TT		0.207	1	0.649	0.04	-0.12 - 0.20
MF x PR		3.999	1	0.046	-0.26	-0.52 - -0.01
TT x PR		0.381	1	.0537	-0.07	-0.30 - 0.16
MF x TT x PR		0.120	1	0.729	-0.03	-0.20 - 0.14
Random Part						
σ^2		0.243				
$\tau_{00, \text{experiment:ID}}$		0.505				
ρ_{01}		-0.784				
$N_{\text{experiment:ID}}$		20				
$ICC_{\text{experiment:ID}}$		0.675				
Observations		400				
R^2		0.834				

Note. touch type (CT-targeted, control), multiplication factor (3, 6). Random slopes: pleasantness rating x multiplication factor.

Table 6.

Results of post-hoc linear mixed model analysis of significant multiplication factor x pleasantness rating x hypothesis guessing interaction. Results of linear mixed model for participants who did not guess hypothesis.

		Return Ratio - Hypothesis not guessed correctly				
		χ^2	df	p	std. Beta	95% CI
Fixed Part						
Multiplication	Factor					
(MF)		357.473	1	< 0.001	-0.56	-0.64 - -0.48
Touch Type (TT)		0.804	1	0.370	0.02	-0.10 - 0.14
Pleasantness rating (PR)		0.255	1	0.614	-0.03	-0.08 - 0.03
MF x TT		3.913	1	0.048	0.08	0.00 - 0.16
MF x PR		11.148	1	< 0.001	0.20	0.08 - 0.32
TT x PR		0.013	1	0.908	0.01	-0.13 - 0.15
MF x TT x PR		4.912	1	0.027	-0.12	-0.23 - -0.01
Random Part						
σ^2		0.223				
$\tau_{00, \text{experiment:ID}}$		0.413				
ρ_{01}		-0.544				
$N_{\text{experiment:ID}}$		109				
$ICC_{\text{experiment:ID}}$		0.649				
Observations		2180				
R^2		0.852				

Note. touch type (CT-targeted, control), multiplication factor (3, 6). Random slopes: pleasantness rating x multiplication factor x touch type.

RESULTS 3

Table 7.

Post-hoc comparison of significant multiplication factor x touch type x pleasantness rating interaction of linear mixed model displayed in Table 6. Results are from combined data from experiment 1 and 2 of participants who did not guess the study's hypothesis and show return ratios predicted by multiplication factor, touch pleasantness rating and their interaction for the CT-targeted touch condition.

	Return Ratio - hypothesis not guessed CT-targeted touch				
	χ^2	df	p	std. Beta	95% CI
Fixed Part					
Multiplication factor	236.976	1	< 0.001	-0.53	-0.65 - -0.41
Pleasantness rating	0.007	1	0.934	-0.06	-0.17 - 0.04
Multiplication factor x Pleasantness rating	2.685	1	0.101	0.12	-0.02 - 0.27
Random Part					
σ^2	0.177				
$\tau_{00, \text{experiment:ID}}$	0.255				
ρ_{01}	-0.231				
$N_{\text{experiment:ID}}$	109				
$ICC_{\text{experiment:ID}}$	0.590				
Observations	1090				
R^2	0.874				

Note. multiplication factor (3, 6). Random slopes: pleasantness rating x multiplication factor.

Table 8.

Post-hoc comparison of significant multiplication factor x touch type x pleasantness rating interaction of linear mixed model displayed in Table 6. Results are from combined data from experiment 1 and 2 of participants who did not guess the study's hypothesis and show return ratios predicted by multiplication factor, touch pleasantness rating and their interaction for the control touch condition.

	Return Ratio - hypothesis not guessed control touch				
	χ^2	df	p	std. Beta	95% CI
Fixed Part					
Multiplication factor	176.559	1	< 0.001	-0.67	-0.80 - -0.54
Pleasantness rating	1.241	1	0.265	-0.11	-0.24 - 0.01
Multiplication factor x Pleasantness rating	13.234	1	< 0.001	0.34	0.16 - 0.52
Random Part					
σ^2	0.274				
$\tau_{00, \text{experiment:ID}}$	0.186				
ρ_{01}	0.575				
$N_{\text{experiment:ID}}$	109				
$ICC_{\text{experiment:ID}}$	0.403				
Observations	1090				
R^2	0.836				

Note. multiplication factor (3, 6). Random slopes: pleasantness rating x multiplication factor.

RESULTS 4

Table 9.

Linear mixed model results from data from experiment 3 of SVO angle predicted by stimulation type, touch pleasantness rating and their interaction.

		SVO angle			
		χ^2	df	p	
					std. Beta 95% CI
Fixed Part					
Stimulation type		5.756	2	0.056	
Control stimulation					0.14 -0.26 - 0.53
CT-targeted stimulation					-0.25 -0.65 - 0.15
visual stimulation					0.11 -0.38 - 0.60
Pleasantness rating		1.164	1	0.281	-0.08 -0.21 - 0.04
Stimulation type	x	2.824	2	0.244	
Pleasantness Rating					
Control stimulation	x				-0.26 -0.66 - 0.14
pleasantness rating					
CT-targeted stimulation	x				0.28 -0.12 - 0.69
pleasantness rating					
visual stimulation	x				0.01 -0.51 - 0.52
pleasantness rating					
Random Part					
σ^2		25.717			
$\tau_{00, ID}$		190.703			
ρ_{01}		-0.885			
N_{ID}		44			
ICC_{ID}		0.881			
Observations		128			
R^2		0.867			

Note. stimulation type (CT-targeted, control, visual). Random slope: pleasantness rating.

CHAPTER 6 - GENERAL DISCUSSION

In the last four chapters I have presented research that has narrowed in on behavioural, psychological, neurofunctional, and clinical aspects of how we learn to trust others. In chapter 2, I reported the first published study that demonstrates that rodent models of basolateral amygdala (BLA) functioning are applicable to human BLA functioning during trust learning. Specifically, the human BLA is critically involved in adjusting one's trust to trustworthy and untrustworthy interaction partners. This process comprises (at least) two crucial components: to identify whether an interaction partner can be trusted (for this we have to learn a stimulus – outcome association), and to translate this knowledge into an appropriate (dis-)trust action (here, response – outcome associations are crucial). From the study described in chapter 2 it is not possible to draw strong conclusions about whether the human BLA is involved in both processes, or only one of them. Rodent research suggests that the BLA is involved in both processes in a non-social reward learning context (Chesworth & Corbit, 2017). However, as the study described in chapter 2 has shown, there seems to be a dissociation in the role of the human BLA in learning from rewards and losses. This is also mirrored in the rodent literature (LeDoux & Daw, 2018), and is in line with an earlier study on Pavlovian threat conditioning in BLA-damaged individuals (Klumpers, Morgan, Terburg, Stein, & van Honk, 2015). Future research can help clarify the exact role of the human BLA in the different processes of non-social and social learning from both positively and negatively valenced outcomes. Another aspect that needs to be addressed by future research regards the replication of the BLA's role in trust learning in healthy subjects. Since the calcification of the brain areas of the BLA-damaged individuals progresses slowly over the lifespan, compensatory brain networks might have developed and taken over the functionality of the lesioned area and its networks. It is not known whether such compensatory brain networks played a role in the behaviour of the BLA-damaged individuals during the study described in chapter 2. Therefore, it is important to replicate the BLA's role in trust learning in healthy individuals with an fMRI study. In such a study the dynamics of the involved brain networks in the processes described above can also be further explored. Despite those open questions, the study majorly increased our understanding of the amygdala in trust learning, and created a strong foundation for research that more precisely maps the functionality of the amygdalar subregions in the field of decision neuroscience. The findings can also be applied to clinical research that focusses on extreme manifestations of trust and distrust, such as with paranoid and borderline personality disorders. A specification of the involved brain areas and networks in these disorders is an important step for the development of precise and personalised medical treatments.

Chapter 3 delineated in how far psychopathic offenders establish trusting relationships through fairness norm adherence. Specifically, anti-social psychopathic offenders (i.e. those scoring high on PCL-R Factor 2) rarely adhere to the fairness norm, but this does not seem to hinder the establishment of a trusting relationship. Thus even though they do not reduce uncertainty about the intentions of their behaviour through norm adherence, this does not lead to any obvious repercussions from the interaction partner and does not seem to affect their trust building. Since the study was primarily designed to establish whether psychopathic offenders build trusting relationships at all, I can only speculate whether anti-social psychopathic offenders use alternative strategies to establish a trusting relationship. Based on the extensive literature on the manipulative nature of psychopathic offenders described in the discussion section of chapter 3, a logical next step in characterizing the anti-social psychopathic offenders' trust building is to explore manipulative strategies. In animals, cooperation between non-kin can be a form of manipulation (Clutton-Brock, 2009). Cooperative strategies can be used to increase the manipulators profits only, and not those of both interaction partners. For example, a so-called pseudo-reciprocity is used to increase the partner's selfish behaviour, because this will lead to positive by-products for the manipulator. Cooperation might also be used to feign positive qualities as an interaction partner, or to exploit tendencies to respond in a habitual manner (Clutton-Brock, 2009). Repeated decision-making paradigms are well suited to identify these kinds of dynamic, time-lagged strategies between interacting partners. However, the design of these paradigms should be adapted to the specific strategy under investigation, so that their detection power is increased. The study described in chapter 3 was limited in this way, as the pre-programmed behaviour of the interaction partner did not specifically include behavioural rules for norm adherence or violations. The task matched trustworthy behaviour of the subjects with pre-programmed trust behaviour from the interaction partner, and vice versa for untrustworthy behaviour. Future research can overcome this limitation with a study design where the trust game is played with real interaction partners. King-Casas et al. (2008) have successfully demonstrated how studying dyads consisting of a healthy individual and one with mental health problems can capture aberrant behavioural dynamics during a repeated trust game. Another option is to include behavioural rules to the strategy under investigation when programming the task script. Besides its value in furthering our understanding of psychopathic offenders' strategic decision-making, the findings of chapter 3 are also important for clinical settings. They quantify the clinical observations and case reports of the norm violating nature of psychopathic offenders, and at the same time specify that this mostly applies to the anti-social subgroup. Such a focused finding in one of the psychopathic offenders' subgroups might help personalize therapy programs.

Chapter 4 described how our behaviour changes towards an interaction partner that we have learned to (dis-)trust. Specifically, after established distrust, our comfortable distance towards distrusted interaction partners increases. Interestingly, after established trust, the comfortable distance towards a trusted interaction partner does not appear to decrease, even though it feels very intuitive that we stand close to trusted others. Research on personal space described in the discussion section of chapter 4 points towards that we include trusted others in our personal space. As a consequence, it is conceivable that by including trusted others in our personal space, we also stand close to them. Therefore I hypothesize that the stable comfortable distance to the trusted interaction partner in the study described in chapter 4, might not be the complete picture. Instead, other processes might have interfered with the distance adjustments to the trusted interaction partner. Specifically, the trust mindset of the subjects, described in the discussion section of chapter 4, might have played a significant role in the stable distance to the trusted interaction partner. Subjects had a positive belief that others behave trustworthy at the beginning of the trust game. This might have already been present during the first interpersonal distance measurement before the trust game interactions. The behaviour of the trusted interaction partner might have confirmed this belief and therefore distance adjustments to this partner were not made. Future studies can address this possibility by manipulating the initial expectations about others and putting subjects into a distrust mindset (for a conceptual description of the two mindsets see Mayo (2015) and the discussion section of chapter 4). The task set-up of the study described in chapter 4 deserves extra attention: the repeated trust game is not used as the main source of outcome measures, but is rather an experimental manipulation to induce trust and distrust. This is a novel task set-up in decision neuroscience, which to my knowledge has only been used once before by Hale, Payne, Taylor, Paoletti, and Hamilton (2018). This ecologically valid way of a (dis-)trust induction opens new paths for research into the consequences of (dis-)trust. As such, an application of the study design to clinical populations might prove very insightful in characterizing mental disorders. For example, individuals with borderline personality disorder prefer a very large interpersonal distance to others (Fineberg et al., 2018). Since they also have difficulties in establishing trusting relationships (King-Casas et al., 2008), it would be interesting to explore how they adjust their interpersonal distance to (dis-)trusted others.

Chapter 5 further explored the relationship between trust and psychological distance. Specifically, the chapter focused on the reversed association between a close physical distance and trust. Here, I detail how CT-fibre activating touch, as a proxy of close physical distance, by itself does not affect our trustworthiness. Even though this so-called affective touch is valuable in bonding with others, it is conceivable that the context in which the touch takes place is crucial for its positive

effects with others (Andersen, Gannon, & Kalchik, 2013). When taking touch as a proxy of close interpersonal distance, it at first appears as if the findings described in chapter 5 are not in line with Bryan, Perona, and Adolphs (2012) study where they show that interpersonal distance affects trustworthiness judgements of faces. However, since our study mainly focussed on the CT-fibre activating effects, the study design is limited in its informational value about the effects of non-CT-fibre activating touch. That is, the two trust game experiments described in chapter 5 did not include a comparison between the effect of touch and the effect of no touch on trustworthy behaviour. The experiment with the measure of (pro)social motivations (i.e. the social value orientation task) did include a comparison between touch and no touch. There, touch compared to no touch did not change the subjects pro-social motivations. This could potentially translate to trust and indicate that touch as a proxy of interpersonal distance does not affect trustworthiness. However, there might be nuances depending on the stability of the formed relationship between the interaction partners: as described in the introduction chapter, when stable (dis-)trust is established between interacting partners, the (dis-)trust is very difficult to change. In this context, an incidental touch seems unlikely to affect the strong beliefs about the trustworthiness of the interaction partner. However, it is unclear whether the same is true for unstable relationships where (dis-)trust still has to be learned. Future research can address this open question with a within-subject touch – no touch study design similar to the set-up described in the third experiment of chapter 5. In general, the question is, of course, whether the analogy of touch as a proxy of close interpersonal distance holds. To test whether touch can be conceptualized as a proxy of close interpersonal distance, it would be very interesting to replicate the findings from chapter 4 with a touch paradigm. If the analogy holds, then touch from a trusted interaction partner should be experienced as more pleasant and less intrusive than touch from a distrusted partner. In addition, it could be explored what effects touch from (dis-)trusted partners have on e.g. the pro-sociality of the subject.

With my thesis I contribute to the ever-growing decision neuroscience literature of trust. When viewing (dis-)trust as a strategy to reduce uncertainty in our social world (as FeldmanHall and Shenhav (2019) have suggested), its purpose and the bigger picture become more accessible. It might help us understand our research findings better, but it also creates many more open questions. For example: why do anti-social psychopathic offenders not adhere to the fairness norm, as this would decrease the interaction partner's uncertainty about the anti-social psychopathic offenders behavioural intention? Is this simply because they have other (e.g. profit maximizing) concerns, or do they maybe process uncertainty concepts differently? And also, are our certain (dis-)trust beliefs rigid, or are there situations where we (temporarily) adjust our beliefs about a (dis-)trusted interaction partner? From our own experiences we know that there are tipping points where we

radically change our (dis-)trust beliefs about an interaction partner. Are there individual differences when this tipping point happens? The research described in the last four chapters triggers many more of these questions in me. My hope is that my thesis also sparks curiosity in the readers for all the unexplored trust topics and that it inspires creative trust research that puts trust in the bigger picture of how we deal with our social world.

REFERENCES

- Andersen, P., Gannon, J., & Kalchik, J. (2013). Proxemic and haptic interaction: the closeness continuum. In P. Schulz & P. Copley (Eds.), *Handbooks of Communication Science* (Vol. 2). Berlin: Walter de Gruyter GmbH.
- Bryan, R., Perona, P., & Adolphs, R. (2012). Perspective distortion from interpersonal distance is an implicit visual cue for social judgments of faces. *PLoS One*, 7(9), e45301. doi:10.1371/journal.pone.0045301
- Chesworth, R., & Corbit, L. (2017). The Contribution of the Amygdala to Reward-Related Learning and Extinction. In B. Ferry (Ed.), *The Amygdala-Where Emotions Shape Perception, Learning and Memories*.
- Clutton-Brock, T. (2009). Cooperation between non-kin in animal societies. *Nature*, 462(7269), 51-57. doi:10.1038/nature08366
- FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nat Hum Behav*, 3(5), 426-435. doi:10.1038/s41562-019-0590-x
- Fineberg, S. K., Leavitt, J., Landry, C. D., Neustadter, E. S., Lesser, R. E., Stahl, D. S., . . . Corlett, P. R. (2018). Individuals with Borderline Personality Disorder show larger preferred social distance in live dyadic interactions. *Psychiatry Research*, 260, 384-390. doi:10.1016/j.psychres.2017.11.054
- Hale, J., Payne, M. E., Taylor, K. M., Paoletti, D., & Hamilton, A. F. (2018). The virtual maze: A behavioural tool for measuring trust. *Q J Exp Psychol (Hove)*, 71(4), 989-1008. doi:10.1080/17470218.2017.1307865
- King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science*, 321(5890), 806-810. doi:10.1126/science.1156902
- Klumpers, F., Morgan, B., Terburg, D., Stein, D. J., & van Honk, J. (2015). Impaired acquisition of classically conditioned fear-potentiated startle reflexes in humans with focal bilateral basolateral amygdala damage. *Soc Cogn Affect Neurosci*, 10(9), 1161-1168. doi:10.1093/scan/nsu164
- LeDoux, J., & Daw, N. D. (2018). Surviving threats: neural circuit and computational implications of a new taxonomy of defensive behaviour. *Nature Reviews Neuroscience*, 19(5), 269-282. doi:10.1038/nrn.2018.22

Mayo, R. (2015). Cognition is a matter of trust: Distrust tunes cognitive processes. *European Review of Social Psychology*, 26(1), 283-327. doi:10.1080/10463283.2015.1117249

CHAPTER 7 - ABSTRACT

An integral part of coordinating our social world is to learn to (dis-)trust others and to adjust our behaviour accordingly. This thesis focuses on the behavioural, psychological, neurofunctional, and clinical aspects of these trust learning processes. All of the reported research utilizes a repeated trust game as the primary method. In this task two interacting people learn through monetary transfers over multiple rounds to trust or distrust each other. In the first study, we conducted research with a sample of basolateral amygdala (BLA) damaged individuals. We demonstrate that rodent models of BLA functioning are applicable to the human BLA during trust learning. Specifically, the human BLA is critically involved in adjusting one's trust to trustworthy and untrustworthy interaction partners. In a second study, we delineate in how far psychopathic offenders establish trusting relationships through fairness norm adherence. Especially anti-social psychopathic offenders (i.e. those scoring high on PCL-R Factor 2) only seldom adhere to the fairness norm, but this does not seem to hinder the establishment of a trusting relationship. A third study describes how our behaviour changes towards an interaction partner that we have learned to (dis-)trust. Specifically, after established distrust, our comfortable distance towards distrusted interaction partners increases. Interestingly, after established trust, the comfortable distance towards a trusted interaction partner does not appear to decrease. Finally, a fourth study details how CT-fibre activating touch by itself does not affect how trustworthy we behave. Even though this so-called affective touch is valuable in bonding with others, it is conceivable that the context in which the touch takes place is crucial for its positive effects with others. Together, these findings are assembled in this thesis and discussed under the umbrella of uncertainty reduction: to successfully navigate a social world full of uncertainty, we use trust and distrust to increase the predictability of our interaction partners' behaviour. By (dis-)trusting someone we assume that the other person will behave in a(n) (un-)trustworthy way, which reduces all the possible actions that the other person can take to one concrete (set of) (un-)trustworthy action(s). This helps us to behave with the best possible outcome in mind.

ZUSAMMENFASSUNG

Ein integraler Bestandteil der Koordinierung unserer sozialen Welt besteht darin, zu lernen Anderen (nicht) zu vertrauen und unser Verhalten entsprechend anzupassen. Diese Dissertation konzentriert sich auf die verhaltensbezogenen, psychologischen, neurofunktionalen und klinischen Aspekte dieser Vertrauenslernprozesse. Die beschriebenen Studien verwenden alle ein wiederholtes Vertrauensspiel (engl. *repeated trust game*) als Hauptmethode. Bei dieser Aufgabe lernen zwei interagierende Personen durch Geldtransfers über mehrere Runden hinweg einander zu vertrauen oder zu misstrauen. In der ersten Studie erforschten wir das Verhalten von Personen mit einer basolateralen Amygdala (BLA)-Schädigung. Wir zeigen, dass Nagetiermodelle der BLA-Funktion auf die menschliche BLA beim Vertrauenslernen anwendbar sind. Insbesondere ist die menschliche BLA kritisch daran beteiligt, das eigene Vertrauen an vertrauenswürdige und nicht vertrauenswürdige Interaktionspartner anzupassen. In einer zweiten Studie beschreiben wir, inwieweit psychopathische Straftäter durch die Einhaltung von Fairnessnormen vertrauensvolle Beziehungen aufbauen. Insbesondere antisoziale psychopathische Täter (d.h. solche, die hohe Werte beim PCL-R-Faktor 2 aufweisen) halten sich nur selten an die Fairnessnorm, aber das scheint den Aufbau einer Vertrauensbeziehung nicht zu behindern. Eine dritte Studie beschreibt, wie sich unser Verhalten gegenüber einem Interaktionspartner verändert, dem wir gelernt haben zu (miss-)vertrauen. Konkret: Nach etabliertem Misstrauen nimmt unser komfortabler Abstand zu misstrauten Interaktionspartnern zu. Interessanterweise scheint sich der komfortable Abstand zu einem vertrauten Interaktionspartner nach dem Aufbau von Vertrauen nicht zu verringern. Schließlich zeigt eine vierte Studie, wie CT-Faser aktivierende Berührung an sich keinen Einfluss darauf hat, wie vertrauenswürdig wir uns verhalten. Auch wenn diese so genannte affektive Berührung für die Bindung mit anderen wertvoll ist, ist anzunehmen, dass der Kontext, in dem die Berührung stattfindet, entscheidend für ihre positiven Auswirkungen auf andere ist. Diese Erkenntnisse werden in der vorliegenden Dissertation zusammengeführt und unter dem Leitthema der Unsicherheitsreduktion diskutiert: Um erfolgreich durch eine soziale Welt voller Unsicherheit zu navigieren, nutzen wir Vertrauen und Misstrauen, um die Vorhersehbarkeit des Verhaltens unserer Interaktionspartner zu erhöhen. Indem wir jemandem (nicht) vertrauen, gehen wir davon aus, dass sich die andere Person (un)vertrauenswürdig verhält, wodurch alle möglichen Handlungen, die die andere Person ausführen kann, auf eine konkrete (Menge) (un)vertrauenswürdige Handlung(en) reduziert werden. Das hilft uns, uns nach den bestmöglichen Ergebnissen zu handeln.

CHAPTER 8 - CV LISA ANNA ROSENBERGER

Graafseweg 59
5213AB Den Bosch
The Netherlands

phone: 0031 – 639204080
email: lisa.anna.rosenberger@univie.ac.at
date of birth: 02 August 1988
place of birth: Plauen, Germany

Work experience

01.2019 – present	Co-project manager at DiDaT, Potsdam Germany
05.2019 – 12.2019	Visiting researcher, Utrecht University The Netherlands & University of Cape Town South Africa
10.2014 – 05.2019	Research associate at Neuropsychopharmacology & Biopsychology Unit, Department of Psychology, University of Vienna Austria
06.2018 – 10.2018	Research developer at Minitopia, Den Bosch The Netherlands
01.2018 – 04.2018	Visiting researcher, Utrecht University The Netherlands
08.2013 – 06.2014	Research assistant at Department of Psychiatry, University of Cape Town South Africa

Teaching experience

04.2020 – 06.2020	Teaching statistical methods master course (Wahlforschungsmodul, Hochschule für Gesundheit Bochum Germany)
12.2019 – 01.2020	Teaching statistical methods master course (Forschungsmodul Gesundheit und Diversity in der Arbeit, Hochschule für Gesundheit Bochum Germany)
03.2019	Teaching 1 lesson psychological methods master course (Forschungsansätze in der Kognitionspsychologie und in den Neurowissenschaften, University of Vienna Austria)
11.2018	Teaching 1 lesson psychological methods master course (Forschungsansätze in der Kognitionspsychologie und in den Neurowissenschaften, University of Vienna Austria)
05.2018	Teaching 1 lesson psychological methods master course (Forschungsansätze in der Kognitionspsychologie und in den Neurowissenschaften, University of Vienna Austria)

09.2017 – 02.2018	Supervising research internship of 1 st year master student (master program: Mei:Cogsci, University of Vienna Austria)
05.2017	Teaching 1 lesson psychological methods master course (Spezialthemen aus der Kognitionspsychologie und aus den Neurowissenschaften, University of Vienna Austria)
04.2016 – 06.2016	Supervising research internship of three 1 st year master students (master program: Mei:Cogsci, University of Vienna Austria)
09.2015 – 12.2015	Supervising research internship of 1 st year master student (master program: Mei:Cogsci, University of Vienna Austria)

Education

2014 – present	PhD student Neuropsychopharmacology & Biopsychology Unit, Department of Psychology, University of Vienna Austria
2011 – 2013	Research master <i>Social and Health Psychology</i> Utrecht University The Netherlands
2011	Summer course <i>Applied Multivariate Analysis</i> Utrecht Summer School The Netherlands
2008 – 2011	Bachelor <i>Psychology</i> , Specialization <i>Social Psychology</i> Utrecht University The Netherlands
2011	Minor <i>Social Neuroscience</i> Utrecht University The Netherlands
2010	Minor <i>Methods and Statistics in the Social and Behavioural Sciences</i> (not completed) Utrecht University The Netherlands

Scholarships

2018	Marietta-Blau grant (18 000 Euro, OEAD & BMWFW): international cooperation with Prof. Jack van Honk, University of Cape Town South Africa
2017	CoBeNe Seed Grant (10 000 Euro, Vienna Doctorate School CoBeNe): interdisciplinary project with J. Goller & M. Stocker, University of Vienna Austria.

Public outreach

2020	Radio interview, Ö1 Punkt Eins Austria
2019	Panel discussion, Ö1 CityScienceTalk Austria
2019	Interview for article, Apotheken Umschau Germany
2018	Podcast with Deutschlandfunk Nova Germany
2018	Interview for blog post, Schrödingers Katze Austria

Other experiences

12.2016 – 05.2019	student representative doctoral school CoBeNe (Cognition, Behaviour, and Neuroscience), University of Vienna Austria
10.2014 – 02.2017	lab manager behavioural lab of Neuropsychopharmacology & Biopsychology unit, Department of Psychology, University of Vienna Austria
01.2016 – 06.2016	member of organizational board of JuWi poster presentation event, Department of Psychology, University of Vienna Austria
10.2014 – 01.2016	member organizational board PhD seminar of Department of Psychology, University of Vienna Austria

Publications

- Rosenberger, L. A., Naef, M., Eisenegger, C., & Lamm, C. (2020). Interpersonal distance adjustments after interactions with a generous and selfish trustee during a repeated trust game. *Journal of Experimental Social Psychology*, 90, 104001.
- Rosenberger, L. A., Eisenegger, C., Naef, M., Terburg, D., Fourie, J., Stein, D. J., & van Honk, J. (2019). The Human Basolateral Amygdala Is Indispensable for Social Experiential Learning. *Current Biology*, 29(20), 3532-3537.
- Rosenberger, L. A., Pfabigan, D. M., Lehner, B., Keckeis, K., Seidel, E. M., Eisenegger, C., & Lamm, C. (2019). Fairness norm violations in anti-social psychopathic offenders in a repeated trust game. *Translational psychiatry*, 9.
- Rosenberger, L. A., Ree, A., Eisenegger, C., & Sailer, U. (2018). Slow touch targeting CT-fibres does not increase prosocial behaviour in economic laboratory tasks. *Scientific reports*, 8(1), 7700.

- Terburg, D., Syal, S., Rosenberger, L., Heany, S., Stein, D. & van Honk, J. (2016). Testosterone abolishes implicit subordination in social anxiety. *Psychoneuroendocrinology*, 72, 205 – 211.
- Montoya, E., Bos, P., Terburg, D., Rosenberger, L. & van Honk, J. (2014). Cortisol administration induces global down-regulation of the brain's reward circuitry. *Psychoneuroendocrinology*, 47, 31-42.
- Terburg, D., Syal, S., Rosenberger, L., Heany, S., Philips, N., Gericke, N., Stein, D. & van Honk, J. (2013). Acute effects of Sceletium tortuosum (Zembrin[®]), a dual 5-HT reuptake and PDE4 inhibitor, in the human amygdala and its connection to the hypothalamus. *Neuropsychopharmacology*, 38, 2708-2716.

