Fachbereich Erziehungswissenschaft und Psychologie

der Freien Universität Berlin

# Processing of self-relevant information

Dissertation

zur Erlangung des akademischen Grades

Doktor der Philosophie (Dr. Phil.)

Doctor of Philosophy (Ph.D.)

vorgelegt von

Christoph W. Korn

M.Sc., B.Sc.

Berlin, 2013

Vorblatt

Erstgutachter:

Prof. Dr. Hauke R. Heekeren

Zweitgutachter:

Prof. Dr. Dr. Henrik Walter

Datum der Disputation:

30.04.2013

## Table of contents

## Acknowledgements

The present thesis would not have been possible without the support of numerous people.

First of all, I would like to thank my supervisor Hauke Heekeren for his excellent support both in questions about data analysis details and questions about science in general. Thank you for motivating me, trusting in me, and for providing a very stimulating working environment.

I am indebted to my second supervisor Henrik Walter who always found time for fruitful discussions and insightful comments.

I am grateful to the members of my dissertation committee for helping me to complete the final step of this dissertation.

I had the great opportunity to work with colleagues in two different countries during my time as a PhD student.

I would like to thank Shihui Han and the members of his lab for the very warm welcome in Beijing and for inviting me to many lab activities. I would like to thank Chenbo Wang for the inspiring discussions and for taking care of all organizational details. Thanks to Stijn Massar for the great time.

I am indebted to Ray Dolan for his support and inspiration. I would like to thank Tali Sharot for drawing my interest to optimism.

I wish to express my appreciation to my colleagues in Berlin. I would like to specifically mention those with whom I have worked as a co-author: Kristin Prehn, Soyoung Park, Yan Fan, Kai Zhang, and Dar Meshi. Thank you for your support and the scientific discussions. Yan and Kai, thank you for your invaluable help in testing Chinese participants in Berlin.

Thanks to all members of the Heekeren lab. Thank you for the numerous scientific and non-scientific discussions. Special thanks to Yulia Oganian, Hannah Brühl, Dorit Kliemann, Nikos Green, and Julia Rodríguez Buritica. I was very lucky to share an office with Gabriela Rosenblau. Thank you for the inspiring discussions and the great time.

I would like to thank my colleagues at the Berlin School of Mind and Brain for the warm atmosphere. I additionally thank the school for the provided funding. Thanks to the helpful administrative staff at the school and in the office of the Heekeren lab.

Finally, I would like to thank Matthias, my family, and my friends, especially Yasemin.

# 1. Summary

People constantly receive self-relevant information. For example, social interaction partners give feedback on character traits (e.g., by telling you that you are polite, tidy, or superficial) and media provide statistical information about the likelihood of experiencing future life events (e.g., by stating the likelihood of living past the age of 80 or getting caner). However, the potential behavioral and neural components of self-relevant information processing are underexplored. In this thesis, I aim at providing the empirical basis for a neurocognitive model of self-relevant information processing. I draw on behavioral research on the self-concept, the social self, and self-related positivity biases as well as on neuroscientific research on the neural processes related to self-judgments, reward, and mentalizing.

Study 1 used behavioral measures and functional magnetic resonance imaging (fMRI) to test how social feedback on character traits changed participants' self-ratings. This social feedback was given by peers within the context of a face-to-face interaction. Study 2 extended the approach of study 1 to a cultural comparison between participants of German and Chinese background. Study 3 investigated potential implications of self-relevant information processing for psychiatry by testing how depressive patients updated their personal estimates of the likelihood of future life events when receiving statistical information about these events.

Healthy participants processed self-relevant information in a positively biased way, i.e., they updated their self-ratings and their estimates of the future more after receiving desirable than after receiving undesirable information. In contrast, positively biased updating about future life events was absent in depressive patients. Culture modulated social conformity, i.e., Chinese participants relied more on social feedback than German participants. Self-relevant information processing comprised a reward component that correlated with neural activity in the ventral striatum and the anterior cingulate cortex/medial prefrontal cortex (ACC/MPFC) and a social comparison component that correlated with neural activity in the mentalizing network including the MPFC, the temporo-parietal junction (TPJ), the superior temporal sulcus (STS), the temporal pole (TP), the inferior frontal gyrus (IFG), and the pre-supplementary motor area (preSMA). Self-related MPFC activity differed between German and Chinese participants.

On the basis of these results, I propose a neurocognitive model of self-relevant information processing. The model supposes that both reward processing and social comparison processing impact on the dynamics of the self-concept. These dynamics are biased toward the positive in healthy individuals. Reward processing involves the ventral striatum and the ACC/MPFC. Social comparison processing involves the mentalizing network. Depression is supposed to disrupt reward processing—resulting in an absence of positivity biases. Cultural differences in self-concepts are supposed to modulate social comparison processing—resulting in cultural differences in social conformity.

In conclusion, this dissertation advances the understanding of self-relevant information processing by combining behavioral research on the self-concept, the social self, and self-related positivity biases with neuroscientific research on reward and mentalizing. The proposed neurocognitve model integrates research on the cultural diversity of human societies, offers a framework for a better understanding of psychiatric disorders, and lends itself to a future adaptation to computational modeling approaches.

Keywords: self-concept, social interaction, positivity bias, mentalizing, reward, culture, depression, medial prefrontal cortex (MPFC)

## 2. Zusammenfassung

Menschen erhalten oft Informationen, die für sie selbst relevant sind. So geben soziale Interaktionspartner häufig Rückmeldungen zu Charaktereigenschaften, zum Beispiel wie höflich, wie ordentlich oder wie oberflächlich jemand ist. In den Medien werden außerdem tagtäglich statistische Informationen über die Eintrittswahrscheinlichkeit von zukünftigen Lebensereignissen veröffentlicht, zum Beispiel die Wahrscheinlichkeit älter als 80 Jahre zu werden oder an Krebs zu erkranken. Die potenziellen Komponenten der Verarbeitung von solchen selbstrelevanten Informationen sind jedoch sowohl auf der Verhaltensebene als auch auf der neuronalen Ebene nicht ausreichend erforscht. Ziel dieser Dissertation ist, eine empirische Basis für ein neurokognitives Model der Verarbeitung von selbstrelevanten Informationen zu schaffen. Ich beziehe mich dazu auf Verhaltensforschung zum Selbstkonzept, zum sozialen Selbst und zu selbstbezogenen positiven Verzerrungen, sowie auf neurowissenschaftliche Forschung zu neuronalen Prozessen, die mit Selbsteinschätzungen, Belohnung und der Inferenz mentaler Zustände (*mentalizing*) zusammenhängen.

In Studie 1 wurde mit der Hilfe von Verhaltensmaßen und funktioneller Magnetresonanztomographie (fMRT) untersucht, wie soziale Rückmeldungen zu Charaktereigenschaften die Selbsteinschätzungen der Versuchsteilnehmer veränderten. Diese sozialen Rückmeldungen wurden von Gleichaltrigen im Kontext einer direkten sozialen Interaktion gegeben. In Studie 2 wurde diese Herangehensweise erweitert, indem kulturelle Unterschiede im Selbstkonzept von deutschen und chinesischen Versuchsteilnehmer verglichen wurden. In Studie 3 wurden mögliche Konsequenzen von selbstrelevanter Informationsverarbeitung für die psychiatrische Forschung untersucht. Depressive Patienten schätzten die Eintrittswahrscheinlichkeit von zukünftigen Lebensereignissen ein und erhielten statistische Informationen zu diesen Ereignissen.

Gesunde Versuchsteilnehmer zeigten eine positive Verzerrung bei der Verarbeitung selbstrelevanter Informationen, das heißt sie veränderten ihre Selbsteinschätzungen und ihre Einschätzungen der Zukunft mehr wenn sie wünschenswerte als wenn sie nicht wünschenswerte Informationen erhielten.

Im Gegensatz dazu zeigten depressive Patienten keine positive Verzerrung bei der Verarbeitung von Informationen über zukünftige Lebensereignisse. Der kulturelle Hintergrund der Versuchsteilnehmer beeinflusste deren soziale Konformität, das heißt chinesische Versuchsteilnehmer integrierten soziale Rückmeldungen in einem stärkeren Ausmaß als deutsche Versuchsteilnehmer. Selbstrelevante Informationsverarbeitung umfasste eine Belohnungskomponente sowie eine soziale Vergleichskomponente. Die Belohnungskomponente korrelierte mit neuronaler Aktivität im ventralen Striatum und im anterioren cingulären Cortex (ACC) beziehungsweise medialen prefrontalen Cortex (MPFC). Die soziale Vergleichskomponente korrelierte mit neuronaler Aktivität im *mentalizing* Netzwerk, welches Aktivität im MPFC, in der temporoparietalen Junktion (TPJ), dem superioren temporalen Sulcus (STS), dem inferioren frontalen Gyrus (IFG) und dem prä-supplementären Motorareal (präSMA) umfasste. Zwischen chinesischen und deutschen Versuchsteilnehmern zeigten sich Unterschiede in der MPFC Aktivität im Zusammenhang mit Selbsteinschätzungen.

Auf der Grundlage dieser Ergebnisse schlage ich ein neurokognitives Modell zur Verarbeitung von selbstrelevanten Informationen vor. Dieses nimmt an, dass sich sowohl die Verarbeitung von Belohnung als auch die Verarbeitung von sozialen Vergleichsprozessen auf die dynamischen Veränderungen des Selbstkonzeptes auswirken. Diese Veränderungen sind in Gesunden zum Positiven hin verzerrt. Belohnungsverarbeitung ist mit dem ventralen Striatum und dem ACC/MPFC assoziiert und die Verarbeitung von sozialen Vergleichsprozessen ist mit dem *mentalizing* Netzwerk verknüpft. Bei depressiven Patienten ist vermutlich die Belohnungsverarbeitung gestört, was mit der Abwesenheit von positiven Verzerrungen einhergeht. Kulturelle Unterschiede im Selbstkonzept wirken sich vermutlich auf die Verarbeitung von sozialen Vergleichsprozessen aus, was sich in kulturell unterschiedlicher sozialer Konformität äußert.

Die vorliegende Dissertation verknüpft Verhaltensforschung zum Selbstkonzept, zum sozialen Selbst und zu selbstbezogenen positiven Verzerrungen mit neurowissenschaftlicher Forschung zu Belohnung und

*mentalizing* und erweitert damit das Verständnis selbstrelevanter Informationsverarbeitung. Das hier vorgeschlagene neurokognitive Model integriert Forschung zur kulturellen Vielfalt menschlicher Gesellschaften und bietet ein Bezugssystem zum vertieften Verständnis psychiatrischer Erkrankungen. Darüber hinaus werden mögliche Erweiterungen des Models durch computationale Modellierungsansätzen diskutiert.

Schlagwörter: Selbstkonzept, soziale Interaktion, positive Verzerrung, *mentalizing*, Belohnung, Kultur, Depression, medialer präfrontaler Cortex (MPFC)

## 3. Introduction

The idea of “self” has fascinated humans throughout history. Notions of self play a central role in everyday life and various scientific disciplines. Philosophers discuss the nature of subjective experience and biologists ask whether the self is uniquely human. Psychologists have conducted thousands of experiments to investigate how the notion of self can explain human behavior and cognition. Folk psychology and scientific disciplines vary in how they characterize what constitutes the self. Many prominent scholars in psychology have not given global definitions of “self” but have instead specified certain self-related phenomena (Myers, 2005; Leary, 2007; Hewstone et al., 2008; Hogg and Vaughan; 2008; Baumeister, 2011). I would like to begin by specifying three perspectives on the self, which have guided theoretical and empirical work: self-concept, social self and self-related positivity biases (**Figure 1**). Based on these perspectives I will formulate the overall aim of this thesis and the research questions of the three empirical studies included in this thesis.

### 1.1. Three perspectives on the self

First, the self-concept consists of a set of cognitive representations that structure and organize information related to somebody’s experience and behavior (*self-concept*; **Figure 1A**) (e.g., Markus and Wurf, 1987; Baumeister, 1998; Myers, 2005; Hewstone et al., 2008; Hogg and Vaughan; 2008; Baumeister, 2011). People’s self-concepts differ in content (subsumed under the notion of self-schemas) and structure (subsumed under the notion of self-complexity). Self-related cognitive representations can be summarized in the form of character traits (e.g. polite, aggressive), perceived physical characteristics (e.g. healthy, beautiful), abilities (e.g. athletic, good in physics), or sets of preferences (e.g. preference for Italian food). People’s self-concepts are relatively continuous over time; people have a sense of who they were in the past (via autobiographical memory) and of who they want to be in the future (via imagination and prospective thinking).

Second, the self is embedded in a social world (*social self*; **Figure 1B**). While theories of the self-concept focus on the minds of individuals, theories of the social self emphasize that individuals live in social groups (e.g., Banaji and Prentice, 1994; Brewer and Hewstone, 2004; Hewstone et al., 2008; Hogg and Vaughan; 2008; Baumeister, 2011; Ellemers, 2012). Within these groups, people have specific social roles (e.g. student) and relations to others (e.g. to a friend). People compare themselves to others and seek self-relevant information when engaging in social interactions (e.g., Banaji and Prentice, 1994; Alicke and Sedikides 2009; Hepper et al. 2011). People's background culture (e.g. whether they live in the West or in East Asia) shapes their social world, and thus their social self (Heine, 2012).

Third, people are motivated to establish or maintain a particular self-concept. The vast majority of research on self-motives has focused on self-related positivity biases such as people's desire to maintain or increase the positivity (or decrease the negativity) of their self-concept or the desire to protect or enhance their self-esteem (*self-related positivity biases*; **Figure 1C**) (e.g., Taylor and Brown, 1988; Leary, 2007; Hewstone et al., 2008; Hogg and Vaughan; 2008; Sedikides and Gregg, 2008; Alicke and Sedikides, 2009). Self-related positivity biases include optimism, i.e., the tendency to underestimate probability of encountering negative events in the future (or to overestimate the probability of encountering positive events) (Weinstein, 1980; Scheier et al., 1994; Sharot, 2011). Self-related positivity biases are relevant for psychiatry since psychiatric patients such as depressed patients often exhibit negative biases (Beck et al., 1979; Gotlib and Joormann, 2010).

These theoretical perspectives focusing on the self-concept, the social self and self-related biases are interrelated in several ways. Theories on self-concept and social self are linked because social roles can be regarded as self-schemas similar to character traits. Self-concept and self-related biases are linked because cognitive representations of "possible selves" (i.e., the ideal self one wants to be and the dreaded self one fears to be) function as evaluative context and incentives for future behavior (Markus and Nurius, 1986). The social self and self-related positivity biases are linked because social interactions often create motivations for self-enhancement (Banaji and Prentice, 1994; Leary, 2007; Hepper et al., 2011).

Thus, the psychological literature has linked theories on the self-concept, the social self, and self-related biases to each other. However, a theoretical framework that integrates these perspectives with respect to information processing and neurobiological mechanisms is lacking. Most previous studies have neglected how people process incoming self-relevant information. Since the brain can be seen as a complicated information-processing system, clarifying the components of self-relevant information processing holds the promise to link the neural components of self-related phenomena to other neurobiological mechanisms (such as reward processing or mentalizing). In addition, understanding self-relevant information processing may bear implications for research on psychiatric diseases (such as depression) (Beck et al., 1979; Gotlib and Joormann, 2010) and for research on the cultural diversity of human cognition (Heine, 2012).

Overall, the three empirical studies of this thesis aim at advancing the understanding of how humans process self-relevant information. Study 1 (Korn et al., 2012) investigated positively biased social feedback processing and the associated related neural activity. Study 2 (Korn et al., submitted) investigated how culture influences behavioral and neural aspects of social feedback processing. Study 3 (Korn et al., in press) investigated whether depressed patients show an absence of optimistically biased processing when receiving information about the future.

Before I summarize and integrate the results of these studies within a neurocognitive model of self-relevant information processing, I will give an overview of the theoretical and empirical foundations.

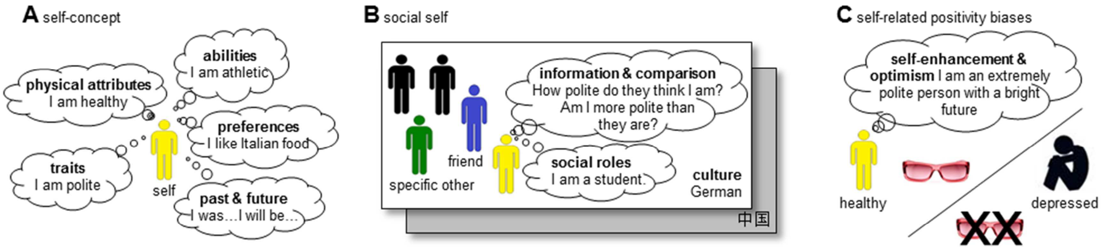

A self-concept
physical attributes I am healthy
abilities I am athletic
preferences I like Italian food
traits I am polite
self
past & future I was…I will be…
B social self
information & comparison How polite do they think I am? Am I more polite than they are?
friend
specific other
social roles I am a student
culture German
中国
C self-related positivity biases
self-enhancement & optimism I am an extremely polite person with a bright future
healthy
depressed

**Figure 1. Three perspectives on the self.**

A. Research on the self-concept focuses on the cognitive representations (depicted as thought balloons) of individuals (Markus and Wurf, 1987; Baumeister, 1998). Self-related cognitive representations can take the form of character traits, perceived physical characteristics, abilities, or sets of preferences. People have a sense of who they were in the past and of who they want to be in the future.

B. Research on the social self emphasizes that the people interact with each other (Banaji and Prentice, 1994; Ellemers, 2012). People have specific social roles and engage in specific relations. Additionally, other people serve as a source of information and a point of reference. Cultural psychology (Heine, 2012) stresses that people's social world is embedded in a cultural context (depicted by the white and grey boxes for Western and East Asian cultures, respectively).

C. Research on self-related positivity biases, especially research on self-enhancement (Taylor and Brown, 1988; Alicke and Sedikides, 2009) and on optimism (Weinstein, 1980; Scheier and Carver, 1994), has shown that healthy individuals tend to put a positive spin (depicted as rose-colored glasses) on self-related cognitive representations. In contrast, psychiatric patients such as depressed patients are characterized by negative cognition (Beck et al., 1979).

## 1.2. Self-related positivity biases

The third perspective outlined above emphasizes that self-related cognition is often biased to the positive in healthy people. A seminal review article (Taylor and Brown, 1988) grouped self-related positivity biases (labeled positive illusions by the authors) into a triad of overly positive self-evaluations, exaggerated perceptions of control, and unrealistic optimism. Of these three, I will focus on positive self-evaluations and optimistic estimates of future life events.

Examples for self-related positivity biases abound. Healthy people tend to rate themselves high on positive and low on negative character traits (Alicke et al., 1995; Leary, 2007; Alicke and Sedikides, 2009). That is, they evaluate themselves more positively than relevant others. Indeed, people tend to rate themselves better than their average peers in contravention of statistical logic (i.e., more than 50% of the participants put themselves into the upper 50% of the distribution). A similar pattern can be found for optimism: People tend to estimate that more positive and less negative events are going to happen to them compared with their average peers (Weinstein, 1980; Alicke et al., 1995). In the same vein, most drivers claim that they are above-average drivers (Svenson, 1981) and most professors think that they are above-average professors (Cross, 1977). Some of this research has been criticized on the grounds that people may have difficulties imagining an average person to whom to compare to (Chambers and Windschitl, 2004; Heine and Hamamura, 2007). But overall positivity biases have been reported across many domains by a substantial amount of studies; many of which are not prone to problems related to comparing themselves to an average person (Alicke et al., 1995; Leary, 2007; Alicke and Sedikides, 2009). For example, people report to experience more positive than negative emotions (Hepach et al., 2011) and they imagine positive future events in more detail—and remember these imagined events better—than negative future events (Sharot et al., 2007; Szpunar et al., 2012). In addition, a prominent research tradition on trait measures has established that most humans tend to show high trait self-esteem and optimism (Rosenberg, 1965; Scheier and Carver, 1992; Scheier et al., 1994; Solberg-Nes and Segerstrom, 2006; Leary, 2007).

Despite the impressive amount of literature on self-related positivity biases, a central proposition of their early conceptualization has been relatively underexplored. Taylor and colleagues (Taylor and Brown, 1988; Taylor et al., 1989) posited that a series of “cognitive filters” distort self-relevant information processing toward the positive. For example, a recent study showed that people expect to receive more positive than negative feedback in social interactions but the authors did not test how participants actually process social feedback (Hepper et al., 2011). Importantly, few studies have used neuroimaging to investigate the neural processes associated with positive self-evaluation (see Beer, 2007; Beer and Hughes, 2010; Hughes and Beer 2010; Somerville et al., 2010 for some notable exceptions) or positive views of the future (Sharot et al., 2007). Using functional magnetic resonance imaging (fMRI), a recent study addressed the behavioral and neural processes which are at play when people’s estimates about the future are challenged by statistical information (Sharot et al., 2011). In line with the idea of “cognitive filters,” participants in that study showed a positive updating bias; they changed their estimates about the likelihood of experiencing adverse life events more toward the statistical likelihood of these events when this statistical information was desirable (i.e., lower than participants’ estimates) than when it was undesirable (i.e., higher than participants’ estimates). On the neural level, the processing of the statistical information was related to activity in the medial prefrontal cortex (MPFC) and inferior frontal gyrus (IFG). The behavioral results of this study have been replicated across two further studies (Sharot et al. 2012a; Sharot et al. 2012b) and form an important basis for the empirical research constituting this thesis.

The notion of bias plays an important role in psychology, behavioral economics, and cognitive neuroscience (e.g., Gigerenzer, 2007; Ariely, 2008; Hogg and Vaughan; 2008; Alicke and Sedikides, 2009; Kahneman, 2011; Sharot, 2011). Since it has been used in different ways, I will outline how I use the term bias in this thesis.

Biases can be defined as violations of rationality. Rationality itself has been defined in various ways (Kacelnik, 2006) but the crucial aspects of rationality in the present context are logic and optimality. One example showing that people's collective behavior defies pure logic has been given above: Not more than 50% of the population can be in the upper 50% of the population (Alicke and Sedikides, 2009). A controversially discussed example is the "Linda the bank teller" problem (see Gigerenzer, 2007; Kahneman, 2011 for introductions into the problem). People tend to say that a certain personality description is less likely to apply to a bank teller than to a feminist bank teller although the category of "bank tellers" includes the category of "feminist bank tellers." Optimality can, for example, be defined in terms of Bayes' law, which describes how information should be integrated (e.g., by sensory systems) (see Friston, 2010 for a discussion of optimality and Bayesian approaches in neuroscience). Bayes' law thus provides a benchmark to empirically test whether human behavior is biased. For example, a noteworthy study has shown that participants adhered quite closely to the Bayesian benchmark when receiving favorable information about their intelligence or their attractiveness—but not when they received unfavorable information (Eli and Rao, 2011).

In this thesis, however, the term bias is not to be understood as a violation of rationality. Instead, in line with much research in psychology, bias means that one condition has a stronger (or a different influence) than another condition. For example, framing biases arise because describing the same event in a positive or negative way influences behavior (see Kahneman, 2011 for a review of the behavioral literature; see De Martino et al., 2006 for a neuroscience example). In-group biases arise because people prefer their in-group over an out-group (Hogg and Vaughan; 2008). The term bias implies that two conditions (e.g., positive/negative framing or in-group/out-group) may be expected to have a similar effect, but there is no independent benchmark to test a violation of rationality in the sense described above. Thus, I use the term positively biased information processing to indicate that participants take information in one condition more into account than in another (i.e., desirable information is taken more into account than undesirable information).

Many scholars have discussed whether positivity biases are adaptive (Taylor and Brown, 1988; Scheier and Carver, 1992; Weinstein and Klein, 1995; Armor and Taylor, 2002; Lovallo and Kahneman, 2003; Haselton and Nettle, 2006). Although the debate may be difficult to settle, several authors tend to agree on a middle position (Leary, 2007; Puri and Robinson, 2007; Alicke and Sedikides, 2009): Mild biases are adaptive for mental health but extreme biases can be detrimental. For example, trait optimism can predict physical and mental health (e.g., Carver, 1989), possibly via effects related to coping (Solberg-Nes and Segerstrom, 2006). But extreme optimism seems to entail overly high risk taking (Puri and Robinson, 2007). In this thesis, I did not directly assess whether positively biased processing confers benefits. Instead, I investigated whether it is absent in depression (see below).

### 1.3. Self-related neural activity

Much research in social neuroscience has taken the first of the three perspectives on the self, which I outlined above, as a starting point and has focused on the neural correlates of how individuals represent their self-concept (for reviews see Northoff and Bermpohl, 2004; Amodio and Frith, 2006; Lieberman, 2007; Lieberman, 2010; Heatherton, 2011; Wagner et al., 2012; see Buckner and Carroll, 2007; Spreng et al., 2009 for a discussions of prospective thinking). In particular, many early fMRI studies on the self were inspired by behavioral research on the self-referential memory advantage, i.e., the mnemonic benefit of linking information to the self-concept (Symons and Johnson, 1997). In these studies, participants typically judge whether or not trait adjectives are self-descriptive. In control conditions, participants indicate whether another set of trait adjectives describes another person (e.g., a friend, a family member, or a public figure such as the current head of state or a famous athlete) and make perceptual judgments about the font in which the adjectives are written (e.g., lower-case/upper-case or italic/bold). Judging the self-relevance of trait adjectives confers a memory advantage since adjectives seen in the self-condition are subsequently better remembered than the adjectives seen in the control conditions (Symons and Johnson, 1997).

When taken to the MRI scanner, the contrast between judgments about the self and about other persons (or about the font of the adjectives) reveals neural activity within cortical midline regions, especially the (ventral) MPFC (e.g., Fossati et al., 2003). The link between ventral MPFC activity and self-related processing has been tightened by a study showing that ventral MPFC activity correlates with the self-referential memory advantage (Macrae et al., 2004). By now, strong meta-analytic evidence has accumulated that confirms the central role of the MPFC in self-related processes across a variety of tasks such as introspecting about one's current traits (as described above) and dispositions, reflecting about oneself in the past and future, or seeing one's face (Northoff et al. 2006; Van Overwalle, 2009; Denny et al., 2012).

Neural activations in self-referential tasks are not limited to the MPFC but include further regions such as the posterior cingulate cortex (PCC), the temporo-parietal junction (TPJ), the anterior insula and the adjacent IFG, as well as the striatum (Denny et al., 2012). Nevertheless, many authors claim that the MPFC plays the most important role because it seems to be the most consistently implicated region in self-referential processing (Amodio and Frith, 2006; Denny et al., 2012; Wagner et al., 2012; Mitchell, 2009). A few lesion studies corroborate the involvement of the MPFC in self-referential processes; patients with lesions in the MPFC show impairments in self-reflection and an absence of the self-referential memory advantage (Philippi et al., 2012; see Wagner et al., 2012 for a discussion).

Within the frontal midline regions, however, different studies have not always used the anatomical labels in consistent ways. Along the anterior-to-posterior axis, some studies distinguish between activity in the superior frontal gyrus, the paracingulate gyrus, and the anterior cingulate gyrus or anterior cingulate cortex (ACC) (e.g., Krienen et al., 2010). Yet, many other studies on self-related neural activity  use the label MPFC to include activity within all of these regions; probably because many clusters seem to encompass both the MPFC proper and the ACC (Denny et al., 2012) (The same holds true for mentalizing. See next section.) For this latter reason, I will also often use the label MPFC activity to refer to activity that extends into ACC. Along the ventral-to-dorsal axis, different naming conventions are in use. Especially, the naming of the most ventral part of the frontal midline (i.e., below z = 0) shows some variation. Studies on reward processing often refer to this region as ventral MPFC, ACC, or medial orbito-frontal cortex (OFC) (Beckmann et al., 2009). Again, I will mostly use the term MPFC.

Several scholars have discussed a ventral-to-dorsal- gradient within the MPFC in relation to the neural correlates of person knowledge (i.e., activity related to different types of other persons) (Amodio and Frith, 2006; Lieberman, 2007; Wagner et al., 2012). As mentioned above, many studies include a condition in which participants judge the traits or dispositions of another person. When contrasted with font judgments, trait judgments of another person reveal activity in MPFC regions that partially overlap with self-related MPFC activity but are slightly more dorsal. Meta-analytic evidence supports the idea of a ventral-to-dorsal- gradient (Denny et al., 2012). More ventral MPFC regions (with the lower border around the MNI coordinate z = 0) seem to be more heavily implicated in self-referential processes and more dorsal MPFC regions seem to be more heavily implicated in processes related to other persons. Interestingly, both judging more similar others (e.g., somebody who shares one's political views versus somebody who does not) and judging more familiar others (e.g., friends versus strangers) seems to elicit greater—and more ventral—MPFC activity (Jenkins, et al. 2008; Krienen et al., 2010). Yet, some controversy remains about which of the two dimensions offers a more parsimonious explanation for the overlap between self- and other-related MPFC activations.

In sum, MPFC activity is consistently involved when participants make trait judgments about themselves (or another person). Yet, only a handful of studies have provided evidence that the role of the MPFC extends to self-relevant feedback processing (Izuma et al., 2008; Somerville et al., 2006; Davey et al., 2010; Eisenberger et al., 2011; Jones et al., 2011).

## 1.4. Mentalizing

When people think about themselves, they often think about themselves in contrast to other people. Consequently, research on the self inherently involves research on the relationships between the self and other people—as can be seen from the fact that almost all studies discussed in the previous section on self-related neural activity involve conditions that implicate another person. The second perspective outlined above, which stresses the role of social interactions for self-related processes, is conceptually linked to research that focuses on processes related to other persons.

This research often takes the question of how people understand other persons as a starting point (see Frith, 2007a; Frith, 2007b; Adolphs, 2009 for introduction). The terms mentalizing and theory of mind are commonly used to refer to the process of inferring another agent's mental state including the other agent's beliefs, goals, or desires. Following Frith and Frith (2012), I will mainly use the term mentalizing since theory of mind suggests that mental state inference is a conscious process, which is not necessarily always the case. Mentalizing implicates that the self attributes a mental state to the other agent or person. In addition, the self often has to realize that the other person's mental state diverges from the self's own mental state. The content of the other person's mental state can refer to the self (e.g., "she thinks that I am chaotic") but in many instances it does not (e.g. "she thinks Tim is nice) and in some instances it does not refer to social agents at all but to physical objects (e.g. "she thinks the ball is in the basket").

Mentalizing has been researched in various disciplines using a variety of tasks. The neuroscientific research on mentalizing is linked to research in at least three different disciplines: biology, developmental psychology, and behavioral economics. First, some forms of mentalizing are not restricted to humans. Chimpanzees and corvids, for example, are able to understand what conspecifics know in the context of food competition (Call and Tomasello, 2008; Emery and Clayton, 2009). Second, (developmental) psychologists have described the temporal trajectory of mentalizing in children and its deficits in autism (Wimmer and Perner, 1983; Frith and Frith, 2003; Low and Perner, 2012). Third, behavioral economics has investigated human interactions—and thereby different mentalizing-related processes—in the framework of game theory (Sanfey, 2007; Walter et al., 2005; Yoshida et al., 2008; Rilling and Sanfey, 2011; Frith and Frith, 2012; see Glimcher et al., 2008 for a wider context), which offers a mathematical description of decision making in well-defined social settings (see Glimcher et al., 2008). In games such as the beauty contest game (Coricelli and Nagel, 2009) or the "matching pennies" game (Hampton et al., 2008), people have to take the mental states of other's into account to obtain the best personal outcome. This outcome is often quantified as monetary gains or losses. In the following, I will discuss different aspects of mentalizing tasks that are relevant for the present thesis. I will then discuss the brain regions implicated in mentalizing.

### 1.4.1. Mentalizing tasks

Research on animals and children has used false belief tasks (often in the form of the "Sally and Ann" task) as the paradigmatic tests for mentalizing (Wimmer and Perner; 1983; see also Amodio and Frith, 2006; Saxe, 2006). In these tasks, participants have to infer that another person has restricted knowledge (e.g., the other person erroneously thinks that the ball is in the basket because she has not seen that the ball had been transferred to the box). In neuroimaging research, false belief tasks have often been presented in the form of short vignettes, which resemble logical puzzles (e.g., Saxe and Powell, 2006), or in the form of comic strips, which depict a sequence of events (Walter et al., 2004; Schnell et al., 2011).

In addition to more or less classic false belief tasks, mentalizing has also been investigated with the help of both very simplified and highly realistic video material. Animations of simple geometric shapes moving in the way of social agents (e.g., a triangle "chasing" a rectangle) can elicit mental state attribution (Castelli et al. 2000). On the other hand, complex videos of real-life social interactions (e.g., a dinner at a friend's home) have been used to probe mentalizing-related activity (Wolf et al., 2010). Mentalizing has also been related to perceiving man-made products (e.g. music written by a composer versus music generated by a computer) (Steinbeis and Koelsch, 2009) as well as to understanding irony which involves inferring that the intended meaning of a statement differs from its literal meaning (Bohrn et al., 2012).

Task derived from game theory involve mentalizing in a specific way: Participants have to incorporate what they think another person plans to do into their own decisions (Walter et al., 2005; Frith and Singer, 2008; Rilling and Sanfey, 2011). While studies using adaptations of the false belief task typically do not involve incentives for correct mental state inference, studies using tasks derived from game theory often quantify the outcomes of participants' decisions in terms of money. In the beauty contest, for example, participants win if they state a number that is equal to a certain fraction (e.g. one half) of the mean of the numbers given by all participants in the experiment (Coricelli and Nagel, 2009). Therefore, participants have to infer which number the other persons are likely to state. In the matching pennies game—a simplified version of rock-paper-scissors—the outcome of the game depends on the simultaneous binary decisions of two players (Hampton et al., 2008). If both players choose the same option player A wins. In contrast, if the two players chose different options player B wins. In iterated rounds of the game, participants can learn about the other person's decisions. Over time participants can thus build a model of the other person—a process that can be described with the help of computational models (for example, derived from reinforcement learning models; see section on reward-related activity) (Behrens et al., 2008; Hampton et al., 2008; Yoshida et al., 2008; Yoshida et al., 2010; for review see Behrens et al., 2009).

In sum, tasks used to study the neural correlates of mentalizing can be classified according to various aspects. First, mentalizing can or cannot entail consequences for the participant (e.g., money won in economic games). Second, tasks can be more or less explicit (e.g., "What does he think?" versus "Choose a number that is close to the half of the mean number given by all participants"). Third, tasks differ in how likely mentalizing is to involve other processes such as logical reasoning (e.g., in some false belief tasks) (e.g., Saxe and Powell, 2006), emotional empathy (Schnell et al., 2011) or reward processing (e.g., in economic games) (e.g., Behrens et al., 2008).

This last point regarding the relation of mentalizing to other processes has often been framed as the pertinent question about how the social aspects of mentalizing relate to its non-social prerequisites (Frith, 2007b; Adolphs, 2009; Adolphs, 2010). Within the context of a specific study, control conditions have to be designed with great care. For example, many studies on economic games include a control condition in which participants play against a computer (see Rilling and Sanfey, 2011 for review). However, this can be problematic given that humans sometimes attribute mental states to machines and programs (e.g., "MATLAB does not understand what I want") or even to geometric shapes as discussed above (Castelli et al. 2000). For this and other reasons, some studies have moved beyond categorical contrasts and have relied on parametric designs (sometimes in conjunction with computational modeling) to disentangle specific components of mentalizing (e.g., Hampton et al., 2008). In a wider context, there has been an important proposition suggesting that social and non-social processes can be described using similar (computational) formalisms (Yoshida et al., 2008; Behrens et al., 2009; Hunt and Behrens, 2011). The hypothesis is that similar formalisms may reflect similar mechanisms on both the behavioral and the neural level.

### 1.4.2. Mentalizing-related neural activity

Two brain regions have played particularly prominent roles in discussions on the neural correlates of mentalizing—the MPFC and the TPJ (e.g., Amodio and Frith, 2006; Saxe, 2006; Bahnemann et al., 2010). As discussed in the previous section, the MPFC is also central for self- and other-related processes such as trait judgments. In line with the proposed ventral-to-dorsal gradient, mentalizing—which per definition is a process related to the mental states of other persons—engages more dorsal parts of the MPFC. In fact, meta-analyses on self- and other-related neural activity have not always drawn a clear-cut line between mentalizing and person knowledge and have often grouped studies on mental state inference and studies on trait attribution together (Denny et al., 2012).

As discussed above for self-related neural activity, many studies reporting mentalizing-related activity use the label MPFC although some clusters encompass the ACC (see Amodio and Frith, 2006). Similarly, the anatomical boundaries of the second prominent region involved in mentalizing, the TPJ, are not clearly defined (Bahnemann et al., 2010). But a recent study has begun to address the structural and functional connectivity of the TPJ (Mars et al., 2012) and suggests that the TPJ can be subdivided into a dorsal cluster (in the middle part of the inferior parietal lobule), and two ventral clusters (one more anterior and one more posterior).

By now, meta-analyses have firmly established the role of the MPFC and the TPJ in mentalizing (Spreng et al., 2009; Van Overwalle, 2009; Bahnemann et al., 2010; Mar, 2011). However, their specific contributions to specific aspects of mentalizing remain a matter of debate. A recent meta-analysis distinguishes between story-based studies (i.e., studies employing false belief tasks) and non-story-based studies (Mar, 2011). Non-story-based studies employ a greater variety of stimuli and tasks including cartoons, videos, and economic games. In particular, studies involving economic games are heterogeneous because they aim at investigating diverse components of social interactions and have therefore mainly been summarized qualitatively rather than quantitatively (Rilling and Sanfey, 2011; Frith and Frith, 2012). The distinction between story-based and non-story-based studies suggests that story-based tasks elicit relatively more TPJ activity whereas non-story-based tasks elicit relatively more MPFC activity (Mar, 2011).

While many early discussions have focused on the MPFC and the TPJ, meta-analyses have firmly established that mentalizing-related processes involve the superior temporal sulcus (STS), the temporal pole (TP), the IFG, especially its orbital part, the PCC, and the pre-supplementary motor area (preSMA) (Van Overwalle, 2009; Mar, 2011). The STS lies ventral and anterior to the TPJ and has been especially implicated in the detection of biological motion (Hein and Knight, 2008). The TP seems to have a specific role when social scripts (e.g., what to do in a restaurant) become relevant (Olson et al., 2007). In contrast to the well-known role of the IFG in linguistic processes, the IFG's role in mentalizing has been somewhat neglected. Its involvement seems to be particularly prominent in non-story-based tasks (Mar, 2011). Similar to the IFG, the PCC has not often been explicitly discussed (Mar, 2011). Activity in the preSMA often forms a contiguous cluster with the dorsal MPFC and tends therefore not to be mentioned specifically (see Mar, 2011).

Taken together, mentalizing comprises a variety of aspects such as detecting that another person has limited knowledge, imbuing geometric shapes with intentions, and interacting with others in an economic transaction. Mentalizing engages a network of brain regions including the MPFC, TPJ, STS, TP, orbital IFG, and the PCC. Although mentalizing and self-related processes seem tightly related on the behavioral and neural level, it remains relatively unclear if mentalizing-related regions play a role when people receive self-relevant social feedback.

## 1.5. Reward-related neural activity

The third perspective on the self, which focuses on self-related positivity biases, suggests that self-related processes are imbued with valence or some kind of rewarding value (e.g., Northoff and Hayes, 2011). Reward has been investigated by an immense literature both in humans and in animals (see Glimcher et al., 2008 for an overview). Within the last years, the study of reward has been central to the interdisciplinary field of neuroeconomics or decision neuroscience (Glimcher et al., 2008). One strand of research within this field has focused on reward processing in non-social contexts (Montague et al., 2006; Rangel et al., 2008; Rushworth and Behrens, 2008; Beckmann et al., 2009; Lee et al., 2012) while another closely related strand of research has focused on reward processing within the context of social interactions (Walter et al., 2005;

Montague et al., 2006; Fehr and Camerer, 2007; Frith and Singer, 2008; Lee, 2008; Rilling and Sanfey, 2011). This latter strand of research often coincides with the research on social interactions within economic games that I introduced above.

Reward processing has the tremendous advantage that its neural correlates can be studied in animals such as macaques, rats, and mice. Animal research on reward processing has often focused on dopaminergic midbrain regions and on the ventral part of the MPFC (often called OFC) (e.g., see Schultz, 2006; Lee et al., 2012 for reviews). A prominent line of research has established that the firing of dopaminergic neurons located in the midbrain track properties of rewarding stimuli such as food (a primary reinforcers) or cues predicting food (secondary reinforcers) (Schultz, 2006). Specifically, the firing of these neurons shows properties that can be described in the framework of reinforcement learning. That is, dopaminergic signals scale with prediction errors, which have at least two basic characteristics (e.g., Tobler et al., 2005; see Schultz, 2006 for review): First, neurons fire when reward delivery is unexpected. Second, when an animal learns that a cue predicts subsequent reward delivery, dopaminergic neurons fire when the cue appears but not when the reward is delivered. Importantly, these processes can be modeled with algorithms derived from reinforcement learning (Montague et al., 2006; Dayan, 2012; Lee et al., 2012).

Research in humans using fMRI has been largely consistent with the neural recordings in animals (Montague et al., 2006; Rushworth and Behrens, 2008; see Glimcher et al., 2008 for a general overview). Due to the nature of the blood oxygen level dependent (BOLD) signal which is the basis for fMRI (Logothetis, 2008), studies on humans have mostly reported BOLD signal changes (commonly referred to as “activity”) in response to receiving reward in the target regions of dopaminergic input—especially in the striatum (and in the ventral part of the MPFC) (see Glimcher et al., 2008 for overview; see e.g., Park et al., 2012 for a study relating activity in the striatum with activity in the MPFC and in the midbrain).

The proposition that the brain converts activity related to different types of rewards into a common currency has been very influential for the research on social reward (see e.g., Sanfey, 2007). The idea is that potential rewards form different sources (e.g., food, money, social feedback) have to be scaled in the same metric in order to allow comparisons between actions leading to different reward types. Studies on the reward-related components of social interactions have often shown activity in regions that overlapped with those found for non-social reward (for reviews see Montague et al., 2006; Fehr and Camerer, 2007; Rushworth et al., 2007; Rilling and Sanfey, 2011)—although only a few studies have directly tested for an overlap of social and non-social reward processing (e.g., Izuma et al., 2008; Zaki et al., 2011). The striatum has for example been implicated in the processing of advice (Behrens et al., 2008; Biele et al., 2011; Meshi et al., 2012), social hierarchy (Zink et al., 2008), trust (King-Casas et al., 2005), and social comparison (Fliessbach et al., 2007). The representation of the value of objects has been consistently associated with activity in the ventral part of the MPFC—and the value of these objects is often modulated by social influences (Erk, et al., 2002; Plassmann et al., 2008; Zaki et al., 2011). Importantly, it has also been suggested that—even in the context of social interactions—activity in the reward circuitry shows aspects akin to prediction errors, which can be described by reinforcement learning models (King-Casas et al., 2005; Behrens et al., 2008).

Thus, as in the case of mentalizing, it is hypothesized that similar formalisms may reflect similar mechanisms of social and non-social reward processing (Fehr and Camerer, 2007; Behrens et al., 2009; Hunt and Behrens, 2011). Yet, the role that reward processing plays for the dynamics of self-concept changes remains underexplored.

## 1.6. Culture

Much research in psychology and cognitive neuroscience assumes to investigate universal aspects of cognition but more than 95% of psychological and neuroscientific studies rely on participants from Western industrialized countries (Henrich et al., 2010). Within the last few years, findings in cultural psychology and in the nascent field of cultural neuroscience have challenged the universality of many aspects of cognition, in particular in the domain of social cognition. By investigating how people's wider sociocultural background influences their cognition (Kitayama and Uskul, 2011; Han and Northoff, 2008; Heine, 2012; Han et al., 2013), this strand of research puts a strong emphasis on social interactions, which are central to the second perspective on the self that I outlined above.

Culture has been defined in many different ways (Heine, 2012; Han et al., 2013). Three aspects are of relevance. First, culture can refer to the fact that humans (and some animals) produce material artifacts such as tools for hunting and farming. Second, culture relates to the variety of social institutions and customs such as different wedding ceremonies. Third, culture refers to the fact that individuals within a given culture share common beliefs, values, and behavioral scripts such as the belief that one should honor one's parents. These three aspects are dynamically interrelated but the last aspect is of special importance for cultural psychology and cultural neuroscience since they aim at elucidating how an individual's cultural background influences this person's cognition (Chiao and Ambady, 2007; Han and Northoff, 2009; Kitayama and Uskul, 2011; Han et al., 2013). Thus, studies often compare individuals from different cultural backgrounds. The preponderance of research has compared Westerners (including North Americans, Europeans, and Australians) with East Asians (including Japanese, Chinese, and Koreans) (Henrich et al., 2010; Heine, 2012). But many studies have also compared individuals from industrialized countries with those from non-industrialized small scale societies, individuals of different religions, or individuals of different social classes (for an overview see Henrich et al., 2010; Heine, 2012).

In terms of research topic, investigations about how people's self-concept varies across different cultures have been especially prominent. Cultural differences in independent versus interdependent self-concepts (often called self-construal in cultural psychology) form the best-researched dimension (Markus and Kitayama, 1991; Oyserman et al., 2002; Triandis and Suh, 2002; Markus and Kitayama, 2010; Heine, 2012; for other important distinctions e.g., in terms of analytic versus holistic cognition see Nisbett et al., 2001; Nisbett et al., 2003; Heine and Buchtel, 2009). That is, individuals with an independent (or individualistic) self-concept construe their selves as relatively distinct from others (**Figure 2A**) while individuals with an interdependent (or collectivistic) self-concept construe their selves as tightly interconnected with close others (**Figure 2B**) (Markus and Kitayama, 2010). Differences in independent versus interdependent self-concepts seem to underlie many of the differences observed in Westerners versus East Asians (Oyserman et al., 2002; Markus and Kitayama, 2010;). People's self-concepts, however, are not supposed to be static with respect to the independent-interdependent dimension. Individuals—especially bi-cultural individuals such as people from Hong Kong—can be primed to change their self-concept dynamically (e.g., by reading stories about individuals or groups or by seeing cultural symbols pertaining to Western or East Asian cultures) (Oyserman et al., 2002; Chiao et al., 2009b; Ng et al., 2010).

Explaining cultural differences by underlying differences in independent and interdependent self-concepts relies on the idea that social interactions vary between cultures (Markus and Kitayama, 1991; Markus and Kitayama, 2010). Cultural differences in social interactions have for example been reported for social support (Kim et al., 2008) and social conformity (Bond and Smith, 1996). Compared with East Asians, Westerners seem to seek for social support in more explicit ways (e.g., by discussing stressful events and disclosing personal feelings of distress). East Asians seem to be reluctant to directly ask for social support from another person because they are concerned about the potential negative consequences for their relationship to the other person (Kim et al., 2008). Regarding social conformity, a meta-analytic analysis of studies using classic Asch-type line judgment task indicates that individuals from more interdependent cultures tend to show higher social conformity, i.e., they rely more on other people's opinion when judging the lengths of two lines (Bond and Smith, 1996; see also Cialdini and Goldstein, 2004; Heine, 2012). This is in accord with evidence suggesting that Westerners and East Asians think differently about "conformity" and "uniqueness" (Kim and Markus, 1999). Conformity tends to have a positive connotation in East Asian cultures while uniqueness tends to be positively valued in Western cultures. In a similar vein, a recent study suggests that priming interdependence undermines the motivation of independent but not of interdependent individuals (Hamedani et al., 2013). Thus, cultural differences in self-concept have been linked to differences in social support and social conformity but it remains unclear whether self-relevant information processing differs across culture.

An important debate in cultural psychology is related to self-related positivity biases, i.e., to the third perspective on the self outlined above. The huge majority of evidence described in the section on self-related positivity biases (see above) has been obtained from Western participants (e.g., Taylor and Brown, 1988; Alicke et al., 1995; Leary, 2007; Alicke and Sedikides, 2009). Some authors claim that—in contrast to Westerners—East Asians do not show self-related positivity biases (see Heine et al., 2001 for an early description). Several meta-analyses have been conducted. Some of them show evidence for East Asian self-enhancement (Sedikides et al., 2003; Sedikides et al., 2007) and some of them show evidence against it (Heine et al., 2007; Heine and Hamamura, 2007). The meta-analyses differ in their definition of self-enhancement and thus in their inclusion criteria. Furthermore, a caveat of some studies on self-related positivity biases—which I mentioned above—plays an important role in this debate. Demonstrations of above average comparisons may be confounded by participants' difficulties to imagine an average person to whom to compare to (see Heine and Hamamura, 2007). Therefore, novel approaches to self-related biases—such as self-relevant information processing—might help to settle the described debate.

Recently, a growing number of studies in the emergent field of cultural neuroscience have investigated cultural differences in neural activity (for reviews see Han and Northoff, 2008; Vogeley and Roepstorff, 2009; Han and Northoff, 2009; Kitayama and Uskul, 2011; Han et al., 2013). Much of this research has taken the findings described in the section on self-related neural activity as a starting point. The first fMRI study suggesting a link of neural activity to cultural differences in self-concept has reported that in East Asians—but not in Westerners—MPFC activity for trait-judgments about self and mother overlapped (Zhu et al., 2007). Since then, a couple of further studies have shown cultural influences on MPFC activity related to trait judgments (Chiao et al., 2009a; Chiao et al., 2009b; Ng et al., 2010; Ray et al., 2010; Wang et al., 2012) or stimuli that are perceived differently across cultures (e.g., Freeman et al., 2009). For example, a recent study showed that cultural modulation of MPFC activity extends to judgments about social roles and physical attributes (Ma et al., 2012).

Taken together, a potential model of self-relevant information processing should be based on data obtained from participants of different cultural origins to avoid that it is restricted to Western samples. Comparing Western (e.g., German) and East Asian (e.g., Chinese) participants has the potential to provide novel evidence for social conformity, positivity biases, and the role of self-related MPFC activity.

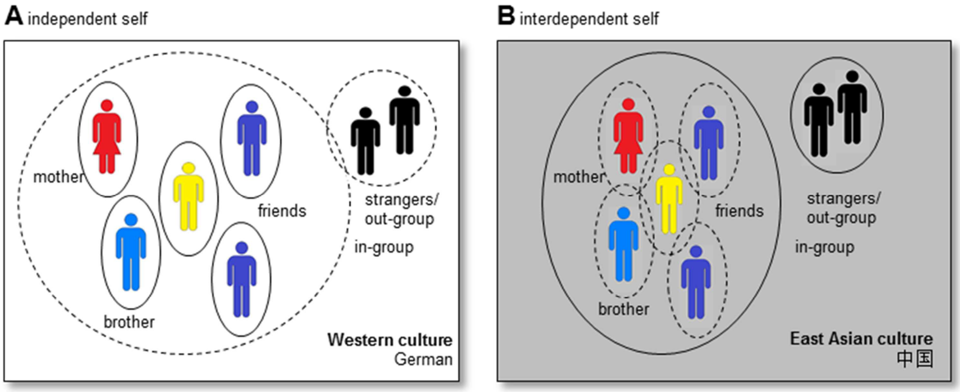


**Figure 2. Cultural differences in self-concept.**

A. Individuals with an independent self-concept tend to see their own and others' self-concept as relatively separate (depicted by small non-overlapping ovals with lines that are not dashed). The difference between in-groups and out-groups is relatively loose (depicted by large overlapping ovals with dashed lines). Independent self-concepts tend to prevail in Western culture.

B. Individuals with an interdependent self-concept tend to see their self-concept as overlapping with the self-concept of close others (depicted by small overlapping ovals with dashed lines). The difference between in-groups and out-groups is relatively clear-cut (depicted by large non-overlapping ovals with lines that are not dashed). Interdependent self-concepts tend to prevail in East Asian culture.

Figure adapted from Markus and Kitayama (2010).

## 1.7. Depression

The third perspective on the self outlined above is relevant for psychiatric disorders such as depression because it emphasizes that healthy humans are motivated to establish or maintain a positive view of themselves and of their future (e.g. Taylor and Brown, 1988; Leary, 2007). Overall, major depressive disorder is classified as an affective disorder characterized by a constellation of physical, emotional, and cognitive symptoms (e.g., psychomotor abnormalities, weight loss, altered appetite, fatigue, sleeping problems, anhedonia, feelings of worthlessness, suicidal ideation, and concentration difficulties) (American Psychiatric Association, 2000). Depression is a highly recurrent disorder with more than 75% of patients experiencing more than one depressive episode. Depression is one of the most prevalent psychiatric disorders with a life-time prevalence of around 15% (Moussavi et al., 2007). The World Health Organization ranks depression as the single most burdensome disease among people in the middle years of life (Murray and Lopez, 1996).

Negative cognitive biases about the self, the world, and the future lie at the heart of prominent cognitive theories of depression such as Beck's cognitive model (Beck et al., 1979; Disner et al., 2011), Seligman's learned helplessness model (Seligman, 1972), or the more recent cognitive neuropsychological model (Clark et al., 2009). Commonly used psychotherapies such as cognitive behavior therapy reflect the pivotal role of negative biases in depression since these therapeutic approaches aim at abolishing maladaptive cognition (e.g., Beck, 2005) More recent approaches suggest interventions on the basis of positive psychology (Sin and Lyubomirsky, 2009) and integrate discussions of neurobiology (Roiser et al., 2012). Depressed patients show negative biases in many aspects of cognition including memory, attention, and executive functions (Mathews and MacLeod, 2005; Gotlib and Joorman, 2010). But not all aspects of information processing seem to be negatively biased (Gotlib and Joorman, 2010). Specific processing aberrances with regard to negative material include increased elaboration, diminished disengagement, and deficient cognitive control. For example, depressive patients remember more negative than positive words, spend more time looking at sad pictures than controls, and have difficulties ignoring irrelevant negative material in specific contexts (see Gotlib and Joorman, 2010 for review).

Overall, it seems to be an open debate whether depressed patients are better characterized by negative biases (i.e., altered responses to negative but not to positive stimuli) or by blunted responses (i.e., an insensitivity to both negative and positive stimuli) (Gotlib and Joorman, 2010). This debate also pervades the literature that links depression to altered reward and punishment processing (Eshel and Roiser, 2010). For example, some studies involving learning from reward and punishment have found evidence suggesting negative biases (e.g., Conklin et al., 2009) while others have found evidence for blunted responses (e.g., Chase et al., 2010). Other studies have reported that depressive individuals show hypersensitivity to uninformative negative feedback in a reversal learning task (Murphy et al., 2003) and hyposensitivity to rewarding feedback—as demonstrated by signal-detection (Pizzagalli et al. 2005; Pizzagalli et al. 2008) and computational reinforcement learning approaches (Huys et al. 2009).

Taken together, strong evidence indicates that depression is characterized by negative biases in memory and executive functions as well as altered learning from reinforcement, although the precise mechanisms are not yet entirely clear. Most of the research discussed above tended to use material that is not directly self-relevant (e.g., lists of emotional words or shapes predicting rewards or punishments). In contrast, some studies have taken a more ecologically realistic approach, for example by asking participants to estimate their likelihood of experiencing positive and negative everyday life events within the next month (e.g. being invited to a party or getting a parking ticket) (Strunk et al., 2006; Strunk and Adler, 2009). After the one month period, participants with high depressive symptoms reported experiencing more positive and less negative events than they had expected, which underscores the that depressive individuals show pervasive pessimism about their future.

Thus, combining aspects from studies on reward and punishment with aspects from studies on pessimism about future live events may provide a powerful means to test whether depressive patients show alterations in self-relevant information processing.

## 2. Research questions and hypotheses

The general aim of this thesis is to advance the understanding of how humans process self-relevant information. The three studies constituting this thesis (Korn et al., 2012;Korn et al., submitted; Korn et al., in press) investigated the following four empirical questions.

Question 1: Is self-relevant information processing positively biased in healthy people? Specifically, does desirable compared with undesirable information lead to a greater change in people's self-concepts about their character traits (Studies 1 and 2) or their future (Study 3)?

Question 2: Is self-relevant information processing linked to neural activity associated with reward and mentalizing? (Studies 1 and 2)

Question 3: Does people's cultural background influence the behavioral and neural components of social information processing? Specifically, do East Asians show higher social conformity and reduced self-enhancement compared with Westerners? Do Westerners show enhanced MPFC activity compared with East Asians? (Study 2)

Question 4: Do depressed patients show an absence of optimistically biased processing of information about their personal future? (Study 3)

To address these questions, self-relevant information processing is supposed to encompass the following three constituents (**Figure 3**):

A. The *self as recipient of self-relevant information* entertains a certain self-concept (i.e., certain prior cognitive representations for example about character traits and possible future life events). Thus, investigating the self-concept of the recipient (and its neural mechanisms) relates to the first perspective on the self as outlined above. Comparing different groups of recipients (e.g., recipients from different cultures or healthy versus depressed recipients) allows to address questions regarding the second and third perspective on the self (i.e., questions related to the social self and self-related positivity biases).

B. Potential *sources of self-relevant information* include feedback by peers or statistical information obtained from media. Investigating social sources of self-relevant feedback is closely related to the second perspective on the self, which emphasizes people's social interactions.

C. The *self-relevant information* itself has certain properties in relation to the recipient's prior cognitive representations such as desirability and similarity. The desirability of the information is especially important to answer questions regarding the third perspective on the self, which emphasizes the role of positivity biases in self-related processes.

Self-relevant information processing implies that recipients track properties of incoming information, and consequently update the representations within their self-concept. Thus, this schema of self-relevant information processing stresses the dynamic nature of the self-concept (Markus and Wurf, 1987).

Question 1 concerns the desirability of self-relevant information. Based on previous research on self-related positivity biases, desirable information was hypothesized to entail greater updates than undesirable information. In studies 1 and 2, peers with whom recipients had engaged in a face-to-face interaction functioned as source of information. Desirable (undesirable) information consisted of trait ratings from peers that were more positive (negative) than the recipients' own trait ratings. In study 3, statistical information on negative life events (such as robbery or Parkinson's disease) functioned as source of information. Desirable (undesirable) information consisted of average probabilities for experiencing negative events that were lower (higher) than the recipients' own estimates. Updates were measured as the differences between recipients’ trait ratings or probability estimates before and after receiving self-relevant information. To test for positively biased updating, we statistically compared the absolute magnitude of updates for desirable and undesirable information.

Question 2 concerns the neural mechanisms of social information processing. Specifically, it concerns the neural mechanisms of processing social feedback on character traits, which were tested in studies 1 and 2. Social information processing was hypothesized to comprise aspects of social reward and mentalizing. To identify brain regions related to reward and mentalizing, my co-authors and I searched for activity that varied parametrically with the reward- and the comparison-related aspects of social feedback in a trial-by-trial fashion. We tested for reward-related activity by searching for activity that correlated positively with the feedback ratings for self, i.e., higher feedback ratings (e.g., an 8 versus a 6 for polite) indicated greater social reward. We required reward-related activity to be self-specific by contrasting activity that correlated with feedback ratings for self with activity that correlated with feedback ratings for another person. We tested for activity within the mentalizing network by searching for activity that correlated on a trial-by-trial fashion with what we call the social comparison component. This component was operationalized as feedback discrepancies, i.e., the absolute differences between recipients’ own trait ratings and the feedback ratings they received. Larger feedback discrepancies (e.g., a feedback rating of 8 versus 6 given that the participant’s own rating was 5) are conceived as a “greater” comparison process between own ratings and the feedback received.

Question 3 extends social information processing to cultural psychology and cultural neuroscience, which emphasize that persons are embedded within their social surrounding (Heine, 2012; Han et al., 2013). In study 2, my co-authors and I compared recipients from Western and East Asian cultural backgrounds. We tested the hypotheses that individuals with a more interdependent self-concept conform more to social feedback (as measured by larger updates) and that they show reduced positively biased updating (as would be indicated by an interaction in updating between feedback desirability and Western versus East Asian participants). Based on previous studies in cultural neuroscience, we hypothesized that self-related feedback would lead to stronger MPFC activity in Westerners compared with East Asians.

Question 4 relates social information processing to psychiatric diseases. Healthy people tend to be optimistic. In contrast, depression is characterized by pessimism (American Psychiatric Association; 2000). My co-authors and I therefore hypothesized that depressed patients would show a reduction of optimistically biased belief updating when receiving self-relevant information about their likelihood of experiencing future life events (as would be indicated by an interaction in updating between feedback desirability and healthy versus depressed participants).

In sum, the empirical studies of this thesis investigated processes related to the self within an information processing framework. By focusing on cultural differences and differences between healthy controls and depressive patients, studies 2 and 3 aimed at corroborating possible components of this information processing framework and at making it relevant for cultural neuroscience and psychiatry.

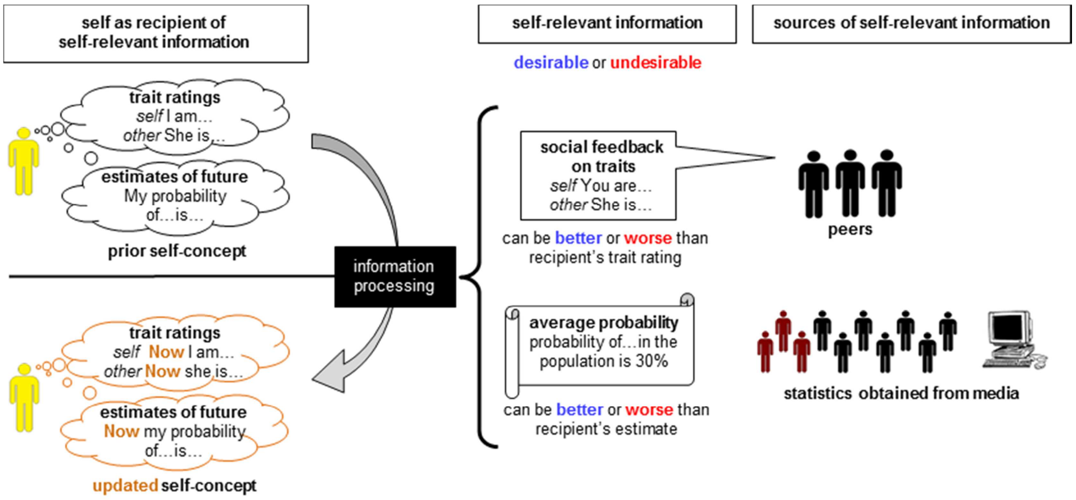

self as recipient of self-relevant information
self-relevant information
sources of self-relevant information
desirable or undesirable
trait ratings
self I am…
other She is…
estimates of future
My probability of…is…
prior self-concept
information processing
social feedback on traits
self You are…
other She is…
can be better or worse than recipient's trait rating
peers
average probability
probability of…in the population is 30%
can be better or worse than recipient's estimate
statistics obtained from media
trait ratings
self Now I am…
other Now she is…
estimates of future
Now my probability of…is…
updated self-concept

**Figure 3. Schematic overview of self-relevant information processing.**

When receiving self-relevant information, people are supposed to update the cognitive representations within their self-concept (for example, ratings of character traits or estimates of the probability of future life events). Self-relevant information can originate from different sources (e.g., social feedback can be obtained from peers within a social interaction and statistics can be obtained from media). Research on self-related positivity biases (Taylor and Brown, 1988; Leary, 2007) suggests that one important property of self-relevant information is its desirability. All constituents of this schema could influence information processing. The empirical studies of this thesis aim at establishing mechanisms of self-relevant information processing by varying properties of the information (desirability: studies 1, 2, and 3) and properties of the recipient (cultural background: study 2; depression: study 3), as well as by investigating two types of sources (peers giving social feedback: studies 1 and 2; statistics about future life events: study 3).

## 3. General methodology

In the following section, I will briefly outline the general methodology of the three empirical studies constituting this thesis (Korn et al., 2012;Korn et al., submitted; Korn et al., in press). In particular, I will focus on the overall task structure, the real-life interaction, and the testing of cultural influences. Please refer to the methods sections of the three studies for a complete description of the respective methodological details.

### 3.1. Task structure

Following the general schema of self-relevant information processing as outlined in **Figure 3**, participants in all three studies first gave an explicit indication of the current representations of their self-concept. In studies 1 and 2, they rated their standing on 40 positive and 40 negative trait adjectives (such as polite or arrogant) on a Likert scale ranging from 1 (trait does not apply at all) to 8 (trait does apply very much). In study 3, participants estimated the average likelihood of experiencing 70 negative life events (e.g., Alzheimer's disease, divorce) during the rest of their lives. Study 3 was adapted from a previously published study (Sharot et al., 2011). Immediately after giving a first rating for a certain trait or after giving a first estimate for a certain event, participants received social information. In studies 1 and 2, participants saw how three peers had on average rated them on the given trait. In study 3, participants saw statistical information about how likely the given event is to happen to a person of the same sociocultural background (i.e., participants saw the base rates of the events). After receiving information for all 80 trait adjectives or all 70 life events, participants rated all 80 trait adjectives or estimated all 70 life events a second time in a separate session. The time lag between the first session (which included the first ratings or estimates as well as the self-relevant information) and the second session (which included the second ratings or estimates) lay in the order of a few minutes. In the fMRI studies, only the first session was scanned. Thus, these studies are agnostic about neural processes occurring during the second ratings.

The 80 trait adjectives were selected on the basis of an extensive list of trait adjectives (Anderson, 1968), which had been used to create stimuli for previous experiments in social neuroscience (Fossati et al., 2003; Izuma et al., 2008), and on the basis of the Berlin Affective Word List (Vo et al., 2006). The 70 life events were adapted from a previous study (Sharot et al., 2011). From the initial list of 80 events, 10 events were excluded because they had a straightforward relationship to depressive symptomatology (e.g., insomnia).

For behavioral analyses, trials were split according to desirability as outlined in the section on the hypothesis regarding the first research question. In studies 1 and 2, desirable (undesirable) information consisted of feedback ratings from peers that were more positive (negative) than participants' own ratings. In study 3, desirable (undesirable) information consisted of average probabilities for experiencing negative events that were lower (higher) than participants' own estimates. Updates were calculated as the differences between participants' trait ratings or likelihood estimates before and after receiving feedback. Absolute mean updates of desirable trials were compared with those of undesirable trials to test for positively biased updating. The numerical differences between participants' first own ratings or estimates and the feedback ratings or statistical numbers they received were calculated and conceptualized as feedback discrepancies (in the case of feedback on character traits) or estimation errors (in the case of statistical information on future life events).

In the fMRI studies, feedback ratings and feedback discrepancies were used to search for reward- and comparison-related brain regions, respectively. That is, the main fMRI analyses were based on parametric modulators that tested for brain activity correlating with feedback ratings and feedback discrepancies on a trial-by-trial basis. Please see the outline of the hypotheses regarding the second question above and the methods sections of studies 1 and 2 for more details.

### 3.2. Real-life interaction

To make social feedback about character traits relevant for the self, studies 1 and 2 included a face-to-face interaction of five peers. So far only a few studies have combined neuroimaging with real-life interactions (Redcay et al., 2010; Cooper et al., 2012) or real-life outcomes (Falk et al., 2011). In many studies, researchers purposefully exclude real-life interactions because the social interactions under investigation should exclusively occur within the setting of an economic game. However, social feedback on a participant's character traits can only be meaningful if given by somebody who has some knowledge about the participant's character. Most previous studies on social feedback have collected questionnaires, photos, or videos from participants and then have told them that unknown others have evaluated them based on this material (Izuma et al., 2008; Davey et al., 2010; Eisenberger et al., 2010; Somerville et al., 2010).

My co-authors and I tried to balance experimental control and ecological realism by combining the experimental task described above with a prior real life interaction, in which participants played the table-top version of a popular board game. The experimental task was adapted to the MRI scanner and allowed us to manipulate the feedback ratings presented. In the real life interaction, each participant had 1h and 15 min to get to know the personality of four peers of the same sex while playing the well-known game Monopoly (Hasbro). We chose this game because it is highly engaging and allows players to show a variety of cooperative and competitive behaviors. Participants were free to speak about whatever topics they wanted (e.g., their current occupation, hobbies, or past experiences). Subsequent to the social interaction, each participant rated three other participants on 80 trait adjectives, i.e., in turn each participant was rated by three other participants. Participants believed that the mean of these three ratings was presented during the experimental task (see above) but in reality these ratings were manipulated during the task to ensure experimental control. To exclude that winning or losing in the board game had an effect on behavior in the task, we tested whether any task-related variable correlated with the rank order in the board game (i.e., the first rank was assigned to the winner and so on). This was not the case.

In addition to making the social feedback self-relevant, the social interaction and the rating of three other persons had two further advantages. First, participants had a direct experience of rating other persons. Second, one of the interaction partners could be chosen for the "other-condition." Thus, in contrast to many other studies (e.g., Zhu et al., 2007), participants did not have to compare themselves to an imagined average person or to a famous person (such as the current head of state, who differs from the average participant in various characteristics such as social status).

### 3.3. Testing cultural influences

Comparing participants from different cultures poses a number of challenges such as choosing cultures which differ along a theoretically relevant dimension and controlling for possible effects of the place where participants are tested (Chiao et al., 2010; Heine, 2012).

First, the aim of cultural comparisons is usually not to simply list cultural differences but to integrate such differences with respect to an underlying theoretical construct (Heine, 2012; Han et al., 2013). In study 2, we aimed at testing the influence of cultural differences in independent and interdependent self-concepts on social feedback processing—specifically on social conformity, self-enhancement, and MPFC activity. Therefore, we chose participants from two cultures known to differ with respect to independence and interdependence: Germans and Chinese (Markus and Kitayama, 1991; Heine, 2012). This follows a common approach, since the majority of research in cultural psychology and cultural neuroscience compares Westerners with East Asians (Oyserman et al., 2002; Heine, 2012). To remedy the fact that we recruited participants on the basis of their background culture, we assessed participants' explicit endorsement of independence and interdependence on the Singelis self-construal scale (Singelis, 1994) and correlated the scores of this scale with measures of social conformity and MPFC activity.

A second problem for cultural comparisons is to control for the place where participants are tested. We tested both Germans and Chinese in Berlin and Beijing. In Berlin, participants were tested in the MRI scanner. In Beijing, participants were tested behaviorally. This allowed us to avoid possible confounds related to using two different MRI scanners (Chiao et al., 2010; Han et al., 2013). But it still allowed us to analyze behavioral data from participants who were tested in their native cultural context. In addition, testing both cultural groups in both places allowed us to explicitly test for possible effects of place. Participants living outside their native culture (called sojourners) might show greater social conformity because living in a foreign culture might trigger a general state of insecurity and meeting compatriots in a foreign country might create strong in-group feelings (Sam and Berry, 2010; Heine, 2012).

## 4. Summary of empirical studies

In this chapter, I will briefly summarize the three empirical studies which constitute this thesis (Korn et al., 2012;Korn et al., submitted; Korn et al., in press).

### 4.1. Study 1

**Positively-biased processing of self-relevant social feedback**

In many everyday interactions, humans receive social feedback about their character traits and have to integrate this feedback into their self-concept. In study 1, my co-authors and I investigated whether healthy participants with a Western cultural background process social feedback from peers in a positively biased way. Research on positivity biases assumes that incoming self-relevant information is distorted in a positive direction (Taylor and Brown, 1988). However, most previous studies have only measured positive self-evaluations and not how self-relevant information impacts on self-evaluations. In addition, the neural mechanisms of social feedback processing are underexplored.

Participants (final n = 27) engaged in a real life social interaction in groups of five and consequently rated each other on 40 positive and 40 negative trait adjectives (e.g., tolerant, selfish). On the following day, participants rated themselves on the same traits while lying in the MRI scanner. Immediately after each self-rating, participants received social feedback in the form of ratings, which they believed three of their interaction partners had given on the previous day. In reality, feedback ratings were manipulated to ensure experimental control. Additionally, participants rated one of their interaction partners in the scanner and received feedback for this person (other-condition). Outside the scanner, participants rated themselves and the other person a second time so that we could assess how much they updated their self- and other-evaluations after receiving social feedback. Importantly, feedback could be desirable (i.e., more positive than participants' own evaluation) or undesirable (i.e., more negative than participants' own evaluation). On the neural level, we searched for activity that correlated with two components of social feedback on a trial-by-trial level. The reward-related component was tested using feedback ratings as parametric modulators and the comparison-related component was tested using feedback discrepancies (i.e., the absolute differences between own ratings and feedback ratings) as parametric modulators.

Our results indicated that participants changed their self- and other-evaluations more toward desirable than toward undesirable social feedback, which indicates a positive updating bias. Control analyses excluded that this bias was driven by effects related to the rating scale (i.e., positive updating was not driven by trials in which participants initially gave the highest or lowest ratings). Furthermore, memory for desirable and undesirable social feedback did not differ and the updating bias did not differ between positive and negative trait adjectives. Our fMRI analyses showed that activity within the frontal midline, including the MPFC, was more pronounced when participants received feedback about themselves than when they received feedback about the other person. Importantly, BOLD signal changes within the ventral striatum and the ACC/MPFC correlated with the rewarding component of social feedback and BOLD signal changes within the mentalizing network (including the MPFC, TPJ, STS, TP, IFG, and preSMA) correlated with the social comparison component. Activity within the mentalizing network has a parsimonious explanation in the context of our task, i.e., activity correlated with the absolute differences between participants' own evaluations and the feedback they received. To identify activity common to both reward and social comparison, we performed a conjunction analysis, which revealed a cluster within the MPFC. Activity in this integration region correlated with the behavioral updating bias across participants.

In sum, the results obtained by my co-authors and me in study 1 identify a positively biased updating mechanism for social feedback on character traits. They underscore the importance of integrating theories on reward and mentalizing for a better understanding of the human self-concept.

## 4.2. Study 2

### Cultural influences on social feedback processing

Cultural differences between independent and interdependent self-concepts have emerged as the key framework for understanding how social aspects of human cognition vary across cultures (Markus and Kitayama, 1991; Oyserman et al., 2002; Triandis and Suh, 2002; Markus and Kitayama, 2010; Heine, 2012). This framework relies on the idea that social interactions differ between cultures that foster independence (such as Western culture) and cultures that promote interdependence (such as East Asian culture). Yet, how culture influences the processing of self-relevant feedback from others has not been investigated—which is surprising given that the relation between self and others lies at the heart of the idea of independent and interdependent self-concepts. Most previous studies on cultural differences tested participants in the solitude of a test cubicle or fMRI scanner. Here, my co-authors and I tested how German and Chinese participants processed social feedback on character traits, which was obtained within the context of a face-to-face interaction. Specifically, we aimed at adding to the literature on social conformity (Bond and Smith, 1996), self-enhancement (Heine and Hamamura, 2007; Sedikides et al., 2007), and self-related neural activity (Wagner et al., 2012; Han et al., 2013).

We compared the behavioral and fMRI data from the German participants obtained for study 1 (final n = 27) with data from three additional groups of participants who performed the same social feedback task. A group of Chinese participants (final n = 28) was scanned in Berlin. Another group of Germans (n = 24) and another group of Chinese (n = 25) were tested behaviorally in Beijing. All participants were tested in their native language. We assessed participants' endorsement of independence and interdependence using the Singelis self-construal scale (Singelis, 1994) and confirmed that in our sample Germans scored higher on independence and lower on interdependence than Chinese.

Our results showed that Chinese conformed more to social feedback than Germans, i.e., Chinese showed higher overall updates. Across all participants, interdependence correlated with social conformity but not with independence or self-esteem. Positively biased feedback processing was evident in both cultural groups and its amount did not differ between them. In addition, participants' initial trait ratings did not differ between the two groups. Whether participants were tested in their native cultural context or not had no effect on social conformity or positively biased updating. This excluded that stress or insecurity related to living abroad influenced social conformity. On the neural level, we tested whether self- and other-related activity differed between German and Chinese participants. We found a significant interaction in a part of the ACC/MPFC. In this region, self-related activity was higher in Germans than in Chinese and correlated with independence in Chinese but not in Germans. In addition, we replicated the neural findings of study 1 for the reward- and comparison-related components in our Chinese sample. The rewarding component of social feedback correlated with activity within the ventral striatum and the ACC/MPFC and the social comparison component correlated with activity within the mentalizing network, i.e., the MPFC, TPJ, STS, TP, IFG, and preSMA. Although not reported in study 2, I would like to mention that the comparison component correlated with activity in a further region of the mentalizing network—the PCC—at a less stringent but corrected threshold ($p < 0.05$ familywise error correction at cluster level; initial threshold $p < 0.0001$ instead of $p < 0.05$ familywise error correction at voxel level; cluster size > 15 voxels). Activity related to the reward- and comparison-related components did not differ between our German and Chinese samples.

Taken together, by testing cultural influences on social conformity, positively biased updating, and self-related activity, my co-authors and I could relate cultural differences in self-concept to the processing of self-relevant information obtained within the context of a social interaction.

## 4.3. Study 3

### Depression is related to an absence of optimistically biased belief updating about future life events

Cognitive theories of major depressive disorder emphasize the role of negative cognition in the onset, development, and treatment of depression (Seligman, 1972; Beck et al., 1979; Clark et al., 2009). Pessimism about the future constitutes an important feature of negative cognition. For example, depressive individuals overestimate the number of negative events and underestimate the number of positive events that they are going to experience (Strunk et al., 2006; Strunk and Adler, 2009). However, it remains elusive how depressed patients update their beliefs when challenged by information about their future. Therefore, in study 3, my co-authors and I investigated how depression relates to the processing of self-relevant information about the statistical likelihood of experiencing adverse future life events.

Depressive patients (final n = 18) and matched healthy controls (n = 19) performed an adapted version of a belief updating task (Sharot et al., 2011). They estimated their likelihood of experiencing 70 adverse life events (e.g. Alzheimer's disease, death before 80). After each estimate, participants saw the average probability of the event occurring to a person living in the same sociocultural environment. This information could be desirable or undesirable with respect to participants' own estimates (i.e., the average probability could be lower or higher than participants' own estimate). To compare how desirable versus undesirable information influenced belief updating, participants estimated their personal probability of experiencing the 70 events a second time.

In line with previous reports (Sharot et al., 2011; Sharot et al., 2012a; Sharot et al., 2012b), healthy participants showed positively biased updating, i.e., they changed their beliefs more toward desirable than toward undesirable information. Importantly, this optimistic bias was absent in depressive patients and the degree to which it was absent correlated with the severity of depressive symptoms. We also replicated previous research by showing that depressive patients were initially more pessimistic than healthy controls (e.g., Strunk et al., 2006; Strunk and Adler, 2009). Because of their pessimistic views, depressive patients received more desirable information than healthy controls in our task. Thus, given that depressive patients had more opportunities to change their beliefs in an optimistic direction, the absence of positively biased updating seems even more striking. In addition, we calculated estimation errors as the numerical differences between participants' initial estimates and the average probabilities presented. The relation of estimation errors and updating differed between the two groups—healthy controls showed a more optimistic pattern. In control analyses, we excluded that differences in updating between the two groups could be influenced by subjective ratings of the events (on vividness, familiarity, prior experience, emotional arousal, negativity, and controllability), by memory for the presented probabilities, or by framing the events as happening or as not happening. Further studies should address the limitations related to our patient sample (8 patients had co-morbid anxiety disorder and 13 patients received medication).

In conclusion, the results of study 3 can be regarded as a proof of principle establishing that research on positively biased processing of self-relevant information in healthy individuals may help to delineate what goes awry in psychiatric disorders.

## 5. General discussion and future directions

In this chapter, I will first outline how the three empirical studies, which I summarized in the previous chapter, have contributed to answer the four main research questions formulated above. I will integrate the empirical results into a potential neurocognitive model for self-relevant information processing and relate this model to previous research. At the same time I will address how this model can guide future research and how some of its shortcomings could be mitigated.

### 5.1. Discussion of research questions

Question 1: Is self-relevant information processing positively biased in healthy people?

For the first research question all three empirical studies of this thesis were of relevance. Processing of self-relevant information was positively biased in healthy participants as indicated by larger updates toward desirable versus undesirable information. The results of the three studies showed positively biased updating for two types of information (feedback on character traits and statistical information about future life events) and for participants of two different cultures (German and Chinese). The findings on future life events replicate previous research (Sharot et al., 2011; Sharot et al., 2012a; Sharot et al., 2012b).

Question 2: Is self-relevant information processing linked to neural activity associated with reward and mentalizing?

The second question was addressed by collecting fMRI data while participants received social feedback on character traits. Activity correlated with two components of social feedback in two partially overlapping neural networks. The reward-related component correlated with activity in the ACC/MPFC and the ventral striatum—regions commonly associated with reward processing across a variety of contexts (Fehr and Camerer, 2007; Rushworth and Behrens, 2008; Beckmann et al., 2009; Rilling and Sanfey, 2011). The comparison-related component correlated with activity in the MPFC, TPJ, STS, TP, IFG, and preSMA. These regions have been consistently related to mentalizing-related tasks by meta-analytic evidence (Spreng et al., 2009; Van Overwalle, 2009; Mar, 2011). Reward- and comparison-related activity did not differ between German and Chinese participants.

Question 3: Does people's cultural background influence the behavioral and neural components of social information processing?

To address the third research question, German and Chinese participants were compared. Cultural differences in self-concept were related to some but not all aspects of social feedback processing under investigation. Chinese updated their trait ratings more than Germans after receiving feedback, i.e., they showed more social conformity. Across all participants, overall updates correlated with the endorsement of an interdependent self-concept. The two groups also differed in ACC/MPFC activity; during feedback presentation, self-related ACC/MPFC activity was higher in Germans than in Chinese. In contrast, the two cultural groups did not differ in the magnitude of the behavioral updating bias. Furthermore, as noted in the previous paragraph, reward- and comparison-related activity did not differ between the two groups.

Question 4: Do depressed patients show an absence of optimistically biased processing of information about their personal future?

To answer the fourth question, depressive patients were compared to healthy controls. Depressive patients did not show positively biased processing of information regarding their future likelihood of experiencing negative life events. Symptom severity correlated with the absence of positively biased updating.

Taken together, the empirical evidence regarding the four main research questions suggests a potential mechanism of self-relevant information processing, which I will describe in the next section.

## 5.2. Neurocognitive model of self-relevant information processing

The proposed model aims to capture the processes which are at play when people receive information that is potentially relevant for their self-concept (**Figure 4**). An underlying assumption of the model is that representations within the self-concept are dynamic, i.e., they are updated during the receipt of self-relevant information such as social feedback on character traits or information about future life events.

Self-relevant information possesses certain properties, which were operationalized in a quantitative fashion in the tasks of this thesis. Information can be dichotomized into desirable and undesirable information (i.e., information that is better or worse relative to the recipient's prior representation). To further characterize self-relevant information, it can be described with respect to two components: First, the reward-related component indexes how positive the information is for the self. (The reward-related component was operationalized as self-related feedback ratings in studies 1 and 2. Please refer to the sections which summarize the research questions, the methodology, and the results for more details.) Second, the social comparison component indexes how close or similar the information is with respect to the recipient's prior representation. (The social comparison component was operationalized as feedback discrepancies in studies 1 and 2 and as estimation errors in study 3.)

These two components are processed in two separable neural networks. Reward processing involves the ACC/MPFC and the ventral striatum—both of which are commonly implicated in processing reward across a variety of contexts (Fehr and Camerer, 2007; Rushworth and Behrens, 2008; Beckmann et al., 2009; Rilling and Sanfey, 2011). Social comparison processing engages activity in the mentalizing network, which includes (dorsal) MPFC, TPJ, STS, TP, IFG, and preSMA (Van Overwalle, 2009; Mar, 2011). The interplay of reward- and comparison-related processes—which seems to be associated with MPFC activity—results in positively biased updating in healthy individuals.

This central part of the proposed model (which has been established on the basis of study 1) was extended by testing self-relevant information processing in two additional groups of recipients. First, depressed individuals, which are characterized by negative cognition and abnormal reward processing (Eshel and Roiser, 2010; Gotlib and Joormann, 2010), show an absence of positively biased updating. Second, cultural differences in self-concept related to interdependence and independence (Markus and Kitayama, 2010) modulate the relationship between the self and other persons. Thus, culture is supposed to influence social comparison processing—resulting in cultural differences in overall updating (i.e., social conformity).

This proposed neurocognitive model integrates the three perspectives on the self outlined in the beginning within a neurobiological framework since dynamics of the self-concept are linked to information obtained from social sources and self-related positivity biases. In the remainder of this chapter, I will discuss the proposed model and its limitations in relation to previous research with a focus on how future studies could provide supporting evidence.

A general theme that will re-occur in the following discussion is that components of the information processing schema (as outlined in **Figure 3**) can be modulated to address further questions. That is, future studies could compare different types of recipients (e.g., participants of different cultures or psychiatric patients as investigated within the empirical studies of this thesis), different sources of information (e.g., information from in-group/out-group members), and different types of self-relevant information (e.g., information on character traits or future life events). Importantly, the general schema should be extended to incorporate actions initiated by the recipient of the information. Currently, the schema depicts that the people receive self-relevant information and change their self-concept. However, in many instances recipients will act upon receiving information. For example, people try to change how they are perceived by others upon receiving social feedback or they try to reduce their risk of encountering certain situations after receiving information about possible future life events. In the same vein, tasks should be developed in which social information is repeatedly exchanged. For example, similar to experiments with iterated economic games (Rilling and Sanfey, 2011; see Glimcher et al., 2008 for an overview), experiments could involve a set-up in which the recipient receives social information from another person and then returns some type of social information to this other person (i.e., both persons alternate in being the recipient and the source of information).

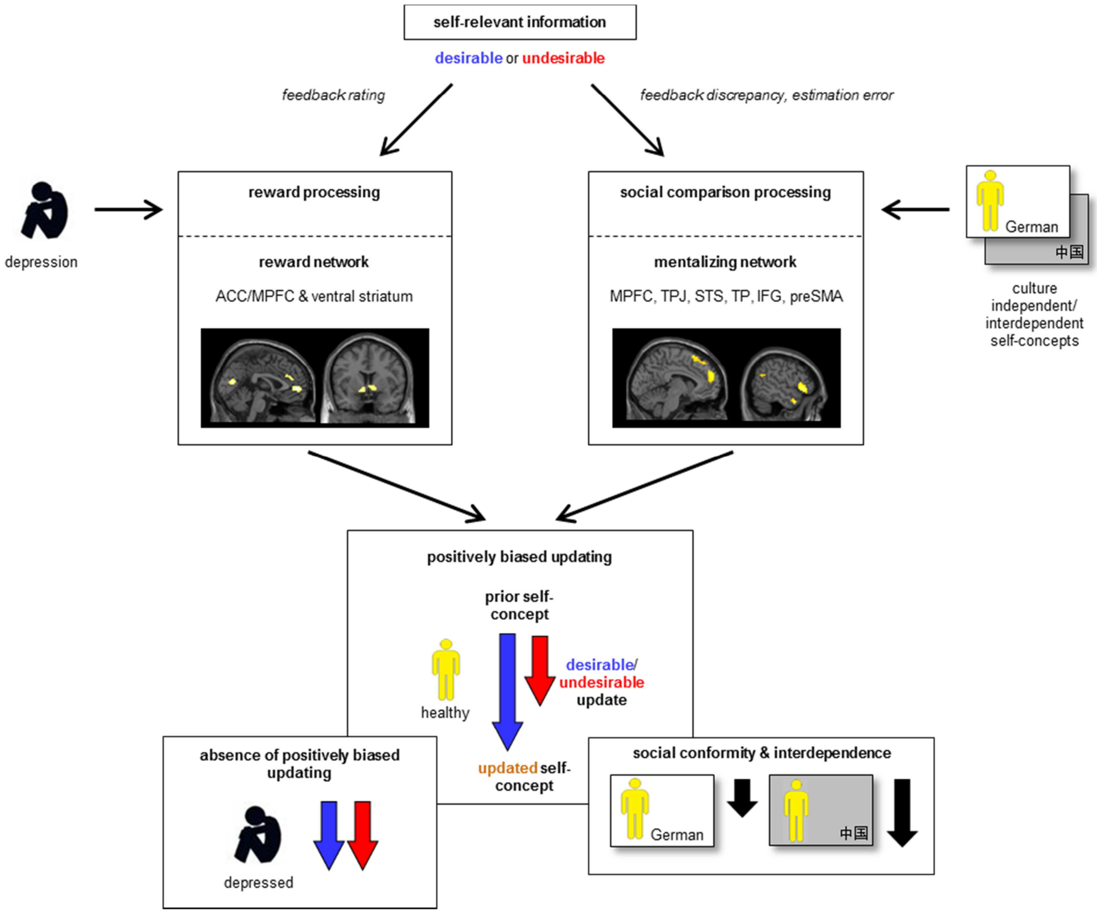

self-relevant information
desirable or undesirable
feedback rating
feedback discrepancy, estimation error
depression
reward processing
reward network
ACC/MPFC & ventral striatum
social comparison processing
mentalizing network
MPFC, TPJ, STS, TP, IFG, preSMA
German
中国
culture
independent/
interdependent
self-concepts
positively biased updating
prior self-concept
healthy
desirable/
undesirable
update
updated self-concept
absence of positively biased updating
depressed
social conformity & interdependence
German
中国

**Figure 4. Potential neurocognitive model of self-relevant information processing.**

**Figure 4. Potential neurocognitive model of self-relevant information processing. (continued)**

Self-relevant information is processed with respect to two separable components: reward (i.e., the positive nature of the information) and social comparison. In the case of social feedback on character traits, the social comparison component is labeled feedback discrepancy (i.e., the difference between information and recipients' prior trait ratings). In the case of statistical information about future life events, the comparison component is labeled estimation error (i.e., the difference between information and recipients' prior probability estimates).

On the neural level, reward processing is mediated by a network encompassing the ACC/MPFC (especially the ventral MPFC) and the ventral striatum. This network has been related to social and non-social reward processing in both humans and animals (Fehr and Camerer, 2007; Rushworth and Behrens, 2008; Beckmann et al., 2009; Rilling and Sanfey, 2011; see Glimcher et al., 2008 for a general overview). The processing of social comparison is reflected by a network previously identified for mentalizing-related processes in humans (Van Overwalle, 2009; Mar, 2011). This network encompasses the MPFC (especially the dorsal MPFC), TPJ, STS, TP, IFG, and the preSMA. (Contrasts are taken from study 2 and depicted at a threshold of $p < 0.05$ familywise error correction at voxel level; cluster size > 15 voxels.)

Reward and social comparison processing influence the updating of representations within the self-concept. Healthy individuals show positively biased updating (i.e., larger updates following desirable compared with undesirable information) for information about character traits and future life events.

Different types of recipients have different influences on the components of the proposed model. Reward processing is supposed to be disrupted by depression; resulting in an absence of positively biased updating. Social comparison processing is supposed to be modulated by culture; resulting in more social conformity (i.e., a larger overall amount of updating) in interdependent compared with independent cultures.

### 5.3. Self-related biases

The proposed neurocognitive model aims at providing a better understanding of self-related positivity biases by integrating them into an information processing framework that combines behavioral and neural aspects. Early conceptualizations of positivity biases have posited that humans engage in self-related positive illusions because cognitive processing mechanisms impose filters on incoming information, distorting it into a positive direction (Taylor and Brown, 1988). The proposed model suggests that this "filtering mechanism" involves reward- and comparison-related processes. The model predicts that specific aspects of the information itself, of the recipient, and of the source (see **Figure 3**) should impact on the posited "filtering mechanism." This also stresses a point made—in a similar way—by earlier conceptualizations of positivity biases: Current states of the self-concept (e.g., overly positive self-evaluations) have to be distinguished from dynamic changes of the self-concept and from influences on such changes (e.g., positively biased feedback processing) (Sedikides and Gregg, 2008, Leary, 2007).

An important aspect that is underspecified in the proposed model concerns the question: What makes information self-relevant? That is, to which instances of information processing can the model be applied? Therefore, in the following, I will discuss how the question of self-relevance bears on different types of information, different types of sources, and different types of recipients (see **Figure 3**).

Of the triad of positivity biases related to self-evaluations, perceptions of control, and estimates of the future (Taylor and Brown, 1988), the proposed model rests on evidence for two of these three processes: processing information about character traits and future life events. Future work should address whether the model can be extended to information about how much a person can be in control of a certain situation. Furthermore, the model should ideally be extended to a couple of phenomena that have been heavily discussed in social psychology such as the self-serving attributions, the false consensus effect, or social conformity. All of these phenomena involve information processing. First, self-serving attributions describe people's tendency to attribute success (i.e., positive information about their performance) internally (i.e., to their abilities) and failure (i.e., negative information about their performance) externally (i.e., to the situation) (see Hewstone et al., 2008; Hogg and Vaughan, 2008 for an overview). A more detailed investigation of how participants process performance feedback (e.g., in a logical puzzle) will help to link self-relevant information processing to previous studies which have operationalized feedback as performance feedback (Alicke and Sedikides, 2009). Second, the false consensus effect describes amongst other things people's tendency to think that—given the same information—others would make similar decisions (Marks and Miller, 1987). Third, some recent studies have used similar designs as used in the empirical studies of this thesis to investigate social conformity (see Cialdini and Goldstein, 2004 for a discussion of the behavioral literature). Participants usually received feedback from others on the properties of something that is not self-relevant (e.g. the attractiveness of the face of an unknown person, or the liking of a previously unknown song) (Klucharev et al., 2009; Campbell-Meiklejohn et al., 2010; Zaki et al., 2011). Opinion changes in these studies were typically unbiased. For example, people were influenced to the same degree when they saw that others judged a face to be more—or less—attractive than they did. However, it remains an open question whether positively biased information processing extends to objects that participants perceive as strongly self-related such as their favorite songs, their clothing, or their cars (Leary, 2007).

In studies 1 and 2, participants received social feedback for another peer, i.e., they received feedback that was not directly self-relevant. Participants rated themselves more favorably than the other person as has been reported in many studies on trait evaluations (Leary, 2007; Alicke and Sedikides, 2009).But there was no difference in positively biased updating for the self and for the other person. It should be noted that participants had direct contact with this person and that the person typically was a fellow student. Therefore, this finding is in general accord with previous research showing that positivity biases are influenced by whether participants have met the other person (Alicke et al., 1995). Manipulating the relationship between the self and the other person (for example by assigning him or her to an out-group) may alter or even abolish positively biased updating and may thus clarify the role of self-relevance for the proposed model.

## 5.4. Self-related activity during feedback processing

A cluster encompassing a huge part of the MPFC was found in a categorical contrast comparing self- versus other-directed feedback. This extends the well-known role of the MPFC in self- and other-judgments (Amodio and Frith, 2006; Denny et al., 2012) to receiving feedback about character traits. Since this activity encompassed a considerable part of the frontal midline at a corrected threshold, it seems difficult to disentangle the role of specific MPFC subdivisions or specific components in feedback processing on the basis of this categorical contrast. Therefore, I did not directly incorporate this result into the proposed neurocognitive model. Research on self-related neural activity will be discussed in the next section in the context of mentalizing.

## 5.5. Mentalizing and social comparison

The neurocognitive model proposed in this thesis associates social comparison processing with the mentalizing network. The label “mentalizing network” refers to the fact that this set of regions has consistently been implicated in mentalizing tasks as shown by meta-analytic evidence (Van Overwalle, 2009; Mar, 2011). That is, social comparison processing and mentalizing involve activity in overlapping brain regions. I do not want to draw the reverse inference based on these fMRI results that social comparison processing equals mentalizing (Poldrack, 2006)—especially not mentalizing as defined by classic false belief tasks (Wimmer and Perner; 1983).

Instead, I would like point out that the changes in neural activity, which were observed in the mentalizing network, have a parsimonious interpretation in the context of the task used in studies 1 and 2. Activity in the mentalizing network tracked the numerical difference between participants' own evaluations and the feedback they received both for their own character and for the character of another person. This finding is in line with the general definition of mentalizing as the process of inferring another agent's mental state (Frith and Frith, 2012) and points to a number of questions for further research. Specifically, I will focus on the relation of mentalizing with self- versus other-related processes, false belief tasks, and computational modeling approaches.

First, the relation between activity related to social comparison processing and activity related to mentalizing suggests that the differences between people's own mental state and other persons' mental states are of relevance—at least when character traits constitute the content of mental states. Much research on mentalizing has investigated instances in which the content of another person's mental state was not about character traits but about the properties of physical objects (e.g., the location of a ball) (for review see Amodio and Frith, 2006). The proposed neurocognitive model suggests a direct link between mentalizing and processes related to judging character traits. Interestingly, self- versus other-directed feedback elicited different overall MPFC activity but not different social comparison-related activity. Given that activity associated with other-judgments varies with the relation of the self to the other person (Jenkins, et al. 2008; Krienen et al., 2010; Denny et al., 2012), future research may address whether comparison-related activity shows differences when participants receive feedback about other persons who vary along dimensions such as similarity, closeness, or in-group/out-group membership. In addition, one could further investigate how differences in the opinion about something that is not directly self-relevant (e.g., faces or pieces music) are related to social comparison processing within the mentalizing network (e.g., Klucharev et al., 2009; Campbell-Meiklejohn et al., 2010; Zaki et al., 2011).

Second, one aspect in which the task of study 1 and 2 differs from classic belief tasks is that participants were not explicitly instructed to infer the mental state of other persons (i.e., they did not have to infer that other persons lack specific knowledge). Instead, they were informed about the content of other persons' mental state in the form of explicitly stated character trait ratings. One could speculate that participants pondered on why the others had given a specific rating and that this process was at play to a higher degree when feedback ratings were more different from participants' own evaluations. This idea could be tested by asking participants to report their thoughts during feedback processing. A more promising approach to relate social comparison processing to mentalizing—as defined by false belief tasks—would be to parametrically modulate the extent to which the other person's belief is false (see Tamir and Mitchell, 2010 for a similar approach). For example, in an adapted version of the Sally and Ann task, one could parametrically vary the spatial distance to which the new location of a physical object differs from its prior location.

Third, many studies have resorted to categorical contrasts to investigate mentalizing (e.g., Walter et al., 2004; Saxe and Powell, 2006). Recently, however, a few studies have used parametric approaches to elucidate specific components of mentalizing (Hampton et al., 2008; Tamir and Mitchell, 2010). In particular, computational models—such as variants of reinforcement learning used within the context of economic games—have been promoted as promising avenues for a better understanding of mentalizing and social cognition in general (Yoshida et al., 2008; Behrens et al., 2009; Hunt and Behrens, 2011; King-Casas and Chiu, 2012). The approach taken in the empirical studies of this thesis occupies a middle ground between categorical contrasts and computational models. That is, in the fMRI studies, my co-authors and I investigated two components of feedback processing, which varied parametrically on a trial-by-trial fashion. This approach goes beyond contrasting a limited number of conditions but does not allow testing different computationally specified models and the relation of model parameters to neural activity. For example, only a linear relationship between social comparison processing (i.e., feedback discrepancies) and neural activity was tested. Further studies should test for nonlinear relationships. To make the tasks employed in the empirical studies of this thesis accessible for the use of computational models such as reinforcement learning models, social feedback should be repeatedly given for the same trait. That is, currently participants receive feedback a single time for 80 different traits or 70 different life events. Reinforcement modeling requires that trait ratings or likelihood estimates are updated multiple times (Sutton and Barto, 1998). This could be implemented in a setting in which participants exchange social information multiple times. Thereby, the actions of one set of participants would influence the actions of another set of participants and vice versa. (In the current fMRI studies, social feedback has no impact on those who gave it.)

Taken together, the proposed lines of research may advance our understanding of whether different regions within the mentalizing network differ with respect to the type of information that is processed (e.g., social or non-social) or with respect to the type of process at play (e.g., belief inference or social comparison).

## 5.6. Reward processing

The definition of social reward used in studies 1 and 2 and by consequence in the proposed neurocognitive model is identical to the definition used in a previous study on social feedback (Izuma et al., 2008): First, more positive feedback for the self indicates higher social reward. Second, social reward is self-specific. Participants in the study by Izuma et al. (2008) received social feedback for themselves or for another person in the form of positive trait adjectives of high or low social desirability. The authors of that study performed an interaction contrast between feedback directed to the self versus the other person and high versus low desirability trait adjectives. My co-authors and I used a parametric approach that is conceptually similar. That is, we searched for activity that correlated more positively with feedback ratings for the self than with feedback ratings for another person. Additionally, we extended the approach used by Izuma et al. (2008) to negative trait adjectives.

Using this parametric approach, we found that reward processing correlated with activity in the ventral striatum and the ACC/MPFC, which replicates the results by Izuma et al. (2008). These findings are consistent with a huge human and animal literature on reward processing in social and nonsocial settings (for reviews see Montague et al., 2006; Fehr and Camerer, 2007; Rushworth et al., 2007; Rilling and Sanfey, 2011). In the social domain, the striatum has for example been implicated in the processing of advice (Behrens et al., 2008; Biele et al., 2011; Meshi et al., 2012), i.e., social information on which of several options to choose. The ventral part of the MPFC has a well-established role in preference statements and in tracking the value of options in choice situations (e.g., in the context of food choices). Social influences can modulate activity in the ventral MPFC (and the ventral striatum) (Plassmann et al., 2008; Zaki et al., 2011).

Several authors have argued that the brain converts different types of reward information into a common currency, which may be reflected by activity in the striatum (Sanfey, 2007) (which in turn might reflect dopaminergic inputs) (Schultz, 2006). The results obtained in studies 1 and 2 support this general notion by corroborating the role of the striatum and the MPFC for reward processing in the social domain. Further evidence could be obtained by pitting social and non-social rewards against each other. For example, future studies could investigate how much money participants are willing to pay for obtaining potentially positive social feedback or how participants decide between the prospect of winning money and the prospect of getting positive social feedback. More importantly, further studies should extend the framework established by decision neuroscience and address how social information processing is influenced by variables such as variance or uncertainty (e.g., Platt and Huettel, 2008; Bach and Dolan, 2012). For example, the uncertainty of social feedback could be modulated by varying the number of peers giving feedback. As outlined above, crucial insights may be gained by adapting computational approaches used for understanding non-social reward processing to investigate social reward processing (Behrens et al., 2008; Hampton et al., 2008; Biele et al., 2011; Jones et al., 2011). Especially algorithms derived from reinforcement learning might provide a powerful means to investigate social reward processing—given their prominent role in understanding non-social reward (for reviews see Montague et al., 2006; Lee, 2008; see O'Doherty et al., 2007 for a description of the methodological approach).

Another open question concerns the relationship between social reward and social exclusion. Clarifying this relationship might be informative for debates in social psychology and decision neuroscience. Theories in social psychology posit that people seek out social reward because it reassures them that they are well integrated into their social environment (Baumeister and Leary, 1995; Eisenberger et al., 2011). In decision neuroscience, an important debate centers on whether processes associated with positive valence (e.g., gains, reward) and processes associated with negative valence (e.g., losses, punishment) constitute opposite endpoints of a continuum or two different dimensions (Montague et al., 2006; see also Glimcher et al., 2008). The proposed neurocognitive model and the results on which it is based do not directly speak to these debates since I have focused on investigating reward-related processes. Studies on social exclusion commonly report activity within the dorsal ACC (at a more posterior location than the ACC activity identified for social feedback processing), anterior insula, and somatosensory cortices among other regions (Eisenberger, 2012). This pattern of activity did not emerge in studies 1 and 2 and we did not find activity that correlated negatively with the reward-related component of social feedback.

Taken together, this seems to suggest that the pattern of activity found for social reward differs from the pattern commonly identified in studies on social exclusion. Social reward and social exclusion might thus constitute two distinct dimensions rather than two poles of one continuum. Further studies should directly address how social reward and social exclusion are linked on a behavioral and neural level.

### 5.7. Functional subdivisions of the MPFC

The MPFC plays a prominent role in different components of the proposed neurocognitive model. To guide further research on the relationship between different roles of MPFC activity, I would like to give a brief outline of the anatomical location of functional MPFC subdivisions identified in studies 1 and 2. The spatial pattern of activity depicted in **Figure 5** is consistent with various strands of previous research.

First, social comparison processing elicited activity in a dorsal part of the MPFC (and the preSMA/SMA)—which is in line with the well-known involvement of this part of the MPFC in mentalizing (Van Overwalle, 2009; Mar, 2011). With regard to the ventral-to-dorsal gradient discussed for self- versus other-related MPFC activity (Denny et al., 2012), the MPFC cluster identified for social comparison processing lies at a more or less middle position.

Second, reward processing was associated with activity in a more ventral part of the ACC/MPFC and an additional cluster in more dorsal parts of the ACC/MPFC. This adds to the substantial literature on the involvement of the most ventral part of the MPFC (often called medial OFC) in reward processing (e.g., Montague et al., 2007; Beckmann et al., 2009). Furthermore, ventral MPFC activity associated with self-related reward on character traits is consistent with previous reports of ventral MPFC (or ventral ACC) activity in the context of valence-related aspects of self-judgments (Beer and Hughes, 2010).

Third, participants’ cultural background influenced self-related activity in a middle part of the ACC/MPFC. Previous cultural neuroscience studies have often found activity within similar parts of the ACC/MPFC with some studies reporting slightly more anterior and/or dorsal clusters (Zhu et al., 2007; Chiao et al., 2009a; Chiao et al., 2009b; Ng et al., 2010; Ray et al., 2010; Ma et al., 2012; Wang et al., 2012). However, these studies have not yet been summarized with the help of meta-analytic procedures.

In study 1, the behavioral updating bias for self correlated with activity in a middle part of the MPFC, in which activity related to reward and social comparison processing overlapped as shown by a conjunction analyses (not depicted in **Figure 5**). However, this finding, which suggests that the MPFC might act as an integration region for social feedback, should be regarded as preliminary since the conjunction but not the correlation could be replicated in the Chinese sample of study 2.

Taken together, understanding MPFC computations might be crucial for linking neurobiological findings on reward and mentalizing with the behavioral literature on positivity biases. Future studies should address how the functional subdivisions of MPFC activity that were identified within the context of self-relevant information processing map onto further MPFC functions discussed in the literatures on cognitive control (Ridderinkhof et al., 2004), on depression (Price and Drevets, 2012), or decision making in general (Rushworth and Behrens, 2008).

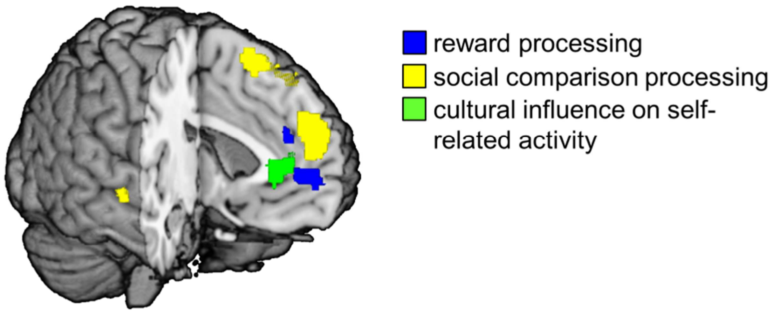

🟦 reward processing
🟨 social comparison processing
🟩 cultural influence on self-related activity

**Figure 5. Functional MPFC subdivisions.**

In line with previous research on reward (Beckmann et al., 2009; Rilling and Sanfey, 2011), mentalizing (Van Overwalle, 2009; Mar, 2011), and on cultural influences of self-related activity (Han et al., 2013), our results support a functional subdivisions of MPFC activity. Dorsal aspects of the MPFC (and the preSMA) are associated with social comparison processing. Dorsal to ventral aspects of the ACC/MPFC are associated with cultural differences in self-related activity. Ventral aspects of the MPFC (and parts of the dorsal ACC) are associated with reward processing. (Contrasts are taken from study 2. Contrasts related to reward and social comparison processing are depicted at a threshold of $p < 0.05$ familywise error correction at voxel level; cluster size > 15 voxels. The interaction contrast related to the cultural difference is small volume corrected within the main effect of self- versus other-directed feedback; initial threshold for interaction: $p < 0.001$, uncorrected.)

## 5.8. Culture

Since culture exerts profound influences on social cognition, study 2 compared participants of German and Chinese origin to avoid that the predictions of the proposed neurocognitive model are limited to Western samples. Comparing Germans to Chinese is relevant for social feedback processing because the members of these two cultures are known to differ in their self-concepts with regard to independence and interdependence (Markus and Kitayama, 1991; Oyserman et al., 2002; Triandis and Suh, 2002; Markus and Kitayama, 2010; Heine, 2012). In the model, cultural differences in self-concept are supposed to impact on social comparison processing but not on reward processing. This is motivated by the behavioral results of study 2, which I will discuss in the following. Chinese participants showed more social conformity than Germans but no difference in the amount of positively biased updating.

Study 2 investigated Germans and Chinese participants to test how differences in independent and interdependent self-concepts influence social comparison processing. Meta-analytic evidence suggests that members of interdependent cultures show more conformity when receiving social information about the lengths of two lines (Bond and Smith, 1996). The results of study 2 extend these previous findings to social conformity about character traits, which—unlike objective physical properties—are directly self-relevant. The cultural difference in social conformity was directly related to individual differences in interdependent self-concept. Across all participants, overall updating correlated significantly with interdependence but not with independence. Previous studies have combined interdependence and independence scores into a composite measure (e.g., Chiao et al., 2009b). In study 2, they were kept separate because the two scores were not correlated and because the initial study that introduced them treated them as two different dimensions (Singelis, 1994). Future research should address whether independence and interdependence are the opposite ends of a continuum or separate dimensions.

There has been a fierce debate about whether East Asians show self-enhancing tendencies to a similar degree as Westerners (Sedikides et al., 2003; Heine et al., 2007; Heine and Hamamura, 2007; Sedikides et al., 2007). The possible absence or reduction of self-enhancement might be a specific feature of East Asian culture that might not directly relate to interdependent versus independent self-concepts. Although rarely discussed (Ryder et al., 2008; Heine, 2012), this debate is interesting in light of the literature on depression since an absence of positivity bias could be described as a "depressive pattern" (see next section). In study 2, the behavior of Chinese participants did not point into such a direction. Both cultural groups showed indistinguishable degrees of positively biased updating and in addition they did not differ in their self-evaluations.

Taken together, the behavioral approach taken in study 2 offers a novel approach to conceptualize self-enhancement, which contributes to the debate on whether East Asians do or do not show self-enhancement (Heine and Hamamura, 2007). Further studies should replicate the results of study 2 with American and Japanese since the majority of previous studies have compared these two cultures and since the original proposition of independent and interdependent self-concepts was based on these two cultures (Markus and Kitayama, 1991; Heine, 2012). Regarding the proposed neurocognitive model, the results of study 2 point toward an influence of culture on social comparison processing but not reward processing.

On the neural level, culture influenced activity in a region of the ACC/MPFC related to self- versus other-directed feedback. This finding extends the role of the ACC/MPFC in cultural differences of self- and other-evaluations (Zhu et al., 2007; Chiao et al., 2009a; Chiao et al., 2009b; Ng et al., 2010; Ray et al., 2010; Ma et al., 2012; Wang et al., 2012). Cultural differences are conceptualized as differences in social interactions between the self and others (Markus and Kitayama, 1991; Markus and Kitayama, 2010). Since participants in study 2 received social feedback within the context of a face-to-face interaction, the reported ACC/MPFC result directly confirms the relevance of this region for cultural differences in social interactions. Culture did not, however, influence activity related to the social comparison component in the sample of study 2. Thus, on the neural level, future studies should provide direct evidence for the proposed link between culture and social comparison processing.

## 5.9. Depression

The potential relevance of self-relevant information processing for psychiatric diseases (especially for affective disorders such as major depressive disorder) has been an important motivation for developing the proposed neurocognitive model. In the model described above, depression is supposed to impact on reward processing but this does not exclude potential influences on social comparison processing. Highlighting the role of reward processing in depression rests upon previous reports (Eshel and Roiser, 2010) and upon the findings of study 3. Depressive patients showed an absence of optimistically biased belief updating, which correlated with symptom severity.

Obtaining fMRI data from depressive patients should help to clarify whether altered reward processes or altered social comparison processes—or both—lead to the described absence of biased updating. A previous fMRI study using the same paradigm in healthy participants identified activity in the IFG and dorsal MPFC (which was related to estimation errors) (Sharot et al., 2011) and a subsequent study showed that transcranial magnetic stimulation (TMS) on the left IFG altered updating beliefs about the future (Sharot et al., 2012b). Within the proposed neurocognitive model, MPFC and IFG are associated with social comparison processing rather than reward processing. Based on the neurocognitive model, depressive individuals might thus be expected to show altered activity within reward-related regions such as the ventral striatum and the ventral MPFC. The neural aspects of the proposed neurocognitive model have been mainly gleaned from the two fMRI studies included in this thesis. Thus, further research should use fMRI to compare how healthy and depressed individuals process information about the future and about character traits. This could specifically inform research on the well-established involvement of altered MPFC activity in depression, which may pertain to reward or comparison processing or both (Disner et al. 2011; Pizzagalli, 2011; Price & Drevets, 2012).

Cognitive models of depression focus on cognitive biases toward the negative such as pessimism or overly negative self-views (for recent reviews see Mathews and MacLeod, 2005; Clark et al., 2009; Gotlib and Joorman, 2010). In contrast, research on healthy participants has usually focused on cognitive biases to the positive such as optimism or overly positive self-views (Taylor and Brown, 1988; Leary, 2007). Many authors have discussed whether being unbiased or whether being biased to the positive confers more benefits for mental health, well-being, and economic success (Taylor and Brown, 1988; Scheier and Carver, 1992; Weinstein and Klein, 1995; Lovallo and Kahneman, 2003; Haselton and Nettle, 2006; Leary, 2007;Puri and Robinson, 2007). Probably, this question cannot be definitely answered because whether a certain behavior is adaptive or not depends very much on the given situation. In the context of specific studies, being unbiased often serves as a benchmark. In empirical studies of this thesis, for example, being unbiased means that participants update their beliefs to a similar degree for desirable and undesirable information. In this sense, depressive patients were unbiased although the correlation with symptom severity implies that more severely depressed patients are actually biased toward the negative. Additionally, depressive pessimism is underscored by the finding that depressed patients estimated their likelihood of experiencing negative events to be higher than did healthy individuals (e.g., Strunk et al., 2006; Strunk and Adler, 2009).

Taken together, the proposed neurocognitive model above lends itself well to testing self-relevant information processing in other psychiatric disorders such as borderline personality disorder for which unstable social relationships are characteristic (American Psychiatric Association, 2000). Patients suffering from Borderline Personality disorder are known to show alterations neural activity related to mentalizing (Dziobek et al., 2011) and to repairing breaches in trust (King-Casas et al., 2008) but the neural mechanisms underlying self-relevant information processing remain elusive.

Currently, psychiatric diseases are categorized on the basis of behavioral symptoms and patients are diagnosed on the basis of clinical interviews (American Psychiatric Association, 2000). It is unclear whether these classifications truly reflect underlying psychopathologies or whether they merely group symptoms that are caused by different mechanisms. That is, are many psychiatric diseases a set of symptoms similar to jaundice, for example, that can develop due to such diverse causes as malaria, hepatitis or gallstones? In the discussion sections on mentalizing and reward, I have argued for extending the approach used in this thesis by putting the proposed model into a computational framework. This is in line with several authors who have recently proposed computational approaches as a means to advance the classification and diagnosis of psychiatric diseases (Maia and Frank 2011; Montague et al., 2012; King-Casas and Chiu, 2012). Advancing computational models of social interactions might be particularly helpful since many psychiatric diseases are accompanied by social deficits (Montague et al., 2012; King-Casas and Chiu, 2012).

## 6. Conclusion

This dissertation advances the understanding of self-relevant information processing by providing a framework that integrates behavioral research on the self and neuroscientific research on reward processing and mentalizing. Dynamic changes of the self-concept are related to the processing of information obtained within the context of social interactions. These changes are shown to be biased toward the positive in healthy people—when they receive social feedback on character traits or statistical information about the likelihood of experiencing adverse life events. The proposed neurocognitive model assumes that people process self-relevant information with respect to two components: reward and social comparison. On the neural level, these components are supposed to correspond to reward-related activity in the ventral striatum and ACC/MPFC and to activity in the mentalizing network including the MPFC, TPJ, STS, TP, IFG, and preSMA. The model incorporates the reported absence of positively biased information processing in depressive patients. In addition, cultural influences on social feedback processing are integrated. I hope that the current thesis will provide a fruitful source for conceiving future studies and that the model proposed in this study will be extended by computational modeling approaches.

## 7. References

Adolphs R (2009) The social brain: neural basis of social knowledge. Annu Rev Psychol 60:693-716.

Adolphs R (2010) Conceptual challenges and directions for social neuroscience. Neuron 65:752-767.

Alicke MD, Klotz ML, Breitenbecher DL, Yurak TJ, Vredenburg DS (1995) Personal contact, individuation, and the better-than-average effect. J Pers Soc Psychol 68:804-25.

Alicke MD, Sedikides C (2009) Self-enhancement and self-protection: What they are and what they do. Eur Rev Soc Psychology 20:1-48.

American Psychiatric Association (2000) Diagnostic and statistical manual of mental health disorders (4th edition, text revision). Washington DC, USA: American Psychiatric Association.

Amodio DM, Frith CD (2006) Meeting of minds: the medial frontal cortex and social cognition. Nat Rev Neurosci 7:268-277.

Armor DA, Taylor SE (2002) When predictions fail: the dilemma of unrealistic optimism. In Gilovich T, Griffin DW, Kahneman D (Eds.) Heuristics and biases: the psychology of intuitive judgment. New York, USA: Cambridge University Press.

Anderson NH (1968) Likableness ratings of 555 personality-trait words. J Pers Soc Psychol 9:272-279.

Ariely D (2008) Predictably irrational. London, UK: Harper Collins Publishers.

Bach DR, Dolan RJ (2012) Knowing how much you don't know: a neural organization of uncertainty estimates. Nat Rev Neurosci 13:572-586.

Bahnemann M, Dziobek I, Prehn K, Wolf I, Heekeren HR (2010) Sociotopy in the temporoparietal cortex: common versus distinct processes. Soc Cog Affect Neurosci 5:48-58.

Banaji MR, Prentice DA (1994) The self in social contexts. Annu Rev Psychol 45:297-332.

Baumeister RF (1998) The self. In Gilbert DT, Fiske ST, Lindzey G (Eds.) Handbook of social psychology (4th edition). New York, USA: McGraw-Hill.

Baumeister RF (2011) Self and identity: a brief overview of what they are, what they do, and how they work. Ann NY Acad Sci 1234:48-55.

Baumeister RF, Leary MR (1995) The need to belong: desire for interpersonal attachments as a fundamental human motivation. Psychol Bull 117: 497-529.

Beck AT (2005) The current state of cognitive therapy: a 40-year retrospective. Archiv Gen Psychiatry 62:953-959.

Beck AT, Rush AJ, Shaw B, Emery G (1979) Cognitive therapy of depression. New York, USA: Guilford Publications.

Beckmann M, Johansen-Berg H, Rushworth MFS (2009) Connectivity-based parcellation of human cingulate cortex and its relation to functional specialization. J Neurosci 29:1175-1190.

Beer JS (2007) The default self: feeling good or being right? Trends Cogn Sci 11:187-189.

Beer JS, Hughes BL. (2010) Neural systems of social comparison and the “above-average” effect. Neuroimage 49:2671-2679.

Behrens TEJ, Hunt LT, Woolrich MW, Rushworth MFS (2008) Associative learning of social value. Nature 456:245-249.

Behrens TEJ, Hunt LT, Rushworth MFS (2009) The computation of social behavior. Science 324:1160-1164.

Biele G, Rieskamp J, Krugel LK, Heekeren HR (2011) The neural basis of following advice. PLoS Biol 9:e1001089.

Bohrn I, Altmann U, Jacobs AM (2012) Looking at the brains behind figurative language—A quantitative meta-analysis of neuroimaging studies on metaphor, idiom, and irony processing. Neuropsychologia 50:2669-2683.

Bond R, Smith PB (1996) Culture and conformity: a meta-analysis of studies using Asch's (1952b, 1956) line judgment task. Psychol Bull 119:111-137.

Brewer MB, Hewstone M (2004) Self and social identity. Oxford, UK: Blackwell Publishing Ltd.

Buckner RL, Carroll DC (2007) Self-projection and the brain. Trends Cogn Sci 11:49-57.

Call J, Tomasello M (2008) Does the chimpanzee have a theory of mind? 30 years later. Trends Cogn Sci 12:187:192.

Campbell-Meiklejohn DK, Bach DR, Roepstorff A, Dolan RJ, Frith CD (2010) How the opinion of others affects our valuation of objects. Curr Biol 20:1165-1170.

Carver CS (1989) Dispositional optimism and recovery from coronary artery bypass surgery: the beneficial effects on physical and psychological wellbeing. J Pers Soc Psychol 57:1024-1040.

Castelli F, Happe F, Frith U, Frith CD (2000) Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. Neuroimage 12:314-325.

Chambers JR, Windschitl PD (2004) Biases in social comparative judgments: the role of nonmotivated factors in above-average and comparative-optimism effects. Psychol Bull 130:813-838.

Chase HW, Frank MJ, Michael A, Bullmore ET, Sahakian BJ, Robbins TW (2010) Approach and avoidance learning in patients with major depression and healthy controls: relation to anhedonia. Psychol Med 40:433-440.

Cialdini RB, Goldstein NJ (2004) Social influence: compliance and conformity. Annu Rev Psychol 55:591-621.

Chiao JY, Ambady N (2007) Cultural neuroscience: parsing universality and diversity across levels of analysis.In Kitayama S, Cohen D (Eds.) Handbook of cultural psychology. New York, USA: Guilford.

Chiao JY, Harada T, Komeda H, Li Z, Mano Y, Saito D, Parrish TB, Sadato N, Iidaka T (2009a). Dynamic cultural influences on neural representations of the self. J Cogn Neurosci 22:1-11.

Chiao JY, Harada T, Komeda H, Li Z, Mano Y, Saito D, Parrish TB, Sadato N, Iidaka T (2009b) Neural basis of individualistic and collectivistic views of self. Hum Brain Mapp 30:2813-2820.

Chiao JY, Hariri AR, Harada T, Mano Y, Sadato N, Parrish TB, Iidaka T. (2010) Theory and methods in cultural neuroscience. Soc Cogn Affect Neurosci 5:356-361.

Clark L, Chamberlain SR, Sahakian BJ (2009) Neurocognitive mechanisms in depression: implications for treatment. Annu Rev Neurosci 32:57-74.

Conklin LR, Strunk DR, Fazio RH (2009) Attitude formation in depression: evidence for deficits informing positive attitudes. J Behav Ther Exp Psychiat 40:120-126.

Cooper JC, Dunne S, Furey T, O'Doherty J (2012) Dorsomedial prefrontal cortex mediates rapid evaluations predicting the outcome of romantic interactions. J Neurosci 32:15647-15656.

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

Cross P (1977) Not can but will college teaching be improved. New Directions for Higher Education 17:1-15.

Davey CG, Allen NB, Harrison BJ, Dwyer DB, Yücel M (2010) Being liked activates primary reward and midline self-related brain regions. HumBrain Mapp 31:660-668.

Dayan P (2012) Models of value and choice. In Dolan RJ, Sharot T (Eds.) Neuroscience of preference and choice. London, UK: Academic Press Elsevier.

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

Denny BT, Kober H, Wager TD, Ochsner KN (2012) A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. J Cogn Neurosci 24:1742-1752.

Disner SG, Beevers CG, Haigh EAP, Beck AT (2011) Neural mechanisms of the cognitive model of depression. Nat Rev Neurosci 12:467-477.

Dziobek I, Preißler S, Grozdanovic Z, Heuser I, Heekeren HR, Roepke S (2011) Neuronal correlates of altered empathy and social cognition in borderline personality disorder. Neuroimage 57:539-548.

Eisenberger NI (2012) The pain of social disconnection: examining the shared neural underpinnings of physical and social pain. Nat Rev Neurosci 3:421-434.

Eisenberger NI, Inagaki TK, Muscatell KA, Byrne Haltom KE, Leary MR (2011) The neural sociometer: brain mechanisms underlying state self-esteem. J Cogn Neurosci 23:3448-3455.

Eli D, Rao JM (2011) The good news-bad news effect: asymmetric processing of objective information about yourself. Am Econ J Microeconomics 3:114-138.

Ellemers N (2012). The group self. Science 336:848-852.

Emery NJ, Clayton NS (2009) Comparative social cognition. Annu Rev Psychol 60:87-113.

Erk S, Spitzer M, Wunderlich AP, Galley L, Walter H (2002) Cultural objects modulate reward circuitry. Neuroreport 13:2499-2503.

Eshel N, Roiser JP (2010) Reward and punishment processing in depression. Biol Psychiatry 68:118-124.

Falk EB, Berkman ET, Whalen D, Lieberman MD (2011) Neural Activity during health messaging predicts reductions in smoking above and beyond self-report. Health Psychol 3:177-185.

Fehr E, Camerer CF (2007) Social neuroeconomics: the neural circuitry of social preferences. Trends Cogn Sci 11:419-427.

Fliessbach K, Weber B, Trautner P, Dohmen T, Sunde U, Elger CE, Falk A (2007) Social comparison affects reward-related brain activity in the human ventral striatum. Science 318:1305-1308.

Fossati P, Hevenor SJ, Graham SJ, Grady C, Keightley ML, Craik F, Mayberg H (2003) In search of the emotional self: an fMRI study using positive and negative emotional words. Am J Psychiatry 160:1938-1945.

Freeman JB, Rule NO, Adams RB, Ambady N (2009) Culture shapes a mesolimbic response to signals of dominance and subordination that associates with behavior. Neuroimage 47:353-359.

Friston K (2010) The free-energy principle: a unified brain theory? Nat Rev Neurosci 11:127-138.

Frith CD (2007a) Making up the mind. Oxford, UK: Blackwell Publishing Ltd.

Frith CD (2007b) The social brain? Philos Trans R Soc Lond B Biol Sci 362: 671-678.

Frith U, Frith CD (2003) Development and neurophysiology of mentalizing. Philos Trans R Soc Lond B Biol Sci 358:459-473.

Frith CD, Frith U (2012) Mechanisms of social cognition. Annu Rev Psychol 63:287-313.

Frith CD, Singer T (2008) The role of social cognition in decision making. Philos Trans R Soc Lond B Biol Sci 363: 3875-3886.

Gigerenzer G (2007) Gut feelings. London, UK: Penguin Books Ltd.

Glimcher PW, Fehr E, Camerer C, Poldrack RA (Eds.) (2008) Neuroeconomics: decision making and the brain. London, UK: Academic Press Elsevier.

Gotlib IH, Joormann J (2010) Cognition and depression: current status and future directions. Annu Rev Clin Psychol 6:285-312.

Hamedani MG, Markus HR, Fu AS (2013) In the land of the free, interdependent action undermines motivation. Psych Sci [epub ahead of print]

Hampton AN, Bossaerts P, O'Doherty JP (2008) Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc Natl Acad Sci USA 105:6741-6746.

Han S, Northoff G (2008) Culture-sensitive neural substrates of human cognition: a transcultural neuroimaging approach. Nat Rev Neurosci 6:646-654.

Han S, Northoff G. (2009) Understanding the self: a cultural neuroscience approach. Prog Brain Res 178:203-212.

Han S, Northoff G, Vogeley K,Wexler BE,Kitayama S, Varnum ME (2013) A cultural neuroscience approach to the biosocial nature of the human brain. Annu Rev Psychol 64:12.1-12.25.

Haselton MG, Nettle D (2006) The paranoid optimist: an integrative evolutionary model of cognitive biases. Pers Soc Psychol Rev 10:47-66.

Heatherton TF (2011) Neuroscience of self and self-regulation. Annu Rev Psychol. 62:363–90.

Hein G, Knight RT (2008) Superior temporal sulcus—It's my area: or is it? J Cogn Neurosci 20:2125-2136.

Heine SJ (2012) Cultural Psychology (1st edition). New York, USA: WW Norton & Company.

Heine SJ, Buchtel EE (2009) Personality: the universal and the culturally specific. Annu Rev. Psychol 60:369-394.

Heine SJ, Hamamura T (2007).In search of East Asian self-enhancement. Pers Soc Psychol Rev 11:4-27.

Heine SJ, Kitayama S, Hamamura T (2007) Which studies test whether self-enhancement is pancultural? Reply to Sedikides, Gaertner, and Vevea, 2007. Asian J Soc Psychol 10:198-200.

Heine SJ, Lehman DR, Ide E, Leung C, Kitayama S, Takata T, Matsumoto H (2001) Divergent consequences of success and failure in Japan and North America: an investigation of self-improving motivations and malleable selves. J Pers Soc Psychol 81:559-615.

Henrich J, Heine SJ, Norenzayan A (2010) The weirdest people in the world? Behav Brain Sci 33:61-83.

Hepach R, Kliemann D, Grüneisen S, Heekeren HR, Dziobek I (2011) Conceptualizing emotions along the dimensions of valence, arousal, and communicative frequency – implications for social-cognitive tests and training tools. Front. Psychol. 2:266.

Hepper EG, Hart CM, Gregg AP, Sedikides C (2011) Motivated expectations of positive feedback in social interactions. J Soc Psychol 151:455-477.

Hewstone M, Stroebe W, Jonas K (2008) Introduction to social psychology (5th edition). Oxford, UK: Blackwell Publishing Ltd.

Hogg MA, Vaughan GM (2008) Social psychology (5th edition). Harlow, UK: Pearson Education Ltd.

Hughes BL, Beer JS (2010) Orbitofrontal cortex and anterior cingulate cortex are modulated by motivated social cognition. Cereb Cortex 22:1372-1381.

Hunt LT, Behrens TEJ (2011) Frames of reference in human social decision making. In Mars RB, Sallet J, Rushworth MFS, Yeung N (Eds.) Neural basis of motivational and cognitive control. Cambridge, MA, USA: MIT Press.

Huys QJM, Vogelstein JT, Dayan P (2009) Psychiatry: insights into depression through normative decision-making models. In Koller D, Schuurmans D, Bengio Y, Bottou L (Eds.) Advances in neural information processing systems. Cambridge, MA, USA: MIT Press.

Izuma K, Saito DN, Sadato N (2008) Processing of social and monetary rewards inthe human striatum. Neuron 58:284-294.

Jenkins AC, Macrae NM, Mitchell JP (2008) Repetition suppression of ventromedial prefrontal activity during judgments of self and others. Proc Natl Acad Sci USA 105:4507-4512.

Jones RM, Somerville LH, Li J, Ruberry EJ, Libby V, Glover G, Voss HU, Ballon DJ, Casey BJ (2011) Behavioral and neural properties of social reinforcement learning. J Neurosci 31:13039-13045.

Kahneman D (2011) Thinking, fast and slow. London, UK: Penguin Books Ltd.

Kacelnik A (2006) Meanings of rationality. In Nudds M, Hurley S (Eds.) Rational Animals? Oxford, UK: Oxford University Press.

Kim H, Markus HR (1999) Deviance or uniqueness, harmony or conformity? A cultural analysis. J Pers Soc Psychol 77:785-800.

Kim HS, Sherman DK, Shelley ET (2008) Culture and social support. Am Psychol 63:518-526.

King-Casas B, Chiu PH (2012) Understanding interpersonal function in psychiatric illness through multiplayer economic games. Biol Psychiatry 15:119-125.

King-Casas B, Tomlin D, Anen C, Camerer CF, Quartz SR, Montague PR. (2005) Getting to know you: reputation and trust in a two-person economic exchange. Science 308:78-83.

King-Casas B, Sharp C, Lomax-Bream L, Lohrenz T, Fonagy P, Montague R (2008) The rupture and repair of cooperation in borderline personality disorder. Science 321:806-810.

Kitayama S, Uskul AK (2011) Culture, mind, and the brain: current evidence and future directions. Annu Rev Psychol 62:419-449.

Klucharev V, Hytönen K, Rijpkema M, Smidts A, Fernández G (2009) Reinforcement learning signal predicts social conformity. Neuron 61:140-151.

Korn CW, Fan Y, Zhang K, Wang C, Han S, Heekeren HR (submitted) Cultural influences on social feedback processing.

Korn CW, Prehn K, Park SQ, Walter H, Heekeren HR (2012) Positively biased processing of self-relevant social feedback. J Neurosci 32:16832-16844.

Korn CW, Sharot T, Walter H, Heekeren HR, Dolan RJ (in press) Depression is related to an absence of optimistically biased belief updating about future life events. Psychol Med.

Krienen FM, Tu P, Buckner RL (2010) Clan mentality: Evidence that the medial prefrontal cortex responds to close others. J Neurosci 30:13906-13915.

Leary MR (2007) Motivational and emotional aspects of the self. Annu Rev Psychol 58:317-44.

Lee D (2008) Game theory and neural basis of social decision making. Nat Neurosci 11:404-409.

Lee D, Seo H, Jung MW (2012) Neural basis of reinforcement learning and decision making. Annu Rev Neurosci 35:287-308.

Lieberman MD (2007) Social cognitive neuroscience: a review of core processes. Annu Rev Psychol 58:259-289.

Lieberman MD (2010). Social cognitive neuroscience. In Gilbert DT, Fiske ST, Lindzey G (Eds.) Handbook of social psychology (5th edition). New York: Wiley.

Logothetis NK (2008) What we can do and what we cannot do with fMRI. Nature 453:869-878.

Lovallo D, Kahneman D (2003) Delusions of success. How optimism undermines executives' decisions. Harv Bus Rev 81:56-63.

Low J, Perner J (2012) Implicit and explicit theory of mind: state of the art. Brit J of Develop Psychol 30: 1-13.

Ma Y, Bang D, Wang C, Allen M, Frith C, Roepstorff A, Han S (2012) Sociocultural patterning of neural activity during self-reflection. Soc Cogn Affect Neurosci [Epub ahead of print].

Macrae CN, Moran JM, Heatherton TF, Banfield JF, Kelley WM (2004). Medial prefrontal activity predicts memory for self. Cereb Cortex 14:647-654.

Maia TV, Frank MJ (2011) From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci 14:154-162.

Mar RA (2011) The neural bases of social cognition and story comprehension. Annu Rev Psychol 62:103-134.

Marks G, Miller N (1987) Ten years of research on the false-consensus effect: an empirical and theoretical review. Psychol Bull 102:72-90.

Markus HR, Kitayama S (1991) Culture and the self: implications for cognition, emotion, and motivation. Psychol Rev 88:224-253.

Markus HR, Kitayama S (2010) Cultures and selves: a cycle of mutual constitution. Pers Psychol Sci 5:420-430.

Markus H, Nurius P (1986) Possible selves. Am Psychol 41:945-969.

Markus H, Wurf E (1987) The dynamic self-concept. Annu Rev Psychol 38:299-337.

Mars RB, Sallet J, Schüffelgen U, Jbadi S, Toni I, Rushworth MFS (2012) Connectivity-based subdivisions of the human right “temporoparietal junction area”: evidence for different areas participating in different cortical networks. Cereb Cortex 22:1894-1903.

Mathews A, MacLeod C (2005) Cognitive vulnerability to emotional disorders. Annu Rev Clin Psychol 1:167-195.

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

Meshi D, Biele G, Korn CW, Heekeren HR (2012) How expert advice influences decision making. PLoS One 7:e49748.

Montague PR, Dolan RJ, Friston KJ, Dayan P (2012) Computational psychiatry. Trends in Cogn Sci 6:72-80.

Montague PR, King-Casas B, Cohen JD (2006) Imaging valuation models in human choice. Annu Rev Neurosci 29:417-448.

Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B (2007) Depression, chronic diseases, and decrements in health: results from the world health surveys. Lancet 370:851-858.

Murphy FC, Michael A, Robbins TW, Sahakian BJ (2003) Neuropsychological impairment in patients with major depressive disorder: the effects of feedback on task performance. Psychol Med 33:455-467.

Murray C, Lopez A (1996) Evidence-based health policy—lessons from the global burden of disease study. Science 274:740-743.

Myers DG (2005) Social psychology (8th edition). New York, USA: Mcgraw-Hill.

Ng SH, Han S, Mao L, Lai JCL (2010) Dynamic bicultural brains: fMRI study of their flexible neural representation of self and significant others in response to culture priming. Asian J Soc Psychol 13:83-91.

Nisbett RE, Masuda T (2003) Culture and point of view. Proc Natl Acad Sci USA 100:11163-11170.

Nisbett RE, Peng K, Choi I, Norenzayan A (2001) Culture and systems of thought: holistic versus analytic cognition. Psychol Rev 108:291-310.

Northoff G, Bermpohl F (2004) Cortical midline structures and the self. Trends Cogn Sci 8:102-107.

Northoff G, Hayes DJ (2011) Is our self nothing but reward? Biol Psychiatry 69:1019-1025.

Northoff G, Heinzel A, de Greck M, Bermpohl F, Dobrowolny H, Panksepp J (2006) Self-referential processing in our brain—a meta-analysis of imaging studies on the self. Neuroimage 31:440- 457.

O’Doherty JP, Hampton A, Kim H (2007) Model-based fMRI and its application to reward learning and decision making. Ann NY Acad Sci 1104:35-53.

Olson IR, Plotzker A, Ezzyat Y (2007) The enigmatic temporal pole: a review of findings on social and emotional processing. Brain 130:1718-1731.

Oyserman D, Coon HM, Kemmelmeier M (2002) Rethinking individualism and collectivism: evaluation of theoretical assumptions and meta-analyses. Psychol Bull 128:3-72.

Park SQ, Kahnt T, Talmi D, Rieskamp J, Dolan RJ, Heekeren HR (2012) Adaptive coding of reward prediction errors is gated by striatal coupling. Proc Natl Acad Sci USA 109:4285-4289.

Philippi CL, Duff MC, Denburg NL, Tranel D, Rudrauf D. (2012) Medial pFC damage abolishes the self-reference effect. J Cogn Neurosci 24:475-481.

Pizzagalli DA (2011) Frontocingulate dysfunction in depression: toward biomarkers of treatment response. Neuropsychopharmacology 36:183-206.

Pizzagalli DA, Iosifescu D, Hallett LA, Ratner KG, Fava M (2008) Reduced hedonic capacity in major depressive disorder: evidence from a probabilistic reward task. J Psychiatric Res 43:76-87.

Pizzagalli DA, Jahn AL, O’Shea JP (2005) Toward an objective characterization of an anhedonic phenotype: a signal-detection approach. Biol Psychiatry 57:319-327.

Plassmann H, O’Doherty J, Shiv B, Rangel A (2008) Marketing actions can modulate neural representations of experienced pleasantness. Proc Natl Acad Sci USA105:1050-1054.

Platt ML, Huettel SA (2008) Risky business: the neuroeconomics of decision making under uncertainty. Nat Neurosci 11:398-403.

Poldrack RA (2006). Can cognitive processes be inferred from neuroimaging data? Trends Cogn Sci 10:59-63.

Price JL, Drevets WC (2012) Neural circuits underlying the pathophysiology of mood disorders. Trends Cogn Sci 16:61-71.

Puri M, Robinson DT (2007) Optimism and economic choice. J Finan Econom 86:71-99.

Rangel A, Camerer C, Montague PR (2008) A framework for studying the neurobiology of value-based decision making. Nat Rev Neurosci 9:545-556.

Ray RD, Shelton AL, Hollon NG, Matsumoto D, Frankel CB, Gross JJ, Gabrieli JDE (2010) Interdependent self-construal and neural representations of self and mother. 5:318-323.

Redcay E, Dodell-Feder D, Pearrow MJ, Mavros PL, Kleiner M, Gabrieli JDE, Saxe R (2010) Live face-to-face interaction during fMRI: A new tool for social cognitive neuroscience. Neuroimage 50:1639–1647.

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

Rilling JK, Sanfey AG (2011) The neuroscience of social decision-making. Annu Rev Psychol 62:23-48.

Roiser JP, Elliott R, Sahakian BJ (2012) Cognitive mechanisms of treatment in depression. Neuropsychopharmacology 37:117-136.

Rosenberg M (1965) Society and the adolescent self-image. Princeton, USA: Princeton University Press.

Rushworth MF, Behrens TE (2008) Choice, uncertainty and value in prefrontal and cingulate cortex. Nat Neurosci 11:389-397.

Rushworth MF, Behrens TE, Rudebeck PH, Walton ME (2007) Contrasting roles for cingulate and orbitofrontal cortex in decisions and social behaviour. Trends Cogn Sci 11:168-176.

Ryder Ag, Yang J, Zhu X, Yao S, Yi J, Heine SJ, Bagby RM (2008) The cultural shaping of depression: somatic symptoms in China, psychological symptoms in North America. J Abnormal Psychol 117:300-313.

Sam DL, Berry JW (2010) Acculturation: when individuals and groups of different cultural backgrounds meet. Pers Psychol Sci 5:472-481.

Sanfey AG (2007) Social decision-making: Insights from game theory and neuroscience. Science 318:599-602.

Saxe R (2006) Uniquely human social cognition. Curr Opin Neurobiol 16:235-239.

Saxe R,Powell LJ (2006) It’s the thought that counts: specific brain regions for one component of theory of mind. Psychol Sci 17:692-699.

Scheier MF, Carver CS (1992) Effects of optimism on psychological and physical well-being: theoretical overview and empirical update. Cogn Ther Res 16:201-228.

Scheier MF, Carver CS, Bridges MW (1994) Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the life orientation test. J Pers Soc Psychol 67:1063-1078.

Schnell K, Bluschke S, Konradt B, Walter H (2011) Functional relations of empathy and mentalizing: an fMRI study on the neural basis of cognitive empathy. Neuroimage 54:1743-1754.

Schultz W (2006). Behavioral theories and the neurophysiology of reward. Annu Rev Psychol 57:87-115.

Sedikides C, Gaertner L, Toguchi Y (2003) Pancultural self-enhancement. J Pers Soc Psychol 84:60-70.

Sedikides C, Gaertner L, Vevea JL (2007) Inclusion of theory-relevant moderators yield the same conclusions as Sedikides, Gaertner, and Vevea (2005): A meta-analytical reply to Heine, Kitayama, and Hamamura (2007). Asian J Soc Psychol 2:59-67.

Sedikides C, Gregg AP (2008) Self-enhancement: food for thought. Persp Psychol Sci 3:102-116.

Seligman ME (1972) Learned helplessness. Annu Rev Med 23:407-412.

Sharot T (2011) The optimism bias. New York, USA: Pantheon Books.

Sharot T, Kanai R, Marston D, Korn CW, Rees G, Dolan RJ (2012b) Selectively altering belief formation in the human brain. Proc Natl Acad Sci USA. 109:17058-17062.

Sharot T, Guitart-Masip M, Korn CW, Chowdhury R, Dolan RJ (2012a) How dopamine enhances an optimism bias in humans. Curr Biol 21:1477-1481.

Sharot T, Korn CW, Dolan RJ (2011) How unrealistic optimism is maintained in the face of reality. Nat Neurosci 4:1475-1479.

Sharot T, Riccardi AM, Raio CM, Phelps EA (2007) Neural mechanisms mediating optimism bias. Nature 450:102-105.

Sin NL, Lyubomirsky S (2009) Enhancing well-being and alleviating depressive symptoms with positive psychology interventions: a practice-friendly meta-analysis. J Clin Psychol 65:467-487.

Singelis TM (1994) The Measurement of independent and interdependent self-construals. Pers Soc Psychol Bull 20:580-591.

Solberg-Nes LS, Segerstrom SC (2006) Dispositional optimism and coping: a meta-analytic review. Pers Soc Psychol Rev 10:235-251.

Somerville LH, Kelley WM, Heatherton TF (2010) Self-esteem modulates medial prefrontal cortical responses to evaluative social feedback. Cereb Cortex 20:3005-3013.

Spreng RN, Mar RA, Kim AS (2009) The common neural basis of autobiographical memory, prospection, navigation, theory of mind and the default mode: a quantitative meta-analysis. J Cogn Neurosci 21:489-510.

Steinbeis N, Koelsch S (2009) Understanding the intentions behind man-made products elicits neural activity in areas dedicated to mental state attribution. Cereb Cortex 19:619-623.

Strunk DR, Adler AD (2009) Cognitive biases in three prediction tasks: a test of the cognitive model of depression. Behav Res Ther 47:34-40.

Strunk DR, Lopez H, DeRubeis RJ (2006) Depressive symptoms are associated with unrealistic negative predictions of future life events. Behav Res Ther 44:861-882.

Sutton RS, Barto AG (1998) Reinforcement learning. Cambridge, MA, USA: MIT Press.

Svenson O (1981) Are we all less risky and more skillful than our fellow drivers? Acta Psychol 47:143-148.

Symons CS, Johnson BT (1997) The self-reference effect in memory: a meta-analysis. Psychol Bull 121:371-394.

Szpunar K, Addis DR, Schacter DL (2012) Memory for emotional simulations: remembering a rosy future. Psychol Sci 23:24-29.

Tamir DI, Mitchell JP (2010) Neural correlates of anchoring-and-adjustment during mentalizing. Proc Natl Acad Sci USA 107:10827-10832.

Taylor SE, Brown JD (1988) Illusion and well-being: a social psychological perspective on mental health. Psychol Bull 103:193–210.

Taylor SE, Collins RL, Skokan LA, Aspinwall LG (1989) Maintaining positive illusions in the face of negative information: Getting the facts without letting them get you. J Soc Clin Psychol 8:114-129.

Tobler PN, Fiorillo CD, Schultz W (2005) Adaptive coding of reward value by dopamine neurons. Science 307:1642-1645.

Triandis HC, Suh EM (2002) Cultural influences on personality. Annu Rev Psychol 53:133-160.

Van Overwalle F (2009) Social cognition and the brain: a meta-analysis. Hum Brain Mapp 30:829-858.

Vo M-L, Jacobs AM,Conrad M (2006) Cross-validating the Berlin affective word list (BAWL). Behav Res Methods 38:606-609.

Vogeley K, Roepstorff A (2009) Contextualising culture and social cognition. Trends Cogn Sci 13:511-516.

Walter H, Abler B, Ciaramidaro A, Erk S (2005) Motivating sources of human actions—Neuroimaging reward and social interaction. Brain Res Bull 67:368-381.

Walter H, Adenzato M, Ciaramidaro A, Enrici I, Pia L, Bara BG (2004) Understanding intentions in social interaction: the role of the anterior paracingulate cortex. J Cogn Neurosci 16:1854-1863.

Wang G, Mao L, Ma Y, Yang X, Cao J, Liu X, Wang J, Wang X, Han S (2012) Neural representations of close others in collectivistic brains. Soc Cogn Affect Neurosci 7:222-229.

Wagner DD, Haxby JV, Heatherton TF (2012) The representation of self and person knowledge in the medial prefrontal cortex. WIREs Cogn Sci 3:451-470.

Weinstein ND (1980) Unrealistic optimism about future life events. J Pers SocPsychol 39:806-820.

Weinstein ND, Klein WM (1995) Resistance of personal risk perceptions to debiasing interventions. Health Psychol 14:132-140.

Wimmer H, Perner J (1983) Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition 13:103-128.

Wolf I, Dziobek I, Heekeren HR (2010) Neural correlates of social cognition in naturalistic settings: a model-free analysis approach. Neuroimage 49:894-904.

Yoshida W, Dolan RJ, Friston KJ (2008) Game theory of mind. PLoS One 4:e1000254.

Yoshida W, Seymour B, Friston KJ, Dolan RJ (2010) Neural mechanisms of belief inference during cooperative games. J Neurosci 30:10744-10751.

Zaki J, Schirmer J, Mitchell JP (2011) Social influence modulates the neural computation of value. Psychol Sci 22:894-900.

Zhu Y, Zhang L, Fan J, Han S (2007) Neural basis of cultural influence on self representation. Neuroimage 34:1310-1317.

Zink CF, Tong Y, Chen Q, Bassett DS, Stein JL, Meyer-Lindenberg A (2008) Know your place: neural processing of social hierarchy in humans. Neuron 58:273–283.

# 8. Supplements

## 8.1. Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt,

- dass ich die vorliegende Arbeit selbstständig und ohne unerlaubte Hilfe verfasst habe,

- dass ich mich nicht bereits anderwärts um einen Doktorgrad beworben habe und keinen Doktorgrad in dem Promotionsfach Psychologie besitze und

- dass ich die zugrunde liegende Promotionsordnung vom 02.12.2008 kenne.

Berlin, 15. Januar 2013

Christoph W. Korn

## 8.2. Research articles

### Study 1

Korn CW, Prehn K, Park SQ, Walter H, Heekeren HR (2012) Positively biased processing of self-relevant social feedback. J Neurosci 32:16832-16844.

The original article is online available at:
http://dx.doi.org/10.1523/JNEUROSCI.3016-12.2012

### Study 2

Korn CW, Fan Y, Zhang K, Wang C, Han S, Heekeren HR (submitted) Cultural influences on social feedback processing.

The article is currently under review in Social Cognitive and Affective Neuroscience (published by Oxford University Press). Attached is a preprint of the article (un-refereed author version).

### Study 3

Korn CW*, Sharot T*, Walter H, Heekeren HR, Dolan RJ (in press) Depression is related to an absence of optimistically biased belief updating about future life events. Psychol Med. *equal contribution

The original article is online available at:
http://dx.doi.org/10.1017/S0033291713001074

**Positively Biased Processing of Self-Relevant Social Feedback**

Christoph W. Korn[1,2,3], Kristin Prehn[3,4], Soyoung Q. Park[1,2,3], Henrik Walter[2,5], and Hauke R. Heekeren[1,2,3,4]

[1]Department of Education and Psychology, Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany

[2]Berlin School of Mind and Brain, Humboldt Universität zu Berlin, Luisenstrasse 56, 10117 Berlin, Germany

[3]Dahlem Institute for Neuroimaging of Emotion, Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany

[4]Cluster of Excellence "Languages of Emotion", Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany

[5]Department of Psychiatry, Division of Mind and Brain Research, Charité University Medicine Berlin, Chariteplatz 1, 10117 Berlin, Germany

Correspondence should be addressed to:

Christoph W. Korn

Habelschwerdter Allee 45, 14195 Berlin, Germany

Phone: 004930838 56226

E-mail: christoph.w.korn@gmail.com

Abbreviated Title: Positively Biased Processing of Social Feedback

Number of pages: 46

Number of figures: 5

Number of tables: 4

Number of words for Abstract: 214

Number of words for Introduction: 627

Number of words for Discussion: 1589

The authors declare that they have no competing financial interests.

### Acknowledgements

This work was supported by the Excellence Initiative of the German Federal Ministry of Education and Research, Deutsche Forschungsgemeinschaft, Grants GSC86/1-2009 and EXC 302.

### Abstract

Receiving social feedback such as praise or blame for one's character traits is a key component of everyday human interactions. It has been proposed that humans are positively biased when integrating social feedback into their self-concept. However, a mechanistic description of how humans process self-relevant feedback is lacking. Here, participants received feedback from peers after a real life interaction. Participants processed feedback in a positively biased way, i.e., they changed their self-evaluation more towards desirable than towards undesirable feedback. Using functional magnetic resonance imaging (fMRI) we investigated two feedback components. First, the rewarding component correlated with activity in ventral striatum and in anterior cingulate cortex/ medio-prefrontal cortex (ACC/MPFC). Second, the comparison-related component correlated with activity in the mentalizing network, including the MPFC, the temporo-parietal junction (TPJ), the superior temporal sulcus (STS), the temporal pole (TP), and the inferior frontal gyrus (IFG). This comparison-related activity within the mentalizing system has a parsimonious interpretation, i.e., activity correlated with the differences between own evaluation and feedback. Importantly, activity within the MPFC that integrated reward- and comparison-related components predicted the self-related positive updating bias across participants offering a mechanistic account of positively biased feedback processing. Thus, theories on both reward and mentalizing are important for a better understanding of how social information is integrated into the human self-concept.

### Introduction

Humans are often confronted with social feedback about their character when interacting with other people and have to integrate this feedback into their self-concept. For example, if somebody tells you that you are polite you weigh this feedback and integrate it into how polite you see yourself. Importantly, people tend to see themselves in a positive light (Leary, 2007) and expect to receive more positive than negative feedback (Hepper et al., 2011). It has been proposed that humans can achieve and maintain a positive self-concept because cognitive-processing mechanisms distort incoming information in a positive direction (Taylor and Brown, 1988). Studying positively biased self-views bears far-reaching implications for psychiatry, health psychology, and policy-making since positivity biases have often been linked to mental health, personal well-being, and professional success (Leary, 2007). The goal of the present study was to determine the information processing mechanisms at play when people receive feedback relevant for their self-concept.

The idea that processing mechanisms distort incoming information in a positive direction suggests that reward should play a central role in social feedback processing. Neuroscientific studies have shown that non-social rewards (e.g. money) and social rewards (e.g. positive feedback on character traits) are processed within shared brain regions, notably the ventral striatum and a region at the border of the pregenual ACC, the ventral MPFC, and the medial orbito-frontal cortex (OFC; Fehr and Camerer, 2007; Fliessbach et al., 2007; Izuma et al., 2008; Beckmann et al., 2009; Rushworth et al., 2011). However, neural activity related to social reward has not been linked to positively biased self-views.

When receiving social feedback about character traits, people compare their own view to the view of others. Self-referential processing, such as judging one’s own personality traits, has been linked to the frontal midline, especially ventral MPFC (Amodio and Frith, 2006; Moran et al., 2006; Northoff et al., 2006; Lieberman, 2007, Wagner et al., 2012). Moreover, inferring the mental states of other’s – known as mentalizing or theory-of-mind – has been reliably associated with a network comprising dorsal MPFC, TPJ, STS, TPs, and orbital IFG (Amodio and Frith, 2006; Gilbert et al., 2006; Saxe, 2006; Van Overwalle, 2009; Bahnemann et al., 2010; Mar, 2011). Activity within the mentalizing network has been observed across a variety of tasks, such as reading stories about false beliefs (Saxe and Powell, 2006), viewing cartoons or videos of social interactions (Walter et al., 2004; Wolf et al., 2010), and engaging in strategic interactions (Behrens et al., 2008; Hampton et al., 2008; Yoshida et al., 2010). Social feedback processing arguably involves inferring other persons' mental state to integrate their views into one's self-concept. However, it has not been tested whether regions associated with mentalizing are implicated in social feedback processing.

Here, we mainly investigated how humans process feedback about their own character traits and were additionally interested in comparing self- versus other-related feedback. We hypothesized that humans process social feedback in a positively biased way and expected feedback processing to include two components. First, we expected a reward-related component to be linked to activity in the ventral striatum and ACC/MPFC. Second, we hypothesized that the comparison between participants' own views and the feedback ratings they received would be reflected in regions previously associated with mentalizing. We expected activity in the MPFC, in particular since distinctive sub-regions of the MPFC have been linked to processes that we expect to be relevant for social feedback processing. First, a region at the border of pregenual ACC, ventral MPFC, and OFC shows involvement in reward and value processing (Beckmann et al., 2009; Rushworth et al., 2011). Second, there is meta-analytic evidence for a spatial gradient with the MPFC with more ventral sub-regions being involved in self-referential processing and more dorsal sub-regions being involved in other-referential processing including mentalizing (Denny et al., 2012).

## Materials and Methods

### Participants

In total, 30 right-handed subjects participated. Three participants had to be excluded (one did not tolerate the scanner environment, another showed excessive head movement (> 8 mm), and data from another subject could not be used due to technical problems) leaving 27 subjects for analyses (14 female, mean age = 24.3 years, standard deviation SD = 2.46). All subjects gave written informed consent.

### Experiment

The experimental procedure is outlined in **Fig. 1**. We wanted participants to believe that they would get realistic feedback on their personality traits from peers with whom they had interacted in real life. We tested how much this feedback changed participants' self-concept by asking them to rate their own personality before and after receiving social feedback. Additionally, each participant rated one other person before and after receiving social feedback for this person. Participants came into the laboratory on two consecutive days. The purpose of the first day was to create a real life interaction among peers so that the social feedback would be meaningful for participants. The purpose of the second day was to assess participants' self-concept change after receiving social feedback.

*Day 1 – Social interaction and rating of 3 players*

On the first day (**Fig. 1A**), participants came into the laboratory in groups of five people of the same sex and got to know each other by playing a table-top version of the popular board game "Monopoly" (HASBRO, Soest, Germany) for 1 h and 15 min. We made sure that participants did not know each other before the experiment. We chose the board game "Monopoly" for the social interaction because it is highly engaging, quite well-known, and allows players to show a variety of cooperative and competitive behaviors. Furthermore, within 1 h 15 min nobody was eliminated from the game. The rules of the game were explained to all participants before the game. The study was introduced as a study about the neural correlates about how people get to know each other. Participants knew before they started to play the game that they were going to be rated by the other players of their group and they believed that their own ratings were going to be shown to the other players in an anonymous fashion. During the game participants were free to talk about whatever topics they wanted. Participants wore name tags and we made sure that participants knew the names of all players after the game. After 1 h 15 min we assessed the ranking of the participants in the game, i.e., assigned the first rank to the winner and so on. After the game, each participant rated three of the four other participants on 80 trait adjectives (**Table 1**; see **Stimuli**) on a Likert scale from 1 (this trait does not apply the person at all) to 8 (this trait does apply the person very much) on a PC using the MATLAB toolbox Cogent 2000 (www.vislab.ucl.ac.uk/cogent.php). Each of the three persons was rated in a separate block. On each trial participants saw one of the 80 adjectives with the first name of the person to rate and had up to 10 s to respond. At the end of day one each participant had rated three other participants and in turn each participant had been rated by three other participants. Participants had not yet rated themselves (depicted in yellow in **Fig. 1A**) and had not yet rated one other player (depicted in green).

*Day 2 – fMRI task and post-fMRI ratings*

On the second day (**Fig. 1B**), participants performed the following fMRI experiment, which was presented using the MATLAB toolbox Cogent 2000. On each trial, participants first saw a cue (1 s) indicating whether the trial was about themselves (self-condition) or about the fourth other participant (other-condition) whom they had not rated on the first day. Then, they saw one of the 80 trait adjectives and had to think about how much that trait applied to themselves or to the other person (imagination phase, 4 s). When the words "How much does this trait apply to you/to this person?" appeared participants had to indicate their rating on an 8-point Likert scale via two button boxes with four buttons each (rating phase, 6 s). After a jittered fixation cross (2, 4, or 6 s) participants saw what they believed to be the mean rating of three other participants from the previous day (feedback phase, 2 s). This mean rating, which served as the feedback rating, was a number with one decimal, ranging from 1.0 to 8.0 in steps of 0.3. The feedback rating was determined by the program during the experiment to reliably create a sufficient number of trials in which participants received desirable and undesirable feedback (see **Task conditions and behavioral analyses** below for a detailed description). After a second jittered fixation cross (1, 3, or 5 s) a new trial began. Participants performed 4 training trials before scanning. The experiment was split up into four blocks with the same 10 positive and the same 10 negative trait adjectives for self and other trials within one block. Trials for self and other were randomly intermixed. Adjectives were randomly assigned to the four blocks for each person.

Immediately after the scanning session participants performed a second rating outside the fMRI scanner on a PC in order to measure how much participants changed their self- and other-ratings after having received social feedback in the scanner. Specifically, they rated themselves and the other person again on all 80 trait adjectives in two separate blocks (one for themselves and one for the other person). These blocks were randomized for order. For each trait adjective participants had up to 6 s to respond.

*Day 2 – Additional behavioral tasks:  memory and individual difference scores*

After rating themselves and the other person a second time, participants were assessed for their memory of the feedback they had received in the scanner. For all 80 trait adjectives participants had to recollect the feedback they had seen in the scanning sessions and had to type in that number, i.e., a number between 1 and 8 with one decimal such as 1.0, 1.3, or 1.7. Participants had to recollect the feedback in two separate blocks (one for themselves and one for the other person), which were randomized for order. They had up to 12 s to respond.

Participants rated how similar they thought the other person was to them on a Likert scale from 1 (not similar at all) to 8 (very similar) and completed the Rosenberg self-esteem scale (Rosenberg, 1965).

### Stimuli

Adjectives were selected on the basis of a comprehensive list of trait adjectives (Anderson, 1968), which had been previously used to create stimuli for social neuroscience experiments (Fossati et al., 2003; Izuma et al., 2008), and on the basis of the Berlin Affective Word List (BAWL; Vo et al., 2006). We selected 40 positive adjectives describing socially desirable traits and 40 negative adjectives describing socially undesirable traits. To assess whether participants really perceived the trait words as positive and negative in the way we had predefined them, participants rated all 80 trait adjectives on social positivity on a scale from 1 (not positive at all) to 8 (very positive) at the very end of the experiment. Mean ratings for positive and negative trait words differed significantly from each other and from the midpoint of the scale (mean rating: positive words = 6.6, SD = 0.49; negative words = 2.4, SD = 0.44; paired-t-test comparing ratings for positive with those for negative words *t*(26) = 29.3, *p* < 0.001; one-sample t-tests comparing ratings to the mid-point of the scale: for positive words *t*(26) = 22.7, *p* < 0.001; for negative words *t*(26) = -25.7, *p* < 0.001).

We used adjectives describing different trait concepts and avoided synonyms or antonyms. Word frequency per million words ranged from 0.09 ("touchy") to 61.32 ("open-minded") with a median frequency of 1.23 ("respectful") as assessed by the lexical database DLEX (Heister et al., 2011; www.dlexdb.de/). See **Table 1** for a list of trait adjectives.

### Task conditions and behavioral analyses

*Task conditions – behavioral analyses*

The main behavioral analyses employed a 2 by 2 design with the within-subject factors feedback target (self/other) and feedback desirability (desirable/undesirable; **Fig. 1C**).

First, feedback was either targeted to the self, i.e., participants rated themselves before and after receiving feedback for themselves, or targeted to one other person, i.e. participants rated one of the other persons he or she had met on the first day before and after receiving feedback for that person.

Second, for each participant trials were classified according to whether feedback was desirable or undesirable. Desirable feedback was defined as feedback ratings that were more "positive" than participants' own initial ratings. For a positive trait, adjective desirable feedback indicated that the feedback rating was numerically *higher* than the initial rating (e.g. a participant's initial rating for "polite" was 6 and the feedback rating was 8). For a negative trait, desirable feedback indicated that the original feedback rating was numerically *lower* than the original initial rating (e.g. a participant's initial rating for "aggressive" was 3 and the feedback rating was 1). Conversely, undesirable feedback was defined as feedback ratings that were more "negative" than participants' own initial ratings. For a positive trait adjective undesirable feedback indicated that the feedback rating was numerically *lower* than the initial rating (e.g. a participant's initial rating for "polite" was 6 and the feedback rating was 4). For a negative trait undesirable feedback indicated that the original feedback rating was numerically *higher* than the original initial rating (e.g. a participant's initial rating for "aggressive" was 3 and the feedback rating was 5).

*Reverse-coding*

Importantly, by the above definition feedback desirability was independent of the valence of the trait word. For all analyses we reverse-coded ratings for negative trait adjectives. Specifically, all ratings were on an 8-point Likert scale ranging from 1 (this trait does not apply the person at all) to 8 (this trait does apply the person very much). Ratings for negative traits were subtracted from 9. For example, if the original rating for a negative trait adjective (e.g. unpleasant) was 1 this number was transformed into 8 for the analyses, i.e., into the rating of the corresponding positive trait adjective (e.g. pleasant).

*Feedback discrepancy*

For each trial (i.e. for each trait adjective; separately for self- and other-conditions) we calculated a "feedback discrepancy" term as the absolute difference between first own ratings and feedback ratings.

(1) feedback discrepancy = I feedback rating – first own rating I

This feedback discrepancy term indicated the social comparison component of receiving social feedback (i.e., the comparison between own ratings and feedback ratings depended on the absolute magnitude of their difference). Since feedback discrepancies were an independent variable of our task we manipulated their magnitude using a random number generator.

*Random number generator for feedback discrepancy*

Feedback discrepancies were determined by a random number generator during the fMRI task to reliably create a similar range of feedback discrepancies across participants and to create a sufficient number of trials with desirable and undesirable feedback. Specifically, on each trial the number of previous trials of the same target condition (self or other) within the same scanning session was determined. These previous trials were classified as either desirable or undesirable according to the definition given above (see *Task conditions*). If the numbers of previous trials with desirable and undesirable feedback differed by more than two trials, the feedback type which had been employed less was chosen (e.g. if there had been 7 trials with desirable feedback and 4 trials with undesirable feedback the feedback of the current trial would by undesirable). Otherwise feedback type was chosen randomly.

Once the feedback type was determined, a random number generator was used to create a feedback discrepancy so that the feedback rating lay between the first own rating on the endpoints of the scale. (For example, a participant had rated herself 6 on “polite” and the feedback should be desirable. In that case the feedback rating had to lie between 6.0 and 8.0. The random number generator determined a feedback rating within that range, i.e., a number between 6 and 8 with one decimal, in steps of 0.3).

To ensure believability of the feedback rating, feedback discrepancies could be zero. These trials were excluded from behavioral analyses (see *Behavioral analyses – ANOVA*).

### *Updates*

To assess how much participants changed their self-concept after receiving social feedback, we calculated an update term quantifying how much participants changed their own ratings.

(2) update = second own rating – first own rating

We expected participants to change their ratings on average towards the feedback ratings. That is, for desirable feedback (i.e. feedback ratings higher than own first rating) participants should increase their ratings (i.e., updates should be positive). For undesirable feedback (i.e. feedback ratings lower than own first rating) participants should decrease their ratings (i.e., updates should be negative).

However, the critical test for positively biased updating is that the change towards desirable feedback (i.e., the increase) is larger than the change towards undesirable feedback (i.e., the decrease). Therefore, trials were split into trials with desirable feedback and trials with undesirable feedback for each participant and both target conditions (self-desirable, self-undesirable, other-desirable, other-undesirable). We first calculated the mean of all signed updates for each participant within each condition and then calculated absolute mean updates. We then scaled absolute mean updates across conditions and participants by the respective mean feedback discrepancies. That is, we obtained relative absolute mean updates for each participant and condition by dividing absolute mean updates by the respective mean feedback discrepancies.

(3) relative absolute mean update = absolute mean update / mean feedback discrepancy

Relative updates can be interpreted in a straightforward way; e.g. a relative update of 0.3 indicates that the change in ratings was on average 30% of the difference between initial own ratings and feedback ratings.

*Behavioral analyses – ANOVA*

For our main behavioral analysis, we performed a 2 (target: self/other) by 2 (desirability: desirable/undesirable) repeated measures ANOVA on relative absolute mean updates. Trials with adjectives for which participants failed to respond in time for the first or second rating were excluded from all analyses (self: mean = 1.7 trials, SD = 1.9; other: mean = 2.2 trials, SD = 2.0). Furthermore, trials with a feedback discrepancy of zero were excluded from behavioral analyses since these trials could not be clearly assigned to either receiving desirable or receiving undesirable feedback (self: mean = 5.5 trials, SD = 2.3; other: mean = 6.4 trials, SD = 2.5).

*Absolute memory errors*

To assess how well participants remembered the feedback presented we asked them to recall all feedback ratings in a separate session. Memory errors were calculated as the absolute differences between the recollected number and the actual feedback rating.

(4) absolute memory error = I feedback rating – recollection of feedback rating I

Mean absolute memory errors were compared in a 2 (target: self/other) by 2 (desirability: desirable/undesirable) repeated measures ANOVA.

### FMRI data acquisition

FMRI were acquired on a 3T scanner (Trio, Siemens, Erlangen, Germany) using a 12-channel head coil. Functional images were acquired with a gradient echo T2*-weighted echo-planar sequence (TR = 2000 ms, TE = 30 ms, flip angle = 70, 64 x 64 matrix, field of view = 192 mm, voxel size = 3x3x3 mm3). A total of 37 axial slices (3 mm thick, no gap,) were sampled for whole brain coverage. Imaging data were acquired in four separate 349-volume runs of 11 min 38 s each. The first five volumes of each run were discarded to allow for T1 equilibration. A high-resolution T1-weighted anatomical scan of the whole brain was acquired (256 x 256 matrix, voxel size = 1x1x1 mm$^3$).

### FMRI data analysis

*Preprocessing*

Image analysis was performed using SPM8 (www.fil.ion.ucl.ac.uk/spm). EPI images were realigned, unwarped, co-registered to the respective participant's T1 scan, normalized to a standard T1 template based on the Montreal Neurological Institute (MNI) reference brain, resampled to 3 mm isotropic voxels, and spatially smoothed with an isotropic 8 mm full width at half maximum (FWHM) Gaussian kernel.

*Modeling of fMRI data – overview*

FMRI time series were regressed onto a general linear model (GLM) containing regressors representing the time periods of the task (**Fig. 1B**): cue (1 s), imagination phase separately for self and other (4 s), rating phase (4 s), feedback phase separately for self and other (2 s), and two motor regressors for button presses with the left and the right hands (0 s). This resulted in 8 regressors per session. The imagination phase regressors for self and other were parametrically modulated by the respective first own ratings. The feedback phase regressors for self and other were modulated by the respective feedback ratings and the respective feedback discrepancies (see *Modeling of fMRI data – parametric modulators* below for more details). This model included trials with feedback discrepancies of zero. The six motion correction parameters estimated from the realignment procedure were entered as covariates of no interest. All regressors and modulators were entered independently into the design matrix, i.e., without the serial orthogonalization used as default in SPM (for a similar approach see Gläscher et al., 2010; Wunderlich et al., 2011). This ensured that only the additional variance that cannot be explained by any other regressor was assigned to the respective effect and thus prevented spurious confounds between regressors. Regressors were convolved with the canonical HRF and low frequency drifts were excluded using a high-pass filter with a 128 s cutoff.

*Modeling of fMRI data – parametric modulators*

For the behavioral analyses we split trials into four categories according to feedback target (self/other) and feedback desirability (desirable/undesirable). In the functional analyses we wanted to investigate trial-by-trial fluctuations in brain activity during the feedback phase, which correlated with two different components of social feedback – reward- and comparison-related components. In our main functional model we therefore split trials according to feedback target (self/other) for each participant and used parametric modulators of feedback ratings and feedback discrepancies to detect activity related to social reward and social comparison, respectively. Thus, we used the full parametric range of feedback ratings and feedback discrepancies across all trials (i.e., across trials with desirable and undesirable feedback).

First, the activity related to the rewarding component of social feedback should correlate positively with the feedback ratings for self. Note that feedback ratings for negative traits were reverse-coded. That is, a high feedback rating indicated high self-relevant social reward (i.e., feedback that a positive trait applied to the self or that a negative trait did not apply to the self) and a low feedback rating indicated low self-relevant social reward (i.e., feedback that a positive trait did not apply to the self or that a negative trait did apply to the self). To make sure that activity related to the rewarding component of social feedback was truly self-specific, we subtracted activity that correlated with the feedback ratings for other.

Second, the activity related to the social comparison component of social feedback should correlate positively with feedback discrepancies defined as the absolute differences between first own ratings and feedback ratings. We defined feedback discrepancies as absolute differences because feedback discrepancies were used to operationalize the social comparison component of social feedback processing; i.e., feedback discrepancies captured how close feedback ratings were to participants' own ratings, regardless of the direction of the differences.

*Follow-up analyses*

To visualize the correlations between neural activity and the parametric modulators (i.e., the betas of the parametric modulators for feedback ratings and the betas of the parametric modulators for feedback discrepancies) we performed follow-up functional region of interest (ROI) analyses. We extracted parameter estimates of the parametric modulators for self and other within the functional ROIs identified in the contrasts used the marsbar toolbox for SPM (marsbar.sourceforge.net/).

Additionally, to analyze activity for desirable and undesirable trials separately in follow-up analyses, we estimated a second GLM to analyze onset activity within functional ROIs defined by the main model described above (see *Modeling of fMRI data – overview* and

*Modeling of fMRI data – parametric modulators)*. Specifically, we estimated a GLM in which regressors for the feedback time period were split up into four conditions in the same fashion as for the main behavioral analysis (self-desirable, self-undesirable, other-desirable, other-undesirable). This follow-up GLM included no parametric modulators.

*Conjunction and statistical inference*

We tested the conjunction null hypothesis using the minimum T-statistic as implemented within SPM8 (Nichols et al., 2005).

All reported activations survived a threshold of $p$ < 0.05 after cluster-wise family-wise error (FWE) correction for multiple comparisons over the entire brain at a cluster-defining threshold of $p$ < 0.0001, uncorrected.

All coordinates are reported in MNI space. Activations are displayed on the standard MNI reference brain. Brodmann areas were manually labeled using the SPM toolbox WFU pick atlas (fmri.wfubmc.edu/software/PickAtlas).

## Results

### Behavioral results – positively biased updating

Participants rated how much 40 positive and 40 negative trait adjectives applied to themselves and to one other person before and after receiving feedback ratings. Importantly, all ratings for negative trait adjectives were reverse-coded so that higher ratings always indicated more positive ratings.

In an initial analysis, we performed a 2 by 2 ANOVA comparing ratings before versus after receiving feedback and ratings targeted to the self versus the other person. Participants rated themselves on average more positively than the other person (main effect: self/other; $F(1,26) = 6.7$, $p < 0.05$, $\eta_p^2 = 0.21$; **Fig. 2A**), indicating a positivity bias towards the self. They also rated themselves and the other person more positively after receiving feedback (main effect: before/after; $F(1,26) = 9.6$, $p < 0.005$, $\eta_p^2 = 0.27$). The interaction was not significant ($p > 0.6$).

In our main behavioral analyses, we tested whether participants showed positively biased processing of social feedback. Specifically, we assessed how participants updated their ratings depending on feedback target (self/other) and feedback desirability (desirable/undesirable; **Fig. 1C**). Desirable feedback was defined as feedback ratings that were higher than participants' first ratings. Conversely, undesirable feedback was defined as feedback ratings lower than participants' first ratings. Participants changed their ratings on average towards the feedback. They increased their ratings for desirable feedback (indicated by positive mean updates significantly different from zero) and decreased their ratings for undesirable feedback (indicated by negative mean updates significantly different from zero; mean update self-desirable = 0.5, SD = 0.35; one-sample t-tests against zero $t(26) = 7.4$, $p < 0.001$; mean update self-undesirable = -0.2, SD = 0.32; $t(26) = -2.7$, $p < 0.05$; mean update other-desirable = 0.6, SD = 0.33; $t(26) = 8.8$, $p < 0.001$; mean update other-undesirable = -0.3, SD = 0.42; $t(26) = -3.0$, $p < 0.01$).

Importantly, the critical test for positively biased updating is that the changes towards desirable feedback are larger than changes towards undesirable feedback (i.e., that absolute mean updates are larger for desirable than undesirable feedback). Additionally, we scaled absolute mean updates by the respective mean feedback discrepancies (i.e., the differences between first own ratings and feedback ratings) to account for possible differences in feedback discrepancies across conditions and participants (relative absolute mean updates: self-desirable = 0.3, SD = 0.23; self-undesirable = 0.1, SD = 0.16; other-desirable = 0.3, SD = 0.20; other-undesirable = 0.1, SD = 0.24). Performing a 2 by 2 ANOVA on relative absolute mean updates comparing self- versus other-directed feedback and desirable vs. undesirable feedback we found that participants showed positively biased processing of social feedback. After receiving desirable feedback participants updated their self- and other-ratings more towards the positive than they updated their ratings towards the negative after receiving undesirable feedback (main effect: desirable/undesirable: $F(1,26) = 12.9$, $p < 0.005$, $\eta_p^2 = 0.33$; **Fig. 2B**). Positively biased feedback processing did not differ between self- and other-directed feedback (main effect: self/other: $p > 0.1$; interaction: $p > 0.6$). In a follow-up analysis we confirmed that similar results were observed, when comparing absolute mean updates that were not scaled by the respective mean feedback discrepancies (main effect: desirable/undesirable: F(1,26) = 15.0, $p < 0.001$, $\eta_p^2 = 0.37$; main effect: self/other: $p > 0.1$; interaction: $p > 0.8$) since the magnitude of mean feedback discrepancies was equal across conditions ($p > 0.1$).

Additionally, we investigated possible influences on the positive updating bias. When participants gave the highest rating possible, they could not receive a feedback rating higher than their own rating and thus feedback could not be desirable. The reverse was true when participants gave the lowest rating possible. To exclude that this relationship between first ratings and feedback compromised our results we tested for positively biased updating only for trials with first ratings in the middle range of the scale (4, 5, and 6). Updating for desirable versus undesirable feedback was still higher when including only trials with first ratings in the middle range of the scale (main effect: desirable/undesirable: $F(1,26) = 14.4$, $p < 0.001$, $\eta_p^2 = 0.36$; main effect: self/other: $p > 0.8$; interaction: $p > 0.5$). This analysis excluded the possibility that positively biased updating was driven by trials in which participants initially rated themselves or the other person on the highest or lowest points of the scale.

Furthermore, we tested whether the valence of the trait adjectives had an effect on updating. We split update scores according to the valence of the trait words and performed a 2 (trait valence: positive/negative) by 2 (feedback target: self/other) by 2 (desirability: desirable/undesirable) ANOVA on absolute mean updates divided by absolute mean feedback discrepancies. Only the main effect of desirability reached significance ($F(1,26) = 13.0$, $p < 0.005$, $\eta_p^2 = 0.33$). Specifically, the interaction between the factors trait valence and desirability did not reach significance ($p > 0.9$), excluding the possibility that trait valence had an effect on positively biased updating in our paradigm.

In sum, our behavioral results establish that humans take desirable feedback more into account than undesirable feedback.

**Behavioral results – control analyses and individual differences**

For an additional control analysis, participants recollected outside the scanner the feedback rating they had seen inside the scanner. Mean absolute memory errors were smaller for self- than for other-related feedback ($F(1,26) = 25.4$, $p < 0.0001$, $\eta_p^2 = 0.49$) but did not differ between desirable and undesirable feedback ($p > 0.1$). Furthermore, we conducted two separate ANCOVAs, one for self and one for other, testing whether the difference between desirable and undesirable updates remained significant when entering additional scores as covariates. These scores were the differences between trials with desirable and undesirable feedback for first ratings, participants' social desirability ratings of the trait adjectives, memory errors, or reaction times on the first or second ratings. The difference between desirable and undesirable updates remained significant when controlling for these scores (self: $F(1,21) = 15.8$, $p < 0.001$; other: $F(1,21) = 8.6$, $p < 0.01$). Moreover, winning or losing in the board game that participants played to get to know each other before receiving feedback, did not have any influence on behavior during the task. Specifically, participants' rank order in the game did not correlate with mean ratings or any update measure using Spearman's correlation coefficient (all $p > 0.1$). Thus, positively biased updating could not be explained by differential memory, first ratings, social desirability ratings of the trait adjectives, valence of the trait words, reaction times, or performance in the board game.

Next, we aimed to establish links between performance in our task and individual differences in trait self-esteem and perceived similarity between self and other. As expected, mean first ratings for self correlated significantly with scores on the Rosenberg self-esteem scale across participants (Pearson's $r = 0.59$, $p < 0.005$) – the higher a participant's trait self-esteem the more positive his or her mean rating across all trait adjectives. Mean first ratings for the other person correlated with perceived similarity to the other person, which was assessed on a Likert scale from 1 (not similar at all) to 8 (very similar; Pearson's $r = 0.51$, $p < 0.01$). Thus, mean ratings in our task were related to inter-subject differences in trait self-esteem and perceived similarity of the other person. In a next step, we explored how first ratings were related to the update bias across participants. Mean first self ratings did not correlate significantly with the updating bias for self (Pearson's r = 0.04, p > 0.8). However, mean ratings for the other person did correlate significantly with the magnitude of the update bias for this other person, i.e., the absolute relative mean update for desirable minus undesirable feedback (Pearson's r = 0.51, $p < 0.01$). This suggests that the higher the other person was rated on average the more pronounced was the positively biased updating pattern. Positively biased updating for self seemed to be unrelated to mean self ratings in our sample.

Behaviorally, participants showed a positively biased updating pattern after receiving feedback. Therefore, we turned to our fMRI data to establish a link between biased updating and neural feedback processing. Specifically, we examined reward- and comparison-related components of feedback processing.

### BOLD signals for self- versus other-related feedback

In an initial step, we examined brain activity during the feedback period to find regions in which activation differed between the processing of self- and other-related feedback. We expected regions previously implicated in self- and other-referential processing, notably the MPFC (Amodio and Frith, 2006). Contrasting the time-point when participants received self-related versus when they received other-related feedback (self > other), we found activity in the medial prefrontal wall (peak voxel in MNI coordinates x, y, z: -3, 59, 28; $p < 0.05$ corrected for multiple comparisons at a cluster-defining threshold of $p < 0.0001$) as well as bilaterally in the orbital part of the IFG extending into the anterior insula (left: -33, 17, -17; right: 30, 20, -17; see **Table 2** for a comprehensive list of activations). The reverse contrast (other > self) revealed among other regions activity in bilateral precuneus (12, -46, 52; $p < 0.05$ corrected for multiple comparisons at a cluster-defining threshold of $p < 0.0001$; **Table 2**).

During the imagination phase, participants in our task rated themselves and another person in a similar way as shown in many previous studies (Northoff et al., 2006; Denny et al., 2012), (**Fig. 1B**). Therefore, we wanted to explore possible differences in activity between feedback phase and imagination phase in a follow-up ROI analysis. We concentrated this analysis to the MPFC since this region has been most consistently been implicated in self-related processing. We extracted parameter estimates during both time points within an ROI that was independently defined based on a recent meta-analysis of self-referential processing (Denny et al., 2012; sphere with a radius of 15 mm centered at the MNI coordinates -6, 50, 4). Parameter estimates were compared in a 2 (imagination/feedback phase) by 2 (self/other) ANOVA. As expected there was a significant main effect of activity for self being higher than for other ($F(1,26) = 126.1$, $p < 0.0001$, $\eta_p^2 = 0.83$). There was also a significant main effect of phase with activity during the feedback phase being higher than during the imagination phase ($F(1,26) = 9.2$, $p < 0.01$, $\eta_p^2 = 0.26$). The interaction was not significant ($p > 0.1$).

These results show that self-relevant feedback implicates the MPFC as has been reliably shown for self-referential processing in general (Amodio and Frith, 2006; Northoff et al., 2006; Denny et al., 2012).

### BOLD signals related to the rewarding component of social feedback

The behavioral analyses showed that participants processed desirable feedback more than undesirable feedback. We hypothesized that neural activity during the feedback phase should mirror two aspects of social feedback processing – a reward-related aspect (operationalized by feedback ratings) and a comparison-related aspect (operationalized by feedback discrepancies). Therefore, in order to identify neural activity related to these two components, we used the full parametric range of feedback ratings and feedback discrepancies. Our model included separate onset regrossors for self- and other-related feedback, which were parametrically modulated by the respective feedback ratings and feedback discrepancies. This model allowed us to search for regions in which these parameters correlated with blood-oxygen-level-dependent (BOLD) signal in a trial-by-trial fashion..

To test for activity correlating with the rewarding component of feedback at the time-point of feedback, we performed a contrast between the two parametric modulators for feedback ratings (feedback ratings for self and feedback ratings for other). First, activity related to reward should correlate positively with feedback ratings for self. That is, the higher the feedback rating the more rewarding was the social feedback (e.g. receiving a self-related feedback rating of 8 is more rewarding than a feedback rating of 7). Note that feedback ratings for negative trait adjectives were reverse-coded so that a higher feedback rating always indicated a more positive feedback. Second, reward-related activity should be self-specific. That is, the trial-by-trial correlation of BOLD signal changes with the feedback ratings for self should be greater than those for other. Contrasting the parametric modulators for the feedback ratings for self versus other, revealed activity in bilateral ventral striatum (left: -15, 2, 11, right: 12, 5, -8) and in a region encompassing anterior cingulate cortex (ACC) and MPFC (3, 32, 25;$p < 0.05$ corrected for multiple comparisons at a cluster-defining threshold of $p < 0.0001$; **Fig. 3A**; see **Table 3** for a comprehensive list of activations).

To better illustrate the correlations between feedback ratings and neural activity we performed a follow-up analysis. We extracted parameter estimates of the parametric modulators for self and other within the functional ROIs identified in the above contrast (**Fig. 3B**). Parametric modulators indicate the correlation (i.e., the slope) between BOLD signals and feedback ratings but give no information about mean onset activity (i.e., the intercept). To additionally illustrate mean onset activity, we estimated a follow-up general linear model. In this follow-up model, trials were separated into four categories according to feedback target (self/other) and feedback desirability (desirable/undesirable) in the same way as in the main behavioral analysis. We extracted parameter estimates of the onset regressors for the four categories within three of the functional ROIs defined by the main model (MPFC and left and right striatum). Plotting these onset regressors illustrates the interaction of feedback target and desirability as defined by the contrast in the main model (**Fig. 3C**). Additionally, mean onset activity showed a significant main effect for self versus other in the MPFC ($F(1,26) = 29.9$, $p < 0.0001$; since we performed an ANOVA within each of the three ROIs, $p$-values were adjusted using a Bonferroni correction for the number of ROIs). In the right striatum the same pattern was observed at trend level ($F(1,26) = 6.1$, $p = 0.06$).

Additionally, we performed the reverse contrast to the one performed above, i.e., we searched for regions that correlated with other-related feedback ratings more than with self-related feedback ratings. This contrast revealed no significant voxels at a threshold of $p < 0.05$ cluster-corrected at a cluster-defining threshold of $p < 0.0001$.

Taken together, the rewarding component of social feedback correlated with activity in ventral striatum and ACC/MPFC, regions previously implicated in processing social and non-social rewards (Izuma et al., 2008; Beckmann et al., 2009).

**BOLD signals related to the comparison component of social feedback**

Having identified activity correlating with the rewarding aspect of feedback, we next tested for BOLD signal changes correlating with the comparison-related aspect on a trial-by-trial basis at the time-point of feedback – both for self- and other-related feedback. Comparison-related activity was operationalized as activity that showed a positive correlation with feedback discrepancies, i.e., with the absolute differences between participants' own ratings and the feedback ratings they received. That is, a larger feedback discrepancy (e.g. a difference between own rating and feedback rating of 2) indicated more "need" for a comparison process than a smaller feedback discrepancy (e.g. 1) regardless of the direction of the difference.

Feedback discrepancies for both self and other correlated positively with activity in MPFC (6, 56, 28), pre-SMA/SMA (9, 17, 64), right STS (51, -25, -8), bilateral IFG (orbital part) extending into anterior insula (left: -36, 20, -23, right: 33, 20, -17), right TPJ (57, -58, 25), left TP (-48, 11, -35), and left cerebellum (-24, -82, -35; $p < 0.05$ corrected for multiple comparisons at a cluster-defining threshold of $p < 0.0001$; **Fig. 4A**; **Table 4**).

As described above for BOLD signal changes related to social reward, we plotted parameter estimates of the parametric modulators for the feedback discrepancies for self and other within the functional ROIs to better illustrate the correlations of feedback discrepancies and BOLD signals (**Fig. 4B**). These parametric modulators indicate the positive correlation (i.e., the slope) between BOLD signals and feedback discrepancies but give no information about the mean onset activity (i.e., the intercept). To extract mean onset activity within seven of the functional ROIs defined by the first model (MPFC, pre-SMA/SMA, right STS, left and right IFG, right TPJ, and left TP), we  conducted a follow-up analysis using a follow-up general linear model, in which trials were separated into four categories according to feedback target and feedback desirability. Mean onset activity showed a significant main effect for self versus other in MPFC (*F*(1,26) = 116.7, *p* < 0.0001), right STS (*F*(1,26) = 9.8, *p* < 0.05), left IFG (*F*(1,26) = 53.6, *p* < 0.0001), right IFG (*F*(1,26) = 77.8, *p* < 0.0001), and left TP (*F*(1,26) = 8.7, *p* < 0.05; since we performed an ANOVA within each of the seven ROIs, *p*-values were adjusted using a Bonferroni correction for the number of ROIs; **Fig. 4C**). Thus, while the relation of feedback discrepancies to BOLD signal changes was the same for self- and other-related feedback (as determined in the first model with the parametric modulators), mean activity was higher for self- versus other-related feedback within these regions (as determined in the second model in which trials were split up into categories for desirable and undesirable feedback).

We also searched for regions in which BOLD signals correlated negatively with the feedback discrepancies for self and other (**Table 4**). BOLD signal changes in no region correlated differentially for self- versus other-related feedback discrepancies, i.e., self > other or other > self, at a threshold of *p* < 0.05 cluster-corrected at a cluster-defining threshold of *p* < 0.0001.

In sum, both in the self- and in the other-condition the difference between participants own views and the feedback they received, i.e., the comparison-related component, correlated with activity in regions previously implicated in mentalizing (Mar, 2011).

### Updating bias for self and activity integrating reward- and comparison-components

Having identified activity that correlated with the rewarding aspect of feedback and activity that correlated with the comparison-related aspect of feedback, we next examined how neural activity was linked to the behavioral update bias for self. We postulated that neural activity mediating the update bias for self should fulfill two requirements. First, candidate regions should integrate activity related to both reward and comparison. Second, activity within this region should correlate with the behavioral update bias for self across participants, i.e., the difference between updates for desirable and undesirable feedback. To address the first requirement, we performed a conjunction analysis testing the conjunction null hypothesis to search for regions which were activated by both reward- and comparison-related activity. The conjunction revealed a region at the border of the MPFC and the ACC (3, 56, 19; *p* < 0.05 corrected for multiple comparisons at a cluster-defining threshold of $p < 0.0001$; **Fig. 5A**; **Table 4**). To address the second requirement, we extracted parameter estimates of self-related absolute feedback discrepancies within this region and tested for a correlation with the behavioral update bias for self. Parameter estimates of self-related absolute feedback discrepancies within the functional ROI defined by the conjunction analysis predicted the behavioral update bias for self (Pearson's r = 0.42, $p < 0.05$, 95% confidence interval [0.05, 0.69]; **Fig. 5B**). Additionally, we extracted parameter estimates of other-related feedback discrepancies errors within the same functional ROI and found no correlation with the behavioral update bias for other (Pearson's r = 0.07, $p > 0.7$). However, we note that the two Pearson's correlation coefficients did not differ significantly (z = 1.3, $p > 0.1$, using the $\bar{Z}_2$* statistic described by Steiger, 1980).

Thus, BOLD signals within the MPFC that integrated reward- and comparison-related components of social feedback predicted individual differences in the self-related positive updating.

## Discussion

After interacting with peers in a real-life setting and then receiving social feedback from them, participants showed positively biased updating of their self- and other-evaluations. Specifically, participants updated their evaluation of themselves and of another peer more towards desirable feedback than towards undesirable feedback. Our fMRI data suggest that neural activity reflects two different components of social feedback. First, activity in the bilateral ventral striatum and in a region encompassing parts of the ACC/MPFC tracked the rewarding component. Second, parts of the mentalizing network tracked the comparison-related component. Changes in activity within the MPFC that integrated reward and comparison-related aspects of feedback predicted the self-related positive updating bias across participants. Our results suggest that a combination of neural signals related to social reward and to the comparison between own views and feedback mediate positively biased feedback processing.

So far only few studies in social neuroscience have investigated social feedback processing (Somerville et al., 2006; Izuma et al., 2008; Davey et al., 2010; Somerville et al., 2010; Eisenberger et al., 2011; Jones et al., 2011). Three crucial aspects of our design allowed us to considerably add to these studies. First, participants in our task engaged in a real life interaction of more than an hour whereas in previous studies participants received feedback which was based on photographs and/or questionnaires. Second, we parametrically modulated the desirability of the feedback and even more importantly we assessed the difference between participants' self-views and the feedback they received. In previous studies feedback was mostly binary and participants did not indicate their own view, e.g. participants just got to know whether they were liked or not or whether a certain trait word applied to them. Third, by assessing how feedback changed self-views we demonstrate a positivity bias in feedback processing.

Positivity biases have been documented across many domains in social cognition (Leary, 2007) and it has been proposed that they arise because cognitive processing mechanisms distort incoming information in a positive direction (Taylor and Brown, 1988). Here, we provide evidence for this idea by showing a striking asymmetry in how humans process self-relevant information about their character traits. A similar approach has been used in the domain of optimism (Sharot et al., 2011). Participants estimated their likelihood of experiencing various negative events in the future. They updated their beliefs more towards the actual statistical likelihood when it was desirable than when it was undesirable. Thus, our results suggest that positivity biases in general may arise due to asymmetric information processing.

Some recent studies investigating social conformity have used designs similar to the present study (Klucharev et al., 2009; Campbell-Meiklejohn et al., 2010; Zaki et al., 2011). In these studies participants make a first evaluation of an object (e.g. an unknown face or song) and then receive feedback from others about this object. Conformity can then be measured as the degree to which participants change their evaluation towards the others' opinion similar to the update measure in our study. However, in these conformity studies participants are unbiased (e.g. they are influenced to the same degree when they see that others judge an unknown song to be better or worse than they do). In contrast, participants in our study processed social information in a positively biased way since the "object of conformity" consisted of participants' own character traits and the character traits of peers.

Behavioral studies have commonly discussed positively biased self-views with relation to theories about the self but not with relation to theories about reward. Here, we specify neural activity related to the rewarding component of positively biased feedback processing. Using a parametric design we show that the ventral striatum and the ACC/MPFC process the self-related reward associated with social feedback. Our results thus replicate the findings by Izuma et al. (2008) and extend them to negative character traits. The striatum and the ACC/MPFC – especially its middle and more ventral parts – are connected and both structures have been linked to reward in social and non-social contexts (Beckmann et al., 2009). Interestingly, activity in the ventral part of the MPFC plays a role in the representation of the value of objects (Rangel et al., 2008; Rushworth et al., 2011) and this activity can be modulated by social influences (Plassmann et al., 2008; Zaki et al., 2011). In sum, our results corroborate that social reward processing can be linked to structures involved in non-social reward processing.

Critically, in addition to the rewarding aspect of social feedback our task also modulated the distance between participants' own views and the feedback they received. This comparison between own views and feedback correlated among others with activity in the MPFC, right STS, bilateral IFG, right TPJ, and left TP. All of these regions are part of the mentalizing network (Amodio and Frith, 2006; Bahnemann et al., 2010). Especially, the MPFC and the TPJ have been most consistently linked to various mentalizing tasks (Van Overwalle, 2009; Mar, 2011). In our study, MPFC activity showed stronger activation than TPJ activity, which is consistent with a recent meta-analysis (Mar, 2011) showing that the MPFC is particularly involved in tasks that are not based on explicit false belief stories as was the case in our task. Furthermore, such non-story based tasks often implicate the orbital IFG (Mar, 2011) and we therefore interpret orbital IFG activity in relation to its plausible role in mentalizing associated processes.

It is important to note that changes in neural activity in the mentalizing network have a very parsimonious interpretation in our task. Activity in the mentalizing network tracked the numerical difference between participants' own evaluations and the feedback they received both for their own character and for the character of another person. Recently, some studies have begun to investigate neural activity related to social cognition by using computational parameters derived from modified versions of reinforcement models or other types of computational models (Behrens, et al. 2008; Hampton et al., 2008; Coricelli and Nagel, 2009; Yoshida, et al., 2010; Biele et al., 2011). These studies provide first steps towards conceptualizing the precise computations underlying activity in the mentalizing network or parts of it (Behrens, et al. 2009). In line with the results of these studies, our results provide a mechanistic account of activity in the mentalizing network for the processing of social feedback.

Our results show that the behavioral updating bias for self is associated with both reward-related and comparison-related components of social feedback. Activity within a region in the MPFC that integrated the two components predicted the amount of positively biased updating for self-related feedback across participants. This was not the case for other-related feedback. Behavioral accounts (Taylor and Brown, 1988) have argued for a filtering mechanism that distorts incoming social information towards the positive. Our results suggest that MPFC activity reflects this filtering mechanism in our task.

The implication of the MPFC in social cognition in general and in self-related processing in particular has been reliably shown by many studies (Mitchell, 2009; Denny et al., 2012; Wagner et al. 2012). Importantly, Moran et al. (2006) have shown that MPFC activity was higher when participants made trait ratings that were self-descriptive compared to when they made self-ratings that were not self-descriptive – independent of trait valence. The MPFC region that integrated reward- and comparison-related components in our task was within the region described by Moran et al. (2006). This suggests that neural processes related to thinking about trait self-descriptiveness and neural processes related to receiving feedback on trait self-descriptiveness might be instantiated in a common MPFC region. The relation of the MPFC to self-related positively biased updating is also concordant with a previous study in which participants received information that they were either liked or not liked by other persons (Somerville et al., 2010). In this study, trait self-esteem correlated with the differential activity towards positive versus negative feedback in a similar region of the MPFC as identified in our task. Importantly, our results are also in line with literature linking different sub-regions of the MPFC to reward processing, self-referential thinking, and mentalizing. Specifically, reward and value processing have been consistently linked to a ventral MPFC region at the border of pregenual ACC and medial OFC (Rangel et al. 2008; Beckmann et al., 2009; Rushworth et al., 2011). Self-referential thinking most consistently involves a ventral part of the MPFC whereas mentalizing involves a more dorsal part (Northoff et al., 2006; Mar, 2011; Denny et al., 2012). The MPFC regions that integrated reward- and comparison-related components of social feedback in our task lay at border position in which there might be some overlap between reward-, self-, and mentalizing-related activity. Our results suggest that this MPFC region seems to be ideally suited for positively biased integration of social information into one's self-concept and that it might be interesting to investigate this region's involvement in reward-, self-, and mentalizing-related processes more closely.

### Conclusions

Many studies have tried to weigh the benefits (e.g. reduced anxiety) and costs (e.g. overly risky decision making) of positivity biases against each other (Taylor and Brown, 1988; Leary, 2007). Positivity biases seem to be generally adaptive but can be detrimental if they are too extreme. To further specify their costs and benefits, it is fundamental to understand the underlying mechanisms. Our results show that positively biased social feedback processing is related to an integration of activity linked to reward and mentalizing. This underscores the importance of integrating theories on reward and mentalizing. By directing the focus towards the interplay between reward processing and mentalizing, we provide an essential step towards a better understanding of how social information is integrated into the human self-concept.

### References

Anderson NH (1968) Likableness ratings of 555 personality-trait words. J Pers Soc Psychol 9:272–279.

Amodio DM, Frith CD (2006) Meeting of minds: the medial frontal cortex and social cognition. Nat Rev Neurosci 7:268–277.

Bahnemann M, Dziobek I, Prehn K, Wolf I, Heekeren HR (2010) Sociotopy in the temporoparietal cortex: common versus distinct processes. Soc Cog Affect Neurosci 5:48–58.

Beckmann M, Johansen-Berg H, Rushworth MFS (2009) Connectivity-based parcellation of human cingulate cortex and its relation to functional specialization. J Neurosci 29:1175–1190.

Behrens TEJ, Hunt LT, Woolrich MW, Rushworth MFS (2008) Associative learning of social value. Nature 456:245–249.

Behrens TEJ, Hunt LT, Rushworth MFS (2009) The computation of social behavior. Science 324:1160–1164.

Biele G, Rieskamp J, Krugel LK, Heekeren HR (2011) The neural basis of following advice. PLoS Biol 9:e1001089.

Campbell-Meiklejohn DK, Bach DR, Roepstorff A, Dolan RJ, Frith CD (2010) How the opinion of others affects our valuation of objects. Curr Biol 20:1165–1170.

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

Davey CG, Allen NB, Harrison BJ, Dwyer DB, Yücel M (2010) Being liked activates primary reward and midline self-related brain regions. Hum Brain Mapp 31:660–668.

Denny BT, Kober H, Wager TD, Ochsner KN (2012) A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. J Cogn Neurosci 24:1742–1752.

Eisenberger NI, Inagaki TK, Muscatell KA, Byrne Haltom KE, Leary MR (2011) The neural sociometer: brain mechanisms underlying state self-esteem. J Cogn Neurosci 23:3448–3455.

Fehr E, Camerer CF (2007) Social neuroeconomics: the neural circuitry of social preferences. Trends Cogn Sci 11:419–427.

Fliessbach K, Weber B, Trautner P, Dohmen T, Sunde U, Elger CE, Falk A (2007) Social comparison affects reward-related brain activity in the human ventral striatum. Science 318:1305–1308.

Fossati P, Hevenor SJ, Graham SJ, Grady C, Keightley ML, Craik F, Mayberg H (2003) In search of the emotional self: an fMRI study using positive and negative emotional words. Am J Psychiatry 160:1938–1945.

Gilbert SJ, Spengler S, Simons JS, Steele JD, Lawrie SM, Frith CD, Burgess PW (2006) Functional specialization within rostral prefrontal cortex (area 10): a meta-analysis. J Cogn Neurosci 18:932–948.

Gläscher J, Daw N, Dayan P, O'Doherty JP (2010) States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66:585–595.

Hampton AN, Bossaerts P, O'Doherty JP (2008) Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc Natl Acad Sci USA 105:6741–6746.

Heister J, Würzner K-M, Bubenzer J, Pohl E, Hanneforth T, Geyken A, Kliegl R (2011) dlexDB - eine lexikalische Datenbank für die psychologische und linguistische Forschung. Psychologische Rundschau 62:10–20.

Hepper EG, Hart CM, Gregg AP, Sedikides C (2011) Motivated expectations of positive feedback in social interactions. J Soc Psychol 151:455–477.

Izuma K, Saito DN, Sadato N (2008) Processing of social and monetary rewards in the human striatum. Neuron 58:284–294.

Jones RM, Somerville LH, Li J, Ruberry EJ, Libby V, Glover G, Voss HU, Ballon DJ, Casey BJ (2011) Behavioral and neural properties of social reinforcement learning. J Neurosci 31:13039–13045.

Klucharev V, Hytönen K, Rijpkema M, Smidts A, Fernández G (2009) Reinforcement learning signal predicts social conformity. Neuron 61:140–151.

Leary MR (2007) Motivational and emotional aspects of the self. Annu Rev Psychol 58:317-344.

Lieberman MD (2007) Social cognitive neuroscience: a review of core processes. Annu Rev Psychol 58:259–289.

Mar RA (2011) The neural bases of cognition and story comprehension. Annu Rev Psychol 62:103–134.

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

Moran JM, Macrae CN, Heatherton TF, Wyland CL, Kelley WM (2006) Neuroanatomical evidence for distinct cognitive and affective components of self. J Cogn Neurosci 18:1586–1594.

Nichols T, Brett M, Andersson J, Wager T, Poline JB (2005) Valid conjunction inference with the minimum statistic. Neuroimage 25:653–660.

Northoff G, Heinzel A, de Greck M, Bermpohl F, Dobrowolny H, Panksepp J (2006) Self-referential processing in our brain – a meta-analysis of imaging studies on the self. Neuroimage 31:440–457.

Plassmann H, O'Doherty J, Shiv B, Rangel A (2008) Marketing actions can modulate neural representations of experienced pleasantness. Proc Natl Acad Sci USA 105:1050–1054.

Rosenberg M (1965) Society and the adolescent self-image. Princeton, NJ: Princeton University Press.

Rangel A, Camerer C, Montague PR (2008) A framework for studying the neurobiology of value-based decision making. Nat Rev Neurosci 9:545–556.

Rushworth MFS, Noonan MP, Boorman ED, Walton ME, Behrens TE (2011) Frontal cortex and reward-guided learning and decision-making. Neuron 70:1054–1069.

Saxe R (2006) Uniquely human social cognition. Curr Opin Neurobiol 16:235–239.

Saxe R, Powell LJ (2006) It's the thought that counts: specific brain regions for one component of theory of mind. Psychol Sci 17:692–699.

Sharot T, Korn CW, Dolan RJ (2011) How unrealistic optimism is maintained in the face of reality. Nat Neurosci 4:1475–1479.

Somerville LH, Heatherton TF, Kelley WM (2006). Anterior cingulate cortex responds differentially to expectancy violation and social rejection. Nat Neurosci 9:1007–1008.

Somerville LH, Kelley WM, Heatherton TF (2010) Self-esteem modulates medial prefrontal cortical responses to evaluative social feedback. Cereb Cortex 20:3005–3013.

Steiger JH (1980) Tests for comparing elements of a correlation matrix. Psychol Bull 87:245–251.

Taylor SE, Brown JD (1988) Illusion and well-Being: A social psychological perspective on mental health. Psychol Bull 103:193–210.

Van Overwalle F (2009) Social cognition and the brain: a meta-analysis. Hum Brain Mapp 30:829–858.

Vo M-L, Jacobs AM, Conrad M (2006) Cross-validating the Berlin affective word list (BAWL). Behav Res Methods 38:606–609.

Wagner DD, Haxby JV, Heatherton TF (2012) The representation of self and person knowledge in the medial prefrontal cortex. WIREs Cogn Sci 3:451–470.

Walter H, Adenzato M, Ciaramidaro A, Enrici I, Pia L, Bara BG (2004) Understanding intentions in social interaction: the role of the anterior paracingulate cortex. J Cogn Neurosci 16:1854–1863.

Wolf I, Dziobek I, Heekeren HR (2010) Neural correlates of social cognition in naturalistic settings: a model-free analysis approach. Neuroimage 49:894–904.

Wunderlich K, Symmonds M, Bossaerts P, Dolan RJ (2011) Hedging your bets by learning reward correlations in the human brain. Neuron 71:1141–1152.

Yoshida W, Seymour B, Friston KJ, Dolan RJ (2010) Neural mechanisms of belief inference during cooperative games. J Neurosci 30:10744–10751.

Zaki J, Schirmer J, Mitchell JP (2011) Social influence modulates the neural computation of value. Psychol Sci 22:894–900.

## Figure legends

**Figure 1.** Task Design – Receiving Social Feedback from Peers after a Real Life Interaction

(A) Participants came to the laboratory in groups of five on two consecutive days. On the first day they got to know each other by playing the board game “monopoly” for 1h 15min. Afterwards, each person rated three of the other players on 40 positive and 40 negative trait adjectives on a Likert scale from 1 (this trait does not apply to the person at all) to 8 (this trait applies to the person very much). On the first day participants did not rate themselves (yellow) and did not rate one of the other players (green). See **Table 1** for a list of the trait adjectives.

(B) On the second day participants performed the following task in the fMRI scanner. They first saw a cue indicating whether the following trial was about themselves or about the other person whom they had not rated on the previous day. They then saw one of 40 positive or 40 negative trait adjectives and had to imagine how much the trait applied to themselves or to the other person. They first gave their own rating and then saw the feedback in form of the mean rating they believed three other participants had given on the previous day. The absolute difference between participants’ own ratings and the feedback ratings they received was conceptualized as feedback discrepancies and manipulated during the experiment. Outside the scanner participants rated themselves and the other player a second time so that we could assess how much they updated their ratings after receiving feedback.

(C) For the main behavioral analyses we employed a 2 by 2 design with the factors feedback target (self/other) and feedback desirability (desirable/undesirable). Desirable feedback was defined as feedback ratings that were higher than participants' own first ratings (e.g. own first rating for "polite" was 6 and feedback rating was 8.0). Conversely, undesirable feedback was defined as feedback ratings lower than participants' first ratings (e.g. own first rating for "polite" was 6 and feedback rating was 4.0). All ratings for negative trait adjectives were reverse-coded. Thus, feedback desirability was independent of the valence of the trait adjective.

**Figure 2.** Positively Biased Updating

(A) Mean first and second ratings for self were significantly higher than for other. Second ratings were significantly higher than first ratings.

(B) Participants changed their ratings more after receiving desirable than after receiving undesirable feedback both for self- and other-related feedback. Trials were split into four conditions (self-desirable, self-undesirable, other-desirable, other-undesirable). For each condition we calculated the mean update (i.e., the mean difference between second and first ratings). Mean updates were positive for desirable feedback (indicating an increase in ratings) and negative for undesirable feedback (indicating a decrease in ratings). To test whether participants updated their ratings more towards desirable than towards undesirable feedback we calculated absolute mean updates (i.e., we compared the magnitude of the increase for desirable feedback with the magnitude of the decrease for undesirable feedback). Additionally, we scaled absolute mean updates by the respective mean feedback discrepancies for each condition and participant. The resulting relative updates indicate by how much participants changed their ratings with respect to the difference between initial own ratings and feedback ratings.

Error bars indicate SEM.

**Figure 3.** BOLD Signals Related to the Rewarding Component of Social Feedback

(A) BOLD signal changes in bilateral ventral striatum and ACC/MPFC correlated with the rewarding component of feedback on a trial-by-trial basis at the time-point of feedback (all clusters are significant at $p < 0.05$ corrected for multiple comparisons at a cluster-defining threshold of $p < 0.0001$). Reward-related activity fulfilled two requirements. First, activity correlated positively with feedback ratings for self since higher feedback ratings for self indicated more rewarding feedback (e.g. a feedback rating of 8.0 on "polite" is more rewarding than a feedback rating of 7.0; feedback ratings for negative trait adjectives were reverse-coded). Second, activity correlated more with the feedback ratings for self than for other since we searched for regions in which reward-related activity was self-specific.

(B) For illustration purposes, we plotted parameter estimates of the parametric modulators for feedback ratings for self and other within functional ROIs.

(C) To explore differences in onset activity, we plotted parameter estimates of the onset regressors within the functional ROIs in a second model that included separate regressors for feedback target and feedback desirability.

Error bars indicate SEM.

**Figure 4.** BOLD Signals Related to the Comparison Component of Social Feedback

(A) BOLD signal changes in the following regions correlated with the comparison-related component of feedback on a trial-by-trial basis at the time-point of feedback: MPFC, preSMA/SMA, bilateral IFG (orbital part) extending into anterior insula, right STS, right TPJ and left TP(all clusters are significant at $p < 0.05$ corrected for multiple comparisons at a cluster-defining threshold of $p < 0.0001$). Comparison-related activity correlated positively with the feedback discrepancies for both self and other, i.e., with the absolute difference between participants' own views and the feedback they received.

(B) For illustration purposes, we plotted parameter estimates of the parametric modulators for feedback discrepancies for self and other within functional ROIs.

(C) To explore differences in onset activity, we plotted parameter estimates of the onset regressors within the functional ROIs in a second model that included separate regressors for feedback target and feedback desirability.

Error bars indicate SEM.

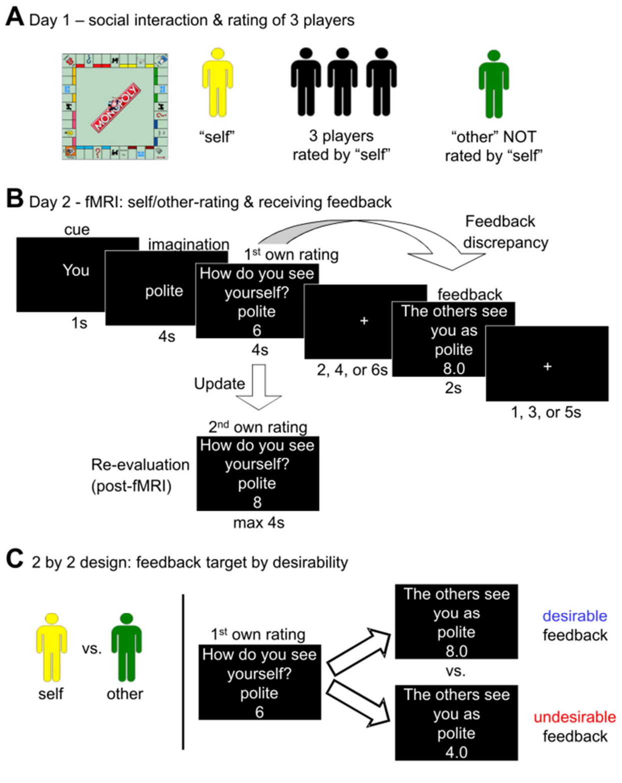
**Figure 5.** Updating Bias for Self and Activity Integrating Reward- and Comparison-Components

(A) Conjunction analysis of activity correlating with the rewarding aspect of feedback, i.e., feedback ratings (see **Fig. 4**) and of activity correlating with the comparison-related aspect of feedback, i.e., feedback discrepancies (**see Fig. 5**;$p < 0.05$ corrected for multiple comparisons at a cluster-defining threshold of $p < 0.0001$).

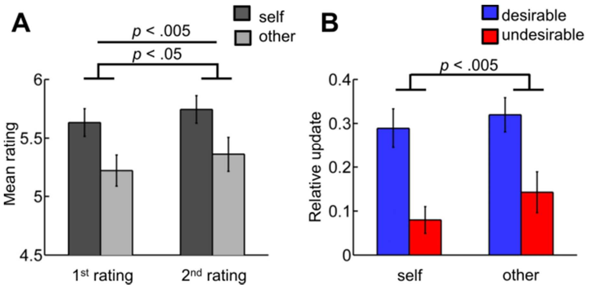
(B) Across participants, parameter estimates of the parametric modulators for self-related absolute feedback discrepancies within this region predicted the behavioral update bias, i.e., the relative mean update for self-related desirable minus self-related undesirable feedback. Each dot represents one participant and the line is the regressions slope.

**Figures**

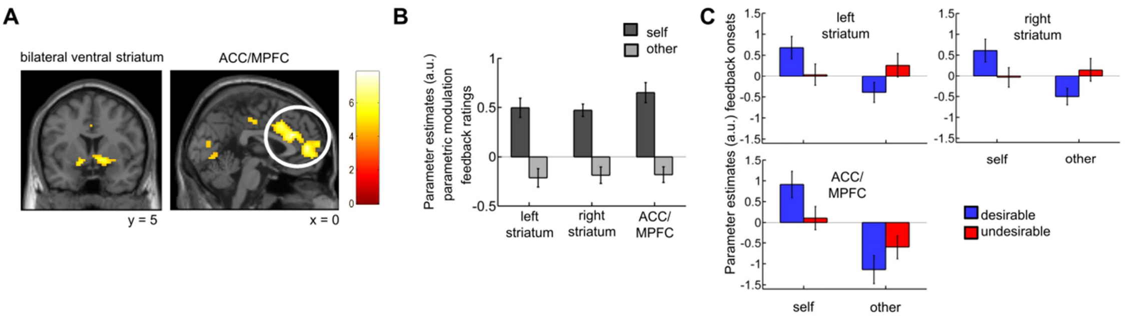
**Figure 1**


A Day 1 – social interaction & rating of 3 players
"self"
3 players
rated by "self"
"other" NOT
rated by "self"
B Day 2 - fMRI: self/other-rating & receiving feedback
cue
You
1s
imagination
polite
4s
1st own rating
How do you see yourself? polite 6
4s
Feedback discrepancy
+
2, 4, or 6s
feedback
The others see you as polite 8.0
2s
+
1, 3, or 5s
Update
2nd own rating
Re-evaluation (post-fMRI)
How do you see yourself? polite 8
max 4s
C 2 by 2 design: feedback target by desirability
self
vs.
other
1st own rating
How do you see yourself? polite 6
The others see you as polite 8.0
vs.
The others see you as polite 4.0
desirable feedback
undesirable feedback

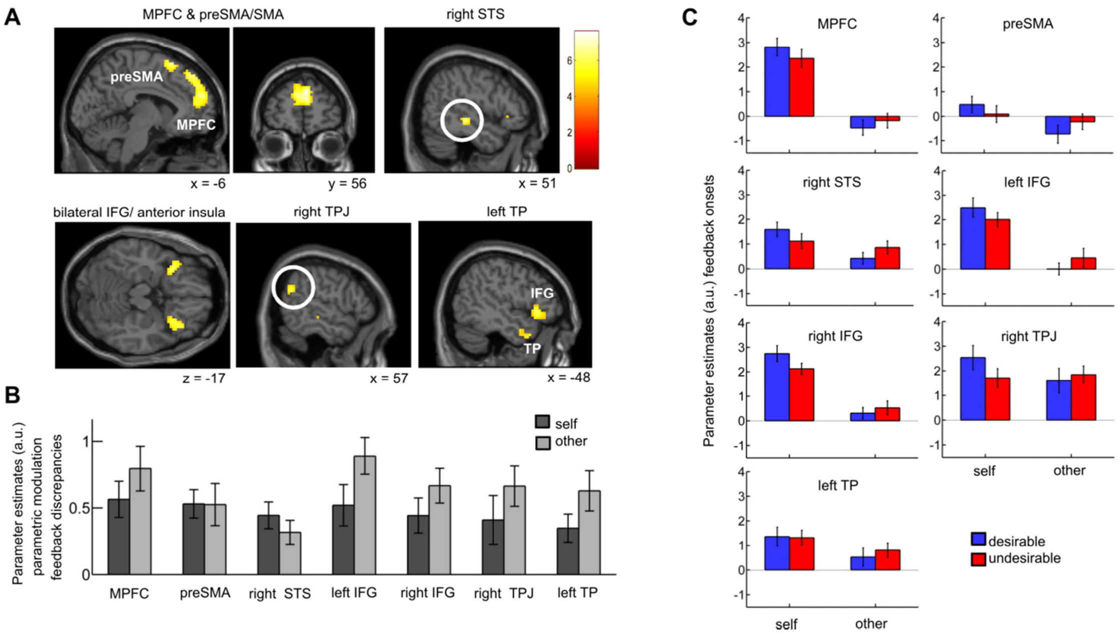
## Figure 2



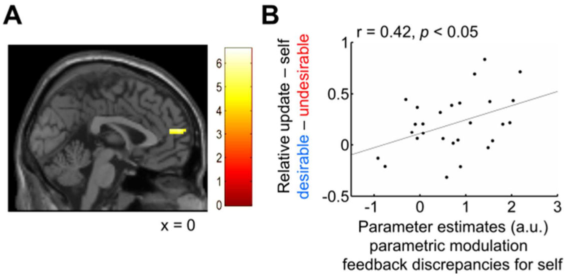
## Figure 3



## Figure 4



## Figure 5



## Tables

### Table 1. List of Trait Adjectives

| German original | English translation |
|---|---|
| Positive trait adjectives | |
| aufrichtig | honest |

| | |
|---|---|
| bescheiden | modest |
| diszipliniert | organized |
| effizient | efficient |
| einfühlsam | empathetic |
| enthusiastisch | enthusiastic |
| fleißig | hard-working |
| freundlich | friendly |
| geistesgegenwärtig | quick-witted |
| gelassen | composed |
| geschickt | skilled |
| gesellig | sociable |
| großzügig | generous |
| hilfsbereit | helpful |
| höflich | polite |
| kompetent | competent |
| kooperativ | cooperative |
| kreativ | creative |
| lebenslustig | fun-loving |
| locker | easy-going |
| loyal | loyal |
| offen | open-minded |
| ordentlich | tidy |
| respektvoll | respectful |
| scharfsinnig | astute |
| schlagfertig | articulate |
| selbstständig | self-reliant |
| sorgfältig | diligent |
| souverän | confident |
| spontan | spontaneous |

| | |
|---|---|
| tatkräftig | dynamic |
| tolerant | tolerant |
| vernünftig | level-headed |
| verständnisvoll | understanding |
| vertrauenswürdig | trustworthy |
| vielseitig | versatile |
| weitsichtig | perspicacious |
| wissbegierig | inquisitive |
| zielstrebig | goal-oriented |
| zuverlässig | reliable |
| Negative trait adjectives | |
| aggressiv | aggressive |
| ängstlich | anxious |
| arrogant | arrogant |
| bieder | overly-conservative |
| chaotisch | chaotic |
| egoistisch | selfish |
| eitel | conceited |
| engstirnig | narrow-minded |
| feige | cowardly |
| gehässig | spiteful |
| großmäulig | loud-mouthed |
| heuchlerisch | two-faced |
| hinterhältig | conniving |
| humorlos | humorless |
| inkonsequent | inconsistent |
| kalt | cold-hearted |
| launisch | moody |

| | |
|---|---|
| leichtsinnig | foolhardy |
| nachtragend | unforgiving |
| naiv | naive |
| oberflächlich | superficial |
| opportunistisch | opportunistic |
| pedantisch | pedantic |
| rücksichtslos | inconsiderate |
| scheu | unassertive |
| stur | stubborn |
| träge | lazy |
| unentschlossen | indecisive |
| ungeduldig | impatient |
| unnahbar | inapproachable |
| unpünktlich | tardy |
| unsicher | insecure |
| unsympathisch | unpleasant |
| verschwenderisch | wasteful |
| voreilig | rash |
| voreingenommen | biased |
| wehleidig | whiny |
| zickig | catty |
| zwanghaft | obsessive |
| zynisch | cynical |
| Adjectives used during the training session | |
| intelligent | intelligent |
| unsportlich | unathletic |

**Table 2.** Significant Activations in Feedback Onsets (all reported clusters are FWE-corrected for multiple comparisons at $p < 0.05$; cluster-defining threshold of $p < 0.0001$)

| | Side | Brodmann area | Peak voxel MNI coordinates (mm) | | | Cluster size (Voxels at *p* < 0.0001) | *p*(cluster FWE corrected) | Peak z score |
|---|---|---|---|---|---|---|---|---|
| | | | x | y | z | | | |
| Feedback onset: self > other | | | | | | | | |
| MPFC | L/R | 10/9/8/6/32/24 | -3 | 59 | 28 | 1602 | < 0.001 | 7.60 |
| IFG (orbital part)/ anterior insula | L | 47/11/13/45/38 | -33 | 17 | -17 | 399 | < 0.001 | 7.25 |
| IFG (orbital part)/ anterior insula | R | 47/11/13/38 | 30 | 20 | -17 | 335 | < 0.001 | 7.18 |
| Cerebellum | R | - | 30 | -82 | -35 | 161 | < 0.001 | 5.98 |
| Cerebellum | L | - | -30 | -85 | -38 | 77 | < 0.001 | 5.41 |
| Midbrain | L/R | - | -12 | -13 | -14 | 381 | < 0.001 | 5.38 |
| Cerebellum | L/R | - | 3 | -55 | -35 | 30 | 0.017 | 4.71 |
| Caudate body | L | - | -9 | 8 | 16 | 58 | 0.002 | 4.45 |
| Feedback onset: other > self | | | | | | | | |
| Precuneus/ postcentral gyrus/ superior temporal gyrus/ supramarginal gyrus | L/R | 7/6/4/1/2/3/5/18/22/40 | 12 | -46 | 52 | 7102 | < 0.001 | 6.89 |
| Middle temporal gyrus | R | 38 | 51 | -64 | 10 | 82 | < 0.001 | 4.86 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Precentral gyrus | R | 4 | 39 | -10 | 58 | 172 | < 0.001 | 4.67 |
| Middle frontal gyrus | R | 9 | 27 | 29 | 40 | 39 | 0.007 | 4.42 |
| Middle frontal gyrus | L | 9 | -30 | 35 | 25 | 27 | 0.023 | 4.19 |
| Middle frontal gyrus | L | 10 | -36 | 50 | 13 | 20 | 0.047 | 4.17 |

**Table 3.** BOLD Signals Related to the Rewarding Component of Social Feedback: Parametric Analysis – Feedback Ratings (all reported clusters are FWE-corrected for multiple comparisons at $p < 0.05$; cluster-defining threshold of $p < 0.0001$)

| | Side | Brodmann area | Peak voxel MNI coordinates (mm) | | | Cluster size (Voxels at *p* < 0.0001) | *p*(cluster FWE corrected) | Peak z score |
|---|---|---|---|---|---|---|---|---|
| | | | x | y | z | | | |
| Feedback rating (trial-by-trial correlation): self > other | | | | | | | | |
| ACC/ mid-cingulate cortex/MPFC | L/R | 32/24/ 9/10 | 3 | 32 | 25 | 414 | < 0.001 | 5.56 |
| Ventral striatum (caudate head and putamen) | R | - | 12 | 5 | -8 | 71 | < 0.001 | 4.73 |
| Thalamus | R | - | 21 | -13 | 22 | 20 | 0.032 | 4.60 |
| Ventral striatum (caudate head and putamen) | L | - | -15 | 2 | -11 | 25 | 0.017 | 4.50 |
| Cerebellum | L | - | -33 | -73 | -23 | 57 | 0.001 | 4.48 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Cerebellum | R | - | 39 | -58 | -26 | 39 | 0.004 | 4.35 |
| Cerebellum | R | - | 12 | -61 | -17 | 47 | 0.002 | 4.33 |
| Lingual gyrus | L | 18 | -3 | -73 | -5 | 26 | 0.015 | 4.31 |
| Calcarine fissure | L/R | 18 | 3 | -82 | 13 | 29 | 0.011 | 4.13 |
| Mid-cingulate cortex | L/R | 24 | 0 | -19 | 43 | 25 | 0.017 | 4.06 |

**Table 4.** BOLD Signals Related to the Comparison Component of Social Feedback: Parametric Analyses – Feedback Discrepancies and Conjunction (all reported clusters are FWE-corrected for multiple comparisons at *p* < 0.05; cluster-defining threshold of *p* < 0.0001)

| | Side | Brodmann area | Peak voxel MNI coordinates (mm) | | | Cluster size (Voxels at *p* < 0.0001) | *p*(cluster FWE corrected) | Peak z score |
|---|---|---|---|---|---|---|---|---|
| | | | x | y | z | | | |
| Feedback discrepancies (positive trial-by-trial correlation): self and other | | | | | | | | |
| MPFC | L/R | 10/9/8/6 | 6 | 56 | 28 | 383 | < 0.001 | 5.47 |
| preSMA/SMA | L/R | 8/6 | 9 | 17 | 64 | 104 | < 0.001 | 4.98 |
| Superior/ middle temporal gyrus – STS | R | 21 | 51 | -25 | -8 | 26 | < 0.001 | 4.93 |
| IFG (orbital part)/ anterior insula | L | 47/45/13 | -36 | 20 | -23 | 181 | < 0.001 | 4.88 |
| IFG (orbital part)/anterior insula | R | 47/13/11 | 33 | 20 | -17 | 117 | < 0.001 | 4.69 |
| Angular gyrus – TPJ | R | 39/40 | 57 | -58 | 25 | 19 | 0.045 | 4.56 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| TP | L | 21/38 | -48 | 11 | -35 | 22 | 0.032 | 4.25 |
| Cerebellum | L | - | -24 | -82 | -35 | 20 | 0.040 | 4.19 |
| Feedback discrepancies (negative trial-by-trial correlation): self and other | | | | | | | | |
| Inferior parietal lobule | L | 40 | -54 | -37 | 46 | 154 | < 0.001 | 5.07 |
| Middle temporal gyrus | R | 21/37 | 60 | -49 | -8 | 53 | 0.002 | 4.54 |
| Superior parietal gyrus | L | 7 | -21 | -49 | 64 | 30 | 0.013 | 4.50 |
| Superior temporal gyrus | L | 22/6 | -54 | -10 | 1 | 37 | 0.007 | 4.40 |
| Inferior parietal lobule | R | 40 | 51 | -37 | 46 | 48 | 0.002 | 4.34 |
| Precentral gyrus/ superior temporal gyrus | R | 6/22 | 54 | 5 | 13 | 30 | 0.013 | 4.19 |
| Conjunction of feedback rating (trial-by-trial correlation): self > other with feedback discrepancies (positive trial-by-trial correlation): self and other | | | | | | | | |
| MPFC/ACC | L/R | 10 | 3 | 56 | 19 | 25 | 0.023 | 5.01 |

## Cultural influences on social feedback processing of character traits

Christoph W. Korn[1,2,3], Yan Fan[1,3,4,5], Kai Zhang[1,3,4,5], Chenbo Wang[6], Shihui Han[6], and Hauke R. Heekeren[1,2,3,4]

[1]Department of Education and Psychology, Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany

[2]Berlin School of Mind and Brain, Humboldt Universität zu Berlin, Luisenstrasse 56, 10117 Berlin, Germany

[3]Dahlem Institute for Neuroimaging of Emotion, Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany

[4]Cluster of Excellence "Languages of Emotion", Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany

[5]Department of Psychiatry, Charité-Universitätsmedizin Berlin, Campus Benjamin Franklin, Hindenburgdamm 30, 12203 Berlin, Germany

[6]Department of Psychology, Peking University, 5 Yiheyuan Road, 100871 Beijing, China

Correspondence should be addressed to:

Christoph W. Korn

Habelschwerdter Allee 45, 14195 Berlin, Germany

Phone: 00493083856226

Fax: 00493083855778

E-mail: christoph.w.korn@gmail.de

Running Title: Cultural influences on social feedback

Number of words: 52

Number of figures: 4

Number of tables: 4

**Conflict of interest**

None.

## Funding

This work was supported by the Excellence Initiative of the German Federal Ministry of Education and Research, Deutsche Forschungsgemeinschaft, Grants GSC86/1-2009 and EXC 302.

## Acknowledgements

We thank G. Müller and Y. Müller-Chiu for translation.

## Abstract

Cultural differences are generally explained by how people see themselves in relation to social interaction partners. While Western culture emphasizes independence, East Asian culture emphasizes interdependence. Despite this focus on social interactions, it remains elusive how people from different cultures process self-relevant feedback from interaction partners. Here, participants of either German or Chinese origin engaged in a face-to-face interaction. Consequently, they updated their ratings of 80 character traits (e.g., polite, pedantic) after receiving feedback from their interaction partners. To exclude potential confounds, we obtained data from German and Chinese participants in Berlin (functional magnetic resonance imaging) and in Beijing (behavior). We tested cultural influences on social conformity, self-enhancement, and self-related neural activity. First, Chinese conformed more to social feedback than Germans (i.e., Chinese updated their trait ratings more). Second, regardless of culture, participants self-enhanced by processing feedback in a positively biased way (i.e., they updated more toward desirable than toward undesirable feedback). Third, changes in self-related medial prefrontal cortex activity were greater in Germans than in Chinese during feedback processing. By relating conformity, self-enhancement, and self-related activity to feedback obtained in a real-life interaction, we provide an essential step towards a unifying framework for understanding the diversity of human culture.

## Keywords

interdependence, independence, social conformity, self-enhancement, medial prefrontal cortex

## Introduction

Culture shapes various aspects of human cognition (Nisbett et al., 2001; Nisbett and Masuda, 2003; Henrich et al., 2010; Heine, 2012). A prominent framework that integrates diverse cultural differences centers on how people relate to those with whom they interact (Markus and Kitayama, 1991; Singelis, 1994; Oyserman et al., 2002; Triandis and Suh, 2002; Markus and Kitayama, 2010): Members of independent (or individualistic) cultures (e.g., Western cultures) construe their selves as distinct from others whereas members of interdependent (or collectivistic) cultures (e.g., East Asian cultures) construe their selves as interconnected with close others. Cultural differences in self-construal have been linked to social conformity (Bond and Smith, 1996), self-enhancement (Heine and Buchtel, 2009), as well as self- and other-related neural activity (Zhu et al., 2007; Han and Northoff, 2008; Ng et al., 2010; Ma et al., 2012). Yet, how culture influences the processing of self-relevant feedback from others has not been investigated—which is surprising given that the relation between self and others defines self-construal. Here, we compared how members of an independent culture (Germans) and from an interdependent culture (Chinese) process social feedback on character traits.

When people receive social feedback on character traits they conform their own view of themselves to feedback provided by others. In independent cultures, the concept of conformity has a negative connotation whereas in interdependent cultures it has a positive connotation (Kim and Markus, 1999). Meta-analytic evidence indicates that members of interdependent cultures conform more to social information in classic Asch-type line judgment tasks (Bond and Smith, 1996). However, the authors of the meta-analysis concede that line judgment tasks limit the concept of conformity to cases where participants can only conform—or not—to objectively incorrect statements about unambiguous physical stimuli (i.e., the lengths of two lines). Conforming to feedback on character traits (i.e., to information that is open to interpretation) differs from conforming to statements about physical stimuli, and directly relates to the concept of self-construal.

Greater conformity to desirable versus undesirable self-relevant feedback can be conceptualized as a self-enhancing tendency. We have previously shown that Germans engage in such positively biased social feedback processing (Korn et al., 2012), which fits with research indicating that Westerners show self-enhancement (Taylor and Brown, 1988). However, it is debated whether East Asians self-enhance (Sedikides et al., 2003; Sedikides et al., 2007) or not (Heine et al., 2007; Heine and Hamamura, 2007). We aimed at adding to this debate by using feedback processing as a novel approach to assess possible cultural differences in self-enhancement.

Self-related processes have also been in the focus of cultural neuroscience (Vogeley and Roepstorff, 2009; Kitayama and Uskul, 2011; Han et al., 2013). Differences in interdependent and independent self-construal have been linked to the anterior cingulate cortex and the medial prefrontal cortex (ACC/MPFC) (Zhu et al., 2007; Chiao et al., 2009a; Chiao et al., 2009b; Ng et al., 2010; Ma et al., 2012), which play an important role in various aspects of social cognition such as judging character traits (Heatherton, 2011; Denny et al., 2012; Wagner et al., 2012) and engaging in mentalizing or theory-of-mind (i.e., inferring the mental states of others) (Mar, 2011; Frith and Frith, 2012). For example, MPFC activity was higher in Westerners compared with East Asians when they judged whether character traits, social roles or physical attributes were self-descriptive (Ma et al., 2012). However, these previous studies have not addressed how MPFC activity is modulated by self-relevant information.

## Materials and Methods

### Participants

In line with common practice in research on cultural comparisons we used nationality as a proxy for cultural group membership but additionally assessed participants' explicit endorsement of independence and interdependence (Singelis,1994; Henrich et al., 2010; de Greck et al, 2012; Ma et al., 2012). Living in a foreign culture could increase conformity due to possible stress and insecurity (Sam and Berry, 2010; Heine, 2012) or in-group/out-group effects (Bond and Smith, 1996). It can also be expected that individuals who move abroad may be more independent (and less interdependent) in general (Kitayama et al., 2006; Kitayama et al., 2012) or that they may be more similar to their host culture. To minimize these potential confounds related to our first hypothesis, we obtained behavioral data from both German and Chinese participants in both Berlin and Beijing (**Table 1**). In Berlin, we collected functional magnetic resonance imaging (fMRI) data to test for cultural differences in ACC/MPFC activity. When tested, these Chinese participants had been sojourners in Germany for less than two years.

Specifically, we recruited participants of German and Chinese cultural origin in Berlin and Beijing via flyers, word-of-mouth, and mailing lists (e.g., by the German Academic Exchange Service; German Berlin: n°=°27, German Beijing: n°=°24, Chinese Berlin: n°=°28, Chinese Beijing: n°=°25) (**Table 1**). Our study employed a 2 (culture: German/Chinese) by 2 (place: Berlin/Beijing) between-subject design. The two groups in Berlin underwent fMRI scanning while the two groups in Beijing were tested behaviorally. Data from the German fMRI group have been reported previously (Korn et al., 2012). See **supplementary information** for **additional participant information**.

### Experiment overview

The experimental procedure has been adapted from our previous study (Korn et al., 2012) (**Figure 1**) and consisted of two sessions. We wanted participants to believe that they would get realistic feedback on their personality traits from peers of the same culture with whom they had interacted in real-life. During the first session (**Figure 1A**) each participant interacted with four other participants for 1°h and 15°min by playing a popular board game. Consequently, each participant rated three of the four other participants on 40 positive and 40 negative trait adjectives. See **supplementary information** for **Stimuli and translation**, **Social interaction and rating of 3 players (first session)**, and **Supplementary Table S1**. In the second session, participants believed that they would receive the mean rating of three other participants on each adjective as feedback. See **Figure 1B** and **Feedback task and re-evaluation task (second session)**. We tested how much participants took this feedback into account by asking them to rate their own personality before and after receiving social feedback (i.e., in the feedback and in the re-evaluation tasks). Additionally, each participant rated one other person before and after receiving social feedback for this person. Since participants in Berlin underwent fMRI scanning while receiving feedback, they performed the two sessions on two consecutive days. Participants in Beijing performed the two sessions on the same day.

### Feedback task and re-evaluation task (second session)

In the second session of the experiment (**Figure 1B**), participants performed the following feedback task, either in the MRI scanner (Berlin groups) or behaviorally on a PC (Beijing groups). The feedback task was presented using the MATLAB toolbox Cogent 2000 (www.vislab.ucl.ac.uk/cogent.php). On each trial, participants first saw a cue (1°s) indicating whether the trial was about themselves ("you") or about the fourth other participant  whom they had not rated during the first main part of the experiment (name of the other person). Then, they saw one of the 80 trait adjectives and had to think about how much that trait applied to themselves or to the other person (imagination phase, 4°s). Afterwards, participants were prompted to indicate their rating on an 8-point Likert scale via two button boxes with four buttons each (rating phase, 6°s). After a jittered fixation cross (2, 4, or 6°s) participants saw what they believed to be the mean rating of three other participants from the first session of the experiment (feedback phase, 2°s). This mean rating, which served as the feedback rating, was a number with one decimal, ranging from 1.0 to 8.0 in steps of 0.3. The feedback rating was determined by the program during the experiment to reliably create a sufficient number of trials in which participants received desirable and undesirable feedback (see **Task conditions and behavioral analyses**). After a second jittered fixation cross (1, 3, or 5°s) a new trial began. Participants performed 4 training trials. The feedback task was split up into four runs with the same 10 positive and the same 10 negative trait adjectives for self and other trials within one run. Trials for self and other were randomly intermixed. Adjectives were randomly assigned to the four blocks for each person.

After the feedback task, all participants performed the re-evaluation task outside the MRI scanner on a PC using the MATLAB toolbox Cogent 2000. Participants gave a second rating so that we could measure how much they changed their self- and other-ratings after having received social feedback in the feedback task. Specifically, they rated themselves and the other person again on all 80 trait adjectives in two separate blocks (one for themselves and one for the other person). These blocks were randomized for order. For each trait adjective participants had up to 6°s to respond. See **supplementary information** for details on a subsequent memory task.

### Task conditions and behavioral analyses

The main behavioral analyses employed a 2°by°2°by°2°by°2 design with the within-subject factors feedback target (self/other) and feedback desirability (desirable/undesirable) (**Figure 1C**) as well as the between-subject factors culture (German/Chinese) and place (Berlin/Beijing) (**Table 1**). For each participant, trials were classified according to whether feedback was desirable or undesirable. For a positive trait adjective, desirable feedback indicated that the feedback rating was numerically higher than the initial rating (e.g., a participant's initial rating for "polite" was 6 and the feedback rating was 8). For a negative trait, desirable feedback indicated that the original feedback rating was numerically lower than the original initial rating (e.g., a participant's initial rating for "aggressive" was 3 and the feedback rating was 1). Conversely, undesirable feedback was defined as feedback ratings that were more "negative" than participants' own initial ratings (e.g., initial rating of 6 and feedback rating of 4 for "polite" or initial rating of 3 and feedback rating of 5 for "aggressive"). Thus, feedback desirability was independent of the valence of the trait word and we reverse-coded ratings for negative trait adjectives. For each trial (i.e. for each trait adjective; separately for self- and other-conditions) we calculated a "feedback discrepancy" term as the absolute difference between first own ratings and feedback ratings. (Trials with adjectives for which participants failed to respond in time for the first or second rating were excluded.)

(1) feedback discrepancy°=°abs(feedback rating – first own rating)

This feedback discrepancy term indicated the social comparison component of receiving social feedback. Since feedback discrepancies were an independent variable of our task we manipulated their magnitude using a random number generator (see Korn et al., 2012 for details). Trials with a feedback discrepancy of zero were excluded from behavioral analyses since these trials could not be clearly assigned to either receiving desirable or receiving undesirable feedback (see **Table 2** for final numbers of trials). To assess how much participants changed their self-concept after receiving social feedback, we calculated an update term quantifying how much participants changed their own ratings.

(2) update°=°second own rating – first own rating

We expected participants to change their ratings on average towards the feedback ratings. That is, for desirable feedback (i.e., feedback ratings higher than own first rating) participants should increase their ratings (i.e., positive updates). For undesirable feedback (i.e., feedback ratings lower than own first rating) participants should decrease their ratings (i.e., negative updates). To test for differences in updating, we first calculated relative mean updates for each participant within each condition by dividing mean updates by the respective mean feedback discrepancies. We then took the absolute value of relative mean updates (i.e., the negative sign of updates following undesirable feedback is changed).

(3) relative absolute mean update°=°absolute mean update / mean feedback discrepancy

Relative absolute updates can be interpreted in a straightforward way; e.g., a relative update of 0.3 indicates that the change in ratings was on average 30% of the difference between initial own ratings and feedback ratings. Overall group differences in relative absolute mean updates indicate group differences in social conformity (i.e., increase in rating for desirable feedback numerically larger than decrease for undesirable feedback). Larger relative absolute mean updates for desirable versus desirable feedback indicate positively biased updating.

**Individual difference scores**

Participants completed the 24-item version of the Singelis self-construal scale (Singelis, 1994) and the Rosenberg self-esteem scale (Rosenberg, 1965). They rated how similar they perceived the other person on a Likert scale from 1 (not similar at all) to 8 (very similar).

**Analysis of fMRI data—overview**

FMRI data were acquired on a 3T scanner (Trio, Siemens, Erlangen, Germany) using standard parameters (see **supplementary information**). FMRI data were preprocessed using standard procedures in SPM8 (www.fil.ion.ucl.ac.uk/spm) (see **supplementary information**) and analyzed using hierarchical random-effects models as implemented in SPM. At the subject-specific first level, fMRI time series were regressed onto a general linear model (GLM) containing regressors which represented the time periods of the feedback task (**Figure 1B**): cue (1°s), imagination phase separately for self and other (4°s), rating phase (4°s), feedback phase separately for self and other (2°s), and two motor regressors for button presses with the left and the right hands (0°s). This resulted in 8 regressors for each of the four scanning runs. The imagination phase regressors for self and other were parametrically modulated by the respective first own ratings. The feedback phase regressors for self and other were modulated by the respective feedback ratings and the respective feedback discrepancies, to investigate the reward- and comparison-related feedback components, respectively (see **supplementary information** for **parametric modulation analyses**). The model included trials with feedback discrepancies of zero. The six motion correction parameters estimated from the realignment procedure were entered as covariates of no interest. Regressors were convolved with the canonical HRF and low frequency drifts were excluded using a high-pass filter with a 128°s cutoff. At the group level, we performed separate flexible factorial designs for the following conditions: feedback phase, imagination phase as well as parametric modulators for feedback ratings and feedback discrepancies. Specifically, we used flexible factorial designs including the following factors: a subject-specific constant, a group factor (culture: German/Chinese), and the interaction of group and condition (feedback target: self/other). All coordinates are reported in MNI space and activations are displayed on the standard MNI reference brain.

## Results

### Updating behavior

In our behavioral data we tested two main hypotheses: First, we tested whether Chinese took social feedback more into account than Germans by comparing relative absolute mean updates.

Second, we tested whether participants of both cultural groups showed positively biased processing of social feedback—operationalized as greater relative absolute mean updates toward desirable compared with undesirable feedback. We found evidence supporting both hypotheses by comparing relative absolute mean updates in a 2 (target: self/other) by 2 (desirability: desirable/undesirable) by 2 (culture: German/Chinese) by 2 (place: Berlin/Beijing) ANOVA (**Figure 2A** and **Table 2**). In support of our first hypothesis, we found a significant main effect of culture with Chinese showing higher updates than Germans (F(1,100)°=°4.64, p°=°0.034, $\eta_p^2$°=°0.04). In support of our second hypothesis, we found a significant main effect of feedback desirability with higher updates following desirable compared with undesirable feedback (F(1,100)°= °107.2, p°<°0.001, $\eta_p^2$°=°0.52). We acknowledge that the effect size of the within-subject test for positively biased updating was larger than the effect size of the between-subject test for cultural differences in updating. Additionally, there was a significant main effect of feedback target with updates for self being higher than for other (F(1,100)°=°4.07, p°=°0.046, $\eta_p^2$°=°0.04). No other main effects or interactions reached significance (all p°>°0.05). See **supplementary information** and **Supplementary Figure S1** for **additional behavioral results** on the direction of updates, trait valence, first ratings, perceived similarity scores, and memory errors.

### Individual differences and overall updating

To test whether updating correlated with individual variability in the endorsement of cultural values, we assessed participants' interdependence and independence scores (Singelis, 1994). The two scores did not correlate with each other (Pearson's r°=°-0.01, p°>°0.9) and therefore we analyzed them separately. As expected, Germans were less interdependent and more independent than Chinese (2 (culture) by 2 (place) ANOVAs: interdependence: F(1,100)°=°29.69, p°<°0.001, $\eta_p^2$°=°0.23; **Figure 2B**; independence: F(1,100)°= °14.44, p°<°0.001, $\eta_p^2$°=°0.13; **Figure 2C**). Participants in Beijing showed more interdependent self-construal compared with those in Berlin (F(1,100)°= °14.90, p°<°0.001, $\eta_p^2$°=°0.13). The interaction of culture and place was at trend level for both scores (interdependence: F(1,100)°=°3.20, p°=°0.077, $\eta_p^2$°=°0.03; independence: F(1,100)°=°2.88, p°=°0.093, $\eta_p^2$°=°0.03).

In addition to trait measures on interdependence and independence, we also collected participants' score on the Rosenberg self-esteem scale (Rosenberg, 1965). Germans showed higher self-esteem than Chinese (F(1,100)°=°5.24, p°=°0.024, $\eta_p^{2}$°=°0.05). Place had no effect on self-esteem (p°>°0.1). Self-esteem scores correlated with independence (r°=°0.32, p°<°0.001, 95% confidence interval (CI) [0.14, 0.49]) but not interdependence scores (r°=°0.02, p°>°0.8; the two correlations differed significantly as assessed by Hotelling's t; t(101)°=°2.23; p°=°0.028).

Interdependence scores correlated significantly with overall relative absolute mean updates (averaged across within-subject conditions; r°=°0.25, p°=°0.012, 95% CI [0.06, 0.42]; **Figure 2D**)—more interdependent participants showed higher updating. The strength of this correlation did not differ between Germans and Chinese as assessed by Fisher's z test (p°>°0.7). To test whether interdependence scores explained additional variance in updating beyond membership to the two cultural groups, we conducted a hierarchical regression on overall relative absolute mean updates including culture and interdependence scores as predictors. Interdependence scores explained additional variance at trend level ($F_{change}$(1,101)°=°3.06, p°=°0.083). The correlations between updating and independence and between updating and self-esteem were not significant (p°>°0.5). The relationship between interdependence and updating remained significant when accounting for independence and self-esteem in a hierarchical regression ($F_{change}$(1,100)°=°6.46, p°=°0.013).

Taken together, participants with more interdependent self-construal took social feedback more strongly into account. Participants in our sample and task showed positively biased feedback processing regardless of cultural group. The place where participants were tested had no effect on updating behavior.

### Cultural difference in neural feedback processing

Based on previous findings, which showed that culture influences ACC/MPFC activity during trait judgments (Zhu et al., 2007; Chiao et al., 2009a; Chiao et al., 2009b; Ng et al., 2010; Ma et al., 2012), we expected cultural differences in ACC/MPFC activity when participants received social feedback on personality traits.

We contrasted the time periods when participants received self- versus other-related feedback. In line with numerous studies on self-referential processing, we found changes in blood oxygen level-dependent (BOLD) signals in the medial prefrontal wall and bilateral inferior frontal gyrus (IFG) (p°<°0.05 family-wise error (FWE) corrected, cluster size°>°15; **Figure 3A**; see **Table 3** for a full list of activations). Consistent with our hypothesis, a cluster in the ACC/MPFC showed a culture (German/Chinese) by feedback target (self/other) interaction (p°<°0.05 small volume corrected within the main effect of self>other; initial threshold for interaction: p°<°0.001, uncorrected; **Figure 3B** and **Table 3**). Parameter estimates of this ACC/MPFC cluster are plotted in **Figure 3C** to visualize the interaction. Note that deactivations with respect to the implicit baseline are commonly observed in self- and other-referential activity (Amodio and Frith, 2006) and we do not draw any conclusions from the fact that in Chinese parameter estimates for self-related activity were around baseline whereas for Germans they were positive.

To relate cultural differences in ACC/MPFC activity to individual variability in self-construal, we correlated parameter estimates for self-related feedback with self-construal scores. We found a significant correlation with independence (r°=°0.36, p°=°0.007, 95% CI [0.10, 0.57]; **Figure 3D**) but not with interdependence (r°=°-0.06, p°>°0.6; the two correlations differed significantly as assessed by Hotelling's t; t(52)°=°2.30; p°=°0.02). Although parameter estimates for self-related feedback correlated with self-esteem (r°=°0.33 p°=°0.016, 95% CI [0.07, 0.54]), the relationship between parameter estimates and independence remained significant when accounting for self-esteem in a hierarchical regression ($F_{change}$(1,52)°= °4.58, p°=°0.037). At trend level, the strength of the correlation between independence and parameter estimates differed between Germans (r°=°0.05, p°>°0.7) and Chinese (r°=°0.53, p°=°0.004, 95% CI [0.20, 0.75]; Fisher's z°=°1.87; p°=°0.061).

In sum, our results extend previous findings by showing that culture influences ACC/MPFC activity during feedback processing. See **supplementary information** and **Supplementary Figure S2** for **additional fMRI results** regarding the imagination phase.

**Neural activity related to reward and social comparison**

We searched for BOLD signals that correlated with reward- and comparison-related components in a trial-by-trial fashion. Activity related to the rewarding component fulfilled two requirements: First, reward-related activity correlated positively with feedback ratings for self (i.e., higher feedback rating for self indicated more rewarding social feedback). Second, reward-related activity was self-specific (i.e., we performed a contrast between the parametric modulators for feedback ratings for self and other). Across all participants, the rewarding component was related to activity in the ACC/MPFC and bilateral ventral striatum; among other regions (p°<°0.05 FWE corrected, cluster size°>°15; **Figure 4A** and **Table 4**).

Activity related to the social comparison component was operationalized as activity that showed a positive trial-by-trial correlation with feedback discrepancies for self- and other-related feedback. That is, we searched for activity correlating with the absolute differences between participants' ratings and the feedback ratings they received (regardless of feedback desirability). We found comparison-related activity in MPFC, bilateral IFG extending into anterior insula, left temporal pole (TP), left temporo-parietal junction (TPJ), right superior temporal sulcus (STS), left cerebellum, and (pre-) supplementary motor area (preSMA/SMA; p°<°0.05 FWE corrected, cluster size°>°15; **Figure 4B** and **Table 4**).

### Cultural comparisons of parametric modulators for reward- and comparison-related components

We explored differences between Germans and Chinese in neural activity associated with reward- and comparison-related components. No voxels were significant in any interaction contrast involving the factor culture (p°<°0.05 FWE corrected, cluster size°>°15). Furthermore, no clusters in any interaction contrast survived small volume correction within the relevant main effect (p°<°0.05 small volume correction; initial contrast threshold for interaction: p°<°0.001, uncorrected). Furthermore, we performed follow-up ROI analyses within the regions identified for reward- and comparison-related components by extracting parameter estimates for self- and other-related parametric modulators. These ROI analyses fulfilled to purposes: First, extracted parameter estimates illustrate the correlations of feedback ratings and feedback discrepancies with BOLD signals (**Supplementary Figure S3A** and **S3B**). Second, extracted parameter estimates were used to test for cultural differences in 2 (culture: German/Chinese) by 2 (feedback target: self/other) ANOVAs. We found no significant main effects or interactions involving the factor culture in any of the ROIs (all p°>°0.1; p-values were adjusted using a Bonferroni correction for the number of ROIs; reward-related activity: 10 ROIs, comparison-related activity: 9 ROIs).

In sum, the rewarding component of social feedback was related to the ACC/MPFC and ventral striatum. The comparison component correlated with activity in the MPFC, IFG, TPJ, STS, and TP. We did not find evidence for a cultural modulation of activity associated with reward- and comparison-related components in our sample and task. See **supplementary information** for **additional fMRI results** regarding reward- and comparison-related activity.

## Discussion

Cultural practices are shaped by social interactions but in most studies in cultural psychology or neuroscience participants perform tasks in the solitude of a test cubicle or fMRI scanner. Our design involves a face-to-face interaction of five peers of the same culture. By investigating how receiving feedback from these peers challenged participants' self-concept, we provide a novel approach to test for cultural differences in social conformity, self-enhancement, and ACC/MPFC activity. We excluded confounds which might arise in the context of a real-life social interaction by testing both cultural groups in both countries.

We found that Chinese—compared with Germans—conformed more to social feedback on their own character traits and those of another person. Across both cultures more interdependent individuals showed higher conformity. When participants received social feedback, MPFC activity differed between Germans and Chinese and correlated with independence scores. However, cultural group membership did not influence positively biased feedback processing in our sample and task. Regardless of culture, participants changed their character trait ratings more toward desirable than toward undesirable feedback. The reward-related component correlated with activity in regions previously implicated in social and non-social reward processing (i.e., ACC/MPFC and ventral striatum) (Fehr and Camerer, 2007; Izuma et al., 2008; Beckmann et al., 2009). The comparison-related component correlated with activity in the mentalizing network (i.e., MPFC, TPJ, STS, IFG, and TP) (Mar, 2011; Frith and Frith, 2012). Culture did not influence reward- and comparison-related components.

Western culture emphasizes that individuals should view their own character traits independently from the opinion of others. In contrast, East Asian culture emphasizes that individuals are interconnected. This difference in cultural values has been used to explain why members of interdependent cultures show higher conformity when receiving social information about the lengths of two lines (Bond and Smith, 1996). Unlike objective physical properties, character traits are open to interpretation and directly self-relevant. Our results relate conformity about social feedback on character traits to differences in interdependence—both on the cultural and on the individual level. Our finding that Chinese changed their trait ratings more than Germans also fits well with the observation that East Asians perceive character traits as more malleable than Westerners (Choi et al., 1999) and with previous research on cultural differences in consensus motives (Fu et al., 2007).

In addition to culture, insecurity and information from an in-group (versus an out-group) can lead to higher conformity (Bond and Smith, 1996). Living in a foreign culture might trigger a general state of insecurity and meeting compatriots in a foreign country might create strong in-group feelings (Sam and Berry, 2010; Heine, 2012). Furthermore, individuals who move abroad tend to be more independent than those who stay in their home country (Kitayama et al., 2006; Kitayama et al., 2012) and may thus not be completely representative of their culture. For these reasons, we obtained behavioral data from both groups in both countries and could directly test for possible influences of the place where participants were tested. Our data did not provide any support that place modulated social conformity.

Our results on social feedback processing provide a novel approach to the extensive debate on whether East Asians do or do not show similar degrees of self-enhancement as Westerners (Sedikides et al., 2003; Heine et al., 2007; Heine and Hamamura, 2007; Sedikides et al., 2007). One of the main arguments centers on how self-enhancement should be measured. Many studies used trait measures (e.g., the Rosenberg self-esteem scale) or compared how participants evaluated themselves and an “imagined” person from a reference group (Heine and Hamamura, 2007). In a few studies, participants received feedback on their performance in a task (e.g., a creativity test) (Heine et al., 2001). Success or failure feedback was then related to persistence on the task. Here, we conceptualized self-enhancement as larger changes in character trait ratings toward desirable versus undesirable feedback. This operationalization confers the following advantages: First, since we analyzed how ratings change and not ratings per se, we reduced possible confounds arising from cultural differences in completing Likert scales (Heine, 2012). Second, participants received feedback from persons with whom they had face-to-face contact and did not have to compare themselves to an “imagined” other person (i.e., reference-group effects were excluded) (Heine, 2012). Third, our approach combines two aspects of previous studies on self-enhancement: self-evaluations and processing of positive versus negative feedback.

We found positively biased updating across both cultural groups in our sample of Germans and Chinese. Nevertheless, we replicate findings showing higher trait self-esteem in Westerners compared with East Asians (Heine and Hamamura, 2007). Thus, our findings suggest that culture affects self-enhancement operationalized as trait self-esteem but not self-enhancement operationalized as biased social feedback processing. Furthermore, our findings are in line with evidence showing that American and Chinese individuals sought similar degrees of self-enhancing and self-improving feedback (Gaertner et al., 2012). Future studies have to corroborate whether our findings extend to Americans and Japanese since studies reporting cultural differences in self-enhancement have mainly compared these two cultures (Heine and Hamamura, 2007).

Our design can be easily adapted to probe various cultural influences on feedback processing. Since close others (e.g., family members, friends, or colleagues) are especially important for interdependent individuals (Markus and Kitayama, 2010), cultural differences in positively biased feedback processing might emerge when feedback is given by close others and not by unrelated peers as in the present study. Furthermore, participants in our study showed positively biased updating when receiving feedback for themselves and when receiving feedback for another person, i.e., one of the peers from the social interaction. Changing the relationship between the self and the other person (e.g., by using an in-group/out-group manipulation) might alter feedback processing. In addition, since modesty has been related to cultural differences in self-enhancement (Cai et al., 2011), future studies should address whether modesty modulates positively biased updating.

Our fMRI results extend previous findings of cultural differences in ACC/MPFC activity during trait judgments (Zhu et al., 2007; Chiao et al., 2009a; Chiao et al., 2009b; Ng et al., 2010; Ma et al., 2012). In East Asians—but not in Westerners—MPFC activity for trait-judgments about self and mother overlapped (Zhu et al., 2007; Wang et al., 2012). The same pattern has been replicated with bicultural individuals from Hong Kong who were primed with Chinese or Western cultural symbols (Ng et al., 2010). General versus contextual trait judgments (e.g., “I am polite” vs. “I am polite when I talk to my mother”) activated the MPFC differently depending on participants’ self-construal (Chiao et al, 2009a)—a result replicated by priming independence or interdependence in bicultural Asian Americans (Chiao et al, 2009b). Importantly, in a recent study (Ma et al, 2012) with a similar sample size as ours self-related MPFC activity was higher in Westerners than in East Asians. Since cultural differences are conceptualized as differences in social interactions between the self and others, ACC/MPFC activity should be especially prominent when individuals receive social information about the self, which is what we found. Germans showed higher self-related ACC/MPFC activity than Chinese during social feedback processing. Individual differences in independence correlated with ACC/MPFC activity. There was a trend which suggested that the strength of the correlation between independence and ACC/MPFC activity might be more pronounced in Chinese than in Germans. Future studies should investigate whether the observed correlation might be higher for East Asians in general or for individuals who live abroad.

We analyzed interdependence and independence scores separately since they did not correlate with each other across participants. Interestingly, interdependence correlated with updating behavior but independence correlated with ACC/MPFC activity. This pattern suggests that the two commonly used trait measures of cultural differences in self-construal might differentially relate to cultural differences on an individual level.

In addition to self-related activity, we explored cultural differences of the reward- and comparison-related components of social feedback processing. We replicated our previous results (Korn et al., 2012). Across all participants the rewarding component of social feedback correlated with activity in ACC/MPFC and ventral striatum; both of which are implicated in reward processing (Fehr and Camerer, 2007; Izuma et al., 2008; Beckmann et al., 2009). The social comparison component correlated with activity in MPFC, IFG, TPJ, STS, and TP; regions previously related to mentalizing (Mar, 2011; Frith and Frith, 2012). Culture did not modulate activity associated with reward- and comparison-related components in our sample and task, suggesting that these components might be processed similarly by members of both cultures.

We acknowledge that obtaining fMRI data only from Chinese participants living in Berlin may have limited the ability to detect cultural differences. Thus, in accord with a previous study on Chinese living in the US, which did not find cultural modulation of self- and mother-related MPFC activity (Chen et al., 2013), our findings suggest that future studies should take a more dynamic approach and investigate longitudinal changes within individuals adapting to a foreign culture.

In conclusion, social interactions are highly complex and differ widely across cultures. By relating social conformity, self-enhancement, and self-related neural activity to the processing of social feedback obtained in a real-life interaction, we provide an essential step towards a unifying framework for understanding human culture.

**References**

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

Beckmann, M., Johansen-Berg, H., & Rushworth, M.F.S. (2009) Connectivity-based parcellation of human cingulate cortex and its relation to functional specialization. Journal of Neuroscience, 29, 1175–1190.

Bond, R., & Smith, P.B. (1996) Culture and conformity: a meta-analysis of studies using Asch's (1952b, 1956) line judgment task. Psychological Bulletin, 119, 111–137.

Cai, H., Sedikides, C., Gaertner, L., Wang, C., Carvallo, M., Xu, Y., et al. (2011) Tactical self-enhancement in china : Is modesty at the service of self-enhancement in East Asian culture? Social Psychological and Personality Science, 2, 59–64.

Chen, P.A., Wagner, D.D., Kelley, W.M., Powers, K.E., & Heatherton, T.F. (2013) Medial prefrontal cortex differentiates self from mother in Chinese: Evidence from self-motivated immigrants. Culture and Brain [in press].

Chiao, J.Y., Harada, T., Komeda, H., Li, Z., Mano, Y., Saito, D., et al. (2009a) Neural basis of individualistic and collectivistic views of self. Human Brain Mapping, 30, 2813–2820.

Chiao, J.Y., Harada, T., Komeda, H., Li, Z., Mano, Y., Saito, D., et al. (2009b) Dynamic cultural influences on neural representations of the self. Journal of Cognitive Neuroscience, 22, 1–11.

Choi, I., Nisbett, R.E., & Norenzayan, A. (1999) Causal attribution across cultures: variation and universality. Psychological Bulletin, 125, 47–63.

de Greck, M., Shi, Z., Wang, G., Zuo, X., Yang, X., Wang, X., et al. (2012) Culture modulates brain activity during empathy with anger. Neuroimage, 59, 2871–2882.

Denny, B.T., Kober, H., Wager, T.D., & Ochsner, K.N. (2012) A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. Journal of Cognitive Neuroscience, 24, 1742–1752.

Fehr, E., & Camerer, C.F. (2007) Social neuroeconomics: the neural circuitry of social preferences. Trends in Cognitive Sciences, 11, 419–427.

Frith, C.D., & Frith, U. (2012) Mechanisms of social cognition. Annual Review of Psychology, 63, 287–313.

Fu, J.H., Morris, M.W., Lee, S.L., Chao, M., Chiu, C.Y., & Hong, Y.Y. (2007) Epistemic motives and cultural conformity: need for closure, culture, and context as determinants of conflict judgments. Journal of Personality and Social Psychology, 92, 191–207.

Gaertner, L., Sedikides, C., & Cai, H. (2012) Wanting to be great and better but not average: on the pancultural desire for self-enhancing and self-improving feedback. Journal of Cross-Cultural Psychology, 43, 521–526.

Han, S., & Northoff, G. (2008) Culture-sensitive neural substrates of human cognition: a transcultural neuroimaging approach. Nature Reviews Neuroscience, 6, 646–654.

Han, S., Northoff, G., Vogeley, K., Wexler, B.E., Kitayama, S., & Varnum, M.E. (2013) A cultural neuroscience approach to the biosocial nature of the human brain. Annual Review of Psychology, 64, 12.1–12.25.

Heatherton, T.F. (2011) Neuroscience of self and self-regulation. Annual Review of Psychology, 62, 363– 390.

Heine, S.J. (2012) Cultural Psychology. New York, NY: W.W. Norton & Company.

Heine, S.J., & Buchtel, E.E. (2009) Personality: the universal and the culturally specific. Annual Review of Psychology, 60, 369–394.

Heine, S.J., & Hamamura, T. (2007) In search of East Asian self-enhancement. Personality and Social Psychology Review, 11, 4–27.

Heine, S.J., Kitayama, S., & Hamamura, T. (2007) Which studies test whether self-enhancement is pancultural? Reply to Sedikides, Gaertner, and Vevea, 2007. Asian Journal of Social Psychology, 10, 198–200.

Heine, S.J., Lehman, D.R., Ide, E., Leung, C., Kitayama, S., Takata, T., et al. (2001) Divergent consequences of success and failure in Japan and North America: an investigation of self-improving motivations and malleable selves. Journal of Personality and Social Psychology, 81, 559–615.

Henrich, J., Heine, S.J., & Norenzayan, A. (2010) The weirdest people in the world? Behavioral and Brain Sciences, 33, 61–83.

Izuma, K., Saito, D.N., & Sadato, N. (2008) Processing of social and monetary rewards in the human striatum. Neuron, 58, 284–294.

Kim, H., & Markus, H.R. (1999) Deviance or uniqueness, harmony or conformity? A cultural analysis. Journal of Personality and Social Psychology, 77, 785–800.

Kitayama, S., Ishii, K., Imada, T., Takemura, K., & Ramaswamy, J. (2006) Voluntary settlement and the spirit of independence: evidence from Japan's "Northern frontier". Journal of Personality and Social Psychology, 91, 369–384.

Kitayama, S., & Uskul, A.K. (2011) Culture, mind, and the brain: current evidence and future directions. Annual Review of Psychology, 62, 419–449.

Kitayama, S., Varnum, M.E.W., & Sevincer, A.T. (2012). The frontier: voluntary settlement and cultural change. In: A. Cohen (ed). New Directions in the Psychology of Culture: Washington, DC: APA Books.

Korn, C.W., Prehn, K., Park, S.Q., Walter, H., Heekeren, H.R. (2012) Positively biased processing of self-relevant social feedback. Journal of Neuroscience, 21, 16832–16844.

Ma, Y., Bang, D., Wang, C., Allen, M., Frith, C., Roepstorff, A., et al. (2012) Sociocultural patterning of neural activity during self-reflection. Social Cognitive and Affective Neuroscience [Epub ahead of print].

Mar, R.A. (2011) The neural bases of social cognition and story comprehension. Annual Review of Psychology, 62, 103–34.

Markus, H.R., & Kitayama, S. (1991) Culture and the self: implications for cognition, emotion, and motivation. Psychological Review, 88, 224–253.

Markus, H.R., & Kitayama, S. (2010) Cultures and selves: a cycle of mutual constitution. Perspectives in Psychological Sciences, 5, 420–430.

Nisbett, R.E., & Masuda, T. (2003) Culture and point of view. Proceedings of the National Academy of Sciences U S A, 100, 11163–11170.

Nisbett, R.E., Peng, K., Choi, I., & Norenzayan, A. (2001) Culture and systems of thought: holistic versus analytic cognition. Psychological Review, 108, 291–310.

Ng, S.H., Han, S., Mao, L., & Lai, J.C.L. (2010) Dynamic bicultural brains: a fMRI study of their flexible neural representation of self and significant others in response to culture priming. Asian Journal of Social Psychology, 13, 83–91.

Oyserman, D., Coon, H.M., & Kemmelmeier, M. (2002) Rethinking individualism and collectivism: evaluation of theoretical assumptions and meta-analyses. Psychological Bulletin, 128, 3–72.

Rosenberg, M. (1965) Society and the adolescent self-image. Princeton, NJ: Princeton University Press.

Sam, D.L., & Berry, J.W. (2010) Acculturation: when individuals and groups of different cultural backgrounds meet. Perspectives in Psychological Sciences, 5, 472–481.

Sedikides, C., Gaertner, L., & Toguchi, Y. (2003) Pancultural self-enhancement. Journal of Personality and Social Psychology, 84, 60–70.

Sedikides, C., Gaertner, L., & Vevea, J.L. (2007) Inclusion of theory-relevant moderators yield the same conclusions as Sedikides, Gaertner, and Vevea (2005): A meta-analytical reply to Heine, Kitayama, and Hamamura (2007). Asian Journal of Social Psychology, 2, 59–67.

Singelis, T.M. (1994) The Measurement of independent and interdependent self-construals. Personality and Social Psychology Bulletin, 20, 580–591.

Taylor, S.E., & Brown, J.D. (1988) Illusion and well-being: a social psychological perspective on mental health. Psychological Bulletin, 103, 193–210.

Triandis, H.C., & Suh, E.M. (2002) Cultural influences on personality. Annual Review of Psychology, 53, 133–160.

Vogeley, K., & Roepstorff, A. (2009) Contextualising culture and social cognition. Trends in Cognitive Sciences, 13, 511–516.

Wagner, D.D., Haxby, J.V., & Heatherton, T.F. (2012) The representation of self and person knowledge in the medial prefrontal cortex. Wiley Interdisciplinary Reviews Cognitive Science, 3, 451–470.

Wang, G., Mao, L., Ma, Y., Yang, X., Cao, J., Liu, X., et al. (2012) Neural representations of close others in collectivistic brains. Social Cognitive and Affective Neuroscience, 7, 222–229.

Zhu, Y., Zhang, L., Fan, J., Han, S. (2007) Neural basis of cultural influence on self representation. Neuroimage, 34, 1310–1317.

## Figure legends

**Figure 1. Experimental design—receiving social feedback from peers after a real-life interaction**

(A) Participants came to the laboratory in groups of either five German or five Chinese participants. In the first session of the experiment, participants got to know each other by playing the board game “monopoly” for 1 h 15 min. Afterwards, each person rated three of the other players on 40 positive and 40 negative trait adjectives (**Supplementary Table S1**) on a Likert scale from 1 (this trait does not apply to the person at all) to 8 (this trait applies to the person very much). Participants did not rate themselves (yellow) and did not rate one of the other players (green).

(B) In the second session, participants in Berlin performed the feedback task in the fMRI scanner and participants in Beijing performed the feedback task on a PC. In each trial, participants first saw a cue indicating whether the trial was about themselves or about the other person whom they had not rated during the first session. They had to imagine how much one of the 80 traits applied to themselves or to the other person. They first gave their own rating and then saw the feedback rating in form of the mean rating they believed three other participants had given during the first session. The absolute differences between participants’ own ratings and the feedback ratings they received was conceptualized as feedback discrepancies and manipulated. Afterwards, all participants performed the re-evaluation task behaviorally on a PC. Participants rated themselves and the other player a second time so that we could assess how much they updated their ratings.

(C) For the main behavioral analyses we employed a design with four factors. There were two within-subject factors (depicted here): feedback target (self/other) and feedback desirability (desirable/undesirable). Feedback was desirable feedback when feedback ratings were higher than participants’ own first ratings and undesirable when feedback ratings were lower than participants’ first ratings. All ratings for negative trait adjectives were reverse-coded. Thus, feedback desirability was independent of the valence of the trait adjective. The two between-subject factors were culture (German/Chinese) and current place of residence (Berlin/Beijing; **Table 1**).

**Figure 2. Behavioral results—cultural difference in overall updating and cultural similarity in positively biased updating**

(A) Overall Chinese showed greater updating than Germans. Positively biased updating was evident across all participants, i.e., updates were higher for desirable compared with undesirable feedback.

(B) Chinese scored higher on interdependence than Germans. Participants living in Beijing scored higher on interdependence than participants in Berlin.

(C) Germans scored higher on independence than Chinese.

(D) Interdependence correlated with overall updates across all participants.

See **Table 2** for further details.

**Figure 3. FMRI results—cultural differences in BOLD signals when receiving feedback**

(A) When participants received social feedback, ACC/MPFC and IFG/anterior insula showed a main effect for feedback target (self > other) across all participants ($p < 0.05$ FWE corrected, cluster size > 15; **Table 3**).

(B) When participants received social feedback, activity in the ACC/MPFC differed between Germans and Chinese. There was a culture (German/Chinese) by feedback target (self/other) interaction ($p < 0.05$ small volume corrected within the main effect shown in (A); initial contrast threshold for interaction: $p < 0.001$, uncorrected).

(C) To illustrate the interaction contrast depicted in (B) we extracted parameter estimates within the ACC/MPFC for self- and other-related feedback separately for Germans and Chinese.

(D) Parameter estimates for self-related feedback onsets within the ACC/MPFC correlated with independence scores. There was a trend indicating that the strength of the correlation between independence and parameter estimates might be stronger for Chinese compared with Germans.

**Figure 4. FMRI results—BOLD signals of reward and comparison-related components of social feedback**

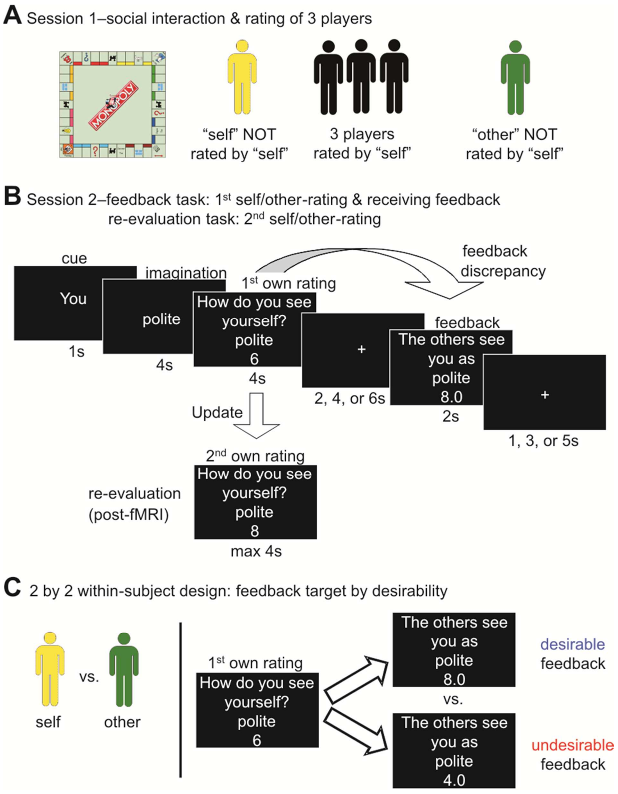
(A) Across all participants BOLD signal changes in ACC/MPFC and bilateral ventral striatum correlated with the rewarding component of feedback on a trial-by-trial basis at the time-point of feedback ($p < 0.05$ FWE corrected, cluster size > 15). See **Table 4** for a full list of activations. Reward-related activity was identified in a contrast between the parametric modulators for feedback ratings for self and other. Thus, reward-related activity correlated positively with feedback ratings for self (e.g., a feedback rating of 8.0 on “polite” is more rewarding than a feedback rating of 7.0; feedback ratings for negative trait adjectives were reverse-coded). Reward-related activity was self-specific since it correlated more positively with feedback ratings for self than with those for other.

(B) Across all participants BOLD signal changes in the following regions correlated with the comparison-related component of feedback on a trial-by-trial basis at the time-point of feedback: MPFC, preSMA/SMA, bilateral IFG (orbital part) extending into anterior insula, left TPJ, left TP and, right STS ($p < 0.05$ FWE corrected, cluster size > 15); **Table 4**). Comparison-related activity correlated positively with the feedback discrepancies for both self and other, i.e., with the absolute difference between participants' own views and the feedback they received.

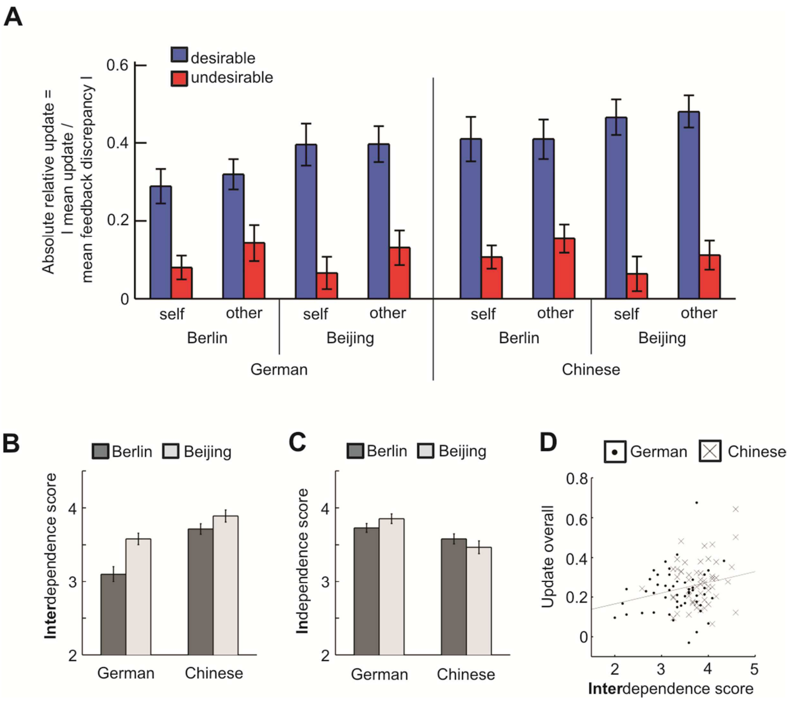
See **Supplementary Figure S3** for further details.

## Figures

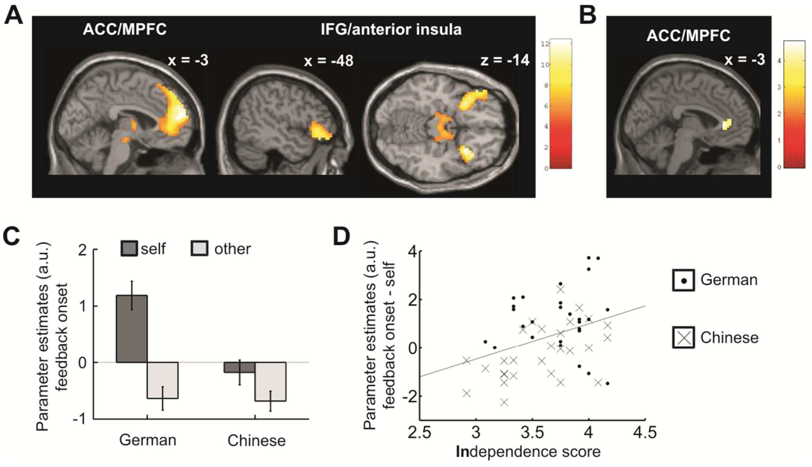
### Figure 1


A Session 1–social interaction & rating of 3 players
"self" NOT rated by "self"
3 players rated by "self"
"other" NOT rated by "self"
B Session 2–feedback task: 1st self/other-rating & receiving feedback
re-evaluation task: 2nd self/other-rating
cue
You
1s
imagination
polite
4s
1st own rating
How do you see yourself?
polite
6
4s
feedback discrepancy
+
2, 4, or 6s
feedback
The others see you as polite
8.0
2s
+
1, 3, or 5s
Update
2nd own rating
How do you see yourself?
polite
8
max 4s
re-evaluation (post-fMRI)
C 2 by 2 within-subject design: feedback target by desirability
self
vs.
other
1st own rating
How do you see yourself?
polite
6
The others see you as polite
8.0
vs.
The others see you as polite
4.0
desirable feedback
undesirable feedback

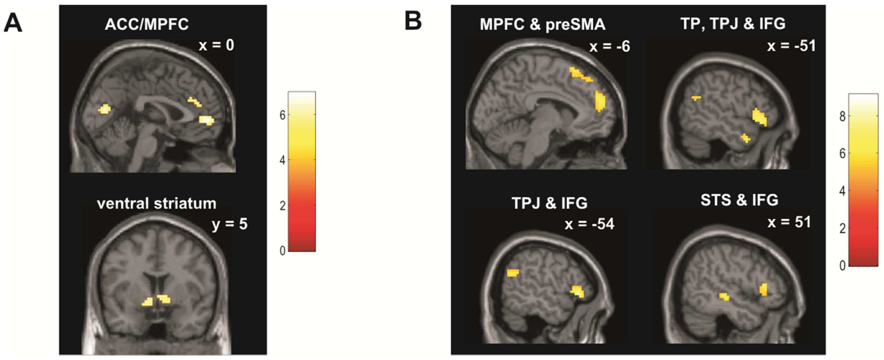
## Figure 2



## Figure 3



## Figure 4



## Tables

**Table 1. Characteristics of participants** (data are given as mean and standard deviation)

| | Germans | | Chinese | |
|---|---|---|---|---|
| | Berlin | Beijing | Berlin | Beijing |
| | fMRI | behavior | fMRI | behavior |
| n | 27 | 24 | 28 | 25 |
| Sex, female | 14 | 10 | 14 | 15 |
| Age, years (y) | 24.3 (2.47) | 24.3 (3.24) | 25.9 (2.53) | 22.7 (1.86) |
| Education, y | 16.1 (2.22) | 16.9 (1.77) | 18.5 (3.13) | 16.1 (2.39) |
| Living without parents, y | 4.6 (2.84) | 4.6 (3.49) | 6.5 (6.08) | 5.8 (4.19) |
| Living abroad, y | - | 0.9 (0.91) | 0.8 (0.45) | - |
| Learning foreign language, y | - | 2.0 (1.72) | 1.7 (2.08) | - |
| Interdependence score | 3.10 (0.53) | 3.58 (0.38) | 3.71 (0.38) | 3.89 (0.41) |
| Independence score | 3.72 (0.30) | 3.85 (0.32) | 3.58 (0.36) | 3.47 (0.43) |

| | | | | |
|---|---|---|---|---|
| Self-esteem score | 23.0 (5.35) | 23.6 (3.57) | 20.6 (5.40) | 21.7 (4.36) |
| Perceived similarity score | 3.67 (1.47) | 4.06 (1.63) | 3.79 (1.77) | 4.24 (1.13) |

**Table 2. Task-related variables**

| | Germans | | | | Chinese | | | |
|---|---|---|---|---|---|---|---|---|
| | Berlin | | Beijing | | Berlin | | Beijing | |
| | fMRI | | behavior | | fMRI | | behavior | |
| | Self | other | self | other | self | other | self | other |
| n trials final[a] | 72.8 | 71.3 | 68.7 | 66.3 | 69.5 | 69.3 | 70.2 | 69.2 |
| | (3.10) | (2.70) | (6.37) | (7.50) | (3.96) | (5.61) | (5.78) | (6.11) |
| n trials excluded: | 1.70 | 2.22 | 4.54 | 6.50 | 3.64 | 3.89 | 3.00 | 2.80 |
| missing answers | (1.92) | (1.97) | (3.40) | (5.33) | (2.63) | (4.00) | (2.65) | (3.16) |
| n trials excluded: | 5.52 | 6.44 | 5.08 | 5.58 | 6.18 | 6.14 | 5.24 | 6.40 |
| zero feedback | (2.29) | (2.50) | (2.06) | (1.93) | (2.68) | (1.99) | (2.07) | (3.00) |
| discrepancies | | | | | | | | |
| first ratings | 5.63 | 5.22 | 5.64 | 5.41 | 5.60 | 5.52 | 5.54 | 5.42 |
| | (0.61) | (0.69) | (0.57) | (0.82) | (0.92) | (0.65) | (0.55) | (0.57) |
| second ratings | 5.74 | 5.36 | 5.84 | 5.60 | 5.76 | 5.64 | 5.82 | 5.65 |
| | (0.61) | (0.76) | (0.59) | (0.87) | (0.93) | (0.59) | (0.63) | (0.58) |
| relative absolute | 0.29 | 0.31 | 0.40 | 0.40 | 0.41 | 0.41 | 0.47 | 0.48 |
| mean update: | (0.23) | (0.20) | (0.26) | (0.23) | (0.31) | (0.27) | (0.23) | (0.21) |
| desirable | | | | | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| relative absolute mean update: undesirable | 0.08<br>(0.15) | 0.14<br>(0.24) | 0.07<br>(0.20) | 0.13<br>(0.22) | 0.11<br>(0.16) | 0.15<br>(0.19) | 0.06<br>(0.22) | 0.11<br>(0.19) |
| absolute memory error: desirable[b] | 1.35<br>(0.34) | 1.46<br>(0.28) | 1.43<br>(0.33) | 1.51<br>(0.29) | 1.35<br>(0.42) | 1.51<br>(0.48) | 1.30<br>(0.34) | 1.51<br>(0.39) |
| absolute memory error: undesirable[b] | 1.18<br>(0.26) | 1.43<br>(0.24) | 1.41<br>(0.42) | 1.52<br>(0.26) | 1.32<br>(0.32) | 1.54<br>(0.31) | 1.21<br>(0.25) | 1.45<br>(0.30) |

[a] Two participants in the German Beijing group, one participant in the Chinese Berlin and two participants in the Chinese Beijing group completed only three out of four feedback runs due to technical problems.

[b] Three participants in the German Beijing group did not complete the memory test and five did not complete the desirability rating of the stimuli due to time constraints.

**Table 3. Significant activations in feedback onsets**. For main effects, clusters are whole-brain FWE-corrected for multiple comparisons at the voxel-level $p < 0.05$, cluster size > 15; for interaction effects, clusters are small volume corrected within main effect: feedback onset: self > other; initial threshold for interaction: $p < 0.001$, uncorrected.

| | Side | Peak voxel MNI coordinates (mm) | | | Cluster size | Peak t score |
|---|---|---|---|---|---|---|
| | | x | y | z | (Voxel) | |
| Main effect: feedback onset: self > other | | | | | | |
| IFG (orbital part)/ anterior insula | R | 33 | 17 | -14 | 217 | 12.42 |
| MPFC/ ACC | L/R | -3 | 56 | 16 | 1198 | 12.19 |
| IFG (orbital part)/ anterior insula | L | -30 | 14 | -17 | 437 | 11.48 |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Cerebellum | R | 27 | -82 | -35 | 146 | 9.71 |
| Cerebellum | L | -30 | -82 | -38 | 106 | 9.20 |
| Midbrain | L/R | 9 | -10 | -14 | 94 | 7.32 |
| Thalamus | L/R | -3 | -4 | 4 | 35 | 6.78 |
|  |  |  |  |  |  |  |
| Main effect: feedback onset: other > self |  |  |  |  |  |  |
| Precuneus/ postcentral gyrus/ superior temporal gyrus/ supramarginal gyrus | L/R | 9 | -55 | 49 | 7946 | 10.48 |
| Middle frontal gyrus—dorso-lateral PFC | L | -36 | 44 | 31 | 76 | 7.30 |
| Middle frontal gyrus—dorso-lateral PFC | R | 24 | 32 | 34 | 51 | 6.33 |
| Precentral gyrus | L | -54 | 5 | 28 | 15 | 5.72 |
|  |  |  |  |  |  |  |
| Interaction: feedback onset: (self > other) X (German > Chinese) |  |  |  |  |  |  |
| MPFC/ ACC | L/R | -3 | 32 | 4 | 44 | 4.58 |

**Table 4. Changes in BOLD signal related to reward- and comparison-related components of social feedback**. Clusters are whole-brain FWE-corrected for multiple comparisons at the voxel-level p < 0.05, cluster size > 15.

|  | Side | Peak voxel MNI coordinates (mm) |  |  | Cluster size (Voxel) | Peak t score |
|---|---|---|---|---|---|---|
|  |  | x | y | z |  |  |
| Feedback rating (trial-by-trial correlation): self > other |  |  |  |  |  |  |
| ACC/MPFC | L/R | 0 | 50 | 1 | 43 | 6.69 |
| Dorsal caudate | R | 21 | -19 | 16 | 62 | 6.78 |
| Calcarine fissure | L/R | 0 | -76 | 13 | 98 | 6.48 |
| Ventral striatum | R | 6 | 11 | -2 | 39 | 6.24 |
| Cerebellum | L | -30 | -73 | -23 | 39 | 6.04 |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Ventral striatum | L | -6 | 5 | -8 | 19 | 5.99 |
| Precuneus | L | -12 | -52 | 22 | 20 | 5.93 |
| Lingual gyrus | L | -9 | -58 | 1 | 19 | 5.84 |
| ACC | L/R | 0 | 41 | 19 | 31 | 5.79 |
| Dorso-lateral PFC | L | -18 | 29 | 49 | 20 | 5.63 |
| Feedback discrepancies (positive trial-by-trial correlation): self and other |  |  |  |  |  |  |
| MPFC | L/R | 9 | 59 | 28 | 253 | 9.11 |
| IFG (orbital part)/ anterior insula | L | -54 | 26 | -2 | 192 | 7.79 |
| TP | L | -51 | 8 | -32 | 24 | 7.04 |
| Angular gyrus—TPJ | L | -60 | -58 | 25 | 50 | 6.92 |
| IFG (orbital part) | R | 54 | 26 | 10 | 54 | 6.89 |
| Anterior insula | R | 33 | 20 | -20 | 26 | 6.87 |
| STS | R | 51 | -25 | -5 | 32 | 6.87 |
| Cerebellum | L | -27 | -85 | -35 | 16 | 6.14 |
| preSMA/SMA | L/R | -6 | 20 | 64 | 83 | 6.13 |
| Feedback discrepancies (negative trial-by-trial correlation): self and other |  |  |  |  |  |  |
| Inferior parietal lobule | R | 54 | -37 | 52 | 114 | 8.12 |
| Middle frontal gyrus | R | 30 | 5 | 58 | 85 | 7.80 |
| Inferior temporal gyrus | R | 57 | -49 | -14 | 24 | 6.80 |
| Inferior parietal lobule | L | -54 | -40 | 43 | 142 | 6.67 |
| IFG | L | -42 | 44 | 10 | 25 | 6.25 |
| Middle occipital gyrus | L | -21 | -61 | 40 | 18 | 5.73 |

## Supplementary information

### Cultural influences on social feedback processing of character traits

Christoph W. Korn, Yan Fan, Kai Zhang, Chenbo Wang, Shihui Han, and Hauke R. Heekeren

## Supplementary methods

### Additional participant information

In the German group in Berlin three of the initial 30 participants had to be excluded (one did not tolerate the scanner environment, another showed excessive head movement (> 8 mm), and data from another subject could not be used due to technical problems). In the Chinese group in Berlin two of the initial 30 participants had to be excluded because of excessive head movement (> 8°). All scanned participants were right-handed.

All German participants spoke German as their mother-tongue. All Chinese participants were fluent in Mandarin and spoke Mandarin or Cantonese as their mother-tongue. All except three German participants had been raised by two German parents; three participants had one German parent and one English, French, or Russian parent. All Chinese participants had been raised by two Chinese parents. In the Chinese group in Berlin four participants were from Hong Kong. German and Chinese participants were recruited, instructed, and tested in German and Mandarin, respectively. All participants gave written informed consent.

Groups did not differ with respect to sex, education, and time living without parents. Chinese and Germans did not differ with regard to age. However, participants in Berlin were older than participants in Beijing (Mann-Whitney U = 840.5, p = 0.001). We accounted for this difference in age by testing whether including age as a covariate had an influence on the ANOVA results. This was not the case. For simplicity we therefore report all analyses without age as covariate. The German participants in Beijing and the Chinese participants in Berlin did not differ in how long they had lived abroad and in how long they had learned the respective foreign language.

### Stimuli and translation

We used 40 positive and 40 negative trait adjectives as described previously (Korn et al., 2012). See **Supplementary Table S1** for a list of trait adjectives. Trait adjectives were translated from German into Mandarin by an accredited court interpreter and double-checked by a native Mandarin speaker. One German and two Chinese authors made sure that German and Mandarin versions of the trait adjectives captured the same meaning. All instructions were translated by two Chinese authors.

To confirm that participants perceived the trait words as positive and negative in the way we had predefined them, participants rated all 80 trait adjectives on social positivity on a Likert scale from 1 (not positive at all) to 8 (very positive) at the very end of the experiment. Five participants in the German Beijing group did not complete this desirability rating due to time constraints. Across all participants, mean ratings for positive and negative trait words differed from the midpoint of the scale as assessed by one-sample t-tests (mean rating: positive words = 6.66, SD = 0.58; t(98) = 36.68, p < 0.001; negative words = 2.63, SD = 0.66; negative words t(98) = -28.12, p < 0.001). Positive trait adjectives were rated similarly by both cultural groups but negative trait adjectives were rated as less desirable by German compared with Chinese participants (independent sample t-test: t(97) = -2.86, p = 0.005). However, this difference did not compromise our findings related to updating behavior since the factor valence did not interact with any other factors (see **Trait valence and updating**).

### Social interaction and rating of 3 players (first session)

The first session aimed at creating a real-life interaction among peers so that the social feedback would be meaningful for participants. For the first session of the experiment (**Figure 1A**), participants came into the laboratory in groups of five people of the same culture and got to know each other by playing a table-top version of the popular board game “Monopoly” (HASBRO, Soest, Germany; HASBRO, Shanghai, China) for 1 h and 15 min. We made sure that participants did not know each other before the experiment.

In the groups in Berlin and in the Chinese groups in Beijing all five participants in a group were of the same sex. German groups in Beijing consisted of members of both sexes since we were unable to recruit enough German participants in Beijing to form same sex groups. Additionally, one of the German participants in Beijing was aware of the experimental manipulation and only participated so that we could form a group of five people. Data from this participant were not analyzed. Therefore, the total number of German participants in Beijing was 24 and not 25.

We chose the board game “Monopoly” for the social interaction because it is highly engaging, quite well-known, and allows players to show a variety of cooperative and competitive behaviors. Furthermore, within 1 h 15 min nobody was eliminated from the game. The rules of the game were explained to all participants before the game. The study was introduced as a study about how people get to know each other. Participants knew before they started to play the game that they were going to be rated by the other players of their group and they believed that their own ratings were going to be shown to the other players in an anonymous fashion. During the game participants were free to talk about whatever topics they wanted. Participants wore name tags and we made sure that participants knew the names of all players after the game. After 1 h 15 min we assessed the ranking of the participants in the game, i.e., assigned the first rank to the winner and so on. Participants’ ranks in the game did not correlate with any behavioral measures on the task as assessed by Spearman correlations (all $p > 0.1$).

After the game, each participant rated three of the four other participants on 80 trait adjectives on a Likert scale from 1 (this trait does not apply the person at all) to 8 (this trait does apply the person very much); for trait adjectives see **Stimuli and translation** and **Supplementary Table S1**. Ratings were given on a PC using the MATLAB toolbox Cogent 2000. Each of the three persons was rated in a separate block. On each trial participants saw one of the 80 adjectives with the first name of the person to rate and had up to 10 s to respond. At the end of the first session of the experiment each participant had rated three other participants and in turn each participant had been rated by three other participants. Participants had not yet rated themselves (depicted in yellow in **Figure 1A**) and had not yet rated one other player (depicted in green).

**Memory task**

After rating themselves and the other person a second time (i.e., after the re-evaluation task) (**Figure 1B**), participants performed a memory task on a PC using the MATLAB toolbox Cogent 2000. For all 80 trait adjectives participants had to recollect the feedback they had seen in the feedback task and had to type in that number, i.e., a number between 1 and 8 with one decimal such as 1.0, 1.3, or 1.7. Participants had to recollect the feedback in two separate blocks (one for themselves and one for the other person), which were randomized for order. They had up to 12 s to respond. Three participants in the German Beijing group did not complete the memory test due to time constraints.

Memory errors were calculated as the absolute differences between the recollected number and the actual feedback rating.

(1) absolute memory error = abs(feedback rating – recollection of feedback rating)

Similar to update scores, mean absolute memory errors were compared in a 2 (target: self/other) by 2 (desirability: desirable/undesirable) by 2 (culture: German/Chinese) by 2 (place: Berlin/Beijing) repeated measures ANOVA.

**FMRI data acquisition**

FMRI data were acquired on a 3T scanner (Trio, Siemens, Erlangen, Germany) using a 12-channel head coil. Functional images were acquired with a gradient echo T2*-weighted echo-planar sequence (TR = 2000 ms, TE = 30 ms, flip angle = 70, 64 x 64 matrix, field of view = 192 mm, voxel size = 3x3x3 mm$^3$). A total of 37 axial slices (3 mm thick, no gap,) were sampled for whole brain coverage. Imaging data were acquired in four separate 349-volume runs of 11 min 38 s each. The first five volumes of each run were discarded to allow for T1 equilibration. A high-resolution T1-weighted anatomical scan of the whole brain was acquired (256 x 256 matrix, voxel size = 1x1x1 mm$^3$).

**FMRI data preprocessing**

EPI images were realigned, unwarped, co-registered to the respective participant's T1 scan, normalized to a standard T1 template based on the Montreal Neurological Institute (MNI) reference brain, resampled to 3 mm isotropic voxels, and spatially smoothed with an isotropic 8 mm full width at half maximum (FWHM) Gaussian kernel. Using the East Asian brain template provided by SPM instead of the standard MNI brain did not result in different clusters in any analyses.

**Parametric modulation analyses**

We investigated trial-by-trial fluctuations in brain activity during the feedback phase, which correlated with two different components of social feedback: reward- and comparison-related components. We split trials according to feedback target (self/other) for each participant.

To detect activity related to social comparison we used parametric modulators of feedback discrepancies. We used the full parametric range of feedback ratings and feedback discrepancies across all trials (i.e., across trials with desirable and undesirable feedback and trials with feedback discrepancies of zero).

Activity related to the rewarding component of social feedback should correlate positively with feedback ratings for self. Note that feedback ratings for negative traits were reverse-coded. That is, a high feedback rating indicated high self-relevant social reward (i.e., feedback that a positive trait applied to the self or that a negative trait did not apply to the self) and a low feedback rating indicated low self-relevant social reward. To make sure that activity related to the rewarding component of social feedback was truly self-specific, we subtracted activity that correlated with the feedback ratings for other.

Activity related to the social comparison component of social feedback should correlate positively with feedback discrepancies, which were defined as the absolute differences between first own ratings and feedback ratings; i.e., feedback discrepancies captured how close feedback ratings were to participants’ own ratings, regardless of the direction of the differences.

All regressors and modulators were entered independently into the design matrix, i.e., without the serial orthogonalization used as default in SPM (for a similar approach see Gläscher et al., 2010; Wunderlich et al., 2011). This ensured that only the additional variance that cannot be explained by any other regressor was assigned to the respective effect and thus prevented spurious confounds between regressors.

### FMRI data—follow-up analyses

We performed follow-up functional ROI analyses to visualize interactions and to visualize correlations between neural activity and the parametric modulators (i.e., the betas of the parametric modulators for feedback ratings and the betas of the parametric modulators for feedback discrepancies). We used the marsbar toolbox for SPM (marsbar.sourceforge.net/) to extract parameter estimates within functional ROIs.

## Supplementary results

### Additional behavioral results—direction of updates, trait valence, first ratings, perceived similarity scores, and memory errors

As expected, participants changed their ratings on average toward the feedback; they increased their ratings for desirable feedback and decreased their ratings for undesirable feedback as indicated by positive and negative relative mean updates, respectively (mean relative updates: self-desirable = 0.39, SD = 0.26; one-sample t-test against zero t(103) = 15.0, p < 0.001; self-undesirable = -0.08, SD = 0.18; t(103) = -4.4, p < 0.001; other-desirable = 0.40, SD = 0.23; t(103) = 17.5, p < 0.001; other-undesirable = -0.14, SD = 0.21; t(103) = -6.6, p < 0.001; see **Table 2** for relative absolute mean updates separated according to group membership).

We explored whether the valence of the trait adjectives had an effect on updating. We split update scores according to the valence of the trait words and included valence as an additional factor in the ANOVA (resulting in a 2 (trait valence: positive/negative) by 2 (target: self/other) by 2 (desirability: desirable/undesirable) by 2 (culture: German/Chinese) by 2 (place: Berlin/Beijing) ANOVA on relative absolute mean updates). The main effects of desirability and of culture were still significant (desirability: F(1,99) = 65.70, p < 0.001, $\eta_p^2$ = 0.40; culture: F(1, 99) = 4.67, p = 0.03, $\eta_p^2$ = 0.05). Additionally, there was a significant main effect of valence (F(1,

99) = 16.66, $p < 0.001$, $\eta_p^2 = 0.14$) with updates for negative trait words being higher than for positive trait words. There were no further significant effects (all $p > 0.05$). Thus, although there was a significant main effect of valence, valence did not significantly interact with any other factor.

We tested for differences in participants' first self- versus other-ratings in a 2 (feedback target: self/other) by 2 (culture: German/Chinese) by 2 (place: Berlin/Beijing) ANOVA. We found a significant main effect of feedback target with self-ratings being higher than other ratings across all participants ($F(1,100) = 7.57$, $p = 0.007$, $\eta_p^2 = 0.07$; **Supplementary Figure S1**; see **Table 2** for mean first ratings separated according to group membership). There were no further significant main effects or interactions (all $p > 0.1$). Thus, we found evidence for a positivity bias towards the self in participants' initial ratings, which did not differ across culture or place in our sample.

In addition to trait measures on interdependence, independence, and self-esteem, participants indicated how similar they perceived the other person on a Likert scale. A 2 (culture) by 2 (place) ANOVA on similarity ratings did not reveal significant effects ($p > 0.1$). We have previously shown for the German participants in Berlin (Korn et al., 2012) that first self-ratings correlated with self-esteem scores and that first other-ratings correlated with how similar participants perceived the other person. These correlations held across all participants (self-ratings and self-esteem: $r = 0.52$, $p < 0.001$, 95% CI [0.36, 0.65]; other-ratings and perceived similarity: $r = 0.46$, $p < 0.001$, 95% CI [0.29, 0.60]).

Outside the scanner participants recollected the feedback rating they had seen inside the scanner. Similar to updates, mean absolute memory errors were subjected to a 2 (target: self/other) by 2 (desirability: desirable/undesirable) by 2 (culture: German/Chinese) by 2 (place: Berlin/Beijing) ANOVA. As expected, we found a significant main effect of feedback target with memory errors being smaller for self- compared with other-related feedback ($F(1,97) = 71.83$, $p < 0.001$, $\eta_p^2 = 0.43$). There was also a significant main effect of culture with memory errors being smaller for Germans compared with Chinese ($F(1,97) = 4.27$, $p = 0.041$, $\eta_p^2 = 0.04$). No other effects reached significance (all $p > 0.05$). Thus, positively biased updating seemed to be unrelated to memory. The cultural difference in memory is unlikely to have influenced cultural differences in updating for three reasons. First, memory performance did not correlate with updating behavior ($p > 0.6$). Second, memory performance did not correlate with interdependence or independence scores ($p > 0.2$). Third, the group with better memory performance should theoretically show higher updating. However, Germans, who had better memory, showed smaller updating.

**Additional fMRI results—imagination phase**

We focused our main fMRI analyses on the time period when participants received social feedback but our task also included a time period when participants made trait judgments of themselves and other persons (imagination phase; **Figure 1B**). Since previous studies have mainly investigated cultural influences on ACC/MPFC activity when participants made trait judgments, we compared both time periods in a follow-up ROI-based approach.

We extracted parameter estimates during both time points within an ROI that was independently defined based on a recent study (Ma et al., 2012) comparing neural activity while Danish and Chinese participants made trait judgments of themselves and a public person (sphere with a radius of 15 mm centered at the MNI coordinate -4, 32, 0). We compared parameter estimates in a 2 (feedback target: self/other) by 2 (time period: feedback/imagination) by 2 (culture: German/Chinese) ANOVA. As expected self-related activity was higher than other-related activity ($F(1,53) = 63.43$, $p < 0.001$, $\eta_p^2 = 0.55$) and the factors culture and feedback target showed a significant interaction ($F(1,53) = 10.97$, $p = 0.002$, $\eta_p^2 = 0.17$; **Supplementary Figure S2**). There was also a significant main effect of culture ($F(1,53) = 5.26$, $p = 0.009$, $\eta_p^2 = 0.12$), a significant time period by culture interaction ($F(1,53) = 10.91$, $p = 0.002$, $\eta_p^2 = 0.17$) as well as a significant three-way interaction ($F(1,53) = 5.50$, $p = 0.023$, $\eta_p^2 = 0.09$). To qualify this three-way interaction we performed two separate 2 (self/other) by 2 (German/Chinese) ANOVAs for the feedback and imagination time periods. As expected from the analyses reported in the main text, the interaction of feedback target and culture was significant for the feedback time period ($F(1,53) = 10.93$, $p = 0.002$, $\eta_p^2 = 0.17$). The same feedback target by culture interaction was also significant for the imagination time period

(F(1,53) = 4.13, p = 0.047, $\eta_p^2$ = 0.07) but at to a lesser degree. Thus, the three-way interaction was qualified by a greater feedback target by culture interaction for the feedback phase compared with the imagination phase.

Taken together, in line with previous studies we found cultural influences on ACC/MPFC activity when participants made trait judgments. Our findings suggest that this cultural effect might be even stronger when participants receive social feedback.

**Additional fMRI results—reward- and comparison-related activity**

For completeness, we performed the reverse contrast to the contrast testing for reward-related activity, i.e., we searched for activity correlating with other-related feedback ratings more than with self-related feedback ratings. This contrast revealed no significant voxels (p°<°0.05 FWE corrected, cluster size°>°15). We also searched for regions correlating negatively with feedback discrepancies (see **Table 4**) and for activity correlating differentially for self- versus other-related feedback discrepancies, i.e., self > other or other > self. These differential contrasts revealed no significant voxels (p°<°0.05 FWE corrected, cluster size°>°15).

For the German fMRI sample (Korn et al., 2012), we have previously shown in a conjunction analysis (i.e., in test of the conjunction null hypothesis) that a region at the border of the MPFC and ACC was activated by both reward- and comparison-related components ($p < 0.05$ FWE corrected at cluster level, cluster-defining threshold of $p < 0.0001$). Using the same threshold, a conjunction for both Germans and Chinese revealed two clusters: a cluster at a similar location as shown before (MPFC/ACC: -3, 50, 10; cluster size 21) and a cluster in a more dorsal part of the ACC (3, 44, 25; cluster size 43). We note that at the more stringent threshold which we used to report clusters in the present study ($p < 0.05$ FWE corrected at voxel level, cluster size > 15), the conjunction revealed no regions of overlap.

In the previous study (Korn et al., 2012), we have shown that for the German fMRI sample the parameter estimates of the self-related absolute feedback discrepancies within the region revealed by the conjunction correlated with the behavioral update bias for self. In the MPFC/ACC region identified in the conjunction across both cultural groups, parameter estimates of the self-related absolute feedback discrepancies correlated with the update bias for self in Germans (Pearson’s r = 0.42, p = 0.029, 95% CI [0.05, 0.70) but not in Chinese (r = -0.02, p > 0.9). The difference in the strength of these correlations approached trend level (Fisher’s z = 1.63; p = 0.103).

### Supplementary references

Gläscher, J., Daw, N., Dayan, P., & O’Doherty, J.P. (2010) States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron, 66, 585–595.

Wunderlich, K., Symmonds, M., Bossaerts, P., & Dolan, R.J. (2011) Hedging your bets by learning reward correlations in the human brain. Neuron, 71, 1141–1152.

### Supplementary figure legends

**Supplementary Figure S1. First ratings**

Across all participants first self-ratings were higher than first other ratings.

**Supplementary Figure S2. FMRI results—cultural differences during the feedback and imagination phases**

We used an independently defined ROI (Ma et al., 2012; sphere with a radius of 15 mm centered at the MNI coordinate -4, 32, 0) to extract parameter estimates during the feedback and imagination phases. During both phases the culture by feedback target interaction was significant.

**Supplementary Figure S3. FMRI results—parameter estimates for reward- and comparison-related components of social feedback**

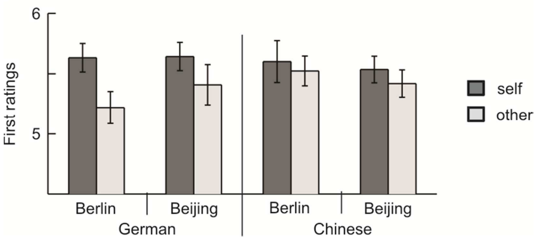
For illustration purposes, we plotted parameter estimates of the parametric modulators for the functional ROIs identified for reward- and comparison related components. No main effects of culture or interactions with the factor culture reached significance (all *p* > 0.1; *p*-values were adjusted using a Bonferroni correction for the number of ROIs)

(A) Reward-related component – Parameter estimates of the parametric modulators of feedback ratings for self and other within functional ROIs. BOLD signals correlated positively with feedback ratings for self but not for other. See **Figure 4A** for ROIs.

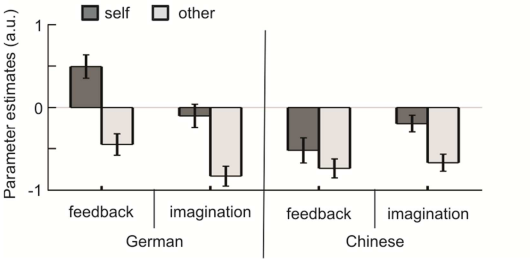
(B) Comparison-related component – Parameter estimates of the parametric modulators of feedback discrepancies for self and other within functional ROIs. BOLD signals correlated positively with feedback discrepancies for self and other. See **Figure 4B** for ROIs.

**Supplementary figures**

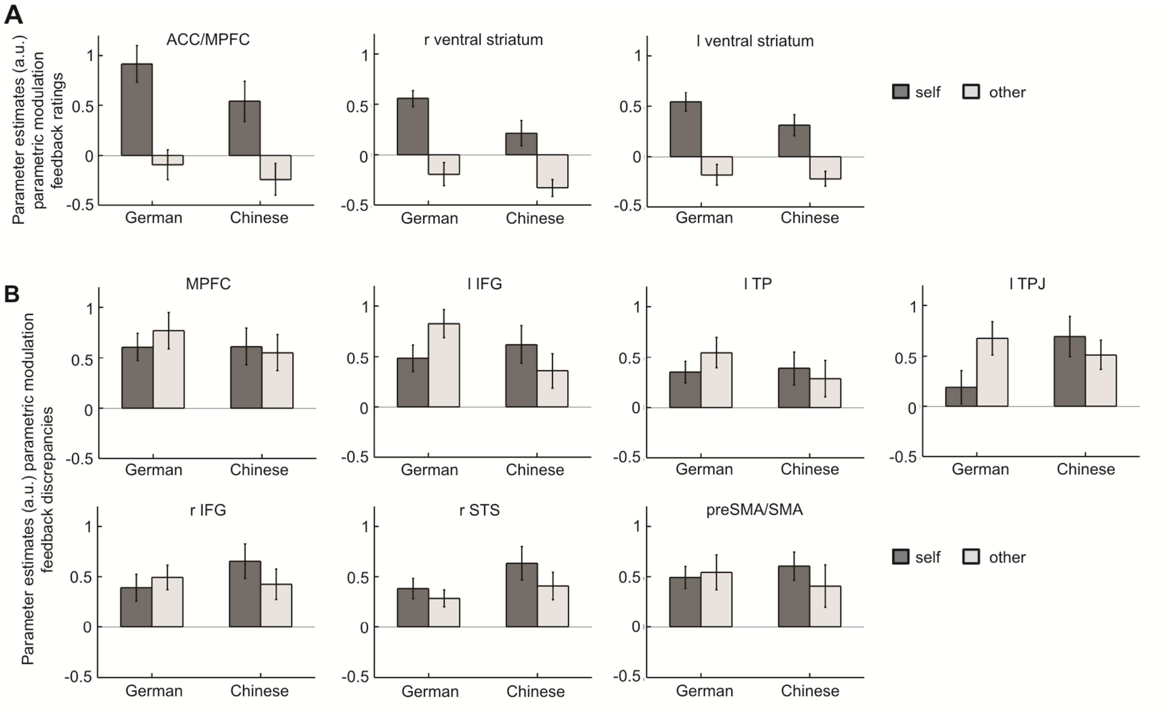
**Supplementary Figure S1**



**Supplementary Figure S2**



## Supplementary Figure S3



## Supplementary Table 1. List of trait adjectives

| German | Chinese | English |
|---|---|---|
| Positive trait adjectives | | |
| aufrichtig | 诚实的 | honest |

| | | |
|---|---|---|
| bescheiden | 谦虚的 | modest |
| diszipliniert | 遵守纪律的 | organized |
| effizient | 有效率的 | efficient |
| einfühlsam | 敏锐的 | empathetic |
| enthusiastisch | 热心的 | enthusiastic |
| fleißig | 努力的 | hard-working |
| freundlich | 友善的 | friendly |
| geistesgegenwärtig | 沉着灵敏的 | quick-witted |
| gelassen | 轻松镇静的 | composed |
| geschickt | 老练的 | skilled |
| gesellig | 好交际的 | sociable |
| großzügig | 大方的 | generous |
| hilfsbereit | 乐于助人的 | helpful |
| höflich | 有礼貌的 | polite |
| kompetent | 有能力的 | competent |
| kooperativ | 愿意合作的 | cooperative |
| kreativ | 有创造力的 | creative |
| lebenslustig | 热爱生活的 | fun-loving |
| locker | 不慌不忙的 | easy-going |

| | | |
|---|---|---|
| loyal | 忠实的 | loyal |
| offen | 坦率的 | open-minded |
| ordentlich | 整齐的 | tidy |
| respektvoll | 尊重人的 | respectful |
| scharfsinnig | 有洞察力的 | astute |
| schlagfertig | 反应敏捷的 | articulate |
| selbstständig | 独立的 | self-reliant |
| sorgfältig | 细心的 | diligent |
| souverän | 很有把握的 | confident |
| spontan | 自发的 | spontaneous |
| tatkräftig | 精力充沛的 | dynamic |
| tolerant | 宽容的 | tolerant |
| vernünftig | 理智 的 | level-headed |
| verständnisvoll | 充分理解的 | understanding |
| vertrauenswürdig | 可信任的 | trustworthy |
| vielseitig | 多才多艺的 | versatile |
| weitsichtig | 有远见的 | perspicacious |
| wissbegierig | 好学的 | inquisitive |
| zielstrebig | 有目标的 | goal-oriented |

| | | |
|---|---|---|
| zuverlässig | 可靠的 | reliable |
| Negative trait adjectives | | |
| aggressiv | 好斗的 | aggressive |
| ängstlich | 胆怯的 | anxious |
| arrogant | 傲慢的 | arrogant |
| bieder | 呆板的 | overly-conservative |
| chaotisch | 乱七八糟的 | chaotic |
| egoistisch | 自私的 | selfish |
| eitel | 虚荣的 | conceited |
| engstirnig | 心胸狭窄的 | narrow-minded |
| feige | 胆小的 | cowardly |
| gehässig | 恶毒的 | spiteful |
| großmäulig | 爱吹牛的 | loud-mouthed |
| heuchlerisch | 虚伪的 | two-faced |
| hinterhältig | 奸猾的 | conniving |
| humorlos | 缺乏幽默感的 | humorless |
| inkonsequent | 前后不一致的 | inconsistent |
| kalt | 冷漠的 | cold-hearted |

|  |  |  |
|---|---|---|
| launisch | 喜怒无常的 | moody |
| leichtsinnig | 漫不经心的 | foolhardy |
| nachtragend | 怀恨在心的 | unforgiving |
| naiv | 天真的 | naive |
| oberflächlich | 肤浅的 | superficial |
| opportunistisch | 机会主义的 | opportunistic |
| pedantisch | 死板的 | pedantic |
| rücksichtslos | 毫无顾忌的 | inconsiderate |
| scheu | 害羞的 | unassertive |
| stur | 固执的 | stubborn |
| träge | 懒散的 | lazy |
| unentschlossen | 犹豫不决的 | indecisive |
| ungeduldig | 不耐烦的 | impatient |
| unnahbar | 不易亲近的 | inapproachable |
| unpünktlich | 不准时的 | tardy |
| unsicher | 缺乏自信的 | insecure |
| unsympathisch | 不讨人喜欢的 | unpleasant |
| verschwenderisch | 浪费的 | wasteful |
| voreilig | 仓促的 | rash |

| | | |
|---|---|---|
| voreingenommen | 先入为主的 | biased |
| wehleidig | 自怜的 | whiny |
| zickig | 爱挑剔的 | catty |
| zwanghaft | 强迫的 | obsessive |
| zynisch | 冷嘲热讽的 | cynical |
| Adjectives used during the training trials | | |
| intelligent | 聪明的 | intelligent |
| unsportlich | 不爱运动的 | unathletic |

# Depression is related to an absence of optimistically biased belief updating about future life events

Christoph W. Korn*[1,2a], Tali Sharot[3a], Henrik Walter[2,4], Hauke R. Heekeren[1,2] and Raymond J. Dolan[5]

[1]Department of Education and Psychology, Freie Universität Berlin, Germany

[2]Berlin School of Mind and Brain, Humboldt Universität zu Berlin, Germany

[3]Department of Cognitive, Perceptual and Brain Sciences, Division of Psychology and Language Sciences, University College London, UK

[4]Department of Psychiatry and Psychotherapy, Division of Mind and Brain Research, Charité Universitätsmedizin Berlin, Germany

[5]Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, UK

* Address for correspondence: Christoph W. Korn, Habelschwerdter Allee 45, 14195 Berlin, Germany, Phone: 004930838 56226, E-mail: christoph.korn@fu-berlin.de

[a] These authors contributed equally.

**Financial support:**

This study was supported by a Senior Investigator Award to RJD (098362/Z/12/Z) , a Max Planck Award to RJD and a Wellcome Trust Career Development Fellowship to TS. The Wellcome Trust Centre for Neuroimaging is supported by core funding from the Wellcome Trust 091593/Z/10/Z.

## Abstract

**Background:** When challenged with information about the future, healthy participants show an optimistically biased updating pattern, taking desirable information more into account than undesirable information. However, it is unknown how patients suffering from major depressive disorder (MDD), who express pervasive pessimistic beliefs, update their beliefs when receiving information about their future.

Here we tested whether an optimistically biased information processing pattern found in healthy individuals is absent in MDD patients.

**Methods:** MDD patients (n=18; 13 medicated; 8 with comorbid anxiety disorder) and healthy controls (n=19) estimated their personal probability of experiencing 70 adverse life events. After each estimate participants were presented with the average probability of the event occurring to a person living in the same sociocultural environment. This information could be desirable (i.e., average probability better than expected) or undesirable (i.e., average probability worse than expected). To assess how desirable versus undesirable information influenced beliefs, participants estimated their personal probability of experiencing the 70 events a second time.

**Results:** Healthy controls showed an optimistic bias in updating, i.e., they changed their beliefs more toward desirable versus undesirable information. Overall, this optimistic bias was absent in MDD patients. Among MDD patients symptom severity correlated with biased updating: Less severely depressed individuals showed an optimistic bias but more severely depressed individuals showed a pessimistic bias in updating. MDD patients also estimated the probability of experiencing adverse life events as higher than healthy controls.

**Conclusion:** Our findings raise the intriguing possibility that optimistically biased updating of expectations about one's personal future is associated with mental health.

**Key words:** optimism, bias, depression, information processing, belief updating

## Introduction

Individuals suffering from major depressive disorder (MDD) process information about the self, the world, and the future in a maladaptive fashion compared with healthy individuals (American Psychiatric Association, 2000). According to prominent cognitive theories of depression such as Beck's cognitive model (Beck *et al*. 1979; Haaga & Beck, 1995; Disner *et al*. 2011), Seligman's learned helplessness model (Seligman, 1972), and the more recent cognitive neuropsychological model (Clark *et al*. 2009;

Roiser *et al*. 2012), maladaptive cognitive biases are central in the development and maintenance of MDD. Beck's cognitive model, for example, emphasizes the role of maladaptive cognitive schemas and has led to the development of cognitive therapy, an effective treatment focused on changing these maladaptive cognitive schemas (Beck, 2005; Beck & Dozois, 2011). Related neuropsychological models of depression emphasize this relationship between maladaptive cognition and vulnerability or resilience to MDD, highlighting that maladaptive cognition may be causal in the progression of depressive symptoms (Clark *et al*. 2009; Roiser *et al*. 2012).

Cognitive theories of MDD often highlight that maladaptive cognition manifests as negativity biases (e.g., Beck *et al*. 1979). However, in some instances the behavior of depressed individuals seems to be better characterized by realism relative to an objective standard or by an absence of a positivity bias relative to healthy individuals (Alloy & Ahrens, 1987; Moore & Fresco, 2012). The general reasoning that MDD may be related to a reduction, absence or reversal of positivity biases relative to mental health is motivated by considerable evidence showing that healthy individuals are characterized by a diverse array of positivity biases including illusions of superiority (Taylor & Brown, 1988; Taylor & Brown, 1994; Leary, 2007), illusions of control (Taylor & Brown, 1988; Thompson et al. 1998), positivity biases in memory (Walker *et al.* 2003) as well as unrealistic optimism about the future (Weinstein, 1980; Taylor & Brown, 1988, Weinstein & Klein, 1995; Armor & Taylor, 2002; Puri & Robinson, 2007; Sharot, 2011). However, the precise relationship between mental health, MDD and cognitive information processing biases (or their absence) has remained surprisingly underexplored.

So far, information processing in MDD patients has been mostly investigated in tasks that are not directly related to positivity biases in healthy individuals (Mathews & MacLeod, 2005; Gotlib & Joorman, 2010). For example, depressed individuals show altered responses to performance feedback in cognitive tasks such as the Tower of London planning task (Elliot *et al*. 1997; Elliot *et al*. 1998) or reversal learning tasks (Murphy *et al*. 2003; Robinson *et al.*, 2012; see Eshel & Roiser, 2010 for review). Furthermore, depressed individuals show altered reward processing as demonstrated by signal-detection analyses (Pizzagalli *et al*. 2005; Pizzagalli *et al*. 2008), computational reinforcement learning approaches (Huys *et al*. 2009), and functional neuroimaging (Tremblay et al. 2005; Steele et al. 2007; Eshel & Roiser, 2010). These studies provide considerably evidence for altered learning and information processing in MDD and discuss whether MDD is better characterized by negative biases (i.e., increased responses to negative compared with positive stimuli) or by a blunting of responses (i.e., an insensitivity to both negative and positive stimuli) (Eshel & Roiser, 2010; Gotlib & Joorman, 2010). However, studies on information processing in MMD typically do not investigate domains in which healthy individuals tend to show positivity biases.

In the current study, we aimed at investigating cognitive biases in processing information about future life events. In contrast to healthy individuals (Sharot, 2011), MDD patients show a pervasive pessimism about their personal future (American Psychiatric Association, 2000). For example, when estimating the likelihood of experiencing positive and negative everyday life events (e.g. being invited to a party or getting a parking ticket) within the next month, individuals with high depressive symptoms expected less positive and more negative events than they eventually experienced while healthy participants showed the opposite pattern (Strunk *et al*. 2006; Strunk & Adler, 2009). Thus, previous studies have highlighted that depressed individuals are pessimistic when predicting their personal future (Cropley & MacLeod, 2003; Strunk *et al*. 2006; Strunk & Adler, 2009) but they do not address how information that challenges these views is incorporated into existing beliefs.

We have recently shown that healthy individuals maintain optimistic expectations as a result of selective updating, i.e. they process information about their personal future in a positively biased way (Sharot *et al*. 2011; Sharot *et al*. 2012a; Sharot *et al*. 2012b). Specifically, when healthy individuals estimated the probability of experiencing various adverse life events (e.g. robbery, Alzheimer's disease), and subsequently received information about how likely these events are to occur to persons living in the same sociocultural environment, they updated their beliefs more in response to desirable information that enforced optimism than to undesirable information that enforced pessimism. That is, participants changed their estimates more when the probabilities of the adverse events were lower than expected compared with when they were higher than expected, indicating a striking asymmetry in belief updating.

Here, we test the hypothesis that unlike healthy individuals, depressed individuals are characterized by a breakdown of a selective updating bias in response to information about their future. To that end, we recruited healthy and MDD participants and quantified their belief changes in response to receipt of both desirable and undesirable information about future life events.

## Method

### *Participants*

Participants were recruited via flyers at the Freie Universität Berlin and from patients at the Charité–Universitätsmedizin Berlin as well as the Schlosspark-Klinik Berlin. Participants were assessed for psychiatric disorders using a structured clinical interview (SCID-I; Wittchen *et al.* 1997) by a cognitive neuroscientist (CWK), who had been trained by a psychotherapist in conducting the SCID-I. For four in-patients an assessment provided by the referring clinical psychiatrist was used instead of the SCID-I. Participants completed the Beck Depression Inventory scale (BDI; Hautzinger *et al.* 1994). To relate task-related variables to participants' trait optimism, they also completed the Life Orientation Test-Revised (LOT-R; Scheier *et al.* 1994). To ensure that healthy controls and MDD patients were matched on IQ, all participants completed a test of verbal IQ (Wortschatztest, WST, a vocabulary test implemented in the HAWIE-R, the German adaptation of the Wechsler Adult Intelligence Scale; Schmidt & Metzler, 1992). Participants with a diagnosis of MDD were assessed on the Hamilton Depression Scale (HAMD-21; Hamilton, 1960) by a cognitive neuroscientist (CWK), who had been trained by a psychotherapist in assessing the HAMD-21. Since patients were only assessed by one person, inter-rater reliabilities could not be calculated. Additionally, to assess possible influences of participants' mood state on our task, they completed the multidimensional mood state questionnaire at the beginning and at the end of the study (Mehrdimensionaler Befindlichkeitsfragebogen, MDBF; Steyer *et al.* 1997). Participants gave informed consent and were paid. The study was approved by the Ethics Committee of the Charité-Universitätsmedizin Berlin.

We recruited two groups of participants: the healthy control group (n=19) included participants with no psychiatric disorders according to SCID-I, a BDI lower than 10 and no history of MDD. The MDD patients group (n=19) included participants with a diagnosis of MDD. One MDD patient was excluded because of a history of alcoholism (final n=18). Comorbid anxiety disorders in the MDD group were not excluded (n=8). Of the 18 MDD patients six patients received a single drug and seven two or more (8 took selective serotonin reuptake inhibitors, 4 selective serotonin and noradrenaline reuptake inhibitors, 3 bupropion, 2 tricyclic antidepressants, 2 pregabalin, 1 melperone, 1 lorazepam, 1 promethazine). See **Table 1** for demographics and clinical characteristics and scores on the mood state questionnaire.

*Stimuli*

Stimuli and task were adapted from our previous studies (Sharot *et al*. 2011; Sharot *et al*. 2012a; Sharot *et al*. 2012b). The original English task and stimuli were translated into German by a native German speaker with English as a second language (CWK). Seventy short descriptions of negative life events were used (e.g., Alzheimer's disease, robbery; see **Table 2** for a complete list of the original English and the translated German stimuli). For each adverse life event, the average probability or frequency of that event occurring at least once to a person living in the same sociocultural environment as the participants was determined based on online resources (e.g., Office for National Statistics, Eurostat, Pubmed). Since the probabilities of the events are *roughly* the same across Western Europe, we used the original event probabilities (from the UK). Participants were told that they would see the probability of the event happening to an average person of a similar background living in the same place. None of the participants reported doubts about this statistical information or about the believability of the reported events. Very rare and very common events were not included, i.e. all event probabilities lay between 10% and 70%. To ensure that the range of possible overestimation was equal to the range of possible underestimation, participants were told that the range of probabilities lay between 3% and 77%. We excluded life events that are clearly related to depressive symptoms such as severe insomnia or anxiety disorder.

*Task*

The task was programmed using the MATLAB toolbox Cogent 2000 (www.vislab.ucl.ac.uk/cogent.php). Participants completed four blocks of stimuli. The 70 adverse life events were split into two lists of 35 events each, which were matched for event probability. One list was used for blocks 1 and 2, the other for blocks 3 and 4. In blocks 1 and 3, participants first estimated their probability of encountering the events and then were presented with the average probability of the events for a demographically similar population (**Fig. 1**). To assess how participants used the information provided in block 1, they were asked to re-estimate their likelihood of encountering the events in block 2. Likewise, the likelihoods of events estimated in block 3 were re-estimated in block 4.

Estimates and average frequencies were framed as either “happening” or “not happening” in order to exclude the possibility that results could be attributed to different processing strategies for high or low numbers. Specifically, in half of the blocks (either blocks 1 and 2 or blocks 3 and 4) participants had to estimate the probability of adverse life events happening to them (and were presented with the probability of the event happening for a demographically similar population) while in the other half they estimated the probability of the events *not* happening to them (and were presented with the probability of the event *not* happening for a demographically similar population). List assignment and order of the aforementioned framing was counter-balanced across participants. Participants completed two training trials before blocks 1 and 3 to get familiar with the task and the change in framing.

On each trial participants were presented with one of the 70 adverse life events for 4 sec (**Fig. 1**) and were instructed to imagine that event happening to them in the future. Then they provided their estimate of how likely the event was to happen (or not to happen) to them in the future. Participants had up to 10 sec to respond using the keyboard. They then saw a fixation cross for 1.2 sec. In the blocks where participants indicated their first estimate (blocks 1 and 3), they were then presented with the average probability of the event happening (or not happening) for a demographically similar population for 3 sec followed by a fixation cross of 1.2 sec. In the blocks in which participants re-estimated their likelihood of encountering the events (blocks 2 and 4), they were not presented with the average frequencies of the events. The order of life events was randomized within each block.

*Memory and subjective scales*

After completing the four blocks of the main task we tested participants’ memory for the information presented. Specifically, we asked them to indicate the average probability of each event happening as previously presented in blocks 1 and 3 (self-paced). Participants then rated all stimuli on six subjective Likert scales (self-paced): vividness (How vividly could you imagine this event? 1 = not vivid at all, 6 = very vivid), familiarity (Regardless if this event has happened to you before, how familiar do you feel it is to you from TV, friends, movies and so on? 1 = not familiar at all, 6 very familiar), prior experience (Has this event happened to you before? 1 = never, 6 = very often), emotional arousal (When you imagine this event happening to you how emotionally arousing is the image in your mind? 1 = not arousing at all, 6 =

very arousing), negativity (How negative would this event be for you? 1 = not negative at all, 6 = very negative) and controllability (How much control do you have over this event? 1 = not at all, 6 = very much).

*Data analysis*

Data were analyzed using MATLAB and SPSS. All estimates and average frequencies in the 'not happen' sessions were transformed into the corresponding numbers of the 'happen' sessions by subtracting the respective number from 100. For each event an estimation error term was calculated as the difference between participants' first estimate and the corresponding average probability presented.

(1) estimation error = first estimate – probability presented

By this definition, estimation errors were positive for overestimations and negative for underestimations. Note that all life events were negative events and that the desirability of the information arose out of whether participants over- or underestimated the probability of the events. When participants initially overestimated the probability of the adverse event relative to the average probability they received desirable information (i.e. the negative event is less likely to happen than estimated; **Fig. 1b**). In contrast, when participants underestimated the probability of the event relative to the average probability they received undesirable information (i.e. the negative event is more likely to happen than estimated; **Fig. 1c**). Therefore, for each participant, trials were classified according to whether the participant initially overestimated or underestimated the probability of the event (i.e., according to whether estimation errors were positive or negative).

To assess how much participants changed their ratings after receiving information, an update term was calculated as the difference between the first and second estimates.

(2) update = first estimate – second estimate

We expected participants to change their estimates on average towards the information presented. That is, for desirable information (overestimations) first estimates should be larger than second estimates (i.e., mean updates for desirable information should be positive). For undesirable information (underestimations) first estimates should be smaller than second estimates (i.e., mean updates for undesirable information should be negative). The critical test for biased updating was whether participants changed their estimates numerically more (or less) towards desirable information than towards undesirable information. Therefore, we compared absolute mean updates for desirable and for undesirable information across participants. To exclude that differences in mean estimation errors across participants and conditions could account for differences in updating, we calculated scaled absolute mean update scores (i.e., we divided absolute mean update scores for each participant and condition by the respective absolute mean estimation errors).

Trials were excluded if (1) participants failed to answer within the allotted time (maximal 10s) in the first or second session (healthy controls: mean = 1.79, SD = 1.18; MDD patients: mean = 1.50, SD = 1.29) or (2) the estimation error was zero (i.e., participants gave a first estimate of their own likelihood that was as exactly the same as the presented probability; healthy controls: mean = 0.68, SD = 1.20; MDD patients: mean = 1.28, SD = 1.78). These trials were excluded because they could not be classified as desirable or undesirable. In both cases the number of excluded trials did not differ between healthy controls and MDD patients as assessed by Mann-Whitney U tests (all $p$>0.1).

To test the strength of association between estimation errors and updates, Pearson correlation coefficients were calculated separately for desirable and undesirable trials within each participant.

Memory errors were calculated as the absolute differences between the frequencies previously presented and participants' recollection of these statistical numbers.

(3) memory error = | probability presented – recollection of probability presented |

Unless, otherwise specified we conducted desirability (desirable/undesirable) by group (MDD/healthy) ANOVAs.

**Results**

Groups did not differ with regard to sex, age, education, and verbal IQ as confirmed by independent $t$-tests, Mann-Whitney U- tests, or $\chi^2$-tests as appropriate (all $p$>0.1; **Table 1**). MDD patients showed lower scores of trait optimism on the Life Orientation Test-Revised (LOT-R) compared with healthy controls [$t(35)=5.6$, $p<0.001$; **Table 1**]. MDD patients also reported having more negative mood, as well as being more tired and more agitated as assessed by significant main effects of group in group (MDD/healthy) by time (pre-task/post-task) ANOVAs of the multidimensional mood state questionnaire scores (**Table 1**). Additionally, there was a significant interaction of group and time on the subscale “awake-tired” of the mood state questionnaire: Pre-task the groups did not differ but post-task MDD patients were more tired than healthy controls.

*Comparison of updating behavior between MDD patients and healthy controls*

Our main hypothesis was that belief updating behavior would differ between MDD patients and healthy controls. We predicted healthy controls would show optimistically biased updating (i.e. we expected them to update their beliefs more in response to desirable than in response to undesirable information regarding adverse life events) and that this optimistic bias would be reduced, absent or reversed in MDD patients. In line with our hypothesis, there was a significant desirability (desirable/undesirable) by group (MDD/ healthy) interaction of absolute mean update scores [$F(1, 35)=6.9$, $p=0.013$, $\eta_p^2=0.17$; **Fig. 2a**; **Table 3**]. This interaction was characterized by an asymmetry in belief updating for healthy participants but not for MDD patients. Specifically, healthy participants updated their beliefs to a greater extent in response to desirable, compared with undesirable, information [$t(18)=3.0$, $p=0.008$]. No difference in updating was evident between desirable and undesirable trials in MDD patients [$t(17)=-0.11$, $p>0.9$]. The significant interaction was further characterized by reduced updating in response to desirable information of MDD patients relative to healthy controls [$t(35)=-2.2$, $p=0.033$] with no significant difference for updating in response to undesirable information [$t(35)=1.4$, $p>0.1$]. The main effect of desirability was significant [$F(1, 35)=7.4$, $p=0.010$, $\eta_p^2=0.17$]. The main effect of group was not significant [$F(1, 35)=0.73$, $p>0.3$, $\eta_p^2=0.02$)].

*Additional analyses: updating behavior*

To exclude the possibility that the observed difference in updating between MDD patients and healthy controls was driven by differences in estimation errors (first estimations minus probabilities presented), we performed an additional ANOVA on scaled absolute mean update scores (i.e., we accounted for differences in mean estimation errors by dividing absolute mean update scores for each participant and condition by the respective absolute mean estimation errors). The desirability by group interaction of scaled absolute mean update scores was significant [interaction desirability by group: $F(1, 35)=6.0$, $p=0.020$, $\eta_p^2=0.15$; main effect desirability: $F(1, 35)=0.67$, $p>0.4$, $\eta_p^2=0.02$; main effect group: $F(1, 35)=0.53$, $p>0.4$].

In our previous study on healthy participants, we analyzed the association between estimation errors and updates since formal learning models suggest that updates rely on error signals (i.e., the differences between expectations and outcomes) (Sharot *et al.* 2011). The strength of the association between estimation errors and updates is indicative of an optimistic bias in healthy individuals. Specifically, for desirable information estimation errors are more closely tied to updates than for undesirable information. Given these previous results in healthy individuals, we expected, similarly as for updates, a desirability by group interaction of the strength of the association between estimation errors and updates. For each participant, we calculated the correlation between estimation errors and updates separately for desirable and undesirable trials. There was a significant desirability by group interaction [mean Fisher-transformed Pearson correlation coefficients revealed an interaction of desirability by group: $F(1, 35)=4.19$, $p=0.048$, $\eta_p^2=0.11$; main effect desirability: $F(1, 35)=26.9$, $p=0.001$, $\eta_p^2=0.43$; main effect group: $F(1, 35)=2.96$, $p=0.094$, $\eta_p^2=0.08$; **Table 3**]. This further suggests that belief updating shows a less optimistic pattern in MDD patients compared with healthy controls.

*Relation of differential updating to MDD symptoms*

Next, we sought to test whether the update bias, i.e. the difference between updates for desirable versus undesirable information, was related to depressive symptoms as measured by BDI scores (**Fig. 2b**). BDI scores correlated negatively with the update bias across MDD patients (Pearson's $r=-0.50$, $p=0.036$), but no significant correlation emerged for healthy controls ($r=-0.001$, $p=0.996$), for which the range of BDI scores was limited. Thus, MDD patients with more severe symptoms showed less optimistic updating and more pessimistic updating (i.e. updating more in response to undesirable information compared with desirable information).

To elucidate whether group differences in update bias are potentially independent from depression severity, we tested whether group differences remain when partialling out BDI scores. We conducted a hierarchical regression analysis across all participants entering update bias as the dependent variable. As independent variables, we entered BDI scores on the first level and group membership on the second. BDI scores significantly predicted update bias [$F(1,35) = 8.41$, $p$=0.006] but group membership explained no additional variance beyond BDI scores [$F_{change}(1,34) = 0.00$, $p$>0.9], suggesting that differential updating is not independent of depression severity.

### *Additional analyses: relation of differential updating to participants' characteristics and task-related variables*

To test whether the update bias was related to demographic characteristics, mood states, task-related variables, or subjective scales, we conducted a step-wise linear regression analysis entering update bias as the dependent variable. As independent variables we entered group membership, BDI, and their interaction, along with age, gender, education, presence of comorbid anxiety disorder, medication status, verbal IQ, LOT-R scores, initial mood scores (**Table 1**), and differential measures (desirable minus undesirable) of memory errors, reaction times, estimation errors, as well as differential scores on all subjective scales (vividness, familiarity, prior experience, arousal, negativity, controllability, **Table 3**). The model that best predicted differential updating only included the interaction of group membership with BDI [$F(1,35) = 8.60$, $p$=0.006]. Demographic characteristics, mood states, task-related variables or subjective scales were not retained in the stepwise regression. That is, the update bias was influenced by whether participants were healthy or depressed and BDI predicted the update bias in MDD patients.

In addition, we specifically analyzed the relationship between update bias and LOT-R scores. In MDD patients, LOT-R scores correlated with the update bias (r=-0.52, $p$=0.027), but LOT-R scores did not explain additional variance beyond BDI in a hierarchical regression [$F_{change}(1, 15)$=1.107; $p$>0.3]. In healthy controls, LOT-R scores did not correlate significantly with the update bias (r=-0.19, $p$>0.4).

*Additional analyses: comparison of initial estimates between MDD patients and healthy controls*

In accord with the general pessimistic tendency of MDD patients, MDD patients initially estimated their overall probability of experiencing adverse life events as greater than healthy controls [(MDD patients: first estimate = 39.9 (8.20); healthy controls: first estimate = 31.5 (7.47); $t(35)=3.3$, $p=0.002$)]. MDD patients overestimated their probability of experiencing negative life events in relation to the average frequencies for a demographically similar population, i.e. mean estimation errors (first estimation minus probability presented) were positive and significantly different from zero [estimation error: mean = 10.4, SD = 8.18 ; $t(17)=5.4$, $p<0.001$]. This was not the case for healthy controls (estimation error: mean = 1.80, SD = 7.40; $t(18)=1.1$, $p>0.3$). Mean estimation errors differed between groups [$t(35)=3.3$, $p=0.002$].

Since MDD patients were more pessimistic overall compared with healthy controls, the number of overestimations (desirable trials) and the number of underestimations (undesirable trials) differed between the two groups. Specifically, there was a significant desirability (desirable/undesirable) by group interaction of the number of trials [interaction desirability by group: $F(1, 35)=13.6$, $p<0.001$, $\eta_p^2=0.28$; main effect desirability: $F(1, 35)=9.40$, $p=0.004$, $\eta_p^2=0.21$; **Table 3**]. For MDD patients the number of trials with desirable information was greater than the number of trials with undesirable information [$t(17)=4.6$, $p<0.001$] because they overestimated the probabilities more often than they underestimated them. For healthy controls the number of trials with desirable and undesirable information did not differ [$t(18)=-0.46$, $p>0.6$]. Furthermore, the interaction was characterized by MDD patients receiving desirable information on more trials and undesirable information on less trials than healthy controls [desirable: $t(35)=3.54$, $p=0.001$; undesirable: $t(35)=-3.80$, $p<0.001$]. There was no group difference in the overall number of trials [main effect group: $F(1, 35)=0.25$, $p>0.6$, $\eta_p^2=0.01$].

Taken together, even though MDD patients received more desirable information than healthy individuals (and thus had more opportunities to change their beliefs in an optimistic direction), MDD patients showed no evidence for optimistically biased updating.

*Additional analyses: memory, subjective rating scales, framing, and reaction times*

Differences in updating are not be explained by differences in memory for the presented probability, by differences in subjective ratings of events, or by differences related to the framing of the presented probability. Specifically, we asked participants to recollect the presented probability of the event happening and computed memory errors for each event as the absolute differences between participants' recollection and the probabilities presented. There was no significant desirability by group interaction for absolute memory errors [interaction desirability by group: $F(1, 35)=1.46$, $p>0.2$, $\eta_p^2=0.04$; main effect desirability: $F(1, 35)=0.52$, $p>0.4$, $\eta_p^2=0.02$; main effect group: $F(1, 35)=0.08$, $p>0.7$, $\eta_p^2=0.00$; **Table 3**]. Additionally, we asked participants to rate all negative events on six scales for vividness, familiarity, prior experience, emotional arousal, negativity, and controllability. There was no significant desirability by group interaction of any of these measures (all $p>0.1$; see **Table 3** for significant main effects).

To control for framing effects (effects due to information being presented in a positive or negative context) we asked participants to estimate how likely the events were to happen on half of the trials and how likely they were *not* to happen on the other half of the trials. In a frame (happen/not happen) by desirability by group ANOVA only the desirability by group interaction reached significance [$F(1, 34)=5.74$, $p=0.022$, $\eta_p^2=0.14$; all other main effects and interactions: $p>0.1$; one healthy participant had to be excluded from this analysis because there were no trials in the desirable-happen condition]. Thus, the framing of the estimates had no effect on updating behavior.

Reaction times for the first estimate did not show significant main effects of desirability or group (all $p>0.1$) but a significant interaction [$F(1, 35)=5.98$, $p=0.020$, $\eta_p^2=0.15$; **Table 3**]. The interaction was characterized by healthy controls being slower for estimates for which they subsequently received desirable information compared with estimates for which they subsequently received undesirable information [$t(18)=-0.46$, $p=0.006$]. This was not the case for MDD patients ($p>0.3$). Reaction times for the second estimate showed no significant main effects or interaction (all $p>0.5$; **Table 3**).

Nevertheless, differential measures (desirable minus undesirable) of memory errors, scores on all subjective scales, and reaction times were included in the step-wise linear regression analysis described above, which revealed that only the interaction of group membership with BDI was retained in the model that best predicted differential updating.

**Discussion**

We show an absence of optimistic bias in belief updating in depressed individuals and this absence correlated with their symptom severity. Healthy individuals updated their beliefs more when presented with desirable information about the likelihood of experiencing adverse life events relative to undesirable information. In contrast, updating from desirable versus undesirable information correlated with symptom severity in MDD patients: Less severely depressed individuals showed a positive bias but more severely depressed individuals showed a negative bias (i.e., they update more from undesirable compared with desirable information). Overall, this resulted in an absence of updating asymmetry across our sample of depressed individuals. Note that both groups were responsive to the information presented in the task and updated their estimates accordingly. The key difference between the groups was that while controls showed a valence-dependent updating bias, the MDD group did on average show an absence of this bias. This lack of biased updating in MDD patients was due to reduced updating from desirable information about the future.

The observed pattern was selective for updating behavior (i.e., memory for the information presented and subjective ratings of the events did not show interactions between the two groups and the desirability of the information). The current results in healthy individuals replicate our previous findings (Sharot *et al*. 2011; Sharot *et al*. 2012a; Sharot *et al*. 2012b) in a German sample and are in line with studies demonstrating that healthy individuals see emotionally laden future events through rose colored spectacles (Sharot *et al*. 2007; Szpunar *et al*. 2012).

In accord with cognitive theories of depression (e.g., Beck *et al*. 1979), depressed individuals exhibited a pessimistic view of the future evident in their inflated estimates of the probabilities of experiencing adverse events relative to controls and to the average probabilities of these events in the population. These results are also in line with previous studies (Strunk *et al*. 2006; Strunk & Adler, 2009) showing that depressed individuals expect more negative events and less positive events within the upcoming month than healthy controls. In our task, MDD patients received more desirable and less undesirable information because of their pessimistic views. Nevertheless, in contrast to controls, MDD patients did not take desirable information more into account than undesirable information but showed an absence of optimistically biased updating despite receiving more information that would warrant such an optimistic bias. It is possible that the more optimistic expectations of healthy compared with depressed individuals are a result of increased responsiveness to desirable information relative to negative information regarding the future; although cause and effect may also be reversed. Taken together, the current study showed that depressed individuals were characterized by pessimistic expectations and the absence of an optimistic updating bias.

Compared to many previous studies that have shown evidence for altered learning and feedback processing (Gotlib & Joorman, 2010; Eshel & Roiser, 2010), our study is more directly related to the research on positivity biases in healthy individuals. That is participants in our study received explicit information about the personal probability of future life events whereas participants in previous studies typically received outcomes in the form of performance feedback, reward, or punishment (e.g., Steele *et al.* 2007; Huys *et al*. 2009; Chase *et al.* 2010; Robinson *et al.* 2012). Using information about future life events, we show that the updating behavior of the MDD patients was less responsive to desirable information relative to controls, but similarly responsive to undesirable information. Therefore, the updating behavior of the MDD patients in our task seems to be better described by a lack of a positivity bias than by notions of general emotional blunting as discussed in previous studies (see Eshel & Roiser, 2010).

Testing whether the updating bias shown by healthy individuals is adaptive indeed is beyond the remit of the present study. However, the relative absence of optimistically biased updating in MDD needs to be considered in the context of previous research suggesting that positivity biases can be adaptive for mental and physical health as well as economic success (Taylor & Brown, 1988; Scheier & Carver, 1992; Taylor & Brown, 1994; Weinstein & Klein, 1995; Peterson, 2000; Armor & Taylor, 2002; Haselton & Nettle, 2006; Leary, 2007; Puri & Robinson, 2007; McKay & Dennett, 2009; Varki, 2009; Johnson & Fowler, 2011; Sharot, 2011). For example, all else being equal it seems that optimists live longer, recover faster from diseases (see Rasmussen *et al.* 2009 for review), and earn more (Puri & Robinson, 2007). This needs to be weighted by evidence that extreme optimists do engage in unhealthy and risky behavior such as smoking and failing to save for retirement (see Sharot, 2011 for review; Puri & Robinson, 2007). Mild to moderate positivity biases may exert an adaptive effect in at least three ways. Positive beliefs can reduce stress and anxiety (Solberg Nes & Segerstrom, 2006). They can enhance a motivation to obtain desired goals. For example, optimists exercise more and work harder (see Sharot, 2011 for review; Puri & Robinson, 2007). Furthermore, positive beliefs enhance exploratory behavior that can enhance individual and group success (see Sharot, 2011 for review). In the context of MDD, recent research suggests that depressed individuals benefit from therapy approaches that focus on inducing positive biases such as positive psychology interventions (Sin & Lyubomirsky, 2009) and cognitive bias modification (Hallion & Ruscio, 2011).

Future studies are needed to determine the generality of our findings in a larger sample that includes more male participants as well as determine possible influences of medication and comorbidity on the observed updating behavior. Importantly, we emphasize that the current study examines biased updating for negative but not for positive life events. Previous studies have shown that a similar updating bias exists in healthy individuals when they learn about positive stimuli (Eil & Rao, 2011; Möbius *et al.*, 2011; Wiswall & Zafar, 2013; Korn *et al.*, 2012), but whether such a lack of bias exists in MDD patients for positive events is an empirical question that needs to be tested.

The optimistic updating pattern described for healthy participants in our task might be associated with resilience to depression. Our study does not address whether altered information processing has a causal role in MDD. Longitudinal studies are needed to establish whether an absence of optimistically biased processing precedes the onset of depressive episodes and whether an increase in optimistic updating predicts treatment effects. Future studies will also be critical in examining whether techniques that enhance updating from desirable information and/or techniques that reduce updating from undesirable information might be beneficial in the treatment of depression. Our results provide a starting point for such investigations by suggesting that an absence of optimistically biased belief updating of information regarding future life events may be relevant for mental health.

## Acknowledgements

This study was supported by a Senior Investigator Award to RJD (098362/Z/12/Z), a Max Planck Award to RJD and a Wellcome Trust Career Development Fellowship to TS. The Wellcome Trust Centre for Neuroimaging is supported by core funding from the Wellcome Trust 091593/Z/10/Z. We thank F. Bermpohl, M. Adli, P. Sterzer and R. Schneibel for help with recruiting participants.

**Declaration of interest**

None.

## References

**Alloy LB, Ahrens AH** (1987). Depression and pessimism for the future: biased use of statistically relevant information in predictions for self versus others. *Journal of Personality and Social Psychology* **52**, 366-378.

**American Psychiatric Association** (2000). *Diagnostic and statistical manual of mental health disorders* (4th edition, text revision). American Psychiatric Association: Washington, DC.

**Armor DA, Taylor SE** (2002). When predictions fail: the dilemma of unrealistic optimism. In *Heuristics and Biases: The Psychology of Intuitive Judgment* (ed. T. Gilovich, D.W. Griffin and D. Kahneman), pp. 334-438. Cambridge University Press: New York.

**Beck AT** (2005). The current state of cognitive therapy: a 40-year retrospective. *Archives of General Psychiatry* **62**, 953-959.

**Beck AT, Dozois DJA** (2011). Cognitive therapy: current status and future directions. *Annual Review of Medicine* **62**, 397-409.

**Beck AT, Rush AJ, Shaw B, Emery G** (1979). *Cognitive therapy of depression.* Guilford Publications: New York.

**Chase HW, Frank MJ, Michael A, Bullmore ET, Sahakian BJ, Robbins TW** (2010). Approach and avoidance learning in patients with major depression and healthy controls: relation to anhedonia. *Psychological Medicine* **40**, 433-440.

**Clark L, Chamberlain SR, Sahakian BJ** (2009). Neurocognitive mechanisms in depression: implications for treatment. *Annual Review of Neuroscience* **32**, 57-74.

**Cropley M, MacLeod A** (2003). Dysphoria, attributional reasoning and future event probability. *Clinical Psychology and Psychotherapy* **10**, 220-227.

**Disner SG, Beevers CG, Haigh EAP, Beck AT** (2011). Neural mechanisms of the cognitive model of depression. *Nature Reviews Neuroscience* **12**, 467-477.

**Eil D, Rao JM** (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics* **3**, 114-138.

**Elliott R, Baker SC, Rogers RD, O'Leary DA, Paykel ES, Frith CD, Dolan RJ, Sahakian BJ** (1997). Prefrontal dysfunction in depressed patients performing a complex planning task: a study using positron emission tomography. *Psychological Medicine* **27**, 931-942.

**Elliott R, Sahakian BJ, Michael A, Paykel ES, Dolan RJ** (1998). Abnormal neural response to feedback on planning and guessing tasks in patients with unipolar depression. *Psychological Medicine* **28**, 559-571.

**Eshel N, Roiser JP** (2010). Reward and punishment processing in depression. *Biological Psychiatry* **68**, 118-124.

**Gotlib IH, Joormann J** (2010). Cognition and depression: current status and future directions. *Annual Review of Clinical Psychology* **6**, 285-312.

**Haaga DAF, Beck AT** (1995). Perspectives on depressive realism: implications for cognitive theory of depression. *Behaviour Research and Therapy* **33**, 41-48.

**Hallion LS, Ruscio AM** (2011). A meta-analysis of the effect of cognitive bias modification on anxiety and depression. *Psychological Bulletin* **6**, 940-958.

**Hamilton M** (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychology* **23**, 56-62.

**Haselton MG, Nettle D** (2006). The paranoid optimist: an integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review* **10**, 47-66.

**Hautzinger M, Bailer M, Worall H** (1994). *Beck-Depressions-Inventar (BDI) Bearbeitung der deutschen Ausgabe.* Huber: Bern.

**Huys QJM, Vogelstein JT, Dayan P** (2009). Psychiatry: insights into depression through normative decision-making models. In *Advances in neural information processing systems 21* (ed. D. Koller, D. Schuurmans, Y. Bengio, L. Bottou). MIT Press, Cambridge, Massachusetts.

**Johnson DDP, Fowler HF** (2011). The evolution of overconfidence. *Nature* **477**, 317-320.

**Korn CW, Prehn K, Park SQ, Walter H, Heekeren HR** (2012). Positively biased processing of self-relevant social feedback. *Journal of Neuroscience* **32**, 16832-16844.

**Leary MR** (2007). Motivational and emotional aspects of the self. *Annual Reviews of Psychology* **58**, 317-44.

**Mathews A, MacLeod C** (2005). Cognitive vulnerability to emotional disorders. *Annual Reviews of Clinical Psychology* **1**, 167-195.

**McKay RT, Dennett DC** (2009). The evolution of misbelief. *Behavioral and Brain Sciences* **32**, 493-561.

**Möbius MM, Niederle M, Niehaus P, Rosenblat TS** (2010). Managing self-confidence: theory and experimental evidence. *Working paper*.

**Moore MT, Fresco DM** (2012). Depressive realism: a meta-analytic review. *Clinical Psychology Review* **32**, 496-509.

**Murphy FC, Michael A, Robbins TW, Sahakian BJ** (2003). Neuropsychological impairment in patients with major depressive disorder: the effects of feedback on task performance. *Psychological Medicine* **33**, 455-467.

**Peterson C** (2000). The future of optimism. *American Psychologist* **55**, 44-55.

**Pizzagalli DA, Iosifescu D, Hallett LA, Ratner KG, Fava M** (2008). Reduced hedonic capacity in major depressive disorder: evidence from a probabilistic reward task. *Journal of Psychiatric Research* **43**, 76-87.

**Pizzagalli DA, Jahn AL, O'Shea JP** (2005). Toward an objective characterization of an anhedonic phenotype: a signal-detection approach. *Biological Psychiatry* **57**, 319-327.

**Puri M, Robinson DT** (2007) Optimism and economic choice. *Journal of Financial Economics* **86**, 71-99.

**Rasmussen HN, Scheier MF, Greenhouse JB** (2009). Optimism and physical health: a meta-analytic review. *Annals of Behavioral Medicine* **37**, 239-256.

**Robinson OJ, Cools R, Carlisi CO, Sahakian BJ, Drevets WC** (2012). Ventral striatum response during reward and punishment reversal learning in unmedicated major depressive disorder. *American Journal of Psychiatry* **196**, 152-158.

**Roiser JP, Elliott R, Sahakian BJ** (2012). Cognitive mechanisms of treatment in depression. *Neuropsychopharmacology* **37**, 117-136.

**Scheier MF, Carver CS** (1992). Effects of Optimism on Psychological and Physical Well-Being: Theoretical Overview and Empirical Update. *Cognitive Therapy and Research* **16**, 201-228.

**Scheier MF, Carver CS, Bridges MW** (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the life orientation test. *Journal of Personality and Social Psychology* **67**, 1063-1078.

**Scheier MF, Carver CS** (1992). Effects of optimism on psychological and physical well-being: theoretical overview and empirical update. *Cognitive Therapy and Research* **16**, 201-228.

**Schmidt K-H, Metzler P** (1992). *Wortschatztest (WST)*. Beltz Test GmbH: Weinheim, Germany.

**Seligman ME** (1972). Learned helplessness. *Annual Review of Medicine* **23**, 407-412.

**Sharot T** (2011). *The Optimism Bias*. Pantheon Books: New York, NY.

**Sharot T, Guitart-Masip M, Korn CW, Chowdhury R, Dolan RJ** (2012a). How dopamine enhances an optimism bias in humans. *Current Biology* **21**, 1477-1481.

**Sharot T, Kanai R, Marston D, Korn CW, Rees G, Dolan RJ** (2012b). Selectively altering belief formation in the human brain. *Proceedings of the National Academy of Sciences U S A* **109**:17058-17062.

**Sharot T, Korn CW, Dolan RJ** (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience* **4**, 1475-1479.

**Sharot T, Riccardi AM, Raio CM, Phelps EA** (2007). Neural mechanisms mediating optimism bias. *Nature* **450**, 102-105.

**Sin NL, Lyubomirsky S** (2009). Enhancing well-being and alleviating depressive symptoms with positive psychology interventions: a practice-friendly meta-analysis. *Journal of Clinical Psychology* **65**, 467-487.

**Solberg Nes L, Segerstrom SC** (2006). Dispositional Optimism and Coping: A Meta-Analytic Review. *Personality and Social Psychology Review* **10**, 235-251.

**Steele JD, Kumar P, Ebmeier KP** (2007). Blunted response to feedback information in depressive illness. *Brain* **130**, 2367-2374.

**Steyer R, Schwenkmezger P, Notz P, Eid M** (1997). *MDBF - Mehrdimensionaler Befindlichkeitsfragebogen*. Hogrefe: Göttingen, Germany.

**Strunk DR, Adler AD** (2009). Cognitive biases in three prediction tasks: a test of the cognitive model of depression. *Behaviour Research and Therapy* **47**, 34-40.

**Strunk DR, Lopez H, DeRubeis RJ** (2006). Depressive symptoms are associated with unrealistic negative predictions of future life events. *Behaviour Research and Therapy* **44**, 861-882.

**Szpunar K, Addis DR, Schacter DL** (2012). Memory for emotional simulations: remembering a rosy future. *Psychological Science* **23**, 24-29.

**Taylor SE, Brown JD** (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin* **103**, 193-210.

**Taylor SE, Brown JD** (1994). Positive illusion and well-being revisited – separating fact from fiction. *Psychological Bulletin* **116**, 21-27.

**Thompson SC, Armstrong W, Thomas C** (1998). Illusions of control, underestimations, and accuracy: a control heuristic explanation. *Psychological Bulletin* **123**, 143-161.

**Tremblay LK, Naranjo CA, Graham SJ, Herrmann N, Mayberg HS, Hevenor S, Busto UE** (2005). Functional neuroanatomical substrates of altered reward processing in major depressive disorder revealed by a dopaminergic probe. *Archives of General Psychiatry* **62**, 1228-1236.

**Varki A** (2009). Human uniqueness and the denial of death. *Nature* **460**, 684.

**Walker RW, Skowronski JJ, Thompson, CP** (2003). Life is pleasant—and memory helps keep it that way! *Review of General Psychology* **7**, 203-210.

**Weinstein ND** (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology* **39***,* 806-820.

**Weinstein ND, Klein WM** (1995). Resistance of personal risk perceptions to debiasing interventions. *Health Psychology* **14**, 132-140.

**Wiswall M, Zafar B** (2013). How Do College Students Respond to Public Information about Earnings? *Federal Reserve Bank of New York Staff Reports* **516**.

**Wittchen H-U, Zaudig M, Fydrich T** (1997). *SKID Strukturiertes Klinisches Interview für DSM-IV.* Hogrefe: Göttingen.

## Figure Legends

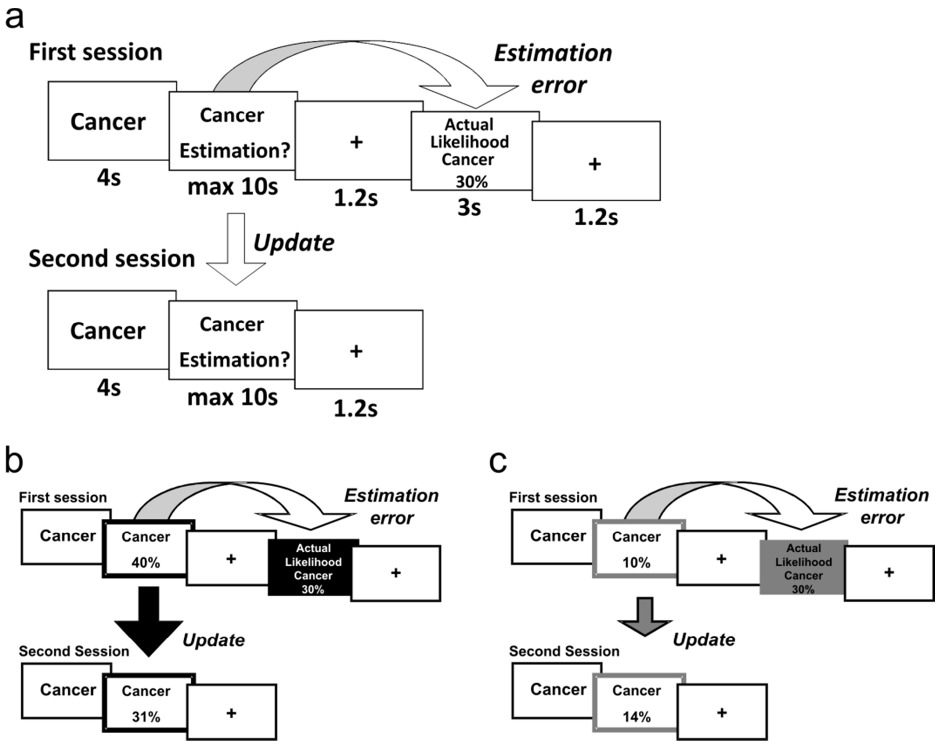


**Fig. 1. Paradigm**

(**a**) On each trial participants were presented with a short description of one of 70 adverse life events and asked to estimate how likely this event was to occur to them in their life time. They were then presented with the average probability of that event occurring to a person living in the same sociocultural environment. The second session was the same as the first session, except that the average probability of the event to occur was not presented again. For each event an update term was calculated as the difference between the participants' first and second estimations. **(b-c)** Examples of trials for which the participant's estimate was (**b**) higher or (**c**) lower than the average probability. Here, for illustration purposes, the thick black and grey frames denote the participant's response (either an overestimation or underestimation, respectively). The black and grey filled boxes denote information that calls for an adjustment in an (**b**) desirable (optimistic) or (**c**) undesirable (pessimistic) direction.

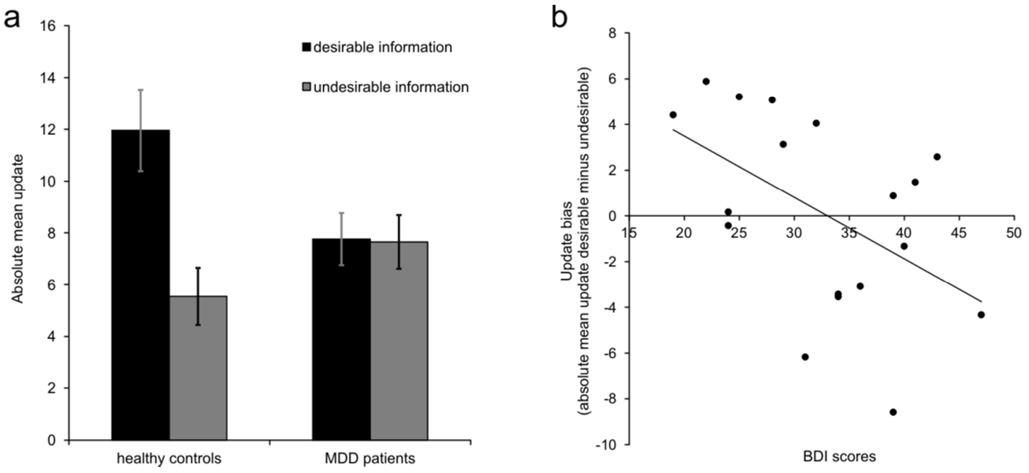


**Fig. 2. Updating behavior**

(a) In the healthy group absolute mean updates were greater on trials where participants received desirable information than on trials where they received undesirable information. This bias was absent in the MDD group.

(b) Relation between BDI scores and update bias (desirable minus undesirable) in MDD patients.

Error bars indicate SEM.

**Table 1. Demographic and clinical characteristics of participants; Multidimensional Mood State Questionnaire (MDBF, Mehrdimensionaler Befindlichkeitsfragebogen)**

| **Characteristic** | **Healthy controls** | **MDD patients** | **Significant effects (*p*<0.05)** |
|---|---|---|---|
| n | 19 | 18 | |
| Female sex | 15 | 12 | |

| | | | |
|---|---|---|---|
| Age, years | 26.5 (6.61) | 29.1 (7.06) | |
| Education, years | 17.1 (3.46) | 15.4 (3.25) | |
| Verbal IQ (WST) | 109.84 (7.42) | 105.5 (11.8) | |
| LOT-R | 16.6 (3.98) | 8.3 (4.89) | g |
| BDI | 4.3 (3.56) | 32.6 (7.96) | g |
| HAMD-21 | - | 24.7 (6.95) | |
| Comorbid anxiety disorder | 0 | 8 | |
| Medication | 0 | 13 | |
| Psychiatric hospitalization | 0 | 11 | |

| | **Healthy controls** | | **MDD patients** | | **Significant effects (*p*<0.05)** |
|---|---|---|---|---|---|
| **Mood state questionnaire*** | **Pre-task** | **Post-task** | **Pre-task** | **Post-task** | |
| Good mood – bad mood | 32.8 (5.94) | 33.4 (5.51) | 19.5 (6.79) | 20.3 (6.49) | g, pre, post |
| Awake – tired | 28.8 (6.41) | 29.0 (7.05) | 28.8 (6.41) | 20.3 (6.68) | g, i, post |
| Calm – agitated | 32. 2 (5.76) | 32.7 (5.40) | 19.8 (6.27) | 19.9 (6.03) | g, pre, post |

Data are given as mean (standard deviation) unless otherwise specified.

Verbal IQ (WST), Wortschatztest, a vocabulary test implemented in the HAWIE-R (German adaptation of the Wechsler Adult Intelligence Scale); LOT-R, Life Orientation Test-Revised; BDI, Beck depression inventory; HAMD-21, Hamilton Depression Scale.

* Data from two healthy controls were not collected post-task on the mood state questionnaire.

[g] Significant group difference or significant main effect: group ($p$<0.05).

[i] Significant interaction between time and group ($p$<0.05).

[pre] Significant difference between group in pre-task condition ($p$<0.05).

[post] Significant difference between group in post-task condition ($p$<0.05).

**Table 2. List of stimuli**

| **English original** | **German translation** |
|---|---|
| abnormal heart rhythm | Herzrhythmusstörungen |
| age related blindness | Altersblindheit |
| Alzheimer's disease | Alzheimer-Erkrankung |
| appendicitis | Blinddarmentzündung |
| arteries hardening (narrowing of blood vessels) | Arteriosklerose (Verkalkung der Blutgefäße) |
| artificial joint | künstliches Gelenk |
| asthma | Asthma |
| autoimmune disease | Autoimmunerkrankung |
| back pain | Rückenschmerzen |

| | |
|---|---|
| being cheated by husband/wife | Ehemann/Ehefrau geht fremd |
| being convicted of crime | für ein Verbrechen verurteilt werden |
| being fired | Gefeuert werden |
| bicycle theft | Fahrraddiebstahl |
| blood clot in vein | Thrombose |
| bone fracture | Knochenbruch |
| cancer | |
| (of digestive system/lung/prostate/breast/skin) | Krebserkrankung (Magen/Darm/Lunge/Prostata/Brust/Haut) |
| car stolen | Autodiebstahl |
| card fraud | Bank-/Kreditkartenbetrug |
| chronic high blood pressure | chronischer Bluthochdruck |
| chronic ringing sound in ear (tinnitus) | Tinnitus (Ohrgeräusche) |
| death before 60 | Tod vor dem 60. Lebensjahr |
| death before 70 | Tod vor dem 70. Lebensjahr |
| death before 80 | Tod vor dem 80. Lebensjahr |
| death by infection | Tod durch Infektion |
| dementia | Demenz |
| diabetes (type 2) | Diabetes (Typ 2) |
| disease of spinal cord | Erkrankung der Wirbelsäule |
| divorce | Scheidung |

| | |
|---|---|
| domestic burglary | Einbruch in Haus/Wohnung |
| drug abuse | Drogenabhängigkeit |
| epilepsy | Epilepsie |
| eye cataract (clouding of the lens of the eye) | Grauer Star (Linsentrübung) |
| fraud when buying something on the internet | Betrug bei Internetkauf |
| gallbladder stones | Gallensteine |
| genital warts | Genitalwarzen |
| gluten intolerance | Glutenunverträglichkeit |
| having a stroke | Schlaganfall |
| having fleas/lice | Flöhe/Läuse haben |
| heart failure | Herzversagen |
| hepatitis A or B | Hepatitis A oder B |
| hernia (rupture of internal tissue wall) | Eingeweide- oder Leistenbruch |
| herpes | Herpes |
| house vandalised | Haus/Wohnung wird mutwillig beschädigt |
| household accident | Haushaltsunfall |
| infertility | Unfruchtbarkeit |
| irritable bowel syndrome (disorder of the gut) | Reizdarm |
| kidney stones | Nierensteine |
| knee osteoarthritis (causing knee pain and swelling) | Kniearthrose |

| | |
|---|---|
| limb amputation | Amputation von Bein oder Arm |
| liver disease | Lebererkrankung |
| migraine | Migräne |
| more than £30000 debts | Schulden über 50,000 Euro |
| obesity | Fettleibigkeit |
| osteoporosis (reduced bone density) | Osteoporose (Knochenschwund) |
| Parkinson's disease | Parkinson-Erkrankung |
| restless legs syndrome | Syndrom der ruhelosen Beine |
| serious hearing problems | schwere Hörprobleme |
| severe injury due to accident (traffic or house) | schwere Verletzungen durch Unfall zu Hause oder im Straßenverkehr |
| severe teeth problems when old | schwere Zahnprobleme im Alter |
| skin burn | extremer Sonnenbrand |
| sport related accident | Sportunfall |
| theft from person | Opfer von Taschendieben |
| theft from vehicle | Diebstahl aus dem Fahrzeug |
| ulcer | Magen-/Darmgeschwür |
| victim of mugging | Opfer eines Überfalls auf der Straße |
| victim of violence at home | Opfer von häuslicher Gewalt |
| victim of violence by acquaintance | Opfer von Gewalt durch einen Bekannten |
| victim of violence by stranger | Opfer von Gewalt durch einen Fremden |

| | |
|---|---|
| victim of violence with need to go to A&E | Gewaltopfer mit Notaufnahmenaufenthalt |
| witness a traumatising accident | Zeuge eines traumatisierenden Unfalls |
| **Events used during the training sessions** | |
| dying before 90 | Tod vor dem 90. Lebensjahr |
| glaucoma | Grüner Star |

**Table 3. Task-related variables, subjective scales, memory, reaction times**

| | **Healthy controls** | | **MDD patients** | | **Significant effects** |
|---|---|---|---|---|---|
| | **Desirable** | **Undesirable** | **Desirable** | **Undesirable** | **(*p*<0.05)** |
| **Task-related variables** | | | | | |
| Number of trials | 32.9 (8.48) | 34.6 (8.23) | 43.0 (8.97) | 24.2 (8.54) | v, i, d, u |
| Number of trials excluded due to missed answers | | 0.68 (1.20) | | 1.28 (1.78) | |
| Number of trials excluded due to estimation errors of zero | | 1.79 (1.18) | | 1.50 (1.29) | |
| Updates | 11.96 (6.84) | 5.55 (4.81) | 7.76 (4.26) | 7.65 (4.37) | v, i, d |

| | | | | | |
|---|---|---|---|---|---|
| Estimation errors | 22.45 (5.65) | 17.92 (2.45) | 25.98 (6.66) | 17.86 (3.33) | v |
| Pearson correlation coefficients (Fisher-transformation): updates and estimation errors | 0.71 (0.31) | 0.28 (0.35) | 0.44 (0.26) | 0.26 (0.32) | v, i, d |
| Memory errors | 15.05 (6.26) | 13.62 (4.01) | 14.68 (7.39) | 15.04 (5.93) | |
| Reaction time 1st estimate, ms | 2483 (561) | 2325 (529) | 2434 (831) | 2520 (739) | i |
| Reaction time 2nd estimate, ms | 1993 (613) | 1917 (592) | 1944 (745) | 1874 (780) | |
| **Subjective scales: (1) low – (6) high** | | | | | |
| Vividness | 3.34 (0.73) | 2.92 (0.72) | 3.72 (0.57) | 3.10 (0.55) | v |
| Familiarity | 3.34 (0.721) | 2.90 (0.79) | 3.78 (0.54) | 3.42 (0.66) | v, g, d, u |
| Prior experience | 1.30 (0.19) | 1.16 (0.14) | 1.59 (0.29) | 1.37 (0.28) | v, g, d, u |
| Emotional arousal | 3.57 (1.00) | 3.40 (0.92) | 3.65 (0.85) | 3.53 (0.64) | v |
| Negativity | 4.26 (0.67) | 4.23 (0.64) | 4.06 (0.59) | 3.98 (0.67) | |
| Controllability | 2.86 (0.85) | 3.13 (0.77) | 2.62 (0.60) | 3.09 (0.63) | v |

Data are given as mean (standard deviation).

[v] Significant main effect: valence (*p*<0.05).

[g] Significant main effect: group (*p*<0.05).

[i] Significant interaction between valence and group (*p*<0.05).

[d] Significant difference between group in desirable condition (*p*<0.05).

[u] Significant difference between group in undesirable condition (*p*<0.05).