BEHAVIORAL AND BRAIN SCIENCES (2011) **34**, 1–56
doi:10.1017/S0140525X10001354

The evolution and psychology of self-deception

William von Hippel
School of Psychology, University of Queensland, St Lucia, QLD 4072, Australia
billvh@psy.uq.edu.au
http://www.psy.uq.edu.au/directory/index.html?id=1159

Robert Trivers
Department of Anthropology, Rutgers University, New Brunswick, NJ 08901
trivers@rci.rutgers.edu
http://anthro.rutgers.edu/index.php?option=com_content&task=view&id=102&Itemid=136

Abstract: In this article we argue that self-deception evolved to facilitate interpersonal deception by allowing people to avoid the cues to conscious deception that might reveal deceptive intent. Self-deception has two additional advantages: It eliminates the costly cognitive load that is typically associated with deceiving, and it can minimize retribution if the deception is discovered. Beyond its role in specific acts of deception, self-deceptive self-enhancement also allows people to display more confidence than is warranted, which has a host of social advantages. The question then arises of how the self can be both deceiver and deceived. We propose that this is achieved through dissociations of mental processes, including conscious versus unconscious memories, conscious versus unconscious attitudes, and automatic versus controlled processes. Given the variety of methods for deceiving others, it should come as no surprise that self-deception manifests itself in a number of different psychological processes, and we discuss various types of self-deception. We then discuss the interpersonal versus intrapersonal nature of self-deception before considering the levels of consciousness at which the self can be deceived. Finally, we contrast our evolutionary approach to self-deception with current theories and debates in psychology and consider some of the costs associated with self-deception.

Keywords: deception; evolutionary psychology; motivated cognition; self-deception; social psychology

Why would people deceive themselves? What is the mental architecture that enables the same person to be both deceiver and deceived? How does self-deception manifest itself psychologically? And how do these three questions interrelate? In this article we address these issues with an evolutionary account of self-deception, according to which self-deception evolved to facilitate deception of others. First, we define what we mean by self-deception and describe how self-deception serves the goal of interpersonal deception. We then discuss the non-unitary nature of the mind and how different types of psychological dualism enable the same person to be both deceiver and deceived. Next we describe different varieties of self-deception and the evidence for these different varieties. We then discuss the interpersonal versus intrapersonal nature of self-deception before considering the levels of consciousness at which the self can be deceived. Finally, we contrast our evolutionary approach to self-deception with current theories and debates in psychology.

1. Deception and self-deception

There are many ways to deceive other people. An obvious choice is to tell an outright lie, but it is also possible to deceive others by avoiding the truth, obfuscating the truth, exaggerating the truth, or casting doubt on the truth. Just as these processes are useful in deceiving others, they can also be useful in deceiving the self. For example, if I can deceive you by avoiding a critical piece of information, then it stands to reason that I can deceive myself in the same manner. Thus, we consider various types of self-deception, including biased information search strategies, biased interpretive processes, and biased memory processes. What marks all of these varieties of self-deception is that people favor welcome over unwelcome information in a manner that reflects their goals or motivations (in this sense, our approach to self-deception is consistent with Kunda 1990; Mele 1997; Pyszczynski & Greenberg 1987). We also consider classic cases of self-deception such as rationalization and convincing the self that a lie is true.

WILLIAM VON HIPPEL, professor of psychology at the University of Queensland, Australia, conducts research in social cognition and evolutionary psychology.

ROBERT TRIVERS, professor of anthropology at Rutgers University, is best known for his theories of reciprocal altruism, parental investment and sexual selection, parental control of offspring sex ratio, and parent–offspring conflict. Trivers is recipient of the Crafoord Prize for "his fundamental analysis of social evolution, conflict and cooperation."

 0140-525X/11 $40.00

Our approach consists of treating self-deception as a variety of different processes that are directly comparable to those involved in interpersonal deception. This approach has the advantage of tying self-deception to processes that have been studied in interpersonal deception. But it has the disadvantage that some behaviors are ambiguous concerning whether they should be classified as self-deception. This sort of ambiguity arises to the same degree, however, in the study of interpersonal deception. For example, consider a man who has returned home late from work because he stopped to talk to a female colleague and who is confronted by his wife, who wants to know why he is late. If he responds by claiming his boss gave him an extra assignment, then he is clearly being deceptive. If he responds by saying that he stopped to talk to this female colleague, then he is clearly being truthful. But if he changes the subject – perhaps by saying that dinner smells great – and thereby distracts his wife from her line of inquiry, then it is not clear whether he was deceptive or simply failed to answer the question. There is no way to know if he is deceiving his wife (by avoiding the truth) without knowing his intent in changing the subject or without knowing more about the relationships among him, his wife, and his colleague. In the same manner, when people avoid unpleasant truths in their own lives, it is often impossible to know if self-deception or some other process is at work without knowing more about their motives or situations. Nevertheless, just as interpersonal deception can be studied in the absence of complete confidence in all classifications, self-deception can also be studied despite the inevitability of such ambiguous cases.

Our approach of treating self-deception as information-processing biases that give priority to welcome over unwelcome information also differs from classic accounts that hold that the self-deceiving individual must have two separate representations of reality, with truth preferentially stored in the unconscious mind and falsehood in the conscious mind (see Gur & Sackeim 1979). As is clear in our review of information-processing biases in section 5, people can deceive themselves by preventing unwanted information from being encoded in the first place. This act can be demonstrably motivational, for example, if people stop gathering information when they like the early returns but keep gathering more information when they do not (Ditto & Lopez 1992). The flexible nature of this information-gathering bias also reveals that people have some awareness that upcoming information may be inconsistent with what they have already discovered. Thus, biases of this sort are consistent with classic definitions of self-deception that emphasize simultaneous knowing and not-knowing, in the sense that the individual consciously knows the welcome information that has been gathered but also has some awareness that unwelcome information could be around the next corner (we refer to this as *potential awareness*). In this case, however, true *knowing* of unwelcome information is precluded because the individual ends the information search before anything unwelcome is ever encountered. Thus, the individual need not have two representations of reality to self-deceive. Rather, people can self-deceive in the same way that they deceive others, by avoiding critical information and thereby not telling (themselves) the whole truth.

It is important to note, however, that not all biases in information processing are self-deceptive. For example, biases can reflect cognitive shortcuts, errors, and differential weighting of prior and new information that have nothing to do with motivational concerns (e.g., Chambers & Windschitl 2004). From our perspective, biases in information processing can be considered self-deceptive only when they favor welcome over unwelcome information in a manner that reflects the individual's goals. For example, if bolstering a person's self-image makes that individual more willing to search out negative information about himself (Trope & Neter 1994), we have some evidence that prior avoidance of negative information about the self was motivated and self-deceptive. Similarly, if people spend more time learning about the negative than the positive features of someone else only when they expect that person will reject them (Wilson et al. 2004), we again have some evidence that focusing on the negative aspects of another was motivated and self-deceptive. But if biases are impervious to such manipulations or irrelevant to such concerns, then the evidence suggests that they are not self-deceptive. For example, people often rely on peripheral cues such as source expertise when processing arguments about issues that do not concern them (Petty & Cacioppo 1986). Such biases do not reflect self-deception, but rather are simply evidence that these individuals are insufficiently motivated to study the arguments carefully and are satisfied to rely on a heuristic that experts tend to be correct. We return to this issue of differentiating self-deceptive from non-self-deceptive biases in section 5.

2. Self-deception in the co-evolutionary struggle between deceiver and deceived

> If (as Dawkins argues) deceit is fundamental in animal communication, then there must be strong selection to spot deception and this ought, in turn, select for a degree of self-deception, rendering some facts and motives unconscious so as to not betray – by the subtle signs of self-knowledge – the deception being practiced. Thus, the conventional view that natural selection favors nervous systems which produce ever more accurate images of the world must be a very naïve view of mental evolution. (Trivers 1976/2006, p. 20)

In the struggle to accrue resources, a strategy that has emerged over evolutionary time is deception. For example, people frequently lie to those on whom they are dependent to receive resources that might not otherwise be provided (DePaulo & Kashy 1998; Steinel & De Dreu 2004). Indeed, approximately half of people's daily deceptions are intended to gain a resource for the self (DePaulo & Kashy 1998). Such deceptive practices instigate a co-evolutionary struggle, because selection favors the deceived evolving new means of detection and the deceiver evolving new means of deception. Self-deception may be an important tool in this co-evolutionary struggle, by allowing deceivers to circumvent detection efforts (Trivers 1976/2006; 1985; 2000; 2009).

In the case of deception among humans, there are at least four general categories of cues (beyond fact-finding itself) that people can use to detect deception in others: Signs of nervousness, suppression, cognitive load, and idiosyncratic sources. Nervousness typically results from consideration of the potential costs of detected deception, and thus cues to nervousness can reveal deception (DePaulo et al. 2003). Nevertheless, nervousness is an

imprecise indicator of deception, in part because many situations induce nervousness independent of deception – including interrogations themselves – as people are often nervous that they will be falsely accused of deception (Bond & Fahey 1987). For this reason, reliance on nervousness to detect deception can lead to a high rate of false positives.

Deception can also be detected via physical indicators of the act of suppression. For example, in an effort to control nonverbal signs of nervousness that might reveal deceptive intent, people try to control their face, trunk, and limbs. This act of muscle control leaves telltale cues, such as increased vocal pitch (DePaulo et al. 2003), which then demand countervailing efforts at muscle relaxation. Deception can be detected when control efforts are disjointed or when efforts at relaxation fail to eliminate signs of suppression.

Cognitive load results when people must maintain two types of content simultaneously in working memory. In the case of consciously mediated deception, this means true information to be suppressed and false information to be promulgated. Deception can be detected by cues associated with cognitive load, such as pausing and simplified sentence structure (Vrij 2004; Vrij et al. 2006). People who are familiar with each other's habits can also detect deception via idiosyncratic signs of nervousness, suppression, and cognitive load, because different individuals reveal their mental states in different ways.

Despite the cues that are available for the detection of deception, research on lie detection suggests that people perform poorly in detecting the deceptions of others. For example, in their review of research conducted with professional lie detectors (e.g., police officers), Vrij and Mann (2005) found an overall detection rate of 55%, barely above chance levels of 50% in these studies. This detection rate is very similar to the 54% that emerged in a huge meta-analysis that included various types of untrained deceivers and detectors (Bond & DePaulo 2006). Many have concluded from these studies that humans are poor lie detectors, which would suggest that the selection pressure on deceptive abilities must be weak. But this conclusion is premature, given that research on detection of deception relies extensively on four conditions that heavily advantage the deceiver.

First, most studies involve deception that is of little consequence to the deceiver. As a result, cues associated with nervousness, cognitive load, and suppression are minimized by the fact that unimportant deceptions do not induce these telltale signs (Mann & Vrij 2006; Vrij 2000; Vrij & Mann 2005). Consistent with this possibility, when police officers evaluated videotapes of corroborated truths and lies told in actual criminal interrogations, accuracy increased to 72% (Vrij & Mann 2005), and signs of cognitive load and suppression appeared to differentiate liars from truth tellers (Mann & Vrij 2006). Meta-analysis also supports this conclusion, given that several cues to deception (e.g., vocal pitch) are more pronounced when people are more motivated to deceive (DePaulo et al. 2003).

Second, most studies do not allow the deceived to question the deceiver. To enhance experimental control and minimize costs, the deceiver is typically presented on videotape, thereby eliminating the possibility of further interaction. This lack of cross-examination minimizes cognitive load, nervousness, and the need for suppression, as the rehearsed lie can be much easier to deliver than a spontaneous answer to an unexpected question.[1]

Third, a substantial portion of the literature on detection of deception has been conducted with the goal of finding cues that reliably identify most deceivers, perhaps because of the potential applied value of general cues to deception. Thus, idiosyncratic signs of cognitive load, suppression failure, or nervousness tend to be missed in such designs.

Fourth, most studies on detection of deception are conducted on people who are unknown to each other. Our ability to detect deception may be poorly suited for this task, because this design feature eliminates the possibility that people can learn to use idiosyncratic cues that might help them determine when a particular individual is lying (DePaulo 1994; Zuckerman et al. 1984). Consistent with this possibility, detection rates increase when people get to know each other (if they feel emotionally close with each other; Anderson et al. 2002) and when lies are told among close friends (DePaulo & Kashy 1998). Such effects are likely to be multiply mediated, but they suggest that cross-examination and idiosyncratic cues could be important in detecting deception.

Consistent with the notion that these four factors reduce detection rates in laboratory studies, diary research suggests that people detect deception at rates that are substantially greater than chance. For example, participants in DePaulo et al.'s (1996) diary study reported that 15–23% of their lies were detected. This rate is already high enough to pose a significant threat – given the loss in reputation and goal attainment when one is perceived as a liar – but DePaulo et al.'s (1996) participants also reported that in an additional 16–23% of the cases they were unsure if their lies were detected.

We suggest that these reports are biased downward, such that people underreport the degree to which others have or might have detected their own deceptions. This downward bias in reporting is likely to emerge due to an asymmetry in the degree to which people accuse others of deception. On the one hand, targets of intended deceptions do not always indicate that they doubt the deceiver even when they do (Jang et al. 2002). To gain an information advantage and to maintain harmony, people who detect or suspect deception in others sometimes pretend to believe them. On the other hand, it seems unlikely that people indicate that they doubt deceivers when they actually believe them, in part because accusations of deception are a serious charge. This asymmetry suggests that people may often think that others believe their deception when in fact they do not, but will only rarely think that others doubt their veracity when in fact they have accepted their lies.[2] Thus, it might be the case that not only do the brunt of these 30–40% of cases reported in DePaulo et al.'s (1996) diary study represent detected deceptions, but a portion of the remaining 60–70% might represent detected deceptions as well. For our purposes, it would also be very useful to know what percentage of these detected deceptions emerged from fact-finding and third-party information, and what percentage were based on information derived directly from the deceiver (e.g., via cross-examination). At this point this remains a topic for future research.

In sum, the literature on deception may have grossly underestimated people's ability to detect deception through

reliance on studies where (a) the deception is of little or no consequence, (b) the deceived has no opportunity to cross-examine the deceiver, (c) deceiver and deceived are strangers to each other, and (d) there are no repeated interactions between deceiver and deceived. If rates of deception detection are, in fact, substantially higher outside the laboratory than in it, we are led back to the notion of a co-evolutionary struggle between deceiver and deceived. Because successful deception can lead to substantial benefits for the deceiver while imposing costs on the deceived (DePaulo 2004), and because unsuccessful deception can lead to substantial costs imposed on the deceiver by people they had endeavored to deceive (Boles et al. 2000; Schweitzer et al. 2006), those who would deceive are in a perennial struggle against those who would not be deceived. We propose that self-deception offers an important tool in this co-evolutionary struggle by allowing the deceiver the opportunity to deceive without cognitive load, conscious suppression, increased nervousness, or idiosyncratic indicators that a deception is being perpetrated. To the degree that people can convince themselves that a deception is true or that their motives are beyond reproach, they are no longer in a position in which they must knowingly deceive others. Thus, *the central proposal of our evolutionary approach to self-deception is that by deceiving themselves, people can better deceive others, because they no longer emit the cues of consciously mediated deception that could reveal their deceptive intent.*

Corollary 1: Cognitive load reveals deception, but it has other costs as well: demands on working memory reduce performance in challenging domains (Schmader & Johns 2003) and disrupt social functioning (von Hippel & Gonsalkorale 2005). When people are forced to maintain both truth and lies in working memory, they are thus likely to show reduced ability to engage in other tasks and access other opportunities. This cognitive load required to maintain conscious deception is difficult to avoid, because many deceptions require the deceiver to keep fact and fiction in mind simultaneously to ensure that the former is hidden and the latter is supported.

Self-deception provides a way to avoid this cognitive load. To the degree that deceivers can convince themselves that their deception is indeed true, they are no longer required to maintain the real facts of the case in mind while they focus on promulgating the fiction. Rather, by believing the fiction that they are expressing to others, they can free their mind to concentrate on other matters. Thus, *the first corollary to our central proposal is that by deceiving themselves, people are able to avoid the cognitive costs of consciously mediated deception*.

Corollary 2: The best-laid plans often go awry, and lies are no exception to this rule; even careful and well-practiced deceptions can be discovered. This ever-present possibility of detection poses a problem for would-be deceivers, as retribution and exclusion are common responses to detected deceptions (Boles et al. 2000; Schweitzer et al. 2006). This retribution appears to have deep evolutionary roots, given that people react with strong feelings of anger and other negative emotions when they realize they are being deceived (Haselton et al. 2005). Such feelings of anger motivate punishment of the offender, even when punishment exacts a cost for the punisher (Fehr & Gächter 2002). It is this anger and subsequent punishment that ensures that detection of deception leads to suffering on the part of deceivers, thereby reducing the likelihood of future deception.

Because there are many legitimate reasons people may fail to behave as desired or expected, one solution to the threat of retribution when an apparent deception is uncovered is to co-opt such reasons by pleading ignorance or ineptitude rather than deception. Attribution of intent is critical in determining whether the deceived feels anger and seeks retribution or is willing to forgive (Schweitzer et al. 2006; Stouten et al. 2006), and thus those who accompany their deception of others with deception of the self are better placed if discovered to avoid retribution. By arguing that they had not intentionally deceived, self-deceivers are more likely than conscious deceivers to avoid retribution. Of course, conscious deceivers can also deceive about their original knowledge and intent, but the cues that are present to give away deception are also present to give away deception about deception. Thus, *the second corollary to our central proposal is that by deceiving themselves, people can reduce retribution if their deception of others is discovered*.

Implication: Our evolutionary hypothesis about the role of self-deception in service of interpersonal deception hinges primarily on the argument laid out earlier – that despite what the research literature might appear to show, people are actually quite good at detecting deception. This possibility is central to our hypothesis regarding a co-evolutionary struggle and the subsequent origins of self-deception. Just as importantly, this possibility is also central to a proper understanding of deception and its place in social life. Thus, a critical issue for future research in deception and self-deception would be to establish how successful people are at detecting important deceptions occurring in naturalistic settings that allow those who are being deceived to gather further information as they see fit. Future research should also consider whether those who deceive in such settings gain an accurate sense of when their deceptions were successful versus unsuccessful, because there are likely to be situations that lead to clear feedback and others that do not. Although studies such as these are likely to be challenging to design and execute, methodologies involving simultaneous experience sampling within groups of closely interacting individuals may be a promising approach. But whatever the hurdles, it should be clear that research on deception must move beyond the closely controlled studies of the sort described earlier if it hopes to answer these fundamental questions about deception and its detection.

3. Self-deception in service of social advancement

Self-deception can also facilitate the deception of others in a more general sense, in that it can help us convince others that we are better (e.g., more moral, stronger, smarter) than we really are. Thus, the benefits of self-deception go beyond convincing others of specific lies, as self-deception can also help us accrue the more general social advantages of self-inflation or self-enhancement.

People are impressed by confidence in others (Price & Stone 2004; Slovenko 1999). Confidence plays a role in determining whom people choose as leaders (Conger &

Kanungo 1987; Shamir et al. 1993), romantic partners (Buss 2009; Miller 2000; Schmitt & Buss 1996), and providers of various social and material services (Brown et al. 1998; de Jong et al. 2006; Westbrook 1980). Confidence is also a determinant of social influence; confident people are believed more, and their advice is more likely to be followed than people who lack in confidence (Penrod & Cutler 1995; Zarnoth & Sniezek 1997). To the degree that people can bolster their image of themselves to themselves and enhance their self-confidence, they thereby increase the chances that they will be able to influence others and will be chosen for socially important roles. For this reason, self-enhancement should be ubiquitous and people should believe their own self-enhancing stories. Evidence supports both of these possibilities.

With regard to ubiquity, self-enhancing biases are evident in a wide variety of domains and strategies among a wide variety of peoples (for a review, see Alicke & Sedikides 2009). Even East Asians, who value humility and harmony over individualistic self-aggrandizement, show self-enhancement in their claims of the superiority of their collectivist qualities (Sedikides et al. 2003; 2005). Furthermore, like Westerners, East Asians who are lower in depression and stress show this self-enhancement to a greater degree than those who have more psychological problems (Gaertner et al. 2008).

People not only self-enhance the world over, but the average person appears to be convinced that he or she is better than average (Alicke & Sedikides 2009). Most of the research on self-enhancement does not allow one to assess whether these aggrandizing tales are self-deceptive or only intended to be other-deceptive, but some of the variables used in this research support the idea that people believe their own self-enhancing stories. For example, in a pair of clever experiments Epley and Whitchurch (2008) photographed participants and then morphed these photographs to varying degrees with highly attractive or highly unattractive photos of same-sex individuals. Epley and Whitchurch then presented participants with these morphed or unaltered photos of themselves under different circumstances. In one experiment participants were asked to identify their actual photo in an array of actual and morphed photographs of themselves. Participants were more likely to choose their photo morphed 10% with the more attractive image than either their actual photo or their photo morphed with the unattractive image. This effect emerged to a similar degree with a photo of a close friend, but it did not emerge with a photo of a relative stranger. Because people often perceive their close friends in an overly positive light (Kenny & Kashy 1994), these findings suggest that people do not have a general bias to perceive people as more attractive than they really are, but rather a specific bias with regard to themselves and close others.

In a second study, participants were presented with an array of photos of other individuals, among which was a single photo of themselves (either their actual photo or a photo morphed 20% with the attractive or unattractive image). Epley and Whitchurch (2008) found that people were able to locate photographs of themselves most rapidly if they were morphed with an attractive photo, at an intermediate speed if they were not morphed, and most slowly if they were morphed with an unattractive photo. These findings suggest that the enhanced photo most closely matches how people see themselves in their mind's eye, suggesting that they are deceiving themselves about their own attractiveness. Were they aware of this inaccuracy, they would be unlikely to claim the attractive photo to an experimenter who has the truth at her disposal, and they would be unlikely to locate their more attractive self more rapidly than their actual self.

Thus far, the evidence from Epley and Whitchurch (2008) suggests that self-enhancement biases appear to be inaccuracies that are believed by the self. But are they really self-deceptive? That is, do we have any evidence that the participants in their experiments have unconscious knowledge of what they really look like, or that they have prevented themselves from gaining accurate self-knowledge via motivated processes? At this point we do not. But it is the case that the magnitude of the self-enhancement effect documented by Epley and Whitchurch (2008) was correlated with participants' implicit self-esteem (as measured by a priming procedure [Spalding & Hardin 1999] and the name letter effect [Nuttin 1985]). This correlation suggests that people systematically distort their self-image not as a function of how much information they have about themselves, but rather as a function of their automatic positivity toward themselves. Nevertheless, the utility of this measure as an indicator of self-deception would be enhanced if it were shown to relate to people's goals or their biases in encoding of information about their appearance (e.g., differential gazing at flattering vs. unflattering images of themselves).

Along with their self-enhancing stories, people also derogate others. Indeed, self-enhancement and other-derogation are opposite sides of the same coin, as people arrive at their self-image via social comparison (Festinger 1954). For this reason all self-evaluations are relative, and the self can be elevated above others either via self-enhancement or other-derogation. Like self-enhancement, derogation of others can also be an offensive tool used in the service of social advancement, as people often derogate their rivals when they are trying to impress (Buss & Dedden 1990; Schmitt & Buss 2001).

As with self-enhancement, derogation of others appears to be both ubiquitous and believed by the self. Some of the best examples of self-deceptive derogation of others can be found in the research of Fein and Spencer. In one of their studies (Fein & Spencer 1997), non-Jewish participants were told either that they did well or poorly on an IQ test. They were then given a chance to watch a videotape of another student being interviewed for a job, and this individual was portrayed as either Jewish or Christian via her ostensible surname and a photo showing her wearing a Jewish star or a cross. When participants watched the individual they believed was Christian, their ratings of this student were not influenced by whether they had ostensibly failed the IQ test. In contrast, when participants thought they were watching a Jewish student, their ratings were influenced by how well they had done on the IQ test. Those who thought they had done well showed no sign of prejudice. Those who thought they had done poorly, however, rated the Jewish student negatively on a composite of personality traits. Furthermore, these individuals also showed a rebound in self-esteem compared to people who thought they had failed but watched the Christian woman. These findings suggest that people

responded to failing an IQ test by denigrating Jews. This denigration appears to have been believed by these individuals, because the more they derogated the Jewish student, the better they felt about themselves. Furthermore, people appeared to have objective information about this person at their disposal, given that they did not rate the Jewish person negatively when portrayed as Christian or when they had not ostensibly failed the test.

The research of Fein and Spencer (1997) suggests that people do not derogate others only to make themselves look better in other people's eyes. Rather, they appear to be self-deceptively making themselves look better in their own eyes as well. This interpretation is corroborated by the findings of their second experiment, in which participants who had reflected on their important values (a process that affirms a sense of self-worth: see sect. 5) evaluated the Jewish student the same as they rated the Christian student even after failure. This sort of threat-induced derogation of others documented by Fein and Spencer (1997) can also take place outside of awareness (Spencer et al. 1998), a finding that further suggests that people truly believe their negative impressions of others when they are led to feel bad about themselves. Because downward social comparison is self-enhancing (Wills 1981), these findings can be considered the flip-side of the bias documented by Epley and Whitchurch (2008). Thus, *our second proposal is that by deceiving themselves about their own positive qualities and the negative qualities of others, people are able to display greater confidence than they might otherwise feel, thereby enabling them to advance socially and materially*.

4. Self-deception and the non-unitary mind

There are a variety of dissociations between seemingly continuous mental processes that ensure that the mental processes that are the target of self-deception do not have access to the same information as the mental processes deceiving the self. For our purposes, these dissociations can be divided into three (overlapping) types: implicit versus explicit memory, implicit versus explicit attitudes, and automatic versus controlled processes. *These mental dualisms do not themselves involve self-deception*, but each of them plays an important role in enabling self-deception. By causing neurologically intact individuals to split some aspects of their self off from others, these dissociations ensure that people have limited conscious access to the contents of their own mind and to the motives that drive their behavior (cf. Nisbett & Wilson 1977). In this manner the mind circumvents the seeming paradox of being both deceiver and deceived.

4.1. Explicit versus implicit memory

Substantial research now indicates that people maintain at least two types of information in memory. People retain information that they can consciously recollect (assessed via explicit measures such as recall) and information for which they have no conscious recollection (assessed via implicit measures such as degraded word identification). This dissociation between types of memories has the potential to serve as a basis for an adaptive form of self-deception, as the conscious memories could be those that are consistent with the fiction that the person wishes to promulgate, whereas the unconscious memories could be the facts as originally encountered. By maintaining accurate information in unconscious memory, the individual would retain the ability to behave in accordance with the truth, as implicitly assessed memories have been shown to influence a variety of behaviors (Coates et al. 2006; Kolers 1976; Lee 2002; Seamon et al. 1995).

How might memories selectively retain or be blocked access to consciousness as a function of their utility in self-deception? Although there are probably numerous ways, one possibility is that deception might begin to replace truth in conscious memory simply through the act of reporting the misinformation (via retrieval-induced forgetting; MacLeod & Saunders 2008). Additionally, because the consequences of being caught lying can be severe, and because lies require important deviations from or omissions to actual events, those who practice to deceive often take their practice literally and (at least mentally) rehearse their lies (Vrij 2000). Rehearsal of misinformation can make the source of this information even more difficult to ascertain, with the result that people often come to believe that the inaccurate depiction of events is veridical (Ceci et al. 1994; Zaragoza & Mitchell 1996). According to this possibility, people initiate a deception knowing that they are promulgating a falsehood, but beginning with their initial transmission of the misinformation, they start to unknowingly convince themselves of the truth of their own lie.

This process of self-inducing false memories can then be enhanced by other factors that typically accompany intentional deception of others. For example, to the degree that deceivers create an elaborate and concrete image of the lie they are telling, they may unintentionally exacerbate their self-deception, because vivid imagining makes false memories more difficult to distinguish from accurate ones (Gonsalves et al. 2004; Slusher & Anderson 1987). Additionally, social sharing of information can lead to selective forgetting of information that is not discussed (Coman et al. 2009; Cuc et al. 2007) and social confirmation of inaccurate information can exacerbate the false memory effect (Zaragoza et al. 2001). These social effects raise the possibility that when people collaborate in their efforts to deceive others, they might also increase the likelihood that they deceive themselves. Thus, one consequence of retrieving, rehearsing, and telling lies is that people may eventually recollect those lies as if they actually happened, while still maintaining the accurate sequence of events in a less consciously accessible form in memory (Chrobak & Zaragoza 2008; Drivdahl et al. 2009; McCloskey & Zaragoza 1985). Therefore, *our third proposal is that the dissociation between conscious and unconscious memories combines with retrieval-induced forgetting and difficulties distinguishing false memories to enable self-deception by facilitating the presence of deceptive information in conscious memory while retaining accurate information in unconscious memory*.

4.2. Explicit versus implicit attitudes

Just as memories can be relatively inaccessible to consciousness, so too can attitudes. And just as inaccessible memories can influence behaviors, so too can attitudes

that are relatively inaccessible to consciousness (Greenwald et al. 2009; Nock et al., 2010). Attitudes that are difficult to access consciously can be measured with implicit procedures, for example, via reaction time tests such as the Implicit Association Test (IAT; Greenwald et al. 1998). The study of implicit attitudes is not as well developed as the parallel research program in the domain of memory, but here again the evidence indicates that people maintain two different types of attitudinal information (Fazio & Olson 2003; Wilson et al. 2000). Sometimes implicit and explicit attitudes overlap substantially, and sometimes they differ dramatically (Nosek et al. 2007). Although dissociations between implicit and explicit attitudes tend to be more common in cases where implicit attitudes are socially undesirable, such dissociations emerge across a variety of domains (Hofmann et al. 2005). As is the case with memory, the coexistence of different implicit and explicit attitudes provides fertile ground for self-deception.

An example of how dual attitudes play a role in self-deception can be found in research that shows that socially undesirable implicit attitudes drive behavior when attributions for behavior are ambiguous, whereas socially desirable explicit attitudes drive behavior when attributions are clear. In a demonstration of this effect, Son Hing et al. (2008) found that white Canadians low in both implicit and explicit prejudice toward Asians did not discriminate between white and Asian job applicants who were equally qualified for the job for which they had applied, regardless of the clarity of their qualifications. When the applicants' qualifications were ambiguous, however, people low in explicit but high in implicit prejudice were more likely to hire the white than the Asian job applicant. In such a manner, low explicit/high implicit prejudice individuals are able to hide their prejudiced beliefs and their discriminatory behavior from self and other. Thus, *our fourth proposal is that the dissociation between implicit and explicit attitudes lends itself to self-deception by enabling people to express socially desirable attitudes while nevertheless acting upon relatively inaccessible socially undesirable attitudes when they can maintain plausible deniability*.

4.3. Automatic versus controlled processes

Controlled processes involve conscious effort, awareness, and intention and can be stopped at will, whereas automatic processes take place in the absence of effort, awareness, and intention and typically run to completion once initiated (Bargh 1994). It is now apparent that automatic and controlled processes make independent contributions to numerous social tasks, and that these processes can be dissociated (Chaiken & Trope 1999). Research on automatic goal activation has demonstrated that a wide variety of goal-directed behaviors that appear to be consciously directed can take place automatically, often with the goal itself outside of conscious awareness (Chartrand et al. 2008). Because these unconscious goals can sometimes run counter to people's consciously stated goals, people can consciously hold goals that are socially desirable and supported by their peers or family while simultaneously holding unconscious alternative goals that are socially undesirable or otherwise unacceptable to peers or family (Chartrand et al. 2007; Fitzsimmons & Anderson, in press). Furthermore, because automatically activated responses to the environment can be executed without awareness (Lakin et al. 2008), people can engage in socially undesirable goal-directed behavior but remain oblivious to that fact. For example, a student whose parents want her to be a physician but who wants to be an artist herself might follow her conscious goal to major in biology and attend medical school, but her unconscious goal might lead her not to study sufficiently. As a consequence, she could find herself unable to attend medical school and left with no choice but to fall back on her artistic talents to find gainful employment. By deceiving herself (and consequently others) about the unconscious motives underlying her failure, the student can gain her parents' sympathy rather than their disapproval. These findings and possibilities suggest *our fifth proposal, that the dissociation between automatic and controlled processes facilitates self-deception by enabling the pursuit of deceptive goals via controlled processing while retaining the automatic expression of actual but hidden goals*.

5. Varieties of self-deception

We begin our review of the evidence for self-deception by describing biases that represent the front end of the information-processing sequence (i.e., information gathering and selective attention). Self-deception at this stage of information processing is akin to failure to tell the self the whole truth. We then discuss varieties of self-deception that represent the middle of the information-processing stream (e.g., memory processes). These are processes that involve obfuscating the truth. We then conclude this discussion with types of self-deception that involve convincing the self that a falsehood is true.

An important question that must be addressed with regard to all of these instances of biased processing is whether they reflect self-deception or some other source of bias. Two manipulations have proven useful in addressing this issue: self-affirmation (Sherman & Cohen 2006; Steele 1988) and cognitive load (e.g., Valdesolo & DeSteno 2008). When people are self-affirmed, they are typically reminded of their important values (e.g., their artistic, humanist, or scientific orientation) or prior positive behaviors (e.g., their kindness to others). By reflecting on their important values or past positive behaviors, people are reminded that they are moral and efficacious individuals, thereby affirming their self-worth. A cornerstone of self-affirmation theory is the idea that specific attacks on one's abilities or morals – such as failure on a test – do not need to be dealt with directly, but rather can be addressed at a more general level by restoring or reaffirming a global sense of self-worth (Steele 1988). Thus, self-affirmation makes people less motivated to defend themselves against a specific attack, as their sense of self-worth is assured despite the threat posed by the attack.

Self-affirmation manipulations can be used to assess whether information-processing biases are self-deceptive. If a particular bias represents unmotivated error, then it will be unaffected by self-affirmation. For example, if people rely on a heuristic to solve a problem for which they do not have the knowledge or interest to use the appropriate algorithm, self-affirmation should not reduce their use of this heuristic. In contrast, if the bias represents a self-deceptive process that favors welcome over

unwelcome information, then this bias should be eliminated or attenuated by self-affirmation (see Correll et al. 2004; Sherman et al. 2009). Such an effect for self-affirmation not only provides evidence for the role of motivation in the information-processing bias but also indicates that the person had the potential to process the information in a less biased or unbiased fashion. In this manner, self-affirmation manipulations can test our self-deception criterion that the person is aware of the welcome information but at the same time also had potential awareness of the unwelcome information.

It should be noted, however, that our perspective on the interpersonal purpose of self-deception suggests that although self-affirmation should attenuate or eliminate many self-deceptive biases, it should only do so when the self-deceptive bias serves the general goal of enhancing the self. That is, when self-deception is in service of social advancement via self-enhancement, self-affirmation should attenuate or eliminate the self-deception because the affirmation itself satisfies the enhancement goal. In contrast, when people self-deceive to facilitate their deception of others on a particular issue, self-affirmation should have no effect. Here the goal of the self-deception is not to make the self seem more efficacious or moral, but rather to convince another individual of a specific fiction that the self-deceiver wishes to promulgate. Self-affirmation is irrelevant to this goal. Unfortunately, this distinction between general self-enhancement and specific deception is not always easily made. Nevertheless, when deception concerns a specific topic, self-affirmation should not influence self-deceptive practices unless the affirmation makes people decide that the deception itself is unnecessary or unimportant (e.g., if reminders of their self-worth make people less motivated to deceive others about their errors or poor behavior).

A second method for evaluating whether an information-processing bias is self-deceptive involves manipulation of cognitive load. Such manipulations typically require people to keep some information in memory (e.g., an eight-digit number) while they are engaged in the primary task of the experiment (e.g., forming an impression). Although manipulations of cognitive load do not address the motivational issues that underlie self-deception, they do address the issues of cost and potential awareness. If cognitive load leads to the elimination or attenuation of a particular bias, then the evidence suggests that the biased processing was actually more effortful than unbiased processing. Such a finding suggests that the individual was potentially aware of the unbiased information but was able to avoid it by engaging in the type of mental gymnastics described in the remainder of this section.

5.1. Biased information search

5.1.1. Amount of searching. There are many situations in daily life in which people avoid further information search because they may encounter information that is incompatible with their goals or desires. For example, on the trivial end of the continuum, some people avoid checking alternative products after they have made a purchase that cannot be undone (Olson & Zanna 1979). On the more important end of the continuum, some people avoid AIDS testing out of concern that they might get a result that they do not want to hear, particularly if they believe the disease is untreatable (Dawson et al. 2006; Lerman et al. 2002). This sort of self-deceptive information avoidance can be seen in the aphorism, "What I don't know can't hurt me." Although a moment's reflection reveals the fallacy of this statement, it is nonetheless psychologically compelling.

Similar sorts of biased information search can be seen in laboratory studies. Perhaps the clearest examples can be found in research by Ditto and colleagues (e.g., Ditto & Lopez 1992; Ditto et al. 2003), in which people are confronted with the possibility that they might have a proclivity for a pancreatic disorder. In these studies people expose a test strip to their saliva and are then led to believe that color change is an indicator of either a positive or negative health prognosis. Ditto and Lopez (1992) found that when people are led to believe that color change is a good thing, they wait more than 60% longer for the test strip to change color than when they believe color change is a bad thing. Studies such as these suggest that information search can be biased in the amount of information gathered even when people are unsure what they will encounter next (see also Josephs et al. 1992). Thus, it appears that people sometimes do not tell themselves the whole truth if a partial truth appears likely to be preferable. We are aware of no experiments that have examined the effect of self-affirmation or cognitive load on amount of information gathered, but we would predict that both manipulations would lead to an elimination or attenuation of the effect documented by Ditto and Lopez (1992).

5.1.2. Selective searching. Information search can also be biased in the type of information gathered. Although one never knows for sure what lies around the next corner, some corners are more likely to yield welcome information than others. Thus, politically liberal people might choose the *New York Times* as their information source, whereas politically conservative individuals might choose Fox News (Frey 1986). In such a manner, people can be relatively confident that the brunt of the information they gather will be consistent with their worldview, even if they do not know what tomorrow's headlines will bring.

Laboratory studies have examined this sort of biased information search, in part by assessing the conditions under which people are interested in learning negative information about themselves. One conclusion from this research is that the better people feel about themselves, the more willing they are to face criticism.[3] For example, Trope and Neter (1994) told participants that they were going to take a social sensitivity test and asked whether they would like feedback on their assets or liabilities. When participants had just ostensibly failed an unrelated spatial abilities test, or had not taken the test, they showed a slight preference for feedback on their assets. In contrast, when bolstered by the experience of ostensibly having performed very well on the spatial abilities test, participants were more interested in learning about their liabilities, presumably in service of self-improvement. In a related vein, Armitage et al. (2008) demonstrated that smokers were more likely to take an antismoking leaflet if they had been self-affirmed by reflecting on their prior acts of kindness (see also Harris et al. 2007). These data

implicate potential awareness of unwanted information by showing that people tend to search for welcome information but are capable of searching for unwelcome information when their self-enhancement goals have been met. Thus, it appears that people are often able to avoid telling themselves the whole truth by searching out those bits of truth that they want to hear, but they are also willing to face uncomfortable truths when feeling secure (Albarracín & Mitchell 2004; Kumashiro & Sedikides 2005).

5.1.3. Selective attention. When information is perceptually available and need not be actively discovered, people can still bias their encoding by selectively attending to aspects of the available information that they would prefer to be true. For example, if a person is at a dinner party where one conversation concerns the dangers of smoking and the other concerns the dangers of alcohol, she can choose to attend to one conversation or the other – and may do so selectively if she is a smoker or a drinker. In such a case she would likely be aware of the general tone of the information she is choosing not to gather, but by not attending to one of the conversations, she could avoid learning details that she may not want to know.

This sort of effect has been documented in a variety of different types of experiments. For example, in a study of proactive coping, Wilson et al. (2004) convinced participants that they might be chosen or were highly unlikely to be chosen for a hypothetical date. When participants believed they might be chosen, they spent slightly more time looking at positive than negative information about their potential partner. In contrast, when they believed that they were highly unlikely to be chosen, they spent more time looking at negative information about their potential partner. Thus, when people faced almost certain disappointment, they directed their attention to information that would make their upcoming rejection more palatable.

Although measures such as reading time provide a good indicator of the amount of information processing, attention can be assessed more directly. Eye-tracking studies provide some of the clearest evidence of where people direct their attention, and such studies have also shown that people are often strategic in their attentional decisions (Isaacowitz 2006). For example, older adults look toward positive stimuli and away from negative stimuli when in a bad mood (Isaacowitz et al. 2008). This attentional bias clearly implicates potential awareness, as some encoding of the negative must take place for preferential attention to be directed toward the positive. This effect did not emerge among younger adults, suggesting that older adults are more likely than younger adults to rely on selective attention for mood repair. In a case such as this, it appears that older adults sacrifice informational content in service of emotional goals. This strategy might be sensible for older adults who have greater immune challenges than their younger counterparts and thus reap greater benefits from maintaining happiness (see sect. 6). As with the strategy of ending information search early, selective attention can allow people to avoid telling themselves the whole truth.

5.2. Biased interpretation

Despite the strategies just described for avoiding unwelcome information, there remain a variety of circumstances in which such information is nevertheless faithfully encoded. Under such circumstances, unwelcome information can still be dismissed through biased interpretation of attitude-consistent and attitude-inconsistent information. In the classic study of this phenomenon (Lord et al. 1979), people who were preselected for their strong attitudes on both sides of the capital punishment debate were exposed to a mixed bag of information about the efficacy of capital punishment. For example, some of the data with which they were presented suggested that capital punishment was an effective crime deterrent, whereas other data suggested that it was not. Given that the findings were new to participants, logic would suggest that they would coalesce at least to some degree in their attitudes. In contrast, people ended the experiment more polarized than they began it.

Lord et al. (1979) discovered that this attitude polarization was a product of biased interpretation of the data. People who were in favor of capital punishment tended to accept the data as sound that supported capital punishment but reject the data as flawed that opposed capital punishment. Those who were against capital punishment showed the opposite pattern (see also Dawson et al. 2002). This selective skepticism appears to be self-deceptive, as it is attenuated or eliminated by self-affirmation (Cohen et al. 2000; Reed & Aspinwall 1998) and cognitive load (Ditto et al. 1998). These findings suggest that people have potential awareness of an unbiased appraisal, given that they appear to be relying on their motivational and mental resources to be differentially skeptical. Thus, selective skepticism appears to be a form of self-deception rather than simply an objective devaluation of new information to the degree that it is inconsistent with a large body of prior experience (see also, Westen et al. 2006).

As a consequence of this selective skepticism, people are able to encounter a mixed bag of evidence but nevertheless walk away with their original beliefs intact and potentially even strengthened. Because they are unaware that a person with a contrary position would show the opposite pattern of acceptance and rejection, they are able to convince themselves that the data support their viewpoint. Thus, it seems that by relying on their considerable powers of skepticism only when information is uncongenial, people are able to prevent themselves from learning the whole truth.

5.3. Misremembering

Even if people attend to unwanted information, and even if they accept it at the time of encoding, this does not guarantee that they will be able to retrieve it later. Rather, information that is inconsistent with their preferences may simply be forgotten or misremembered later as preference-consistent or neutral. Thus, a person might have great memory for the details of his victory in the championship tennis match but very poor memory for the time he lost badly. Indeed, this latter memory might also be distorted to implicate his doubles partner or the unusual talents of his opponent. In section 4, we discussed theoretical mechanisms by which deceptive information might remain in consciousness and truthful information might be relegated to unconsciousness, but what evidence is there that memory is selective in this sort of fashion? Unfortunately, we know of no evidence showing how

intending to deceive others can lead to self-deception as described in section 4, but there is evidence that other motivational sources can lead to selective forgetting processes.

First, when people put effort into self-improvement, but the improvement does not materialize, they can manufacture the gains they wish they had made by misremembering how they used to be. For example, Conway and Ross (1984) demonstrated that after taking a study skills class, people misremembered their prior study skills as lower than they rated them originally, thereby supporting their belief that their skills have improved. They then later misremembered their subsequent course performance as better than it was to maintain the fiction of improvement. Through processes such as these, people are able to purge their memories of inconvenient truths, thereby preventing themselves from knowing the whole truth, even if they accurately encoded it in the first instance.

Health information can be similarly distorted in memory (Croyle et al. 2006). In Croyle et al.'s research, participants were given cholesterol screening, and one, three, or six months later tested for their memory of their results. Respondents showed highly accurate memory of their risk category (89% correctly recalled this information), and this accuracy did not decay over six months. Nevertheless, even in the context of this apparently easy memory task, respondents were more than twice as likely to recall their cholesterol as being lower rather than higher than it really was.

This sort of memory bias can also be seen in recollection of daily experiences, whereby people have better recall of their own good than bad behavior but do not show this bias in their recall of the behaviors of others (D'Argembeau & Van der Linden 2008). This self-enhancing recall bias is also eliminated by information that bolsters people's self-image (in this case, doing well on a test; Green et al. 2008), suggesting that people have potential awareness of both positive and negative information about the self. Thus, people's memories appear to be self-enhancing, sometimes containing information that is biased to be consistent with preferences and sometimes just failing to contain the whole truth.

5.4. Rationalization

Even if one's prior misdeeds are accurately recalled by self and others, it is still possible to avoid telling oneself the whole truth by reconstructing or rationalizing the motives behind the original behavior to make it more socially acceptable. For example, after eating a second helping of cake that leaves none for those who have not yet had dessert, a person could explain that he had not noticed that there was no other cake, or that he thought more cakes were available elsewhere. Here it is not memory of the misdeed that is critical, but interpretation of the motivation that underlies that deed.

Again, laboratory evidence supports this sort of rationalization process. For example, von Hippel et al. (2005) demonstrated that when cheating could be cast as unintentional, people who showed a self-serving bias in another domain were more likely to cheat, but when cheating was clearly intentional, self-serving individuals were no more likely to cheat than others. These data suggest that some types of self-serving biases involve rationalization processes that are also common to some types of cheating. Indeed, people also cheat more when they are told that free will is just an illusion (Vohs & Schooler 2007), suggesting that they rationalize their cheating in these circumstances as caused by life situations rather than their own internal qualities.

More direct evidence for this sort of rationalization can be found in the hypocrisy research of Valdesolo and DeSteno (2008). In their study participants were given the opportunity to (a) choose whether to give themselves or another individual an onerous task or (b) randomly assign the onerous task to self versus other. When given this opportunity, nearly all participants chose to give the onerous task to the other participant rather than rely on random assignment. Observers were not asked to make the choice themselves but rather watched a confederate make this same self-serving choice. When asked how fair the choice was, observers rated the act of choosing rather than relying on random assignment as less fair than it was rated by those who had actually made this choice. This hypocrisy in self-ratings shown by those who chose to assign the onerous task to another was eliminated by cognitive load, suggesting that participants have potential awareness of the unfairness underlying their judgments.

Research on misattribution reveals evidence of similar rationalization processes. A classic example can be found in Snyder et al.'s (1979) research on avoidance of disabled people. In Snyder et al.'s experiment, participants chose a seat from two options – one next to a person who was disabled and one next to a person who was not disabled. In front of each empty seat there was a television, and the two televisions sometimes showed the same program and sometimes a different program. Snyder et al. (1979) found that when the televisions were showing the same program, the majority of participants sat next to the disabled person, presumably to demonstrate to self and other that they were not prejudiced against disabled people. In contrast, when the televisions were showing different programs, the majority sat away from the disabled person. These data show that people only avoided the disabled person when they could rationalize their behavior as caused by external factors.

Similar effects have been documented in differential helping rates for African versus white Americans, with a meta-analysis showing that whites are less likely to help African Americans than fellow whites, but only when there are actual situational impediments to helping, such as distance or risk (Saucier et al. 2005). When people cannot attribute their non-helping to such situational factors, they apparently feel compelled to help African Americans at equal rates to whites. In cases such as these, people are not denying or misremembering their cheating, self-serving choices, avoidance, or lack of helping. Rather, they are denying the socially undesirable motives that appear to underlie their behaviors by rationalizing their actions as the product of external forces.

5.5. Convincing the self that a lie is true

The classic form of self-deception is convincing oneself that a lie is true. This sort of self-deception can be difficult to verify, as it is difficult to know if the person believes the lie that they are telling others, given that situations that motivate lying to the self typically motivate lying to

others. Nevertheless, there are examples of experiments in which this sort of process has been shown. Most of these examples rely on research paradigms in which the experimenter knows the truth, so the participant has little or nothing to gain interpersonally by lying and often has much to lose, given that lying makes the individual look vain, foolish, or deceptive.

5.5.1. Self-deception accompanied by neurological damage. A clear example of this sort of self-deception can be found in the split-brain research of Gazzaniga and colleagues, which relies on participants who have had their corpus callosum severed and are thereby unable to communicate directly between the two hemispheres. Gazzaniga (1997) described a series of experiments with split-brain patients that suggest that the left hemisphere confabulates when necessary to explain one's own behavior. In one such study a split-brain patient was presented with an apparatus that displayed a chicken foot to the left hemisphere (via the right visual hemifield) and a snowy scene to the right hemisphere (via the left visual hemifield). The participant was then asked to point with each hand at the picture that most closely matched what was seen. The left hemisphere directed the right hand to point at a chicken head, and the right hemisphere directed the left hand to point at a shovel.

When Gazzaniga asked the participant why his left hand was pointing at the shovel, he created a quandary. Because the right hemisphere does not initially have access to speech centers, and thus is functionally mute in early split-brain patients, the participant could not accurately answer this question. The left hemisphere, which has access to speech, did not know why the left hand was pointing at a shovel. Nevertheless, rather than responding that he did not know why he was pointing at the shovel, the participant invented an answer – in this case, the plausible story that chickens make a lot of waste and the shovel is necessary to remove this waste. This study reveals an individual who self-deceives only to avoid the uncertainty caused by his lack of awareness of the source of his own behavior (a situation that is likely to be very common; see Nisbett & Wilson 1977). Nevertheless, this motivation appears to be sufficient to cause the person to invent a reason for his behavior and then apparently convince himself of the accuracy of this fiction.

Other examples from the neuropsychological literature can be found in various types of brain and body damage, in response to which the individual is motivated to maintain certain beliefs that are at odds with reality. For example, anosognosia is a prototypical self-deceptive disorder, in which people who have sustained an injury to some part of their body deny the reality of their injury. Ramachandran (2009) described an anosognosic woman who denied that her left arm was paralyzed. He wrote:

> An intelligent and lucid patient I saw recently claimed that her own left arm was not paralyzed and that the lifeless left arm on her lap belonged to her father who was "hiding under the table." Yet when I asked her to touch her nose with her left hand she used her intact right hand to grab and raise the paralyzed hand – using the latter as a "tool" to touch her nose! Clearly somebody in there knew that her left arm was paralyzed and that the arm on her lap was her own, but "she" – the person I was talking to – didn't know. (Ramachandran 2009)

As can be seen in this description, the patient had awareness that her left arm was paralyzed, as indicated by her use of her right arm to move it, but she appeared to suffer from lack of awareness as well, suggesting self-deception. Consistent with this interpretation of anosognosia, a recent study (Nardone et al., 2007) presented disabled individuals with neutral and threatening words relevant to their immobility (e.g., walk). Those who were aware of their disability showed rapid disengagement from the threatening words, as indicated by more rapid response to a dot presented elsewhere on the screen in the presence of threatening than neutral words. In contrast, anosognosic individuals were less able to disengage from the threatening words, showing a slower response to the dot probe when paired with threatening versus neutral words. These data highlight implicit awareness of the disability in anosognosia, despite its explicit denial.

Cases such as these involve damage to the right hemisphere, which appears to prevent individuals from recognizing the logical inconsistency in their belief systems. Similarly, cases such as the one documented by Gazzaniga also require damage to the brain (the corpus callosum) for the individual to be denied access to information in some parts of the brain that is available to others. In the case of selective access of information to consciousness, however, individuals with no damage to the brain also experience this situation regularly, as reviewed earlier. Thus, there are also studies of individuals deceiving themselves when they are neurologically intact. We turn now to such evidence.

5.5.2. Self-deception unaccompanied by neurological damage. A ubiquitous variety of self-deception can be found in research on perceptions of control. Perceptions of control appear to be necessary for the maintenance of psychological and physical health (Cohen 1986; Glass & Singer 1972; Klein et al. 1976). When people are deprived of actual control, they often endeavor to regain a sense of control. In a self-deceptive example of this effect, Whitson and Galinsky (2008) found that when people are led to feel low levels of personal control, they perceive illusory patterns in random configurations and are more likely to endorse conspiracy theories to explain co-occurring world events. Importantly, these effects did not emerge when people had self-affirmed, suggesting potential awareness of the absence of patterns and conspiracies. Similar findings have been documented by Kay et al. (2008), who argued that beliefs in a controlling God and a strong government serve people's need for control. Consistent with their reasoning, differences in the percentage of people who believe in God between countries can be predicted by the insecurities of existence within countries (e.g., availability of health care, food, and housing), with increased insecurity associated with increased religiosity (Norris & Inglehart 2004). Such a finding suggests the possibility of self-deception on a worldwide scale.

Another example of individuals deceiving themselves can be found in the research of Epley and Whitchurch (2008) reviewed earlier, in which people more rapidly located photos of themselves when the photo had been morphed to be more attractive than when the photo was unaltered. This finding suggests that people's self-image is more attractive than their actual one, as the enhanced self provides a quicker match to their internal template than the actual self. Finally, experiments in cognitive

dissonance also suggest that people are facile at lying to others and then coming to believe their own lies. For example, when they believe that they have freely chosen to tell another person that a tedious task is actually interesting, people soon believe that the task really is interesting (Festinger & Carlsmith 1959), and again this effect is eliminated by self-affirmation (Steele & Liu 1983).

6. Who is the audience for self-deception?

Thus far we have argued that self-deception evolved to facilitate deception of others, but the examples of self-deception described in section 5 appear to be directed primarily toward the self. If self-deception evolved to deceive others, why is there so much evidence for self-deception that appears to be intended only for the self? There are three answers to this question.

First and foremost, although Trivers (1976/2006) originally suggested that self-deception might have evolved to facilitate the deception of others over 30 years ago, this suggestion has not been taken seriously in the empirical literature. Rather, the tradition in psychology has been to treat self-deception as a defensive response to an uncongenial world (a point to which we return in sect. 7). As a consequence, to the best of our knowledge no one has examined whether self-deception is more likely when people attempt to deceive others. Thus, the theoretical possibility of self-deception in service of other deception remains just that, and the evidence described in section 5 stems primarily from studies in which the motives appear to be more intrapersonal than interpersonal.

Second, to the degree that self-deception allows the individual to carry on with greater equanimity and confidence, it serves the general interpersonal goal of self-enhancement described in section 3. Most of the cases of apparently self-directed self-deception from section 5 would fit under this explanatory rubric. For example, if individuals selectively gather information in a manner that enables them to deny to themselves that they are at risk for a health disorder (as in Ditto & Lopez 1992), then they are better positioned to convince others that they would make reliable and vigorous sexual or coalitional partners. That this pattern of self-deception makes them less capable of dealing with an impending health threat might have been a relatively small price to pay in an ancestral environment where there was little that could be done in any case. A similar interpersonal logic might underlie self-deceptions that help people maintain conviction in their beliefs (e.g., Lord et al. 1979) and give them a sense of control over the world (e.g., Whitson & Galinsky 2008).

Other classic examples of intrapersonally oriented self-deception also have important interpersonal consequences. For example, people tend to be unrealistically optimistic about their future (Armor & Taylor 1998; Taylor & Brown 1988; Weinstein 1980). This optimism can have self-deceptive origins; for example, people high in self-deception – measured via Sackeim and Gur's (1979) Self-Deception Scale – have been shown to be more likely than people low in self-deception to perceive an upcoming onerous task as a challenge (i.e., within one's capabilities) rather than a threat (i.e., overwhelming one's capabilities; Tomaka et al. 1992). In this sense, self-deceptive optimism appears to create a self-fulfilling prophecy, as confidence in eventual success leads optimists to greater perseverance in the face of difficulties (Carver & Scheier 2002; Solberg Nes & Segerstrom 2006). As a consequence of these processes, optimists gain numerous social and financial benefits over their less optimistic counterparts (Assad et al. 2007; Brissette et al. 2002; Carver et al. 1994; Segerstrom 2007).

Finally, it is also possible that this system of self-deception that evolved to deceive others becomes applied in intrapersonal domains because of the good feelings that it brings the individual. By analogy, consider masturbation, another intrapersonal activity that has origins in an interpersonal system. Masturbation presumably emerged in primates because we evolved to enjoy copulation (thereby facilitating reproduction), but with the later evolution of hands rather than hooves or claws, we found a way to experience that enjoyment when circumstances conspire against sharing it with others. Self-directed self-deception might be analogous to masturbation in the sense that self-deception evolved for interpersonal purposes, but people found a way to use it to enhance happiness when circumstances conspire against other methods. So long as the consequences of this self-deception are fitness neutral or biologically affordable, self-deception might have leaked into a variety of intrapersonal domains that have important consequences for human happiness.

Indeed, the analogy to masturbation remains apt in this regard, as masturbation might be fitness neutral or even fitness positive among males, because it might enhance sperm quality by shedding older sperm (Baker & Bellis 1993). Thus, primates appear to have found a way to enhance their happiness with little if any cost to inclusive fitness. Similarly, although the spread of self-deceptive practices in the pursuit of happiness would clearly bear a cost to the degree that people sacrifice information quality for hedonic gains, such practices might also be fitness neutral – or even fitness positive – to the degree that people reap the gains brought about by happiness itself (Fredrickson 1998; 2001).

As with optimism, happiness has important interpersonal consequences; people experience increased social and financial success when they are happy (Boehm & Lyubomirsky 2008; Fredrickson et al. 2008; Hertenstein et al. 2009; Lyubomirsky et al. 2005). Others are attracted to happy people and put off by sad people, for a variety of reasons (Bower 1991; Frijda & Mesquita 1994; Harker & Keltner 2001; Keltner & Kring 1998). As the aphorism goes, "laugh and the world laughs with you, cry and you cry alone." Because humans are a social species wherein individuals achieve most of their important outcomes through coordinated or cooperative actions with others, attracting others to oneself and one's causes is an important achievement with notable fitness consequences (Fredrickson 1998; 2001). Thus, to the degree that people generally maintain a sunny disposition, they are likely to be more effective in their goal pursuits.

Happiness is also important for physical well-being, given that there are immune benefits to feeling happy and immune costs to feeling sad (Cohen et al. 2006; Marsland et al. 2007; Rosenkranz et al. 2003; Segerstrom & Sephton 2010). Because threats to health loom larger late in life, particularly from cancers and parasites (World Health Organization 2009), happiness may be even more important in late than early adulthood. Consistent with

this possibility, older adults are more likely than younger adults to focus on and remember positive rather than negative information (Mather & Carstensen 2005; Mather et al. 2004). Thus, it appears that late life happiness is maintained in part by the knowledge avoidance form of self-deception, as older but not younger adults look away from negative information and toward positive information when they are in a bad mood (Isaacowitz et al. 2008). Enhanced immune functioning may offset the informational costs of this aging positivity effect.[4]

7. At what level of consciousness is the self deceived?

At the beginning of this article we rejected the classic claim that self-deception must involve two separate representations of reality, with truth preferentially stored in the unconscious mind and falsehood in the conscious mind. Rather, our biased processing perspective suggests that the individual can self-deceive in a variety of ways, some of which prevent even unconscious knowledge of the truth. Nevertheless, such a possibility does not preclude classic forms of self-deception, and here we consider the question of how the truth is represented in different forms of self-deception. We begin with types of self-deception in which conscious and unconscious knowledge are aligned and inaccurate, that is, cases in which individuals believe consciously and unconsciously in the veridicality of the deceptive information. We propose that this sort of self-deception should occur in two types of situations.

First, self-deception should exist at both conscious and unconscious levels when individuals prevent themselves from ever encoding the unwelcome truth. For example, Ditto and Lopez (1992) documented how individuals who are pleased with the early returns often stop gathering information before they encounter unwanted information, and thus their self-deception has prevented unwanted information from entering either conscious or unconscious knowledge. Such individuals might be aware that their information gathering strategy could have influenced the nature of their knowledge, but such awareness is likely to be rare, as (a) there is no set standard for how much information a person should gather in most settings, and (b) there is often no specific reason to believe that the next bit of information would have been contrary to that which had already been gathered.

Second, self-deception should also exist at both conscious and unconscious levels in many (although probably not all) cases of self-enhancement. In such cases individuals have a lifetime of gathering and processing information in a manner that favors the self, and it would probably be difficult if not impossible for them to parse out the impact of their long-term processing strategies on their understanding of the world. Support for this possibility can be found in two types of effects. First, as Epley and Whitchurch (2008) demonstrated, people are faster to identify their more attractive self than their actual self in an array of faces. This finding suggests that the enhanced version of the self is likely to be represented in memory below consciousness awareness, as unconscious processes tend to be more rapid than conscious ones (e.g., Neely 1977), and conflict between conscious and unconscious processes leads to slower rather than more rapid responses (e.g., Greenwald et al. 1998). Second, research on the convergence of implicit and explicit self-esteem reveals that individuals who show high implicit and high explicit self-esteem appear not to be defensive, whereas individuals who show high explicit but low implicit self-esteem appear to be the most defensive and narcissistic (Jordan et al. 2003). Because defensive and narcissistic individuals tend not to be well liked (Colvin et al. 1995; Paulhus 1998), this research suggests that the interpersonal benefits of self-enhancement are most likely to be realized if people believe their self-enhancing stories at both conscious and unconscious levels.

In contrast to these situations are classic cases of self-deception in which the conscious mind is deceived but the unconscious mind is well aware of the truth. This should be common in cases of self-deception intended to facilitate deception of others on specific issues. Under such circumstances individuals often have a limited time frame in which to convince the self of the truth of the deception that they wish to promulgate, and thus the types of memory processes outlined in section 4 are likely to lead to conscious recollection of the deception but implicit memory of the truth as originally encountered. Thus, it seems likely that self-deception can vary in the depth to which the self is deceived, with some processes leading to deception of conscious and unconscious aspects of the self and other processes leading to deception of only conscious aspects of the self.

8. Contrasting our approach with other approaches

Our evolutionary approach to self-deception is based on the premise that self-deception is a useful tool in negotiating the social world. According to this viewpoint, self-deception is best considered an *offensive* strategy evolved for deceiving others. In contrast to this viewpoint, most prior research on self-deception considers it a *defensive* strategy, adopted by individuals who are having difficulty coping with a threatening world. In this research tradition, the hedonic consequences for the self-deceiver are considered to be the primary outcome of self-deception. From an evolutionary perspective, however, hedonic consequences are not an important endpoint themselves, but only a means to an end, such as when they lead to enhanced immune functioning or greater interpersonal success, as described in section 6.

The most important consequence of this prior emphasis on hedonic consequences is that the field has been focused on what we would regard as a secondary aspect of self-deception. From our perspective, the study of self-deception as if it is a "psychological immune system" (Gilbert et al. 1998; Wilson & Gilbert 2003) or a suite of "positive illusions" (Taylor & Brown 1988) intended to enhance or restore happiness is somewhat akin to the study of masturbation as if this is the purpose for which the sexual system was designed. Such an approach to sexuality would lead to worthwhile findings about some of the affective consequences of sexual behavior, but by ignoring the interpersonal purpose of sexuality, this approach would miss most of the important questions and answers.

Thus, in many ways the arguments outlined in this article call for a fundamental change in our approach to

the problem of self-deception. The research described in sections 4 and 5 provide ample evidence for the psychological processes involved in self-deception and for the manner in which they protect people's beliefs and desires from a contrary reality. But because self-deception has been viewed as a defensive response to an uncongenial world, there is virtually no research that considers the interpersonal opportunities that might lead to self-deception. For example, research suggests that heterosexual men enhance their self-presentation when they meet attractive women (Buss 1988), and women rate other women as less attractive when they are themselves ovulating (Fisher 2004).[5] The current approach suggests that men should actually believe at least some of their self-enhancement, and women should believe their more negative evaluations of others as well. That is, these strategies should be more effective if people are not just posturing but actually accept their own self-inflating and other-deflating stories. To test these possibilities, one could assess whether the presence of attractive women causes men to show greater self-enhancement in their information-processing biases and whether ovulation causes women to show greater derogation of rivals in their biases as well.

Future research could also address the utility of self-deception in specific acts of deception of others. For example, people could be put in a situation in which they must lie or otherwise deceive and their use of various self-deceptive strategies reviewed in section 5 could be examined. In such a setting they could be assessed for the degree to which they stop their search early when they encounter initial information that is consistent with their deceptive goals, avoid information that appears potentially inconsistent with an upcoming deception, show biased interpretation of information that is consistent versus inconsistent with an upcoming deception, and show better recall for deception-consistent than deception-inconsistent information. Importantly, the utility of such strategies could be assessed by examining the degree to which these information-processing biases are associated with increased effectiveness in deceiving others. As noted earlier, similar biases should also be invoked by situations that enhance the interpersonal need to appear confident and efficacious. And if these information-processing strategies reflect self-deception, then the degree to which people show biases under these circumstances may be moderated by individual differences in the tendency to self-deceive. For example, these effects should be stronger among people who show greater self-enhancement in the Epley and Whitchurch (2008) paradigm than among people who do not show such self-enhancement.

The current approach to self-deception is also relevant to a series of debates that have been percolating for some time in psychology, in which self-enhancement (self-deception) is pitted against self-verification (reality orientation). In the first of these debates, researchers have asked whether self-enhancement motives are stronger than self-verification motives (see Swann, in press). From an evolutionary perspective this is not a meaningful question to ask, because it is akin to asking whether thirst is stronger than hunger. Just as the desire to eat versus drink is jointly determined by the current needs of the individual and the quality of resources available to satisfy hunger versus thirst, the desire to self-verify versus self-enhance should be jointly determined by current need states and the quality of available verifying and enhancing opportunities. Indeed, the research on self-affirmation described in section 5 highlights the fact that when people's self-enhancement needs are met, they become more willing to self-verify in negative domains. Thus, both enhancing and verifying needs are likely to be present nearly all of the time, and their relative strength will vary as a function of the social and material benefits they can accrue for the individual at any particular moment. Self-verification allows people to accurately gauge their abilities and behave accordingly, and self-enhancement allows people to gain an edge beyond what their actual abilities provide them.

In the second of these debates (e.g., Sedikides et al. 2003 vs. Heine et al. 1999), researchers have argued about whether self-enhancement is pan-cultural or a uniquely Western phenomenon. From our perspective, this debate has gotten off track because of its emphasis on the role of positive self-regard rather than interpersonal goals as the source of self-enhancement. Because self-enhancement induces greater confidence, our perspective suggests that it will emerge in every culture in which confidence brings social and material gains. If a culture exists in which people benefit in various social and material domains by being meek and self-doubting, then individuals in such a culture should not self-enhance in those domains. From an evolutionary perspective, however, people in all cultures should benefit from enhancing those qualities that are necessary to win confrontations and secure mates. Because individual differences in abilities and proclivities enable some individuals to win conflicts and mates by physical means, others to win them by intellectual means, and still others by displays of kindness, artistic abilities, and so forth, one would expect that self-enhancement should similarly emerge in different forms for different individuals, but should do so in every culture on earth.

This debate has also become caught up in issues of measurement focused on how to properly assess self-enhancement in different cultures. Because the rules of social engagement vary widely across different cultures, it should come as no surprise that in some cultures people are unwilling to claim to be better than others. Again, however, the self-deceptive goal of self-enhancement is to *believe* one is slightly better than one really is, and thus explicit claims are less important than implicit beliefs. Indeed, self-enhancing claims are often likely to be evidence of bad manners or poor socialization, whereas exaggerated beliefs in one's abilities in domains that are necessary to win confrontations and secure mates should be universal.

In the third of these debates, people have asked whether enhancing self-views or accurate self-views reflect greater mental health (e.g., Colvin & Block 1994 vs. Taylor & Brown 1994). This debate has shades of the second, given that again a balance of the two and knowing when each is appropriate should reflect the best strategy for goal achievement, and thus should be reflective of mental health. This debate has also gotten bogged down in the misperception (noted by Taylor & Brown 1994) that if a little self-enhancement is a good thing, then more can only be better. Biological systems rely on balance or homeostasis, and too much of even a good thing disrupts this balance. Thus, from an evolutionary

perspective, it is obvious that the benefits of self-enhancement depend on proper dosage. Too much self-enhancement might not only be hard to believe oneself but might strike others as preposterous or mentally unbalanced (Colvin et al. 1995; Paulhus 1998) and might also lead to complacency and poor preparation in the face of real problems. This is not evidence, as Colvin et al. (1995) and Paulhus (1998) suggested, that self-enhancement is socially maladaptive. Rather, this finding simply highlights the fact that self-deception remains in permanent tension with veridical perception.

Finally, similar to the second debate, our interpersonal perspective on self-deception suggests that the key to understanding whether self-enhancement is a sign of good versus poor mental health is whether people believe their own self-enhancing story. Self-enhancement is useful only to the degree that it is self-deceptive, because only when it is believed by the self will others accept the enhanced self as genuine. If the claims ring hollow or if the claimants appear to be trying unsuccessfully to convince themselves, then this is a sign of failed self-deception and is likely to be an indicator of poor social functioning and poor mental health. This perspective suggests that much of the evidence showing that self-enhancement and other "positive illusions" are a sign of poor mental health comes from studies in which no differentiation is being made between people who are convinced versus unconvinced by their own self-enhancing claims.

As noted earlier, people who are high in explicit but low in implicit self-esteem are the most defensive and narcissistic (Jordan et al. 2003), and thus it seems likely that self-enhancement only brings benefits to the degree that it is believed both consciously and unconsciously. For this reason, we would predict that individuals who show convergence in their explicit and implicit measures of self-enhancement will confirm Taylor and Brown's (1988; 1994) claims that self-enhancement is a sign of good mental health. Nevertheless, there may still be the occasional individual who shows extreme but convergent self-enhancement, and from an evolutionary perspective, this is not likely to be a successful interpersonal strategy and therefore unlikely to be a sign of good mental health. Such individuals may not come across as defensive, but they are likely to appear deluded.

9. Costs versus benefits

Finally, it is worth considering the costs versus benefits of self-deception. Because self-deception requires a mental architecture that sometimes favors falsehoods, those features of our mental landscape that allow us to self-deceive are likely to have attendant costs. Although we have focused thus far on the possibility that self-deception can be beneficial, it is also worth considering the costs. The most obvious cost of self-deception is loss of information integrity – with the resulting potential for inappropriate action and inaction – but there are likely to be other costs as well. For example, consider the case of memory discussed in section 4.1, in which we outlined how source confusions can facilitate self-deception in people's efforts to deceive others. Research suggests that some individuals are more likely than others to have difficulty differentiating the source of their memories, with the result that they are more susceptible to the implantation of false memories (Clancy et al. 2000). Such individuals suffer the cost of greater memory failures and therefore poorer foresight and greater manipulation by others. But they should also be more capable of self-deception in their efforts to deceive others, as retrieval of their lies should make it particularly difficult for them to differentiate the source of the false information. In this manner the costs associated with their poor source-monitoring capabilities might be offset by the gains associated with their greater likelihood of deceiving others.

At a more general level, the overarching fact that the mind might have evolved to self-deceive – and therefore to be tolerant of inconsistencies between information in the conscious and unconscious mind – raises the possibility that this weapon of deception might be capable of being turned upon the self. That is, self-deception might be imposed by others. For example, an abusive spouse who rationalizes his abuse as a product of his partner's failings might also convince her that she is to blame for his abuse. Consistent with this sort of possibility, system-justification theory as elaborated by Jost and colleagues (Jost et al. 2004; Jost & Hunyady 2005), argues that there are a variety of motivational reasons why people support the status quo, even when they are clear losers in the current system with very little likelihood of improving their situation. Such system-justifying beliefs among those at the bottom of the social ladder serve the purposes of those at the top of the ladder, in part by preventing agitation for social change. This argument suggests that system justification might be considered a variety of self-deception imposed onto low-status individuals by high-status people who benefit when those beneath them accept their low status as legitimate and do not struggle against it. This argument also suggests that the consequences of self-deception might be wide ranging, as a process that evolved to facilitate the deception of others appears to have effects that manifest themselves from an intrapersonal all the way to a societal level.

ACKNOWLEDGMENTS

Preparation of this manuscript was supported by fellowships at the Institute for Advanced Study, Berlin, and by grants from the Australian Research Council and the Biosocial Research Foundation. We thank Michael Anderson, Marzu Banaji, Pablo Briñol, David Buss, Bella DePaulo, Tony Greenwald, Wilhelm Hofmann, Geoffrey Miller, Srini Narayanan, Steve Pinker, and Constantine Sedikides for helpful comments on an earlier draft of the manuscript.

NOTES

1. Cross-examination can also cause deceivers to seem more honest (Levine & McCornack 2001), but further research is needed on the effects of cross-examination in detecting lies that are consequential for deceivers if discovered.

2. It should be noted in this regard that there is evidence for the contrary possibility that people overestimate the degree to which others have detected their lies (Gilovich et al. 1998). However, this research relied on the usual paradigm of people telling trivial lies to strangers. There are few cues in such cases about whether others believe our lies, and thus deceivers might be particularly inclined to assume that whatever cues they are aware they are emitting are equally obvious to observers.

3. People with low self-esteem also seek out negative information about themselves even when not feeling good about themselves. This information search strategy appears to involve self-verification strivings, or people's desire to be known by others as they see themselves (Swann, in press). Such self-verification leads most people to seek out positive information, however, because most people feel positive about themselves.

4. Although selection pressure diminishes in late life, grandparents historically played an important role in the survival of their grandchildren (Lahdenperä et al. 2004), and thus their continued good health and survival is important for their inclusive fitness.

5. It should be noted that women are more attractive when ovulating (Thornhill & Gangestad 2009), so this change in ratings of others could reflect an unbiased relative judgment.

Open Peer Commentary

Self-deception: A paradox revisited

doi:10.1017/S0140525X10002499

Albert Bandura

Department of Psychology, Stanford University, Stanford, CA 94305.
bandura@psych.stanford.edu
www.stanford.edu/dept/psychology/abandura

Abstract: A major challenge to von Hippel & Trivers's evolutionary analysis of self-deception is the paradox that one cannot deceive oneself into believing something while simultaneously knowing it to be false. The authors use biased information seeking and processing as evidence that individuals knowingly convince themselves of the truth of their falsehood. Acting in ways that keep one uninformed about unwanted information is self-deception. Acting in selectively biasing and misinforming ways is self-bias.

Von Hippel & Trivers (VH&T) present the case that self-deception evolved because it enables individuals to be good deceivers of others. By convincing themselves that their fabrication is true, people can enjoy the benefits of misleading others without the intrapsychic and social costs. The authors review a large body of evidence on biased information processing on the assumption that this is the means by which individuals unknowingly convince themselves of the truth of their falsehood.

A major challenge to a functional analysis of self-deception is the problematic nature of the phenomenon itself. One cannot deceive oneself into believing something while simultaneously knowing it to be false. Hence, literal self-deception cannot exist (Bok 1980; Champlin 1977; Haight 1980). Attempts to resolve the paradox of how one can be a deceiver fooling oneself have met with little success (Bandura 1986). These efforts usually involve creating split selves and rendering one of them unconscious.

Awareness is not an all-or-none phenomenon. There are gradations of partial awareness. Hence, self-splitting can produce a conscious self, various partially unconscious selves, and a deeply unconscious self. In this line of theorizing, the unconscious is not inert. It seeks expression in the intrapsychic conflict. The deceiving self has to be aware of what the deceived self believes in order to know how to concoct the self-deception. Different levels of awareness are sometimes proposed as another possible solution to the paradox. It is said that "deep down" people really know what they believe. Reuniting the conscious and unconscious split selves only reinstates the paradox of how one can be the deceiver and the one deceived at the same time.

VH&T propose another variant of the split-self solution to the paradox. They posit a dissociated mental dualism in which the deceived mind is disconnected from the unconscious mind that knows the truth. Given the richly interconnected neuronal structure of the brain, is this Cartesian physical dualism at odds with what is known physiologically about the brain? The dissociated dualism removes not only the paradox, but also any commerce between the two minds. While the conscious mind is biasing information processing to bolster the self-deception, there is no mention of what the veridical unconscious mind is doing. If it is dissociated, how can it affect the conscious mind?

A multitude of neuronal systems is involved in the processing of input information. However, when it comes to action, there is only one body. The diverse systems have to generate a coherent action. There is diversity in information processing but unity of agency in action (Bandura 2008; Korsgaard 1989). Contradictory minds cannot simultaneously be doing their own thing behaviorally. The authors do not explain how the disconnected conflicting minds can produce a coherent action.

There are other epistemological issues, including the verifiability of the central thesis, that need to be addressed. The article presents an excellent review of biased information processing, but it leaves a lot to be desired in conceptual specification. How does one know what the unconscious mind knows? How does one assess the unconscious knowledge? By what criteria does one substantiate truth? Why is the unconscious mind veridical in self-enhancement but self-deceptive in other spheres of functioning? How does one gauge the benefits of self-deception, whether in the short term or the long term? Given the evolutionary billing of the article under discussion, what were the ancestral selection pressures that favored self-deception? Claiming that a given behavior has functional value does not necessarily mean it is genetically programmed. People are selective in their information seeking, often misconstrue events, lead themselves astray by their biases and misbeliefs, and act on deficient knowledge. However, biased information seeking and processing are not necessarily self-deception. VH&T cite, as an example of self-deception, conservatives favoring Fox News and liberals favoring MSNBC. By their selective exposure, they reinforce their political bias. But that does not mean that they are lying to themselves. The same is true for some of the other forms of biased information processing that are misconstrued as self-deception.

In genuine self-deception people, avoid doing things that they have an inkling might reveal what they do not want to know. Suspecting something is not the same as knowing it to be true, however. As long as one does not find out the truth, what one believes is not known to be false. Keeping oneself uninformed about an unwanted truth is the main vehicle of genuine self-deception. By not pursuing courses of action that would reveal the truth, individuals render the knowable unknown (Haight 1980). Acting in ways that keep one uninformed about unwanted information is self-deception. Acting in selectively biasing and misinforming ways is a process of self-bias. These are different phenomena. The truth is not harbored in a dissociated unconscious mind. It exists in the information available in the avoided reality. The disconnected unconscious mind cannot know the truth if the evidential basis for it is avoided.

VH&T emphasize the benefits of self-deception but ignore the social costs to both the self-deceiver and deceived. Being misled is costly to others. Therefore, the deceived do not take kindly to being led astray. Human relationships are long-term affairs. There are limits to how often one can mislead others. After a while, the victims quit caring about whether the deception was intentional or carried out unknowingly. If done repeatedly, the short-term gains of misleading others come with the cost of discrediting one's trustworthiness. Undermining one's ability to exercise social influence does not have adaptive value.

Human behavior is regulated by self-sanctions, not just by social ones. Unless self-deceivers are devoid of moral standards, they have to live with themselves as well as with the social reactions of others to deceptive conduct. The maintenance of positive self-regard while behaving harmfully is a strong motivator for self-exoneration. Self-deception serves as a means of disengaging moral self-sanctions from detrimental conduct. The social cognitive theory of moral agency specifies a variety of psychosocial mechanisms by which individuals convince themselves that they are doing good while inflicting harm on others (Bandura 1999). By these means, otherwise considerate people can behave inhumanely without forfeiting their sense of self-worth.

Is social interaction based on guile or honesty?

doi:10.1017/S0140525X10002621

Matthew L. Brooks and William B. Swann, Jr.

Department of Psychology, University of Texas, Austin, TX 78712-0187.

mattlbrooks@gmail.com **swann@mail.utexas.edu**

http://homepage.psy.utexas.edu/homepage/faculty/swann/

Abstract: Von Hippel & Trivers suggest that people enhance their own self-views as a means of persuading others to adopt similarly inflated perceptions of them. We question the existence of a pervasive desire for self-enhancement, noting that the evidence the authors cite could reflect self-verification strivings or no motive whatsoever. An identity negotiation framework provides a more tenable approach to social interaction.

We were impressed with many of von Hippel & Trivers's (VH&T's) specific arguments. For example, they build a convincing case that self-deception may facilitate the manipulation of others and that it can be used to effectively mask clues as to one's intent. Yet when it comes to the contention that the deception of the self and others represents a core ingredient of human social interaction, we respectfully disagree.

Some of our reservations regarding VH&T's central "self-deceptive self-enhancement" thesis are purely conceptual. At one point the authors assert that "People are impressed by confidence in others" (sect. 3, para. 2) and then proceed to list the supposed benefits of confidence. Fair enough. But they then conclude that the premium humans place on confidence suggests that "self-enhancement should be ubiquitous and people should believe their own self-enhancing stories" (sect. 3, para. 2). Whereas *confidence* is often based on actual abilities or achievements, self-deceptive self-enhancement is thought to produce *over*confidence. Surely the consequences of confidence and overconfidence are very different. In fact, we could readily imagine a complementary paragraph that lists the hazards of overconfidence and concludes that self-enhancement should be maladaptive and hence rare.

Empirical evidence of self-verification strivings provides an even stronger basis for concluding that self-enhancement plays a relatively modest role in social interaction. For example, contrary to VH&T's contention that people preferentially seek and embrace positive evaluations, there is strong evidence that people with negative self-views preferentially seek negative evaluations and interaction partners and even divorce spouses who perceive them in an overly positive manner (e.g., Swann 1983; in press).

VH&T dismiss this literature by suggesting that self-verification is merely another motive that exists alongside self-enhancement. What they fail to recognize, or at least acknowledge, is that their argument loses force insofar as self-verification overrides self-enhancement. Recent evidence suggests that this may indeed be a problem for the authors' formulation. A meta-analysis of studies that pitted self-enhancement against self-verification indicated that self-verification effects were at least as strong as self-enhancement effects (Kwang & Swann 2010). Moreover, when the response class "cognitive reactions" (which included selective attribution, attention, recall, overclaiming bias, and perceived accuracy) was examined specifically, the overall pattern favored self-verification over self-enhancement. This is noteworthy because it is precisely this response class that VH&T emphasize as the province of self-enhancement.

At the very least, evidence that self-verification strivings often trump self-enhancement strivings challenges the notion that there exists a pervasive desire for self-enhancement. More tellingly, however, such evidence also calls into question the most appropriate interpretation of much of the evidentiary basis for self-enhancement. Consider that there is convincing evidence that roughly 70% of the people in the world have positive views of themselves (Diener & Diener 1995), presumably because most people enjoy secure attachment relationships (Cassidy 1988; Sroufe 1989) and are able to systematically seek and engineer success experiences (e.g., Bandura 1982). The pervasiveness of positive self-views, in conjunction with evidence that people work to verify their negative and positive self-views, means that evidence that people seek and embrace positive evaluations could reflect either self-verification or self-enhancement strivings. Of particular relevance here, studies in which most participants embrace positive feedback may reflect a tendency for the 70% of participants with positive self-views to seek feedback that is, for them, self-verifying.

To concretize our claim, consider the Fein and Spencer (1997) article that the authors cite repeatedly. Because no measure of self-views was included in this research, it is possible that the tendency for negative feedback to amplify participants' derogation of out-group members was driven by a tendency for those with positive self-views (who presumably constituted most of the sample) to work to maintain their positive self-views. If so, then the findings may reflect self-verification rather than self-enhancement strivings. Of course, the authors also cited a study by Epley and Whitchurch (2008), which *did* include a measure of global self-esteem. Yet expecting the measure of self-esteem to act as a moderator in this study is problematic because there is little reason to believe that feedback regarding physical attractiveness should be moderated by global esteem (which is influenced by numerous factors besides attractiveness; for a discussion of this specificity matching problem, see Swann et al. 2007).

The authors also rely heavily on a series of studies that purport to show that people routinely claim that they are "better than average." Whether this body of work should be viewed as evidence of self-enhancement, however, is debatable. For example, careful analyses have shown that participants are rather non-discriminating when it comes to endorsing the above-average option, even claiming that an unknown stranger performs better than average (Klar & Giladi 1997). The most straightforward interpretation of these findings appears to be that people have a very dim understanding of what an average score means and what it means to assert that they are better than average. If so, the results of such studies can hardly be regarded as evidence of self-enhancement (for a review, see Chambers & Windschitl 2004).

In short, although there are surely instances in which deception is personally beneficial, we believe that the research literature provides little evidence that such activities offer an apt characterization of human social conduct. Instead, a more tenable model of human social interaction may be offered by the identity negotiation formulation (e.g., Swann 1987; Swann & Bosson 2008). When people begin interacting, the argument goes, their first order of business is to determine "who is who." Once each person lays claim to an identity, they are expected to honor it henceforth; failure to do so will be disruptive to the interaction and could even trigger termination of the relationship. So, people are not only motivated to seek subjectively

accurate (i.e., self-verifying) feedback, but also their success in eliciting such feedback represents the interpersonal "glue" that holds their relationships together. Within this framework, social relationships are maintained through transparency and mutual understanding rather than deceit and obfuscation, and it is allegiance to truth that enables people to enjoy healthy, prosperous relationships.

Domains of deception

doi:10.1017/S0140525X10002682

David M. Buss

Department of Psychology, University of Texas, Austin, TX 78712.
dbuss@psy.utexas.edu
www.davidbuss.com

Abstract: The von Hippel & Trivers theory of self-deception will gain added traction by identifying psychological design features that come into play in different domains of deception. These include the domains of mating, kinship, coalition formation, status hierarchy negotiation, parenting, friendship, and enmity. Exploring these domains will uncover psychological adaptations and sex-differentiated patterns of self-deception that are logically entailed by their theory.

The von Hippel & Trivers (VH&T) theory provides a powerful explanation of the several evolved functions of self-deception. Their theory provides a compelling account of the information processing mechanisms – such as dissociations between conscious and unconscious memories, biased information search, and automatic versus controlled processes – that plausibly explain how self-deception in general can be implemented. I suggest that a more comprehensive theory of self-deception will require identifying specific psychological design features relevant to different domains of deception. These domains will be defined, in large part, by predictable regions of social conflict, such as sexual conflict, intrasexual rivalry, parent-offspring conflict, and coalitional conflict.

Consider sexual conflict in the mating arena. It has been documented that men and women attempt to deceive members of the opposite sex in different ways and for different goals (Buss 2003; Haselton et al. 2005). Men, for example, sometimes deceive women about the depth of their feelings of emotional involvement for the goal of short-term sexual access. According to the VH&T theory of self-deception, men should self-deceive in this domain, truly believing that they feel more deeply about the woman than they actually do, prior to sexual consummation, in order to better carry out the deceptive strategy – a prediction yet to be tested.

Studies of personal ads that have checked self-reported qualities with objectively measured qualities find that men tend to deceive about their height and women about their age and weight. Men report that they are taller and women that they are younger and weigh less than objective verifications reveal. Do men really believe their deceptively reported stature? Do women really believe their deceptively reported youth and shaved pounds? And if so, do these self-deceptions better enable deception of opposite sex on these qualities? These are examples of sex-differentiated empirical predictions from the VH&T theory of self-deception that require empirical study. If verified, they would imply that men's and women's psychology of self-deception contain somewhat different content-specific design features.

Deception within families differs from the exaggeration of mating qualities in the service of mate attraction. Do children delude themselves about their hunger in order to better deceive their parents about their true level of need? Do they believe their lies of the physical pain inflicted by their siblings in order to better manipulate their parents? Do parents truly believe that they love all their children equally, when they clearly do not, in order to deceive their children for the goal of minimizing costly sibling conflict? Do stepparents truly believe that they love their stepchildren as much as their genetically related progeny? These within-family domains of deception point to psychological design features that differ from those within the mating domain, and they require specification in any comprehensive theory of deception and self-deception.

Wrangham (1999) hypothesized that males in warfare coalitions deceive themselves about the probability of victory, particularly in battles – a positive illusion bias. His hypothesis is that this form of self-deception functions to increase the likelihood of successful bluffs. A complementary hypothesis, I suggest, is that leaders self-deceive about the likelihood of their success to better unify and motivate other males within their coalition, with the goal of increasing the likelihood of success in battle. These hypotheses point to psychological design features associated with coalitional conflict that differ from those that occur in the domains of sexual conflict or within-family conflict. They also suggest a specific psychology of deception and self-deception present in men, but absent in women – specifications required for any comprehensive theory of deception and self-deception.

Analogous arguments can be made in other domains, such as intrasexual rivalry conflict, tactics of hierarchy negotiation, equity negotiations with friends, and tactics to deter enemies. The VH&T theory of self-deception provides a compelling information-processing foundation from which a more comprehensive theory can be built. Exploring deception in domain-specific and sex-differentiated ways, with the recognition that different adaptive problems often require somewhat different information-processing solutions, opens avenues for discovering a rich array of psychological adaptations that accompany the functional implementation of specific forms of deception and self-deception.

Get thee to a laboratory

doi:10.1017/S0140525X10002530

David Dunning

Department of Psychology, Uris Hall, Cornell University, Ithaca, NY 14853.
dad6@cornell.edu
http://cornellpsych.org/sasi/index.php

Abstract: von Hippel & Trivers's central assertion that people self-deceive to better deceive others carries so many implications that it must be taken to the laboratory to be tested, rather than promoted by more indirect argument. Although plausible, many psychological findings oppose it. There is also an evolutionary alternative: People better deceive not through self-deception, but rather by not caring about the truth.

In their thoughtful and stimulating essay, von Hippel & Trivers (VH&T) assert a number of intriguing proposals, none more thought-provoking than their central one that self-deception evolved in order to facilitate the deception of others.

My overall reaction to this central assertion is favorable. It is a well-formed hypothesis that readers easily grasp and that resonates with their intuition. The hypothesis, however, lacks one characteristic I wish it had more of – data. That is, the hypothesis is not completely new, having been forwarded, in some form or another, over that last quarter-century (Trivers 1985; 1991), and so it could profit now from direct data that potentially support it rather than from any additional weaving of indirect arguments and findings such as those the authors have spun here. It should be relatively easy to construct empirical studies to see if people engage in self-deception more eagerly when they must persuade another person of some proposition. Similarly, it should be easy to create experiments to see if people are more persuasive to others to the extent they have persuaded themselves of some untruth first.

Such empirical evidence, ultimately, is essential for two reasons. First, although intuitively compelling, VH&T's hypothesis already faces an empirical headwind. Many research findings oppose it. For example, VH&T suggest that people need self-deception to become convincing liars because others would be so good at catching their lies otherwise. However, one can reasonably read the literature on lie-detection to suggest that people are not very good at detecting lies (e.g., Bond & DePaulo 2006; Ekman 1996), so this pressure does not really exist. Moreover, people do not seem to be especially skilled at lie detection under circumstances in which they arguably should be, given the authors' assertions. For example, people are not much better at detecting lies among those they know well (Anderson et al. 2002) or among those they are interacting with directly rather than merely overhearing (Bond & DePaulo 2006).

In addition, one could argue from extant data that evolution would not have selected for self-deception, and the lying it supports, in the social setting associated with most of human evolution – one in which humans huddled together in small, interdependent groups, banded together against the potentially fatal dangers of nature and other competitive social factions (Brewer & Caporael 2006). Recent arguments suggest that this small group setting boosted the survival value of groups whose members cooperated and cared for one another over groups whose members were more egoistic and selfish (e.g., Boyd & Richerson 2009). One can presume that truth telling would be one of those behaviors so selected – and that self-deceptive lying within the small group would place any of its practitioners at a disadvantage.

Consider the fate of people in small groups who actively pursue one form of self-deception that VH&T discuss at length – people who engage in self-deceptive self-enhancement that allows them to display more confidence about themselves than their reality actually warrants. According to VH&T, this self-deceptive confidence is associated with many social advantages, but modern-day empirical data suggest the opposite – that this self-deception does not work well in small-group settings. At least two studies have found that people who boast with confidence about their talents and character are initially well-liked in small groups. However, over time, these individuals become the most disliked and least valued within those small social groups (John & Robins 1994; Paulhus 1998). Presumably, their self-deceptions are eventually found out, and whatever advantages they obtain initially are ones the group increasingly withholds as time goes on. This leads to a paradox. Perhaps self-deception in the service of deceiving others may plausibly work in contemporary social life, which is marked by a rather anonymous, ever socially shifting world. In the modern day, one can deceive and then move on to deceive other strangers. But what about a human evolutionary past in which people did not move on, but rather woke up each morning to deal with the same small group of individuals for most of their mortal lives?

Finally, VH&T's intuitive central assertion must be put to empirical test because there is an equally intuitive alternative. People may become more persuasive not because they deceive themselves of some illegitimate fact, but they instead decide that the facts just do not matter. They lay aside the truth and are unconcerned about it, making whatever claims they think will be the most convincing to the other person. This technique of simply not caring about the truth has been labeled by the philosopher Harry Frankfurt as *bullshitting* (Frankfurt 2005), and it potentially is a strategy that would make self-deception unnecessary under the authors' current framework. Thus, if evolution crafted the best liars among humans, perhaps it did so not via the route of self-deception but rather by creating individuals (or a species) who could dismiss any worry about the truth-value of what they were saying as an active consideration as they said it. Laboratory work could examine this. Are people better persuaders when they self-deceive themselves into some untruth? Or are they better persuaders when they strike any consideration of truth or falsity from their minds? Current social cognitive techniques seem tailor-made to tackle this question.

All these observations lead me to ask of the authors – or any interested reader: Get thee to a laboratory. The central contention guiding this essay is too broad and deep in its implications not to deserve direct and extensive empirical study.

Self-deception is adaptive in itself

doi:10.1017/S0140525X10002542

Louisa C. Egan

Kellogg School of Management, Northwestern University, Evanston, IL 60208.
louisa-egan@kellogg.northwestern.edu
http://www.kellogg.northwestern.edu/faculty/directory/egan_louisa.aspx

Abstract: Von Hippel & Trivers reason that the potential benefits of successfully deceiving others provide a basis for the evolution of self-deception. However, as self-deceptive processes themselves provide considerable adaptive value to an individual, self-deception may have evolved as an end in itself, rather than as the means to an end of improving other-deception.

Von Hippel & Trivers (VH&T) argue that the suite of motivated processes associated with self-deception (SD) evolved to facilitate other-deception (OD). VH&T contend that a capacity for SD may increase individuals' success in specific instances of OD by helping them conceal deception-revealing "tells," such as those due to nervousness and cognitive load. A capacity for SD also produces self-serving biases and self-enhancement, which help individuals appear more confident and "better" to others than they really are, and thus increases one's chances of reaping interpersonal rewards. Furthermore, the optimism and happiness associated with self-serving biases and self-enhancement may produce interpersonal gains. Although it is true that downstream interpersonal benefits of a capacity for SD may increase evolutionary fitness, the immediate benefits of SD alone increase potential for reproductive success more directly. It is therefore unclear that selective pressures arising from the need to be effective at OD are necessary to explain the evolution of SD.

An ability to succeed in OD has clear benefits. Individuals who can secure undeserved resources possess a significant advantage over individuals who cannot, in terms of potential for reproductive success. To address whether a capacity for SD leads to greater effectiveness at OD in a given context, VH&T propose that researchers ought to examine whether individuals are most likely to engage in SD when they are motivated to deceive others. This test would indeed inform the discussion of whether SD assists in OD and is worthy of serious consideration. A related test is whether individuals are most likely to engage in self-enhancement on traits for which it might be particularly useful to succeed in OD. The appearance of morality is one dimension in which success in OD would be particularly helpful: If Sally can deceive Anne into thinking that Sally is moral, Anne will not only possess a generally more positive view of Sally, but Anne will also likely lower her guard against future deceptions from Sally. Deceiving others into believing that one is moral should be a useful tactic for everybody, as evidence suggests that vigilance against cheaters is a human universal (Sugiyama et al. 2002). Therefore, if SD evolved in the service of OD, one might reasonably expect that all individuals self-enhance on the dimension of morality. Contrary to this prediction, people with interdependent mind-sets are less likely to self-enhance in the moral or altruistic domains than are people with independent mind-sets (Balcetis et al. 2008). This effect holds controlling for culture. This finding casts doubt on VH&T's model of SD's existence for the purpose of facilitating OD.

Considerable evidence indicates that a capacity for motivated cognition, a variety of SD, helps us to attain our goals. Objects we desire may appear larger (Bruner & Goodman 1947) or closer (Balcetis & Dunning 2010) to us, and individuals predict that desired, randomly determined outcomes will occur (e.g., Babad 1997; but see Krizan & Windschitl 2009). Such desire-biased perceptions and predictions can cause us to frame situations in terms of possible gains, making us more likely to pursue courses of action that will allow us to achieve desirable outcomes (e.g., Bandura 1989, p. 1177; Sternberg & Kolligan 1990). Conversely, individuals who are motivated to avoid certain entities (such as arachnophobes toward spiders) exhibit greater vigilance for those things (Riskind et al. 1995). The concrete results of our goal-oriented behaviors (e.g., food or safety, both of which were often uncertain in our evolutionary history) provide adaptive value, with interpersonal adaptive value possibly as mere epiphenomena.

VH&T argue that self-serving biases and self-enhancement cause us to be optimistic and happy. Overly positive thinking manifested as optimism and happiness contributes to good health, goal attainment, and resilience in the face of adversity (e.g., Taylor & Armor 1996; Taylor & Brown 1988). VH&T discuss optimism and happiness as possible products of SD with interpersonal benefits. Optimism, in VH&T's formulation, causes individuals to confidently persevere and occasionally succeed in difficult tasks, resulting in dividends that translate into interpersonal currency. (Underscoring optimism's adaptive value, Aspinwall and Richter (1999), have demonstrated that optimism is also associated with abandoning an impossible task within a shorter timeframe.) VH&T point out that both optimism-driven confidence and happiness draw others to an individual, leading to greater potential for success on cooperative tasks. However, although VH&T acknowledge that sunny outlooks are associated with positive effects aside from interpersonal concerns, they discount the adaptive value of such outcomes and instead present optimism and confidence as means to interpersonal success.

Self-serving rationalizations and interpretations of our own behavior allow us to preserve our rosy views of ourselves and our prospects. Among other rationalizations, individuals reduce cognitive dissonance to uphold their views of themselves as moral and competent (e.g., Festinger & Carlsmith 1959; Steele & Liu 1983). Individuals from preschool-aged children to adults are motivated to rationalize even blind choices (e.g., Egan et al. 2010; Johansson et al. 2005). People engage in moral hypocrisy, whereby they convince themselves that their selfish behaviors are acceptable (e.g., Batson & Thompson 2001) and judge their own immoral behavior as less egregious than that of others (e.g., Valdesolo & DeSteno 2007). Thus, our views of ourselves are consistently bolstered by means of self-serving biases and self-enhancement throughout our everyday existence. SD does not require demands from OD to have evolved – it is self-sustaining, and certainly provides considerable adaptive advantages independently. Although a capacity for SD may enhance one's skills in OD, it seems more likely that OD enjoys a commensal relationship with SD than that it drives the evolution of SD.

Conscious thinking, acceptance, and self-deception

doi:10.1017/S0140525X10002554

Keith Frankish

Department of Philosophy, The Open University, Walton Hall, Milton Keynes, Buckinghamshire MK7 6AA, United Kingdom.

k.frankish@gmail.com

http://www.open.ac.uk/Arts/philos/frankish.htm

Abstract: This commentary describes another variety of self-deception, highly relevant to von Hippel & Trivers's (VH&T's) project. Drawing on dual-process theories, I propose that conscious thinking is a voluntary activity motivated by metacognitive attitudes, and that our choice of reasoning strategies and premises may be biased by unconscious desires to self-deceive. Such biased reasoning could facilitate interpersonal deception, in line with VH&T's view.

In their target article, von Hippel & Trivers (VH&T) invoke a dual-process framework, arguing that self-deception is facilitated by dissociations between implicit and explicit memories, attitudes, and processes. However, they focus on work in learning and social psychology and say relatively little about dual-process theories of reasoning and judgment. Such theories are, however, highly relevant to VH&T's project, and in this commentary I add some further supporting considerations, drawing on work in this area.

There is now considerable evidence for the existence of two distinct but interacting types of processing in human reasoning and decision making: type 1, which is fast, effortless, automatic, unconscious, inflexible, and contextualized, and type 2, which is slow, effortful, controlled, conscious, flexible, and decontextualized (e.g., Evans 2007; Evans & Over 1996; Kahneman & Frederick 2002; Sloman 1996; Stanovich 1999; 2004; for surveys, see Evans 2008; Frankish & Evans 2009; Frankish 2010). Beyond this core agreement there is much debate, concerning the further properties of the two processes, the relations between them, and whether they are associated with distinct neural systems (see, e.g., the papers in Evans & Frankish 2009). Here, however, I wish to focus on a specific proposal about the nature of type 2 processing.

The proposal is that type 2 processing is best thought of as an internalized, self-directed form of public argumentation, and that it is a voluntary activity – something we do rather than something that happens within us (Frankish 2004; 2009; see also Carruthers 2006; 2009a; Dennett 1991). It might involve, for example, constructing arguments in inner speech, using sensory imagery to test hypotheses and run thought experiments, or interrogating oneself in order to stimulate one's memory. On this view, type 2 thinking is *motivated*; we perform the activities involved because we desire to find a solution to some problem and believe that these activities may deliver one. (Typically, these metacognitive beliefs and desires will be unconscious, implicit ones.) I have argued elsewhere that this view provides an attractive explanation of why type 2 thinking possesses the distinctive features it does (Frankish 2009).

On this view, we also have some control over our conscious mental *attitudes*. If we can regulate our conscious thinking, then we can decide to treat a proposition as true for the purposes of reasoning and decision making, committing ourselves to taking it as a premise in the arguments we construct and to assuming its truth when we evaluate propositions and courses of action. Such premising commitments constitute a distinct mental attitude, usually called "acceptance" (Bratman 1992; Cohen 1992; Frankish 2004). When backed with high confidence, acceptance may be regarded as a form of belief (Frankish 2004), but it can also be purely pragmatic, as when a lawyer accepts a client's innocence for professional purposes. Such pragmatic acceptance will, however, be functionally similar to belief, and it will guide inference and action, at least in contexts where truth is not of paramount importance to the subject.

If this is right, then it points to further powerful avenues of self-deception, involving biased reasoning, judgment, and acceptance. Our reasoning activities and premising policies may be biased by our self-deceptive goals, in pursuit either of specific ends or of general self-enhancement. We may be motivated to display greater effort and inventiveness in finding arguments for conclusions and decisions we welcome and against ones we dislike. And we may accept or reject propositions as premises based on their attractiveness rather than their evidential support. Of course, if we accept a claim in full awareness that we are doing so for pragmatic reasons, then no self-deception is involved; our attitude is like that that of the lawyer. Self-deception enters

when we do not consciously admit our aims, and engage in biased reasoning and the other forms of self-deception described by VH&T (biased search, biased interpretation, etc.) in order to support the accepted claim.

I have previously set out this view of self-deception at more length (Frankish 2004, Ch. 8). However, I there assumed that the function of self-deceptive acceptance was primarily defensive. I referred to it as a "shielding strategy" designed to protect one from consciously facing up to an uncomfortable truth. But biased reasoning and acceptance could equally facilitate interpersonal deception, in line with VH&T's view. To accept a proposition as a premise is, in effect, to simulate conscious belief in it, both inwardly, in one's conscious reasoning and decision making, and outwardly, in one's behavior (so far as this is guided by one's conscious thinking). In doing this, one would display the signals of genuine belief to others whom one might wish to deceive. Moreover, these signals of belief might have an influence upon oneself as well, being taken by unconscious belief-forming processes as evidence for the truth of the accepted proposition (a sort of self-generated testimony) and thus fostering implicit belief in it. In this way, a deception that begins at the conscious level may later extend to the unconscious one, thereby eliminating any unconscious signals of deceptive intent.

Biased conscious thinking and acceptance are closely related to the information-processing biases discussed by VH&T, and they should be detectable by similar means. In particular, where they serve the goal of self-enhancement, they should be reduced when prior self-affirmation has taken place (provided, that is, that the deception has not taken root at the unconscious level, too). Experimental manipulations of cognitive load might also be employed to detect self-deceptive bias. There are complications here, however. For although biased conscious thinking will be effortful and demanding of working memory, it will not necessarily be *more* demanding than the unbiased sort, which, on the view we are considering, is also an effortful, intentional activity. However, when self-deception involves specific deviations from established reasoning strategies and premising policies, it will require additional self-regulatory effort, and in these cases, manipulations of cognitive load should affect it. "Talk-aloud" and "think-aloud" protocols, in which subjects are asked to verbalize and explain their thought processes, should also be useful in helping to identify self-deceptive biases in conscious thinking.

The evolutionary route to self-deception: Why offensive versus defensive strategy might be a false alternative

doi:10.1017/S0140525X10002645

Ulrich Frey and Eckart Voland

Zentrum für Philosophie und Grundlagen der Wissenschaft, Universität Giessen, D-35394 Giessen, Germany.

Ulrich.Frey@phil.uni-giessen.de Eckart.Voland@phil.uni-giessen.de
http://www.uni-giessen.de/cms/fbz/zentren/philosophie/ZfP/wiss_mitarbeiter/frey
http://www.uni-giessen.de/cms/fbz/zentren/philosophie/ZfP/biophil

Abstract: Self-deception may be the result of social manipulation and conflict management of social in-groups. Although self-deception certainly has offensive and defensive aspects, a full evolutionary understanding of this phenomenon is not possible until strategies of other parties are included into a model of self-perception and self-representation.

Von Hippel & Trivers (VH&T) develop a complex picture of selective information, processing and present a wealth of different evidence concerning self-deception. This complexity is already one reason to doubt that self-deception is either a purely offensive or a defensive strategy. The evidence cited supports self-deception both enhancing *and* decreasing fitness. Examples for the latter are wrong body perceptions of anorexic people, denials of being addicted, or putting a gloss on violent relationships. Even if self-deception enhances fitness – as part of an offensive strategy – obvious costs arise. Therefore, the benefits have to be substantial to overcome this barrier.

One problem might be that self-deceptions are not just taken at face value by others but are verified. Therefore, follow-up costs are high. Take as an example self-assessment. If individuals overestimate their own capabilities in physical contexts, this will invariably lead to serious injuries or death. In social contexts, rivals and allies alike will frequently challenge these alleged qualities, which will then break down and produce negative consequences. In mating contexts, overestimating one's qualities and, as a result, courting superior mates will lead to rejections, given that discriminating abilities in mating contexts are highly developed (not only in humans, but also in many species).

For that reason, we would like to suggest a modification of self-deception as offensive strategy: Its continued use should depend on its success. If a particular deception is successful, then self-deception can be incorporated via the subconscious processes suggested by VH&T, because discrepancies to external perception are apparently not too large. If, on the other hand, such bluffs are called, self-deception should no longer be used in these kinds of situations making way again for an accurate self-perception.

Deception and self-deception are, furthermore, especially hard to keep up in stable groups over longer periods. Problems include intimate knowledge of others and high costs of discovered deception. However, there is an evolutionary mechanism for this problem: costly signals. They ensure that even in a world of egoistic individuals, honest and reliable communication can be effected (Zahavi & Zahavi 1997). This implies, however, that each communication should be treated as possibly deceptive by default if not backed up by an honest signal.

Costly signaling weakens the argument that "Self-enhancement is useful only to the degree that it is self-deceptive, because only when it is believed by the self will others accept the enhanced self as genuine." (sect. 8, para. 9). Because signals can and are indeed faked, it follows that individuals should and do in fact rely on their *own* knowledge to evaluate the truth of any signal. Genuine signals require at least agreement between the signal of the other individual *and* external cues, as well as the knowledge of the receiver. If there are discrepancies – which would often be the case if self-enhancement is faked – such signals should be discarded as dishonest or be inspected more closely.

Self-deception as defensive strategy is implausible, too. We completely agree with VH&T that self-deception as a means to cope with a threatening world confuses means with ends from an evolutionary perspective – hedonistic rewards per se are not the ultimate target of selection.

Given that the evidence presented does not favor self-deception as either a purely offensive or defensive strategy, we would like to put forward a third model. Here, self-deception is seen as incongruence between self-perception and perception by others. Selective information processing is then used as a strategy to keep that incongruence.

It is essential to keep in mind that children do not have autonomy when constructing their self-perception. Humans develop their self-perceptions in light of others, through a process of attribution primarily by members of the kin group and not through "objective" introspection (Carruthers 2009b; Prinz 2008; Voland 2007). Coupled with the fact that humans are "cooperative breeders" with all the corresponding strategies of cognitive networking in the group (Hrdy 2009), it follows that selfishly motivated individuals may influence others. This is particular

true for kin groups – for example, getting children to adapt certain roles like "the hero," "the helper," and so forth, to enhance the fitness of members of the kin group, irrespective of the fitness of the children themselves. This was first pointed out by Trivers (1974) and labeled with the term "parent–offspring conflict."

Kurland and Gaulin (2005, p. 453) could be right when they point out that "some humanists have found in our peculiarly intensive family ecology the source of all neurosis, psychosis, and the world's troubles." Implications of this family-conflict model include self-deception as well, because the parent–offspring conflict is not only an investment conflict (Salmon 2007), but may also become a pronounced role conflict.

If role expectations are actually accepted by the persons concerned, this could mean that the parent–offspring conflict has been won by members of the kin group and explain why such "wrong" self-perceptions are not corrected, even if self-deception is costly. The obvious reason is that it is costly, too, to avoid self-deception. Such costs are attested by psychological studies. Evidence suggests that self-perception has to be without inner contradictions. Personality disorders like schizophrenia attest to this claim. Many biases ensure a whole personality (e.g., hindsight bias, confirmation bias, attributional biases, etc.; see Frey 2010; Gilovich 1991; Plous 1993). Cognitive dissonance is stressful (Taylor 1989/1995), and correcting cognitive dissonance is also stressful because it means conflict with the manipulating party. Stress, however, is costly (Flinn 2007). Self-representation, which includes questions of self-deception, therefore constitutes a typical trade-off problem. Thus, it may be fitness enhancing for individuals to let themselves be manipulated.

To conclude, our alternative answers the question posed by VH&T: "If self-deception evolved to deceive others, why is there so much evidence for self-deception that appears to be intended only for the self?" Other parties with different interests must be included in an analysis of self-deception.

Reviewing the logic of self-deception

doi:10.1017/S0140525X10002566

Ellen Fridland

Berlin School of Mind and Brain, Humboldt University of Berlin, 10099 Berlin, Germany.

ellenfridland@yahoo.com

http://sites.google.com/site/ellenfridland/

Abstract: I argue that framing the issue of motivated belief formation and its subsequent social gains in the language of self-deception raises logical difficulties. Two such difficulties are that (1) in trying to provide an evolutionary motive for viewing self-deception as a mechanism to facilitate other-deception, the ease and ubiquity of self-deception are undermined, and (2) because after one has successfully deceived oneself, what one communicates to others, though untrue, is not deceptive, we cannot say that self-deception evolved in order to facilitate the deception of others.

The argument that self-deception evolved in order to facilitate the deception of others relies on the assumption that, pace all empirical evidence, people really are good lie detectors. Von Hippel & Trivers (VH&T) claim that "despite what the research literature might appear to show, people are actually quite good at detecting deception. This possibility is central to our hypothesis regarding a co-evolutionary struggle and the subsequent origins of self-deception." VH&T argue that empirical studies showing that people are poor lie detectors are fundamentally flawed. As such, the authors posit the opposite conclusion – that, in fact, people are very good at detecting lies, especially when they are properly motivated and dealing with close friends and family.

However, the problem is that *if* the hypothesis that people are good lie detectors is true, then the facility with which people are able to lie to themselves is undermined. In fact, one reason the authors offer to explain the misleading empirical results is that most studies of deception are conducted with strangers – VH&T say that we are much better at detecting the lies of people with whom we are close. But, and this is the kicker, if people get better at detecting lies as their relationship to the liar becomes closer and closer, then deceiving oneself should be hardest of all. After all, there is no one closer to us than ourselves. Thus, in trying to provide an evolutionary motive for viewing self-deception as a mechanism to facilitate other-deception, VH&T undermine the conditions required to make self-deception common and easy.

The relationship between skill in detecting lies and lying to oneself must be inversely proportional: The better one is at lie detection, the harder it will be to self-deceive. The fact remains that we must be fairly poor lie detectors if we are able to lie to ourselves with ease. Unfortunately, on the account that VH&T have forwarded, the commonness and facility of self-deception becomes difficult, if not impossible, to explain.

One of the virtues of VH&T's account of self-deception is that it moves away from the standard model of self-deception, which requires two simultaneous, contradictory representations of reality. VH&T highlight the fact that self-deception does not require one person to possess contradictory beliefs, but rather, self-deception can be the result of biases in information gathering, which reflect one's own goals or motives. As VH&T state, "if I can deceive you by avoiding a critical piece of information, then it stands to reason that I can deceive myself in the same manner."

The problem, however, is that if one sincerely believes that p, then when one expresses that p to someone else, that expression is not an instance of deception, despite the fact that p is false. After all, if I sincerely believe that Canada is in Eastern Europe, and I communicate to you that Canada is in Eastern Europe, although I am wrong, I am not lying. In order to lie, I must know that what I am communicating is false.

So, arguably, we can get a real instance of other-deception on the classical view of self-deception. We can say that one knows that p is false somewhere in one's unconscious, and so, when one expresses that p is true, one is being deceptive. However, if a person never came to believe that p is false, even though she *could* have come to know that p is false (given more search, etc.), then when she communicates that p, she is not lying.

Given that in VH&T's view, the person who expresses falsehoods believes them to be truths, that person cannot be involved in other-deception. Again, other-deception requires knowledge of the falsity of what one is communicating. As such, self-deception could not have evolved to facilitate other-deception, because after one has successfully deceived oneself, what one communicates to others, though untrue, is not deceptive.

These points are not meant to show that VH&T's general orientation is misguided but rather that the language of self-deception has not done them any favors. VH&T have successfully illustrated that our beliefs often develop in ways that are relative to our motives and goals. Also, they have done well to argue that these biased beliefs result in a whole host of social gains.

Rather than pursue an evolutionary account of self-deception, however, it seems that the real revelation is in the following point that VH&T cite but do not develop: "Thus, the conventional view that natural selection favors nervous systems which produce ever more accurate images of the world must be a very naive view of mental evolution" (Trivers 1976/2006). For, it is this assumption that creates the dichotomy between "normal" beliefs that correspond to facts and "self-deceptive" beliefs that are the result of motivated biases. Rather than looking at these two kinds of practices as normal and abnormal, the real sea change would be to think of beliefs, in general, as developed in motivated settings, relative to an agent's abilities, goals, environmental conditions,

and intersubjective situation. The real revelation would be to question the assumption that beliefs are most useful when true.

Directions and beliefs of self-presentational bias

doi:10.1017/S0140525X10002086

David C. Funder

Department of Psychology, University of California, Riverside, CA 92521.
funder@ucr.edu
http://rap.ucr.edu

Abstract: The target article tends to conflate self-deception and self-enhancement, but biased self-presentation can be negative as well as positive. Self-deceiving self-diminishers may be depressed and non-self-deceiving self-diminishers may project false modesty. The article's otherwise brilliant argument for the advantages of self-deceptive self-enhancement for deceiving others may underemphasize the risks it entails for poor decision making.

This brilliant paper goes a long way to resolve a long-standing conflict between ancient Delphic wisdom that urges "know thyself" and the somewhat more recent evidence, most prominently offered by Taylor and colleagues (Taylor & Brown 1988; 1994), that self-deception, in a self-enhancing direction, may incur some advantages. Misleading yourself can help you to mislead others, and a little bit of extra confidence in one's own attractiveness or abilities can enhance one's perceived attractiveness and competence. I was once a peripheral participant in the debate over the adaptiveness of self-enhancement (Colvin et al. 1995). Speaking only for myself, I can say that von Hippel & Trivers' (VH&T's) argument has convinced me that a modicum of self-enhancement can indeed, in some circumstances – and especially when accompanied by self-deception – be adaptive.

However, as ambitious as it is, the target article's exposition of the mechanisms and implications of self-deception still neglects three key points. First, self-enhancement is not is not the only direction in which self-presentation can be biased. People often present themselves as better than they are, but they also sometimes present themselves as worse than they are. In one study, while about 35% of the participants showed a self-enhancement bias, 15% showed the opposite, self-diminishment bias, and the remaining 50% were fairly accurate (John & Robins 1994). I look forward to VH&T's evolutionary explanation of what those 15% were up to.

Second, the observation that biased self-presentation can occur in either direction highlights the separateness of self-deception and self-enhancement, which the target article tends to conflate. Consider the following 2×2 table. The cell entries are simplified.

The target article is mostly about the upper left cell. The two right-hand cells are not acknowledged at all. The lower left cell is acknowledged, implicitly, when VH&T imply that that the maladaptiveness and mental unhealthiness of self-enhancement is limited to those cases in which people "are... unconvinced by their own self-enhancing claims."

Table 1 (Funder). *Implications of self-enhancement versus self-diminishment*

	Self-enhancement	Self-diminishment
Self is deceived	Confidence	Depression
Self is not deceived	Bombastic narcissism	False modesty

Third, and despite the target article's implication, it is easy to imagine many cases in which the confidence engendered by sincerely believed self-enhancements could be harmful. Misleading oneself about one's own abilities can lead to years of wasted effort as a failed artist, writer, or premed student. Misleading oneself about one's attractiveness can lead to the pursuit of unattainable mates at the expense of perfectly suitable mates one could otherwise achieve. Misleading oneself about one's own physical strength can be dangerous or even fatal if it leads to a fight with someone who really is stronger. In other words, exaggeration of your positive attributes can lead to unfortunate consequences even if you believe it. These obvious points are perhaps implicit in VH&T's characterization of the "proper dosage" for self-deception and their brief acknowledgment of the costs of losing "information integrity," but are surprisingly underemphasized and underdeveloped in an otherwise carefully nuanced argument.

But these are matters of emphasis. The present point is simply that in the enthusiasm to develop a fascinating and creative argument for the otherwise paradoxical advantages of self-enhancing self-deception, we should not neglect how self-presentational biases can run in two directions and that in some important ways the oracle at Delphi was correct. To know yourself, accurately, can definitely be useful.

Understanding self-deception demands a co-evolutionary framework

doi:10.1017/S0140525X10002578

Steven W. Gangestad

Department of Psychology, University of New Mexico, Albuquerque, NM 87131.
sgangest@unm.edu
http://www.unm.edu/~psych/faculty/sm_gangestad.html

Abstract: The foundational theme of the target article is that processes of self-deception occur in a functional context: a social one through which self-deceptive processes enhance fitness by affecting an actor's performances. One essential component of this context not addressed explicitly is that audiences should have been selected to resist, where possible, enhancements falsely bolstered by self-deception. Theoretical implications follow.

Self-deception is an intrapsychic phenomenon, and explanations of it have typically been in terms of what intrapsychic outcomes it achieves (e.g., through self-deception, one can be happier with oneself or about prospects to come). The foundational theme of von Hippel & Trivers's (VH&T's) target article is a simple one, albeit one with profound implications: Any sensible approach to understanding an intrapsychic process – one consistent with modern evolutionary biology –.must ultimately connect that process with effects on the world. That is, intrapsychic processes, to evolve, must have either have had effects on fitness achieved through those effects or have been by-products of some process that did. The primary line of thinking that VH&T specifically pursue is one introduced by Trivers (1985) 25 years ago: Through self-deception, agents are able to control the inferences that other organisms make about them in ways that enhance the effectiveness of those agents to achieve desired ends.

As VH&T emphasize, the insistence that self-deception be understood in an *interpersonal* context has broad and deep consequences for how it should be conceptualized and studied. In what ways does self-deception enable other-deception, and through what processes does it do so? When is self-deception most effective at enabling other-deception? Do individuals

accordingly engage in self-deception when its beneficial interpersonal consequences are most substantial? If so, through what means do individuals "know" that self-deception is likely to be effective? More generally, how has selection shaped the processes that regulate self-deceptive tactics in ways that, ancestrally, enhanced net fitness benefits? Though self-deceptive processes have been studied by psychologists for decades, we will lack a deep understanding of them until researchers take to heart VH&T's fundamental point.

I take up here one component of self-deception's functional context that VH&T are surely aware of but do not explicitly discuss in any detail. In the interpersonal context in which self-deception operates, other-deception via self-deception may be in the interest of the actor, but *not* being deceived by others is typically in the interest of all target perceivers. That is, selection should favor perceivers whose inferences are not readily manipulated in the interests of others. Hence, self-deception not only must be understood in an evolutionary framework, but it also must be appreciated in the context of a *co-evolutionary* framework in which the targets of other-deceit must be assumed to have been subject to selection to avoid being deceived (e.g., Rice & Holland 1997).

In the 30-plus years since the concept of self-deception was introduced, biologists have developed sophisticated theories pertaining to communication between organisms, now collectively referred to as signaling theory (e.g., Searcy & Nowicki 2005). One core component of signaling theory is the concept of honest signaling. The idea is that, for any signaling system to evolve, both senders and receivers must benefit, and for receivers to benefit, the signal must contain accurate or "honest" information. If the signal is dishonest (e.g., the size of peacocks' tails reveal nothing about the quality of their bearers), receivers should evolve to ignore it – that is, the signaling system should collapse (or fail to evolve in the first place).

In light of this notion, how have self-deceptive processes aimed to deceive others been maintained by selection? If perceivers suffer from attending to performances rendered deceptive by self-deception, why has selection not led them to ignore such performances? Perhaps most notably, why should perceivers be fooled by false confidence bolstered by self-deception? Several possibilities come to mind.

First, most performances may well be honest portrayals of earned or honest confidence. A signaling system that is basically honest (honest on average) can tolerate some level of dishonesty (e.g., Searcy & Nowicki 2005).

Second, in many circumstances it may be difficult for individuals to detect the difference between a performance backed by earned confidence and one enabled by self-deceived confidence. An implication of the co-evolutionary nature of signaling systems is that perceivers should be attentive to cues of false confidence – for instance, through utilization of multiple cues and reliance on reputation based on past performance as well as current performance. They should furthermore not tolerate false confidence even when the actor is unaware of its nature. (One reason why narcissistic individuals have unstable interpersonal relations is because their unearned arrogance leads others to reject them.) Yet one may be able to successfully perform with false confidence under conditions in which other information is lacking (such as one-shot interactions). This possibility once again underscores the need for researchers to examine the contexts in which self-deception affects performance, including the audiences of those performances in light of their desire not to be fooled.

Third, confidence bolstered by self-deception may not, ultimately, be other-deceptive. Individuals who have earned confidence may nonetheless benefit by carrying off that confidence by self-deceiving (e.g., not attending to their own shortcomings). In this view, individuals who can best afford to self-enhance through self-deception are those who are viewed positively by others in any case. Ironically, in this view, self-deception facilitates honest, not deceptive, social performance.

These possibilities are not mutually exclusive. And there may be others.

More generally, the fact that social performances enhanced by self-deception must be understood in the context of co-evolved audience resistance to falsely enhanced performance has implications for how self-deception should be studied. And beyond that, at a basic theoretical level, it suggests ways in which an appreciation for the evolutionary processes that have shaped self-deception should be deepened.

Culture of deception

doi:10.1017/S0140525X10003122

Gregory Gorelik[a] and Todd K. Shackelford[b]

[a]*Department of Psychology, Florida Atlantic University, Boca Raton, FL 33431;* [b]*Department of Psychology, Oakland University, Rochester, MI 48309.*

ggorelik@fau.edu

shackelf@oakland.edu

http://www.toddkshackelford.com/

Abstract: We examine the self-deceptive aspects of religion and nationalism. By embracing various religious or political ideals, regardless of their truth, our ancestors could have enhanced their confidence, solidified their social ties, and manipulated their reproductive rivals. This use of culture as one's extended phenotype may increase the spread of misinformation and create global webs of deception and self-deception.

If humans have evolved a capacity to deceive themselves so as to better deceive others, then human technologies, languages, ideas, and traditions might display cultural manifestations of deceptive and self-deceptive adaptations. Deceiving oneself may be easier if others are complicit in the deception. Collective self-deception is manifested as groupthink and deindividuation, and it is likely mediated and enabled by various cultural elements. Von Hippel & Trivers (VH&T) briefly discuss the social reinforcement of individual-level self-deception, but they do not elaborate upon the full implications of the cultural aspects of self-deception. We discuss the ways in which self-deception may be expressed collectively in religious and political contexts, and we present several possibilities for how gene-culture co-evolution has affected human deception and self-deception.

According to Dawkins's (1982) concept of the extended phenotype, genes are selected for how well they code for an organism's ability to manipulate its environment. An organism's environment includes other organisms, of both the same and different species. Therefore, organisms may be selected for how well they can manipulate other organisms, effectively using them as extended phenotypes of their own selfish genes. If humans have competed with one another over reproductively relevant resources throughout their evolutionary history, then selection pressures may have sculpted adaptations by which humans manipulate and deceive their reproductive rivals. In addition, given the human capacity for non-genetic transfer of information (i.e., culture), many cultural phenomena may display design features indicative of their use in deceiving oneself and others. Therefore, human genes may be selected for how well they code for psychological programs that use cultural information to deceive other humans. In effect, culture is part of our extended phenotype and is an integral part of the environment to which our genes have evolved.

Following this line of thought, we can investigate human culture for features that enable its use during deception of oneself and others. Organized religion and nationalism display several exemplar features. In most ancestral contexts, religious or political self-deception may have benefited individual

members, but there was a risk of exploitation if some individuals accepted the benefits of membership without paying the costs of helping other members. In such instances, the institution in question could have been used as a tool by which some individuals manipulated others. If manipulators benefited by their manipulation, then manipulative traits may have proliferated throughout human populations (until the costs of manipulation outweighed the benefits). At the same time, the cultural tools that manipulators used to express their manipulative traits might have been refined and passed down the generations alongside the genetically coded, manipulative psychological programs. In this way, genes and culture depend on each other for the evolution and expression of deceptive and self-deceptive adaptations.

Various design features of religious and political institutions may be indicative of their role in deception and self-deception. As described by VH&T (sect. 5.5.2, para. 1), insecure societies display higher rates of religious belief, because belief in God may provide individuals with a sense of control over their lives. Assuming that this sense of control was advantageous for our ancestors because it enabled the manipulation of reproductive rivals, it should then be no surprise that humans are willing and able to accept as true certain fantastic doctrines and dogmas. Likewise, religion and nationalism exhibit a strength-in-numbers effect that facilitates collective self-deception. The costs of religious or political misinformation may not offset the benefits of joining and supporting such institutions. Therefore, the deception of individual members is made easier by the pervasiveness of self-deception within these institutions.

There are other features of organized religion and nationalism that portray self-deceptive qualities. The avoidance of information that threatens or could weaken a religious or political institution is ubiquitous. This is seen when totalitarian regimes limit the types of media that are available to the public, or when religious followers avoid being exposed to competing doctrines or scientific facts (i.e., evolution by natural selection). If exposed to threatening information, followers may attempt to rationalize away whatever threat they were exposed to or be skeptical of this information. In this way, patriots from one nation may doubt the veracity of a rival nation's messages and ideas by calling them propaganda. Likewise, creationists sometimes tie themselves into psychological knots in attempting to explain away the evidence for evolution (when they do not deny or ignore this evidence altogether).

Derogation of others and enhancement of oneself are also common features of nationalism and religion. Some examples of this include the American motto "one nation, under God," or the belief that one is a member of the "chosen people" or of the "master race," while dehumanizing members of other nations or religions. Furthermore, optimism about the future is pervasive within religious and political circles. This optimism can lead to a self-fulfilling prophecy if one is motivated to action by the promise of a political utopia or a heavenly paradise, but it also can be used to manipulate members into acting against their own interests. Likewise, such cultural modes of self-enhancement may increase one's confidence and lead to social solidarity with one's community, but they also may bring about social conflict and war.

According to VH&T, convincing oneself that a lie is true while knowing that it is false at some psychological level is the most extreme form of self-deception. Religion, in particular, may use the consequent cognitive dissonance to its advantage by pointing to this internal conflict as evidence of its veracity. The constant struggles to retain one's faith or to remain spiritual amid the onslaught of secularism seem to be essential features of modern Judaeo-Christian practices. In this way, religion may be an especially useful cultural tool by which individuals manipulate their rivals by imposing self-deception upon them.

Deceiving ourselves about self-deception

doi:10.1017/S0140525X1000227X

Stevan Harnad

Institut des sciences cognitives, Université du Québec à Montréal, Montreal, QC H3C 3P8, canada; School of Electronics and Computer Science, University of Southampton, SO17 1BJ Southampton, United Kingdom.

harnad@ecs.soton.ac.uk

http://www.ecs.soton.ac.uk/people/harnad

Abstract: Were we just the Darwinian adaptive survival/reproduction machines von Hippel & Trivers invoke to explain us, the self-deception problem would not only be simpler, but also nonexistent. Why would *unconscious* robots bother to misinform *themselves* so as to misinform others more effectively? But as we are indeed conscious rather than unconscious robots, the problem is explaining the causal role of consciousness itself, not just its supererogatory tendency to misinform itself so as to misinform (or perform) better.

Von Hippel & Trivers (VH&T) are deceiving themselves – with the help of adaptivist psychodynamics and a Darwinian Unconscious. They have not proposed an adaptive function for self-deception; they have merely clad adaptive interpersonal behaviour in a non-explanatory mentalistic interpretation: *I can persuade you more convincingly that I am unafraid of you (or better fool you into thinking that the treasure is on the right rather than the left, where it really is) if I am unaware of – or "forget" – my own fear (or the fact that the treasure is really on the left rather than the right).*

Sure. But then in what sense am I afraid at all (or aware where the treasure really is)? If I *feel* (hence act) afraid, then you detect it. If I don't feel the fear (or the sinistroversive urge), then I don't act afraid, and you don't detect any fear (because there is nothing there to detect).

So in what sense am I "self-deceived"? (Ditto for left/right.) Is it always self-deception not to feel afraid (or not to remember that the treasure's on the right), when I "ought to" (or used to)?

The same is true of "self-enhancement": Yes, I am more convincing to others, hence more influential on their behaviour, if I behave as if I expect to succeed (even when I have no objective grounds for the expectation). But in what sense am I self-deceived? In feeling brave and confident, when I "ought to" be feeling fearful and pessimistic? Shouldn't organisms all simply be behaving in such a way as to maximize their adaptive chances?

In fact, what does what organisms *feel* have to do with any of this at all (apart from the entirely unexplained fact that they do indeed feel, that their feelings are indeed correlated with their adaptive behaviour, and that their feelings do indeed feel causal to them)? The feelings themselves (i.e., consciousness) are much harder to situate in the adaptive causal explanation – unless you believe in telekinesis (Harnad 2000)! (Hence, I feel that VH&T have bitten off a lot more here, phenomenally, than they can ever hope to chew, functionally.)

The treasure is the best example of all, because that is about *facts* (data) rather than just feelings: Suppose I did indeed "know" at some point that the treasure was on the left – in the sense that if at that point I could have reached for it without risk of being attacked by you, I would have reached for it on the left. But, according to VH&T, it was adaptive for me to "forget" where the treasure really was, and to believe (and behave as if) it was on the right rather than the left, so as to deceive you into heading off to the right so I could eventually grab the treasure on the left and dart off with it.

But isn't the true adaptive design problem for the Blind Watchmaker – apart from the untouched problem of how and why we feel at all (Harnad 1995) – a lot simpler here than we are making it out to be (Harnad 2002)? And are we not deceiving ourselves when we "adapt" the adaptive explanation so as to square with our subjective experience?

All that's needed for adaptive cognition and behaviour is *information* (i.e., data). To be able to retrieve the treasure, what I (or rather my brain) must have is reliable data on where the treasure really is – on the left or the right. Likewise, in order to get you to head off toward the right, leaving the treasure to me, I need to be able to behave exactly as if I had information to the effect that it was on the right rather than the left (or as if I had no information at all). Adaptive "mind-reading" (*sensu* Premack & Woodruff 1978), after all, is just behavioural-intention-reading and information-possession-reading. It is not really telepathy.

Nor does it need to be. Insofar as the putative adaptive value of self-deception in interpersonal interactions is concerned, an adaptive behaviourist (who has foolishly – and falsely – deceived himself into denying the existence of consciousness) could easily explain every single one of VH&T's examples in terms of the adaptive value of mere deception – behavioural deception – of other organisms.

And when it comes to true "self-deception": do I really have to *forget* where the treasure actually is to successfully convince either you or me of something that is adaptive for me? Well, there the only reason VH&T would seem to have a leg up on the blinkered adaptive behaviourist is that VH&T do *not* deceive themselves into denying consciousness (Harnad 2003). But what VH&T completely fail to do is to explain (1) what causal role consciousness (feeling) itself performs in our adaptive success, let alone (2) what second-order causal role consciousness might need to perform in the kind of peek-a-boo game individual organisms sometimes seem to play with themselves. Both remain just as unexplained as they were before, and the first (1) is the harder problem, hence the one that needs to be solved first (Harnad & Scherzer 2008).

Might self-deception rather be a form of *anosognosia*, where our brains are busy making do with whatever informational and behavioural resources they have at their disposal, with no spare time to deceive us (inexplicably) into feeling that we're doing what we're doing because we *feel* like it?

Apart from that, it's simple to explain, adaptively, why people lie, cheat, and steal (or try to overachieve, against the odds, or avoid untoward data): It's because it works, when it works. It is much harder to explain why we *don't* deceive, when we don't, than why we do, when we do. We usually distinguish between the sociopaths, who deceive without feeling (or showing) any qualms, and the rest of us. Have sociopaths deceived themselves about what's right and wrong, confusing true and false with being whatever it takes to get what one wants, whereas the rest of us are keeping the faith? Or are they just better method actors than the rest of us?

Evolutionary explanations need to account for cultural variation

doi:10.1017/S0140525X10002669

Steven J. Heine

Department of Psychology, University of British Columbia, Vancouver, BC V6T 1Z4, Canada.

heine@psych.ubc.ca

http://www2.psych.ubc.ca/~heine/index.html

Abstract: Cultural variability in self-enhancement is far more pronounced than the authors suggest; the sum of the evidence does not show that East Asians self-enhance in different domains from Westerners. Incorporating this cultural variation suggests a different way of understanding the adaptiveness of self-enhancement: It is adaptive in contexts where positive self-feelings and confidence are valued over relationship harmony, but is maladaptive in contexts where relationship harmony is prioritized.

I applaud von Hippel & Trivers (VH&T) for seeking an evolutionary account for a phenomenon as pervasive and intriguing as self-deception. They have made a compelling case for the adaptiveness of self-deception in certain contexts; however, it would be more persuasive if they had taken seriously the problem of cultural variability in self-enhancement.

In offering a compelling evolutionary account for any phenomenon, it is critical to consider evidence from a broad enough array of contexts to allow for confident generalizations. Nearly all of the empirical citations from this article derive from what we call WEIRD (Western, educated, industrialized, rich, democratic) samples (Henrich et al. 2010). This would not be so problematic if the data from such samples yielded a similar pattern to that from other samples, but they do not; this is particularly the case for self-enhancement (Heine et al. 1999; Mezulis et al. 2004).

VH&T claim that self-enhancement emerges "in every culture on earth" (sect. 8, para. 6). However, these claims stand in stark conflict with the cross-cultural evidence. A meta-analysis of cross-cultural studies of self-enhancement (Heine & Hamamura 2007), yielded a pronounced effect for Westerners ($d = 0.87$), and a non-existent effect for East Asians ($d = -0.01$). Cultural differences emerged for 30 of the 31 different methods, with the one exception being the self-esteem IAT (Implicit Association Test) measure (Greenwald & Farnham 2000). It remains unclear what the self-esteem IAT assesses, given that it has the least validity evidence of any of the IAT measures (Hofmann et al. 2005), and it does not correlate reliably with other implicit or explicit measures of self-esteem or external validity criteria (Bosson et al. 2000; Buhrmester et al., in press; Falk et al. 2009). Further, studies of self-enhancement that employ hidden behavioral measures find equally pronounced cultural differences as those with explicit measures (Heine et al. 2000; 2001), indicating that these differences extend to people's true beliefs. The only studies that reliably yield self-enhancement among East Asians employ the better-than-average-effect (BTAE) method (average d values are 1.31 and 0.38 for Westerners and East Asians, respectively; Heine & Hamamura 2007). However, as VH&T acknowledge in their citation of Chambers and Windschitl (2004), the BTAE incorporates a few cognitive biases, which results in exaggerated estimates of self-enhancement (Klar & Giladi 1997; Krizan & Suls 2008; Kruger 1999). The effects are inflated for both cultures by a magnitude of approximately $d = 0.60$ (Heine & Hamamura 2007).

VH&T note that "even East Asians, who value humility and harmony over individualistic self-aggrandizement, show self-enhancement in their claims of the superiority of their collectivist qualities," (sect. 3, para. 3), and they cite two articles by Sedikides et al. (2003; 2005). In those articles, Sedikides et al. argue that self-enhancing motivations are universal but expressed differently: Westerners enhance themselves in domains that are important to them (e.g., individualism), while East Asians enhance themselves in domains that are important to them (e.g., collectivism). The evidence for this from those articles derives from the BTAE method. The other 11 methods that have addressed this identical question (i.e., the false-uniqueness bias, actual-ideal self-discrepancies, manipulations of success and failure, situation sampling, self-peer biases, relative-likelihood and absolute-likelihood optimism biases, open-ended self-descriptions, automatic self-evaluations, social relations model, and a corrected BTAE) all yield an opposite pattern of results: East Asians do *not* self-enhance more in domains that are especially important to them (Falk et al. 2009; Hamamura et al. 2007; Heine 2005b; Ross et al. 2005; Su & Oishi 2010). A meta-analysis including all of the published studies on this topic finds no support for the claim that East Asians self-enhance more in important domains ($r = -0.01$), although Westerners do ($r = 0.18$; Heine et al. 2007a; 2007b). The meta-analyses by Sedikides et al. (2005; 2007) find different results because they excluded most of the studies that yielded contrary findings. Further, the evidence that East Asians enhance in

collectivistic/important domains using the BTAE appears to be specifically the product of methodological artifacts of this measure (Hamamura et al. 2007). In sum, contrary to VH&T's claims, the evidence does not support the universality for self-enhancement or that East Asians self-enhance in particular domains. If instead of considering data almost exclusively from WEIRD samples, they had instead only considered East Asian data, VH&T would not have proposed their evolutionary account for self-enhancement; there would not have been any self-enhancement effect in need of an explanation.

Given this cross-cultural variability, how might we consider how self-enhancement evolved? Like VH&T, I think it is important to consider the costs and benefits of self-enhancement. Benefits of self-enhancement include positive self-feelings and confidence (Taylor & Armor 1996; Taylor & Brown 1988). It feels good to self-enhance, and it leads people to expect that they will do well on future tasks, and these relations appear to hold across cultures (Heine 2005a). On the other hand, a cost of self-enhancement is that it can strain interpersonal relations; self-enhancers risk attracting the scorn of others (Colvin et al. 1995; Exline & Lobel 1999; Paulhus 1998; Vohs & Heatherton 2001). People are often alienated by self-enhancers, especially over long-term relationships (Robins & Beer 2001). Likewise, positive self-presentations can lead to less liking by others (Godfrey et al. 1986; Tice et al. 1995).

This analysis suggests that in cultural contexts where people value positive feelings and self-confidence, yet are not overly concerned about maintaining harmonious relationships, self-enhancement is more adaptive. In contrast, in cultures where positive feelings and self-confidence are valued less, but the maintenance of smooth interpersonal relationships is prioritized, self-criticism and a concern for face is more adaptive. Compared with Westerners, East Asians are not as concerned about positive self-feelings (Suh et al. 1998), and they often perform better when they *lack* confidence (Heine et al. 2001; Oishi & Diener 2003; Peters & Williams 2006). Further, much research reveals a greater concern among East Asians about relationship harmony (Markus & Kitayama 1991).

In sum, a compelling account of the adaptiveness of self-enhancement needs to account for why cultures differ in their tendencies to self-enhance. The existence of cultural variation in self-enhancement can help to elucidate the contexts in which self-enhancement should be most adaptive.

The selfish goal: Self-deception occurs naturally from autonomous goal operation

doi:10.1017/S0140525X1000258X

Julie Y. Huang and John A. Bargh

Department of Psychology, Yale University, New Haven, CT 06520.

julie.huang@yale.edu, john.bargh@yale.edu

www.yale.edu/acmelab

Abstract: Self-deception may be a natural consequence of active goal operation instead of an adaptation for negotiating the social world. We argue that because autonomous goal programs likely drove human judgment and behavior prior to evolution of a central executive or "self," these goal programs can operate independently to attain their desired end states and thereby produce outcomes that "deceive" the individual.

We agree with von Hippel & Trivers (VH&T) that motivation plays a key role in self-deception. There are reasons to believe, however, that self-deception is part and parcel of normal goal operation, instead of constituting a specific adaptation for negotiating the social world. To support this argument, we discuss the selfish goal model of behavior (Bargh & Huang 2009), which holds that human judgment and behavior were driven (unconsciously) by goal pursuit programs prior to the evolution of a central executive self. Next, we review research suggesting that all goals – even conscious ones – maintain this ability to operate autonomously and thus are capable of producing effects that appear, on the surface, to "deceive" the individual self.

Evolutionary theorists have argued that consciousness and strategic, intentional mental processes were relatively late arrivals in human evolutionary history (e.g., Corballis 2007; Donald 1991). If so, then another, unconscious, system must have directed hominid behavior in adaptive ways prior to the evolution of consciousness.

Indeed, evolutionary biologists and psychologists view motivations as the crucial link between genetic influences and adaptive behavior (Dawkins 1976; Tooby & Cosmides 1992). Because of constantly changing and shifting environmental conditions, coupled with the very slow rate of genetic change, direct genetic controls over behavior tend to be inflexible and unable to adapt quickly enough to changes in the environment. Genes program the individual with generally adaptive motivations, which are translated as "goal programs" within the nervous system (Mayr 1976).

These goals had to guide the individual toward evolutionarily adaptive outcomes in the local environment without the guidance of an executive self – in other words, they must have been capable of autonomous operation. When conscious goal-pursuit processes then came on-line, they likely took advantage (made use) of the already-existing autonomous goal structures (Bargh & Morsella 2008). As a consequence, conscious goals retain some features of nonconscious goal pursuits, including the capability to operate independently from the "self" or central executive. Thus, we argue, any goal pursuit, conscious or unconscious, operates autonomously to an extent and can thus produce effects that can be considered "deceptive" from the perspective of the self.

How do goals operate independently from individual guidance while still prodding that individual to achieve the goal's end state? Research suggests that both conscious and nonconscious goals, once active, exert temporary downstream effects upon the individual's information processing and behaviors in ways that facilitate successful pursuit of that goal. An active goal's systematic influence can be considered "selfish" because it is geared toward attaining its desired end state, regardless of whether the consequences of goal pursuit (e.g., temporary valuation of stimuli) are consistent with the values of the stable self-concept. For instance, participants perceive goal-facilitating stimuli as bigger (Veltkamp et al. 2008), closer (Balcetis & Dunning 2010), and more likable (Ferguson 2008) when that goal is active than when it is not. These influences can be considered deceptive because as perceptions of reality, they are shifted in inconsistent, sometimes inaccurate ways.

Because they also operate autonomously (once intentionally activated), even consciously pursued goals can produce effects which deceive the individual. Bargh, Green, and Fitzsimons (2008) tested the hypothesis that all goal pursuits, conscious and unconscious alike, operate autonomously and so can produce consequences unwanted at the level of the self. In their experiments, participants watched a videotape of two people in an office with the explicit, conscious goal of evaluating one of the people for a job. Some participants were told the job in question was a restaurant waiter; others were told it was a newspaper crime reporter position. During the interview, the two conversation partners were interrupted by a person who behaved in either a rude and aggressive or a polite and deferential manner.

Note that the desired personality characteristics of a waiter and a crime reporter are mirror opposites: The ideal crime reporter is tough and aggressive, whereas the ideal waiter is polite and deferential. After viewing the video, participants were asked how much they liked not the job candidate (on whom they had been consciously focused), but this interrupter. Unsurprisingly,

participants in the control and the waiter-job conditions liked the polite interrupter more than the rude interrupter. However, participants in the reporter-goal condition (for whom rudeness and aggressiveness are desired traits) liked the rude interrupter more than the polite interrupter. Because the interrupter's traits matched the desired qualities of the currently active goal, he was evaluated positively by people prepared to evaluate a crime reporter – though as the control condition indicates, he would not have been liked at all in the absence of this active goal. Thus, even consciously pursued goals can lead to self-deception by producing effects contrary to the individual's (i.e., self's) preferences.

Moreover, both conscious and unconscious goal pursuits can turn off, independently from self-direction or awareness, thus producing potentially self-deceptive effects. When a goal is completed it temporarily deactivates, inhibiting the mental representations involved in the pursuit of that goal (Förster et al. 2005). The goal's downstream influence on the person's cognition and behavior evaporates, now allowing the production of behavior that is inconsistent with previous actions. For example, participants given the opportunity to disagree with blatantly sexist remarks were ironically more likely afterward to recommend a man for a stereotypically male job than if they had not had the counterarguing opportunity (Monin & Miller 2001). Similarly, when supporters of then-candidate Barack Obama were given a chance to express that support, afterwards they were counterintuitively more likely to rate a job opening as more suitable for Whites than for Blacks (Effron et al. 2009). In these studies, successfully asserting egalitarian values temporarily completed participants' self-valued goals to appear egalitarian; this goal completion caused the production of behaviors counter to the participants' self-professed values – in other words, it produced self-deception at the behavioral level.

Instead of evolving to facilitate the deception of others, self-deception may be a natural consequence of the autonomous goal operation that characterized our pre-conscious past. Goal structures predated the evolution of a central self and so must have operated autonomously to guide the individual toward their specific desired end states. Research suggests they continue to do so today.

It takes a thief to catch a thief

doi:10.1017/S0140525X10002098

Nicholas Humphrey

London School of Economics (Emeritus Professor).

humphrey@me.com

www.humphrey.org.uk

Abstract: Von Hippel & Trivers (VH&T) dismiss in a couple of pages the possible *costs* of self-deception. But there is a downside to self-deception that they do not consider. This is the loss of psychological insight into deceit by others that blindness to deceit by oneself is likely to entail.

In discussing the various strategies by which people may be able to pull the wool over their own eyes, Von Hippel & Trivers (VH&T) acknowledge implicitly that the default mode is *self-knowing* rather than self-deception. That is to say, other things being equal, people have a remarkable degree of insight into their own behavior and have to take active steps to avoid this if and when it does not suit them. There are good reasons why it should be so. Self-awareness brings substantial benefits. Not least, as I have argued, self-awareness underlies people's ability to develop a theory of human mind. Just to the extent that people know from the inside how they themselves truly felt and thought in a given situation, they can imagine what the same situation will be like for someone else – and so will be able to simulate how others are likely to behave. It is precisely the evolved capacity for conscious insight – *veridical* insight – that has allowed human beings to become what I have called "natural psychologists," with an unparalleled ability to predict and manipulate the behavior of other members of their species (Humphrey 1978).

However, if this is so, it clearly makes difficulties for VH&T's – otherwise convincing – theory that there are situations where it is better for an individual to be blind to the psychological reality; or, at any rate, it means VH&T are seeing only half of the picture. Given that in general self-knowing brings with it a greater understanding of others, the corollary has to be that self-deception brings about lesser understanding. In particular, if and when someone fails to recognize that he himself has behaved in a mendacious way, he is less likely to recognize the mendacity of others. Thus, although it may well be the case that people who deceive themselves when they cheat are less likely to be caught, such people are also more likely to be duped by others. As I remarked in my book *Consciousness Regained*, "It takes a thief to catch a thief and an intimate of his own consciousness to catch the intimations of consciousness in others" (Humphrey 1983, p. 63).

Evidence that this is in fact how things play out has been provided by Surbey (2004). In a study with Rankin, Surbey gave subjects a self-deception questionnaire and also a personality test to assess Machiavellianism – the ability to get the better of others through psychological manipulation. It turned out that people with strong Machiavellian tendencies got low scores on the self-deception test. "They're more consciously aware of selfish motivations than others," Surbey is quoted as saying, "and they're projecting their selfish motivations on others." (Motluk 2001).

VH&T's theory would predict that self-deception – having little downside – should be a universal human trait. However, Surbey and Rankin found big individual differences, with some people being highly self-deceptive and others hardly at all. But this is just what we should expect if self-deception does have this downside, that is, if it makes people better thieves but poorer detectives. For this means there will have been balancing selection in the course of human evolution. Assuming that self-deceivers will have won out when few people suspected deceit, but suspicion will have won out when most people were self-deceivers, selection will have resulted in a mix of strategies in the human population – a classic balance between doves and hawks.

Choice blindness and the non-unitary nature of the human mind

doi:10.1017/S0140525X10002591

Petter Johansson,[a] Lars Hall,[b] and Peter Gärdenfors[b]

[a]*Division of Psychology and Language Sciences, University College London, United Kingdom;* [b]*Lund University Cognitive Science, Lund University, Sweden.*

petter.johansson@lucs.lu.se **lars.hall@lucs.lu.se**
peter.gardenfors@lucs.lu.se
http://www.lucs.lu.se/petter.johansson/
http://www.lucs.lu.se/lars.hall/
http://www.lucs.lu.se/peter.gardenfors

Abstract: Experiments on choice blindness support von Hippel & Trivers's (VH&T's) conception of the mind as fundamentally divided, but they also highlight a problem for VH&T's idea of non-conscious self-deception: If I try to trick you into believing that I have a certain preference, and the best way is to also trick myself, I might actually end up having that preference, *at all levels of processing.*

The classic paradox of self-deception is how the self can be both deceiver and deceived. Von Hippel & Trivers (VH&T) solve this

conundrum by appealing to the separation of implicit and explicit processes in the mind; I cannot knowingly deceive myself, but the non-conscious part of my mind can "deceive" me by pursuing goals that are contradictory to my consciously stated ambitions. VH&T identify and draw support from three different areas of research: explicit versus implicit memory, explicit versus implicit attitudes, and controlled versus automatic processes. None of these processes are inherently self-deceptive. Instead, as VH&T write: "These mental dualisms do not themselves involve self-deception, but each of them plays an important role in enabling self-deception" (sect. 4, para. 1).

We suggest adding a fourth set of related studies: work on choice blindness – that is, the failure to detect mismatches between a choice made and the outcome received (Johansson et al., 2005). Choice blindness is an experimental paradigm inspired by techniques from the domain of close-up card magic, which permits a surreptitious manipulation of the relationship between choice and outcome that the participants experience. The participants in Johansson et al. (2005) were asked to choose which of two pair-wise presented faces they found most attractive. Immediately after, they were also asked to describe the reasons for their choice. Unknown to the participants, on certain trials, a double-card ploy was used to covertly exchange one face for the other. Thus, on these trials, the outcome of the choice became the opposite of what they intended. Remarkably, in the great majority of trials, the participants were blind to the mismatch between choice and outcome, while nevertheless being able to offer elaborate reasons for their choices. The two classes of reports were analysed on a number of different dimensions, such as the level of effort, emotionality, specificity, and certainty expressed, but no substantial differences between manipulated and non-manipulated reports were found (Johansson et al. 2006). The lack of differentiation between reasons given for an actual and a manipulated choice shows that there is probably an element of confabulation in "truthful" reporting as well. In addition to faces and abstract patterns (Hall & Johansson 2008), choice blindness has been demonstrated for taste and smell (Hall et al. 2010 in press), as well as for moral and political opinion (Hall et al., in preparation).

Experiments on choice blindness support VH&T by providing a dramatic example of the non-unitary nature of the mind; we may have far less access to the reasons for our actions than we think we do. But experiments on choice blindness also highlight a possible problem lurking in VH&T's conception of self-deception. Is it really possible to maintain two separate sets of conscious and non-conscious goals as a technique to deceive oneself in order to better deceive someone else? For example, in one version of the experiment described earlier, the participants had to choose between the same pairs of faces a second time, as well as separately rate all the faces at the end of the experiment. This procedure revealed that the manipulation induced a pronounced, but to the participants unknown, preference change, because they came to prefer the originally non-preferred face in subsequent choices, as well as rate the face they were led to believe they liked higher than the one they thought they rejected (Hall et al., in preparation). This result is of course in line with a long tradition of studies showing the constructive nature of preferences, i.e. that we come to like what we think we like (see Ariely & Norton 2008; Bem 1967; Festinger 1957; Lichtenstein & Slovic 2006).

The crucial point is that if it is possible to get people to reverse their initial preferences by making them publicly endorse an outcome they believe they prefer, then using self-deception as a means to deceive others might result in fundamental changes to the self as well. If I try to trick you into believing that I prefer *a* over *b*, and the best way to do that is to also trick myself into believing that I prefer *a* over *b*, I might actually end up preferring *a* over *b*, *at all levels of processing*. In such a case, it would be the conscious parts of the self that makes the unconscious parts change, and in a process more akin to self-persuasion than self-deception. The apparent ease with which the participants in choice blindness experiments confabulate reasons in favor of a previously rejected alternative indicates that this form of self-persuasion is something that comes quite naturally to us.

A single self-deceived or several subselves divided?

doi:10.1017/S0140525X10002517

Douglas T. Kenrick and Andrew E. White
Department of Psychology, Arizona State University, Tempe, AZ 85282.
douglas.kenrick@asu.edu aewhite7@asu.edu
http://douglaskenrick.faculty.asu.edu/?q=node/10

Abstract: Would we lie to ourselves? We don't need to. Rather than a single self equipped with a few bivariate processes, the mind is composed of a dissociated aggregation of subselves processing qualitatively different information relevant to different adaptive problems. Each subself selectively processes the information coming in to the brain as well as information previously stored in the brain.

Von Hippel and Trivers (VH&T) drive home a point psychologists often miss – a functional explanation cannot begin and end inside a person's head – people do not strive to "feel good" for its own sake, they feel good when they act in ways that, on average, increased their ancestors' chances of survival and reproduction. VH&T's target article underscores the theoretical functions of interdisciplinary work – broadening the significance of a generation of experimental studies (previously interpreted as a random array of apparently senseless information processing biases) while simultaneously grounding the fuzzy philosophical problem of self-deception in solid empirical findings. But their view raises two questions for us: First, is there really a "self" to be deceived? Second, are we really talking about "deception" or simply division of labor between mental modules?

VH&T do not go far enough in applying recent views of modularity. They focus on bivariate cognitive processes such as implicit versus explicit memory. But from an adaptationist perspective, important cognitive subdivisions cut along lines of content rather than process – different adaptive problems require *qualitatively* different sets of decision mechanisms. How a person's brain crunches information depends critically on whether he or she is thinking about attracting a mate, avoiding a fistfight, seeking status, making a friend, or caring for a child. Understanding those differences requires us to think about content and to think about divisions larger than two.

Thinking about the mind as composed of several motivational subselves, each dealing with different classes of problem content, has already begun to build bridges between research on social cognition and ideas in evolutionary biology, as well as generating a host of novel empirical findings (e.g., Kenrick et al. 2010). For example, people in whom a self-protective motive is activated are more likely to remember angry faces, especially on male members of out-groups (who are otherwise homogenized in memory; Ackerman et al. 2006) and to encode a neutral facial expression as anger (but only when it is expressed by an out-group male; Maner et al. 2005). Consistent with Trivers's classic theories about parental investment and sexual selection, males (but not females) in a mating frame of mind are more likely to interpret an attractive females' facial expression as expressing sexual interest, and mating-conscious males are likely to think more creatively and independently and to conspicuously display in other ways (Griskevicius et al. 2006; Maner et al. 2005; Sundie et al., in press).

In a brilliant article in the inaugural edition of *Personality and Social Psychology Review*, titled "Subselves," Martindale (1980) described how mental dissociations could be understood in rigorous cognitive terms. Building on cognitive concepts such as lateral inhibition and state dependent memory, Martindale described how the brain accomplishes parallel processing without attentional overload. Only a small portion of the information available to the brain can be consciously processed at any given time, requiring mechanisms for suppressing most of what is going on up there. At the level of single neurons, there is lateral inhibition; at the level of the whole functioning brain, Martindale proposed that we have different subselves – executive systems with preferential access to different memories and different action programs. Because it is simply impossible to have conscious access to all our memories, attitudes, and ongoing experiences, an implication of Martindale's analysis is that we are all, in a sense, dissociative personalities.

My colleagues and I have linked Martindale's analysis with the idea of functional modularity to propose a set of fundamental motivational systems – each of which serves functional priorities by linking different affective and motor programs to adaptively relevant environmental events. These functional motivation systems can lead to biased information processing that spans many of the "varieties of self-deception" proposed by VH&T. For example, men and women both selectively search for attractive members of the opposite sex, but later, uncommitted men and committed women misremember a greater frequency of attractive women (Maner et al. 2003). When looking at pictures of disfigured and healthy others, people selectively attend to photographs of disfigured others, but later confuse them with one another and do not remember them very well – a disjunction between attention and memory that may be functional because disfigurement (unlike an angry facial expression) is an invariant threat cue – it will still be there next time you encounter the person (Ackerman et al. 2009). In addition to biases in attention and memory, people may be biased to interpret neutral expressions on goal-relevant social targets in functional ways. In one set of experiments, activating a mate search goal led men to selectively perceive sexual arousal in the neutral expressions of attractive members of the opposite sex, whereas activating a self-protection goal led men and women to selectively perceive anger in the neutral expressions of out-group males (Maner et al. 2005). Taken together, these findings reveal the importance of examining functional motivation systems and goal-relevant content, especially when considering biased information processing.

Thinking about cognitive-processing limitations and adaptive motivational systems demystifies the concept of self-deception. At the acquisition phase, we gather what is important to the currently active subself and discard what is not. At the encoding phase, we label what is important to the currently important subself and ignore what is not. When later dipping back into those memory bins, we ignore most of what is in there, and dig out what is functionally relevant to the currently active subself, to deal with the functional problem that is currently most salient. It's not deception, just selectivity.

The weightless hat: Is self-deception optimal?

doi:10.1017/S0140525X10002670

Elias L. Khalil

Department of Economics, Monash University, Clayton, Victoria 3800, Australia.

elias.khalil@monash.edu

www.eliaskhalil.com

Abstract: There are problems with the thesis of von Hippel & Trivers (VH&T): (1) It entails that self-deception arises from interpersonal deception – which may not be true; (2) it entails that self-deception is optimal – which is not necessarily so; and (3) it entails that interpersonal deception is optimum – which may not be true.

Von Hippel & Trivers (VH&T) argue that self-deception arises from interpersonal deception: Agents deceive themselves as a tactic in order to suppress the biological cues that may betray them when they deceive others. If so, self-deception is optimal. As VH&T's corollary 1 states, by deceiving themselves, agents are able to "avoid" the cognitive costs of "consciously mediated deception." As corollary 2 states, "by deceiving themselves, people can reduce retribution if their deception of others is discovered."

I find five problems with the thesis:

1. The thesis implies that interpersonal deception necessarily precedes self-deception, in the logical sense. So, Robinson Crusoe does not resort to self-deception. Or, if he does, it must be a leftover trait given that his social past is not too far back. However, from casual empiricism, many acts of self-deception do not involve or entail society (Khalil, submitted). In the insightful Aesop's fable of the fox and sour grapes, there is no society. Agents simply go to enormous effort of "face saving" in order to look good to themselves – i.e., where no other humans are involved. One who makes a commitment to avoid, for example, eating chocolate cake, but succumbs to the temptation of eating a chocolate cake while on a visit to a foreign country, may say to himself: "I was not thinking – I was too absorbed watching the scenery."

2. If self-deception is optimum, why would societies develop religious rules and encourage ethics of integrity, and why would even individuals appeal to their "conscience" and enact general rules in order to avoid and detect self-deception? Actually, the phenomenon of general rules – for example, one should not eat chocolate cake – has puzzled Adam Smith (1759/1982, pp. 156–61) who asked why individuals living independently of societies, such as Robinson Crusoe, would seek and want general rules? Smith (see Khalil 2009) argued that agents can easily fall prey to self-deception (what he calls "self-deceit"). Smith observed that people often undertake the wrong (i.e., suboptimal) actions, when they "honestly" think they are following moral rules (Khalil 2010). The impartial spectator (one who resides in one's "breast," in Smith's words, and decides what the correct action is) might become partial – without the full awareness of the agent. That is, one can eat chocolate cake without one noticing that one is doing something wrong. This could be the case because one can come under the influence of the moment, such as listening to a speech, reading a book, and so on. Such self-serving bias is long known in the literature (e.g., Rosenhan & Messick 1966; Babcock & Loewenstein 1997). Agents, in turn, according to Smith, resort to self-deception to cover up their mistakes, that is, their suboptimal actions. In this sense, the deception of others, which the agent is trying to cover, is already suboptimal – given that the retaliation of the other would offset any expected benefit from cheating the other.

Smith proceeded to argue that agents, to check their tendency to fall victim to a partial spectator and hence hurt themselves, erect general rules that express the opinion of the impartial spectator. Smith's "general rules" act as what economists call "precommitments" or "self-control" (see Khalil 2010). Examples of precommitments include checking oneself into a weight-loss spa to lose weight, avoiding bars if one wants to avoid drinking, and burning the bridges of retreat to prevent troops from running away in the face of an upcoming battle. But there is one minor difference: While precommitments are enacted to avoid temptations, Smith's general rules are enacted to avoid self-deception.

If this analysis is granted, self-deception is suboptimal because the action that the agent tries to cover up is, to start with, suboptimal.

3. For self-deception to be an optimum tactic, interpersonal deception must, to start with, be optimum. But why should

we suppose that interpersonal deception is optimum? For example, if interpersonal deception can cost the person some "conscience agony," and the agent decides that the benefit from cheating friends is not worth it, then cheating will be suboptimal. If so, self-deception is not only suboptimal, but also a "bad," that is, a negative good because it obstructs the agent from seeing the suboptimal decision.

4. Corollary 1 states that agents want to avoid the cognitive cost of deception and hence resort to self-deception. But self-deception can be very costly, where agents reinvent the past, keep telling the fiction to themselves, and suppress other versions, sometimes with violence. Even when self-deception is costlier than frank deception, agents might still resort to self-deception – and hence would be irrational. This would be contrary to VH&T's thesis. But if agents avoided self-deception in such a case, they would be rational. That is, they would only resort to self-deception when the cognitive cost of self-deception was lower than the cognitive cost of frank deception. If agents did such calculation, even unconsciously, the phenomenon of self-deception would not exist – and hence the raison d'être of the target article does not exist.

5. Corollary 2 states that agents resort to self-deception because they can "play dumb" and reduce the intensity of retribution. But from casual empiricism, if an embezzler or a politician apologizes, that is, "comes out clean," public scorn will be much less. In fact, the legal system of *all* countries and cultural norms in *all* societies give lenient punishment to people who express remorse and admit guilt.

If self-deception were a weightless hat – which everyone one can see while the person who wears it cannot – it could not be optimal. It would always be better to carry a mirror at all times, that is, a conscience or the "man within the breast" to use the expression of Adam Smith (1759/1982, p. 130), in order to undertake ruthless self-examination rather than face retribution by injured others.

Belief in God and in strong government as accidental cognitive by-products

doi:10.1017/S0140525X10002104

Peter Kramer and Paola Bressan

Dipartimento di Psicologia Generale, Università di Padova, 35131 Padova, Italy.

peter.kramer@unipd.it **paola.bressan@unipd.it**
www.psy.unipd.it/~kramer **www.psy.unipd.it/~pbressan**

Abstract: Von Hippel & Trivers (VH&T) interpret belief in God and belief in strong government as the outcome of an active process of self-deception on a worldwide scale. We propose, instead, that these beliefs might simply be a passive spin-off of efficient cognitive processes.

Von Hippel & Trivers (VH&T) define self-deception as a collection of biases that prioritize welcome over unwelcome information. They argue that self-deception is an active process (biased information searching, misinterpreting, misremembering, rationalizing, convincing oneself that a lie is true) that is related to individual motivations and goals and serves to facilitate deception of others. They further claim that differences between countries in both belief in God and belief in strong government suggest "self-deception on a worldwide scale." We propose that rather than reflecting motivations and goals, these beliefs may ensue automatically from efficient memory, attentional, and cognitive processes.

Subjects attempting to produce random sequences avoid repetitions too much. The reason is that repetitions look meaningful and thus unlikely to arise by chance. It is difficult to see why people should deceive themselves in a task so simple and so unrelated to their specific interests. Yet, repetition avoidance turns out to predict belief in extrasensory perception (Bressan 2002; Brugger et al. 1995).

We have presented a theory that such a "meaningfulness belief", rather than resulting from self-deception, can emerge as a side effect of schema-based processing of the stimulus (Bressan et al. 2008). A schema is an abstract memory representation built up by concrete past experience (Bartlett 1932; Rumelhart 1984). Schemata include constants for those characteristics of the stimulus that do not change over time, variables for those that do, and constraints that encode regularities of changes. The better a stimulus fits an existing schema, the more it activates this schema. When a schema is activated, those aspects of the stimulus that tend to remain constant are simply retrieved from memory, leaving more resources available to process the aspects that vary. Schemata, thus, speed up the processing of stimuli and render it more efficient (Bartlett 1932; Minsky 1975; Shank & Abelson 1977).

If a stimulus fits a schema in many but not all respects, then the schema is activated but also violated. In this case, the stimulus captures attention and does so in proportion to the schema's strength (Horstmann 2002; Meyer et al. 1991). Schemata can be seen as expectations or beliefs. Whereas each of us can entertain both strong, unchangeable beliefs and weak, flexible ones, different individuals are likely to be differently inclined to maintain strong or weak beliefs. As a measure of schema strength, attentional capture may thus predict the degree of a tendency to maintain either strong or weak beliefs in general.

To test this hypothesis, we performed (Bressan et al. 2008) one of the simple reaction-time experiments of Niepel et al. (1994). Our subjects were to press, as fast as possible, one key if a dot appeared above two words and another if the dot appeared below them. After 32 similar trials, known to establish a strong schema, the 33rd trial presented one of the words in black on white, rather than in the usual white on black. This schema-violating event captured attention, and we found that the attentional capture correlated with the belief that coincidences have meaning. In other words, schema strength (as measured by attentional capture) correlated with meaningfulness belief.

Schemata provide order and, by relating present to past events, also meaning. Schemata that are too strong provide too much order and meaning, to the point that even coincidences can be considered nonaccidental. Belief in a controlling God provides meaning where it might be lacking, too. Consistent with this idea, we found that meaningfulness belief was indeed highly correlated with religious belief. Thus, belief in God and belief in extrasensory perception might be incidental products of efficient memory and attentional processes and need not result from self-deception.

VH&T cite evidence that people led to feel low levels of control are more likely to see illusory patterns in random configurations and to endorse conspiracy theories than people who are self-affirmed. We contend that, rather than implicating self-deception among the former group, the illusions and conspiracy beliefs could simply result from reduced motivation to perform an accurate analysis of the true state of affairs. Those who are likely to gain control over a situation may benefit from efforts to assess that situation well. Those who are unlikely to gain that control may be better off saving themselves the trouble. Because the employment of schemata saves effort at the expense of accuracy, the illusions and conspiracy beliefs could be a by-product of activated schemata and need not result from self-deception.

Like religious belief, political preferences associated with belief in strong government can be predicted on the basis of performance in a simple discrimination experiment. Amodio et al. (2007) asked subjects to respond, as fast as possible, upon seeing an M and withhold their response upon seeing a W, a task that – we believe – is unlikely to involve self-deception. The M was presented much more frequently than the W. The strength of the expectation (the schema) that, after many Ms,

the next trial would also show an M differed from person to person. Amodio et al. found that this variation was systematic and that liberals were better able to withhold their response to a W than conservatives. As conservatives tend to believe in strong government, the latter belief too appears to be related to the efficient processing of stimuli via schemata and need not result from self-deception.

The likely evolution of self-deception is, indeed, at odds with the naïve, conventional view that "natural selection favors nervous systems which produce ever more accurate images of the world." Also at odds with this view is the even more likely evolution of an adaptive system that creates and maintains cognitive schemata, whose by-product is the inclination to ascribe order and meaning to the world even when it has neither. To explain belief in God and belief in strong government, it may not be necessary to assume that self-deception is involved; assuming mental efficiency via the use of schemata might, by itself, be enough.

ACKNOWLEDGMENT

This work was supported in part by a grant from the University of Padova (Progetto di Ricerca di Ateneo CPDA084849) to Paola Bressan.

Two problems with "self-deception": No "self" and no "deception"

doi:10.1017/S0140525X10002116

Robert Kurzban

Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104.

kurzban@psych.upenn.edu

http://www.psych.upenn.edu/~kurzban/

Abstract: While the idea that being wrong can be strategically advantageous in the context of social strategy is sound, the idea that there is a "self" to be deceived might not be. The modular view of the mind finesses this difficulty and is useful – perhaps necessary – for discussing the phenomena currently grouped under the term "self-deception."

I agree with a key argument in the target article, that the phenomena discussed under the rubric of "self-deception" are best understood as strategic (Kurzban, in press; Kurzban & Aktipis 2006; 2007). For a social species like humans, representations can play roles not just in guiding behavior, but also in manipulating others (Dawkins & Krebs 1978). If, for example, incorrect representations in my head (about, e.g., my own traits) will contribute to generating representations in your head that I am a valuable social partner, then selection can act to bring about mechanisms that generate such incorrect representations, even if these representations are not the best estimate of what is true (Churchland 1987).

This is an important idea because generating true representations has frequently been viewed as the key – indeed only – job of cognition (Fodor 2000; Pears 1985). True beliefs are obviously useful for guiding adaptive behavior, so claims that evolved computational mechanisms are designed to be anything other than as accurate as possible requires a powerful argument (McKay & Dennett 2009). Indeed, in the context of mechanisms designed around individual decision-making problems in which nature alone determines one's payoff, mechanisms designed to maximize expected value should be expected because the relentless calculus of decision theory punishes any other design (Kurzban & Christner, in press). However, when manipulation is possible and a false belief can influence others, these social benefits can offset the costs, if any, of false beliefs.

Despite my broad agreement with these arguments, I have deep worries about the implicit ontological commitments lurking behind constructions that animate the discussion in the target article, such as "deceiving the self, "convincing the self," or "telling the self." Because I, among others, do not think there is a plausible referent for "the self" used in this way (Dennett 1981; Humphrey & Dennett 1998; Kurzban, in press; Kurzban & Aktipis 2007; Rorty 1985), my concern is that referring to the self at best is mistaken and at worst reifies a Cartesian dualist ontology. That is, when "the self" is being convinced, what, precisely, is doing the convincing and what, precisely, is being convinced? Talk about whatever it is that is being deceived (or "controlled," for that matter; Wegner 2005) comes perilously close to dualism, with a homuncular "self" being the thing that is being deceived (Kurzban, in press).

So, the first task for self-deception researchers is to purge discussions of the "self" and discuss these issues without using this term. *Modularity*, the idea that the mind consists of a large number of functionally specialized mechanisms (Tooby & Cosmides 1992) that can be isolated from one another (Barrett 2005; Fodor 1983), does exactly this and grants indispensable clarity. For this reason, modularity ought to play a prominent role in any discussion of the phenomena grouped under the rubric of self-deception. Modularity allows a much more coherent way to talk about self-deception and positive illusions that finesses the ontological difficulty.

Consider the modular construal of two different types of self-deception. In the context of so-called "positive illusions" (Taylor 1989), suppose that representations contained in certain modules – but not others – "leak" into the social world. For such modules, the benefits of being *correct* – that is, having the most accurate possible representation of what is true in these modules – must be balanced against the benefits of persuasion (sect. 9). If representations that contain information about one's traits and likely future will be consumed by others, then errors in the direction that is favorable might be advantageous, offsetting the costs of error. For this reason, such representations are best understood not as illusions but as cases in which some very specific subset of modules that have important effects on the social world are designed to be strategically wrong, – that is, they generate representations that are not the best estimate of what is true, but what is valuable in the context of social games, especially persuasion.

Next, consider cases in which two mutually inconsistent representations coexist within the same head. On the modular view, the presence of mutually inconsistent representations presents no difficulties as a result of informational encapsulation (Barrett & Kurzban 2006). If one modular system guides action, then the most accurate representations possible should be expected to be retained in such systems. If another modular system interacts with the social world, then representations that will be advantageous if consumed by others should be stored there. These representations might, of course, be about the very same thing but differ in their content. As Pinker (1997) put it, "the truth is useful, so it should be registered somewhere in the mind, walled off from the parts that interact with other people" (p. 421). One part of the mind is not "deceiving" another part; these modular systems are simply operating with a certain degree of autonomy.

The modular view also makes sense of another difficulty natural language introduces into discussion of self-deception, the folk concept of "belief" (e.g., Stitch 1983). If it is true that two modular systems might have representations about the very same thing, and that these two representations might be inconsistent, then it makes no sense to talk about what an agent "really," "genuinely," or "sincerely" believes. Instead, the predicate "believe" attaches to modular systems rather than people or other agents (Kurzban, in press). This has the added advantage of allowing us to do away with metaphorical terms like the "level" on which something is believed (sect. 7), and we can substitute a discussion of which representations are

present in different modules. Again, this undermines the folk understanding of what it means to "believe" something, but such a move, taking belief predicates away from agents as a whole, is required on the modular view and helps clarify that the belief applies to modules, that is, parts of people's minds, rather than a person as a whole.

Generally, trying to understand self-deception with the conceptual tool of evolved function is an advance. Trying to understand self-deception without the conceptual tool of modularity is needlessly limiting.

Self-deceive to countermine detection

doi:10.1017/S0140525X10002529

Hui Jing Lu and Lei Chang

Department of Educational Psychology, The Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR, People's Republic of China.

luhuijing@cuhk.edu.hk **leichang@cuhk.edu.hk**

http://www.fed.cuhk.edu.hk/eps/people/changl.html

Abstract: Having evolved to escape detection of deception completely, self-deception must respond to social conditions registering different probabilities of detection. To be adaptive, it must have a mechanism to keep truthful information temporarily from the self during deception and retrieve it after deception. The memory system may serve this mechanism and provides a paradigm in which to conduct research on self-deception.

Self-deception has been studied mainly as an intrapersonal process, representing personality traits (Paulhus & John 1998; Paulhus & Reid 1991), motivational biases in information processing (Mele 1997; Balcetis 2008), or inconsistencies between implicit and explicit self-systems (Greenwald 1997). The target article (also see Trivers 1976/2006; 1985; 2000) is among the first to treat self-deception as an interpersonal process by which humans deceive themselves to deceive others. However, the evidence von Hippel & Trivers (VH&T) use to make the interpersonal argument is mainly intrapersonal, given the lack of existing relevant empirical studies. We present interpersonal evidence to augment VH&T's argument. In doing so, we emphasize that, resulting from the interpersonal "arms race" between deception and deception detection, self-deception must respond to social conditions registering detection-varying probabilities. Social status of the deceived and the number of detectors are examples of such social conditions that may also shape the evolution of morality. We also argue that by keeping fitness-enhancing information away from both self and others, self-deception as an adaptation must cease to operate in most instances once the goal of deception is achieved so that truthful information can be retrieved to benefit the self. Such information manipulation makes memory a good target to be co-opted to execute self-deception. Memory research thus provides a good paradigm within which to conduct empirical research on self-deception.

Social status of deceived. The target article (also see Trivers 2000) suggests that by keeping the self unaware of ongoing deception, self-deception may have evolved to avoid detection. Based on this logic, we speculate that self-deception should be sensitive to situations registering different probabilities of detection with individuals being more likely to self-deceive when sensing a higher chance of detection. Social status of the deceived and the number of people to be deceived are examples of variables that affect the chance of being detected and thus the likelihood and activation of self-deception. According to social dominance theory (Cummins 1999), low-status individuals are more motivated to deceive and high-status individuals are more motivated to detect deception, because the latter have more resources to lose, and the former have more to gain in a successful deceptive ploy. However, high-status individuals who have more access to accurate information are more able to detect deception from low-status individuals who are less able although more motivated to deceive. The arms race between deception and deception detection is thus likely to have played out between the low-status individuals as deceivers and the high-status individuals as detectors, and this competition is likely to lead to individuals deceiving themselves to better deceive high- rather than low- or equal-status others. Results from our ongoing research show that low-status individuals unconsciously withhold information from themselves when experimentally motivated to deceive high- but not equal-status others, supporting our view that individuals deceive themselves by keeping information from the consciousness when they sense higher probabilities of detection from the target of deception (Lu & Chang 2010).

Multiple detectors and morality. In addition to status, the number of targets to be deceived also affects the probability of detection. It is generally easier to deceive one rather than multiple targets because each target adds detective pressure to elevate the probability of detection. Using this logic, self-deception is adaptive because, by deceiving one target (the self), it successfully deceives multiple targets (others). We speculate that self-deception is more likely to be used when deceiving multiple targets or groups of individuals than one individual. That is, individuals self-deceive to better group-deceive. A good example of self-deceiving to group-deceiving is self-enhancement in competence or morality, which has so far been studied as intrapersonal processes (Paulhus & John 1998). From an evolutionary point of view, self-enhancement is a form of self-deception that is interpersonally oriented with the deception target being the largest group possible – the public. This group notion about self-deception provides a direction for the understanding of the evolution of morality. In a moral context, self-deception can be seen as an unintentional or unconscious presentation of a falsely moralistic self to avoid public detection of deception. This is in contrast to impression management, which can be seen as straight deception whereby social deceivers intentionally or consciously present a false socially acceptable or laudable self to appease the public. As public pressure for detection eases, for example, when religious congregations or political rallies disperse, the self-deceiver may become more in touch with his or her moral weaknesses and become less critical about moral transgression. Using icons of eyes that have previously been used to represent an audience (Haley & Fessler 2005) to manipulate fear of detection, we found that individuals were less likely to think of moral issues as the number of eyes was reduced.

The memory system. Either to self-deceive high-status or multiple targets, self-deception may be a temporary state of mind rather than permanent or long-lasting self-ignorance. It has been argued that self-deception cannot be adaptive because one deceives her- or himself by being unaware of the truth, which, being intentionally kept away from others, is fitness enhancing or beneficial to oneself (Leeuwen 2007). However, self-deception could be adaptive if it functioned only when deception was ongoing and ceased to operate after the deceiving goal had been achieved. Kept in the unconscious mind, truthful information does not get retrieved while the self-deceiver is deceiving others. When deception ceases, the hidden truth resurfaces so that the self-deceiver benefits from the accuracy of information. The memory system, which has been shown to be a direct target of selection for survival (Nairne et al. 2008), may help to achieve the state of the aforementioned self-deception. By keeping truthful information from the conscious self (i.e., concealing to deceive) or by distorting encoded material (fabricating to deceive), the memory system helps one to honestly offer null or false information to others. This part of self-deception does not achieve a net fitness gain. By later retrieving the truthful information, the memory system helps to complete self-deception to achieve net fitness enhancement. Memory research thus provides a paradigm within which to conduct research on self-deception.

Protesting too much: Self-deception and self-signaling

doi:10.1017/S0140525X10002608

Ryan McKay,[a] Danica Mijović-Prelec,[b] and Dražen Prelec[c]

[a]*Department of Psychology, Royal Holloway, University of London, Egham TW20 0EX, United Kingdom;* [b]*Sloan School and Neuroeconomics Center, Massachusetts Institute of Technology, Cambridge, MA 02139;* [c]*Sloan School and Neuroeconomics Center, Department of Economics, and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139.*

ryantmckay@mac.com mijovic@mit.edu dprelec@mit.edu
http://homepage.mac.com/ryantmckay/

Abstract: Von Hippel & Trivers (VH&T) propose that self-deception has evolved to facilitate the deception of others. However, they ignore the subjective moral costs of deception and the crucial issue of credibility in self-deceptive speech. A *self-signaling* interpretation can account for the ritualistic quality of some self-deceptive affirmations and for the often-noted gap between what self-deceivers say and what they truly believe.

> *The lady doth protest too much, methinks.*
> —Hamlet, Act 3, scene 2, 222–230

> *Like every politician, he always has a card up his sleeve; but unlike the others, he thinks the Lord put it there.*
> —Bertrand Russell (2009, p. 165), citing Labouchere on Gladstone

The notion that overly vehement avowals and overly emphatic behaviors betray knowledge of a disavowed reality is not new. In *Hamlet*, the lady's vow of fidelity to her husband is so passionate and insistent as to arouse suspicion. One possibility is that she is a pure hypocrite, attempting to deceive her audience while knowing full well that her feelings are otherwise. A less cynical observer, however, might conclude that she is only attempting to deceive herself.

For von Hippel & Trivers (VH&T), self-deception and other-deception are not mutually exclusive possibilities. Their evolutionary claim is that the former has evolved in order to facilitate the latter. As they acknowledge, this claim has received surprisingly little attention in the empirical literature (but see McKay & Dennett 2009), which makes the hypothesis almost entirely speculative but not for that reason any less interesting.

The aspect that we focus on here is the psychological architecture that enables self-deception. Although VH&T endorse a "non-unitary mind," defined by separate mental processes with access to privileged information, they resist treating these processes as fully fledged subagents with distinct interests, decision roles, and modes of interaction. Consequently, their theory leaves unresolved the crucial issues of author, audience, and credibility in self-deceptive speech.

Observe, first, that for VH&T the benefits of self-deception are defined as performance enhancement: The "self-deceived deceiver" puts on a smoother show and makes fewer slips that might give the game away. What seems to be ignored in this performance-centered account is the moral dimension of deception. One may ask why psychopathy is not a universal condition if glib performance is so valuable from an evolutionary standpoint.

An alternative interpretation is available, namely, that the benefits of self-deception are realized in the internal moral economy of the self-deceiving individual: The conveniently self-deceived deceivers are absolved from the burden of dealing with unpleasant awareness of their own treachery (Elster 1999). Like Russell's Gladstone, they have license to deceive others without any attendant loss of self-esteem.

On this interpretation, therefore, the motive to self-deceive arises from a desire to perceive oneself as a moral agent. There remains the question of whether the desire will be satisfied, whether ostensibly self-deceptive judgments and affirmations will achieve their goal (Funkhouser 2005). This issue of *self-credibility* can be assessed if we view self-deceptive speech as a form of *self-signaling*, the attempt to convince ourselves that we possess some desired underlying characteristic or trait (Mijović-Prelec & Prelec 2010). If the self-signaling attempt does succeed, and the characteristic is also socially desirable, then guilt-free deception of others may follow as a collateral benefit. However, even if it fails, and fails repeatedly, that need not remove the compulsion to self-signal. Ritualistic affirmations may remain in force, even as they fail to convince.

The prediction emerges once we conceptualize self-signaling by analogy to signaling between individuals. In theoretical biology, signaling refers to actions taken by a *sender* to influence the beliefs of *receivers* about the sender's unobservable characteristics, for example, reproductive quality (Grafen 1990). The sender plays offense by emitting signals that exaggerate his qualities, and the receiver plays defense by discounting or ignoring the messages altogether. The tug of war between offense and defense encourages futile but costly signaling. Even if senders with inferior characteristics do succeed in perfectly emulating the signals emitted by their superiors, the receiver, according to theory, will take this into account and will discount the value of the signal accordingly. The signaling equilibrium is a losing proposition all round; what makes it stick is the fact that failure to send the mandated signal immediately brands the deviant as undesirable.

With *self-signaling*, this entire dynamic is internalized, and messages conveying desired characteristics are reinterpreted as messages to *oneself* (Bodner & Prelec 2003; Quattrone & Tversky 1984). The details of this approach are spelled out elsewhere (Mijović-Prelec & Prelec 2010), but the basic assumption, with respect to psychological architecture, is that there is a division of labor between a sender subsystem responsible for authoring signals, and a receiver subsystem responsible for interpreting them. It is crucial that the two subsystems cannot share information internally, but only through externalized behavior.

What determines whether attempted self-deception is successful? As in the interpersonal case, it all hinges on the credulity of the receiver. If the receiver takes the sender's signal at face value, not discounting for ulterior motives, then attempted self-deception will succeed and we have the "Gladstone" mode. However, the receiver may also discount the signal. This might occur because the receiver has some prior expectation of an ulterior sender motive, or because the deceptive sender misjudges the signal strength. Interestingly, however, discounting may not eliminate the sender's motive to self-signal, because self-serving and pessimistic statements may be discounted asymmetrically (the latter lack an obvious ulterior motive). In such cases, self-deceptive speech becomes mandatory not because it is believed but because deviating from the self-deceptive norm could lead to a catastrophic loss in self-esteem. Self-signaling can therefore lead to ritualistic expression that appears self-deceptive on the surface but that may not truly reflect what a person feels. There will be a mismatch, often noted in the psychotherapeutic literature (Shapiro 1996), between beliefs-as-expressed, for example, about one's self-esteem, sexuality, future prospects, family relationships, and so forth, and beliefs as actually experienced.

If VH&T's evolutionary story is right, then individuals who cannot deceive themselves will be poor at deceiving others. This would not, however, preclude occasional dissociations between self-deception and the deception of others. Some individuals with crushing self-doubts may fail to conceal these doubts from themselves yet manage to maintain an external façade of confidence. Others, with sufficiently credulous receiver subselves, may manage to convince themselves of their self-worth; if, however, their self-aggrandizing statements ring hollow to others, they may be suspected – and accused – of protesting too much.

ACKNOWLEDGMENTS

The first author was supported by grants from the European Commission ("Explaining Religion") and the John Templeton Foundation ("Cognition,

Religion and Theology Project"), both coordinated from the Centre for Anthropology and Mind at the University of Oxford.

Self-deception: Adaptation or by-product?

doi:10.1017/S0140525X10002281

Hugo Mercier

Philosophy, Politics, and Economics Program, University of Pennsylvania, Philadelphia, PA 19104.

hmercier@sas.upenn.edu

http://sites.google.com/site/hugomercier/

Abstract: By systematically biasing our beliefs, self-deception can endanger our ability to successfully convey our messages. It can also lead lies to degenerate into more severe damages in relationships. Accordingly, I suggest that the biases reviewed in the target article do not aim at self-deception but instead are the by-products of several other mechanisms: our natural tendency to self-enhance, the confirmation bias inherent in reasoning, and the lack of access to our unconscious minds.

In their target article, von Hippel & Trivers (VH&T) defend the hypothesis that many psychological biases are by nature self-deceptive. Their rationale is the following: People get caught lying because of "signs of nervousness, suppression, cognitive load, and idiosyncratic sources." In order to make deception detection less likely, these superficial cues should be reduced or eliminated. Given that these cues all stem from the fact that we have to keep in mind the truth and the lie – which we know when we lie – it would make sense for people to actually believe the lies they tell – to self-deceive. However, VH&T fail to take into account that one of the most important cues to deception is lack of consistency (DePaulo et al. 2003). When people are confronted with communicated information, they evaluate its internal consistency as well as its consistency with their previously held beliefs (Sperber et al. 2010). Any benefit gained by lying to ourselves in terms of suppression of superficial cues compromises our ability to make up lies that will pass this consistency test. VH&T also suggest that self-deception could be adaptive because it makes it easier for deceivers to maintain that they had no deceptive intent (their "second corollary"). However, here again self-deception has the potential to backfire. When we know we lied, we can recognize that we did it and feel guilty, apologize, try to make amends, and so forth. These can be essential to the maintenance of trust (Kim et al. 2004; Schweitzer et al. 2006). If we do not even realize that we are trying to deceive, any accusation – however well founded – is likely to be received with aggravation. Thus, by suppressing any common ground between self and audience, self-deception critically endangers the maintenance of trust.

The costs of self-deception weaken the principled case for its adaptiveness. But how are we, then, to account for the evidence that VH&T present in support of their hypothesis? In what follows, I will argue that this evidence can be better explained as the by-product of other mechanisms. Many results presented in the target article show that people have a strong tendency to self-enhance, and that we often do so without even realizing it. This claim would be hard to dispute. For these results to support VH&T's hypothesis, the lack of more veridical information processing must stem from the adaptive character of self-deception. But it is more plausible that the lack of veridical information processing is a simple result of the costs it would entail. It is possible here to make an analogy with other systematically biased mechanisms. For instance, following a simple cost-benefit analysis, it is reasonable to surmise that a mechanism aimed at the detection of poisonous food should be systematically biased toward the "poisonous" verdict. The lack of a less biased information processing requires no explanation beyond this cost-benefit analysis. If a given degree of self-enhancement is adaptive in and of itself, then this is enough to explain why less biased mechanisms would be superfluous. Contrary to what VH&T claim, the fact that we can sometimes engage in more veridical processing does not show that the mechanisms have a self-deceptive purpose. By analogy, our poisonous food detector could also be more or less biased – depending on the individual who is providing us with the food, for instance – without having self-deception as its goal.

The authors' case rests not only on our ability to sometimes turn off our biases and engage in veridical processing, but also on the conditions that trigger veridical processing. More specifically, they claim that because self-affirmation or cognitive load manipulations can make us less biased, then any bias that is otherwise present is likely to be self-deceptive. But these findings can also be explained by the effect of these manipulations on the use of high-level processing – in particular, reasoning. Self-affirmation manipulations can be understood as belonging to a larger group of manipulation – including self-esteem and mood manipulations (e.g., Raghunathan & Trope 2002) – that reduce our tendency to engage in some types of high-level processing (Schwarz & Skurnik 2003). Likewise, cognitive load will automatically impair high-level processing. Reasoning is one of the main mechanisms that can be affected by these manipulations, and the confirmation bias exhibited by reasoning is the source of many of the biased results described by VH&T (Nickerson 1998). It is therefore not surprising that self-affirmation or cognitive load manipulations should make us appear less biased. However, it has been argued that the confirmation bias does not have a self-deceptive function and that it is instead the result of the argumentative function of reasoning (Mercier & Sperber, in press). Accordingly, when reasoning is used in a natural setting (such as group discussion), the confirmation bias does not systematically lead to biased beliefs (Mercier & Landemore, in press). Thus most of the results used by the authors can be accounted for as a by-product of a confirmation bias inherent in reasoning that does not have a self-deceptive function.

Finally, a case can also be made against the authors' interpretation of the dual-process literature. According to VH&T, "these dissociations [between, e.g., implicit and explicit memory] ensure that people have limited conscious access to the contents of their own mind and to the motives that drive their behavior." For this statement to be correct, conscious access to the content of our own mind would have to be a given from which it can sometimes be useful to deviate. But this is not the case. Being able to know the content of our own minds is a very costly process. In fact, it is sometimes speculated that there was little evolutionary advantage to be gained by knowing ourselves, and that this ability is a mere by-product of our ability to understand others (e.g., Carruthers 2009b). If *not* knowing ourselves – or knowing ourselves very imperfectly – is the baseline, then dissociations between conscious and unconscious processes require no further explanation. These dissociations cannot *ensure* us against a self-knowledge that we have no reason to possess in the first place.

Trying to elucidate the ultimate function of our cognitive biases is a very worthwhile endeavor that is bound to lead to a much deeper understanding of human psychology. However, for VH&T's specific hypothesis to be truly convincing, they would need to provide stronger evidence, such as the direct experimental tests – whose absence they repeatedly deplore – of their theory.

Representations and decision rules in the theory of self-deception

doi:10.1017/S0140525X1000261X

Steven Pinker

Department of Psychology, Harvard University, Cambridge, MA 02138.

pinker@wjh.harvard.edu **http://pinker.wjh.harvard.edu**

Abstract: Self-deception is a powerful but overapplied theory. It is adaptive only when a deception-detecting audience is in the loop, not when an

inaccurate representation is invoked as an internal motivator. First, an inaccurate representation cannot be equated with self-deception, which entails *two* representations, one inaccurate and the other accurate. Second, any motivational advantages are best achieved with an adjustment to the decision rule on when to act, not with a systematic error in an internal representation.

> If . . . deceit is fundamental to animal communication, then there must be strong selection to spot deception and this ought, in turn, to select for a degree of self-deception, rendering some facts and motives unconscious so as not to betray – by the subtle signs of self-knowledge – the deception being practiced.

This sentence, from Robert Trivers's foreword to *The Selfish Gene* (Trivers 1976/2006), might have the highest ratio of profundity to words in the history of the social sciences. Von Hippel & Trivers's (VH&T's) elaboration and empirical grounding of that offhand comment in the target article is a substantial and highly welcome development.

For all its explanatory power, the adaptive theory of self-deception is often applied too glibly in the social psychology literature. The theory always had two apparent problems. The first is the paradox (or at very least, puzzling redundancy) in which the self is both deceiver and deceived. The second is the claim that selection systematically favored inaccurate representations of the world. The claim that self-deception is a weapon in an arms race of deception and deception-detection appears to resolve these problems, and it is what makes the theory so interesting. The insertion of a second party – the audience for self-presentation – into the deceiver–deceived loop resolves the various paradoxes, and VH&T lay out this logic convincingly.

But many psychologists who invoke self-deception (including, occasionally, VH&T) dilute the force of the theory by applying it to phenomena such as happiness, optimism, confidence, and self-motivation, in which the loop is strictly internal, with no outside party to be deceived. I see two problems with this extrapolation.

The first is that it is essential to distinguish *errors and biases*, on the one hand, from *self-deception*, on the other. Just because a computational system is tuned or designed with inaccurate representations, that does not mean that it is deceiving itself. If my thermostat is inaccurate, and I set the temperature at a higher level than what I want in order to get what I want, or if my car works better when I set the fuel-air ratio to a different value than is "optimal" according to the manufacturer, it seems gratuitous to describe this as self-deception.

For the counterintuitive and apparently profligate concept of self-deception to be useful, the following condition must be met: The system must have *two* representations of some aspect of reality, one of them accurate and the other systematically inaccurate, and the part with access to the accurate information (the self-deceiver) must have control over the information available to the other part (the deceived self). I agree with VH&T that the deception-detection arms race offers a convincing explanation of why this seemingly odd arrangement should have evolved (the deceived self is there to present an inflated self-image designed to fool other parties; the deceiving self is there to keep the entire person from losing all touch with reality). But many putative examples of self-deception (such as being over-optimistic in order to fire up one's own motivation, or being over-impressed with one's own assets to enhance self-confidence) require only a one-level representation with error or bias, not a two-level representation, one inflated and one accurate. In such cases, the theory of self-deception is superfluous. For example, in Epley and Whitchurch's (2008) experiment on inflated self-images, is there any evidence that a more accurate representation of the self's appearance is registered somewhere in the brain? Or that it is actively suppressed?

The second problem is that the adaptive explanation of self-deception, when there is no external audience in the loop, does not work. Prima facie, any computational system ought to be accurately rather than inaccurately tuned to the world. Any need to behave in a way that differs from reading out an accurate representation and acting accordingly ought to be accommodated by changing the *decision rule* that is fed by the information, not by adding noise or bias to the information. After all, it is only the output of the decision rule in real behavior that is ultimately adaptive or not; the internal route to the optimal behavior is not, by itself, visible to selection. If every day I look at the thermometer and end up dressing too warmly, the optimum response is not to reprogram my thermometer to display a too-warm temperature (e.g., display 70° when it is really 65°); it is to change my rule on how to dress for a given temperature (e.g., "put on a sweater when it is 60° out" rather than "put on a sweater when it is 65° out"). The reason that this is the optimum is that if you jigger with the representation rather than the decision rule, then any *other* decision rule that looks at that information readout will now make an undesired error. In this example, if you want to bring in your potted plants when there's a danger of freezing, your jiggered thermometer will now read 35° when it is really 30°, fooling you into leaving the plants outside and letting them die. As long as there is more than one decision rule that accesses a given piece of information, an adjustment toward optimal behavior should always change the decision rule, not the information representation. (If there is only a single decision rule that looks at the representation, there does not need to be a separate representation at all; one could compile the representation and decision rule into a single stimulus-response reflex.)

Now, one could always plead that the human brain is not designed optimally in this regard – but without the external benchmark of optimal design against which to compare the facts of human psychology, one is in just-so-story land, pleading that whatever the facts are had to be the way they are. VH&T escape this problem with the deception-detection arms-race rationale for self-deception (because of the intrusion of an audience whose ultimate genetic interests diverge from those of the self), but such an explanation does not go through when it comes to the putative internal motivating function of self-deception involving happiness or optimism.

Consider the suggestion, common in the literature on positive illusions, that people are overly optimistic because of the adaptive benefit of enhancing their motivation. The problem with this explanation is as follows. Instead of designing an organism with unrealistically optimistic life prospects and a too-conservative motivational rule, why not design it with *realistic* life prospects and a slightly more *liberal* motivational rule, which would avoid the pitfalls of having tainted information floating around in the brain where it might cause mischief with other processes? Consider the situation in which a person is faced with the choice of engaging in a risky game or venture. It is hard to see the adaptive advantages of having a mind that works as in situation (a), which is the common assumption in the positive-illusion and overconfidence literature, rather than as in situation (b):

(a) The objective chance of success is 35%. The self only engages in a venture if it thinks the chance of success exceeds 50%. Taking this particular risk, however, is an adaptively good bet. Therefore, the self is deluded into believing that the chances of success are 70%.

(b) The objective chance of success is 35%. The self only engages in a venture if it thinks that the chance of success exceeds 30%. Taking this particular risk is an adaptively good bet. Therefore, the self accurately represents its chances and engages in the venture.

For any adaptive explanation of self-deception to be convincing, it would have to demonstrate some kind of design considerations that would show why (a) is optimal a priori, rather than just that it is what people tend to do. That seems unlikely.

VH&T are admirably cautious in applying the theory of self-deception. For the theory to stand as a coherent rather than a glib adaptive explanation of human error, the psychologists invoking it must be explicit as to whether they are positing a single-representation *bias* or a double-representation *self-deception*, and

whether they are positing an inaccuracy in the *representation* or a bias in the *decision rule*.

Self-deception, social desirability, and psychopathology

doi:10.1017/S0140525X10002487

Antonio Preti[a] and Paola Miotto[b]

[a]*Centro Medico Genneruxi, Cagliari, Italy, and Chair of Clinical Psychology, University of Cagliari, Italy;* [b]*Department of Mental Health, ULSS 7, Conegliano (TV), 31015, Italy.*

apreti@tin.it

miottopaola@yahoo.it

Abstract: Social desirability can be conceived as a proxy for self-deception, as it involves a positive attribution side and a denial side. People with mental disorders have lower scores on measures of social desirability, which could depend on cognitive load caused by symptoms. This suggests that self-deception is an active strategy and not merely a faulty cognitive process.

Von Hippel & Trivers (VH&T) have interestingly expanded past speculations by Trivers on self-deception (Trivers 1976/2006; 2000). At the core of this theory is the idea that "by deceiving themselves, people can better deceive others, because they no longer emit the cues of consciously mediated deception that could reveal their deceptive intent." Self-deception also helps the deceiver to accrue "the more general social advantages of self-inflation or self-enhancement." However, we think that the role of mental health in self-deception deserves a better place in their model.

There are three levels of deception: denying the truth (as saying that something true is false), advocating the false (as saying that something false is true), and withholding information about truth or falsehood (as in a secret).

Self-deception can occur only by withholding information about truth or falsehood, because this can occur without the subject's being conscious about what she or he is doing: Indeed, throughout the target article, VH&T hold that by "deceiving themselves, people are able to avoid the cognitive costs of consciously mediated deception," that is, self-deception is based on some mechanism that operates at a non-conscious level.

The process of withholding information about truth or falsehood could occur by cognitive malfunctioning, while still achieving the result of the subject's deceiving others.

Social desirability can be conceived as a proxy for self-deception, given that it involves a positive attribution side (attributing to themselves rare but socially appreciated qualities) and a denial side (denying to have the negative qualities that are common in the general population) (Crowne & Marlowe 1964; Ramanaiah et al. 1977; Paulhus 1991).

In the past 20 years, a series of studies showed that people with symptoms of mental disorders have statistically lower scores on measures of social desirability (Lane et al. 1990). In particular, we found that adolescents or young adults scoring higher on measures of depression (Miotto et al. 2002), psychosis proneness (Preti et al., 2010), or suicidal ideation (Miotto & Preti 2008) reported concurrently lower scores on the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe 1960), in particular on the denial subscale of that questionnaire.

Lower scores on social desirability measures could be a result of the cognitive load produced by ruminations on sad thoughts in depression (Lane et al. 1990; Miotto et al. 2002) or be caused by hallucinations and delusions in psychosis (Preti et al., 2010). As pointed out by VH&T, a role for cognitive load in self-deception might depend on self-deception being an active strategy, and not merely a faulty cognitive process resulting in the wrong withholding of information about truth or falsehood. Therefore, the investigation of the links between self-deception and psychopathology plays a role in discovering the neuropsychological basis of self-deception.

How can it be that a mechanism implying a loss of information integrity can effectively favor adaptation? As Stich (1990, p. 62) put it, "natural selection does not care about truth; it cares only about reproductive success," and this is the case when an unbiased system is more detrimental to fitness than a system characterized by occasionally mistaken evaluations.

In the past we proposed that cheaters are part of the mechanism that challenges the subjects' fitness. Because the cognitive abilities leading to cheater's detection might prove useful in all kinds of cooperative exchange (Stevens & Hauser 2004), groups of discriminative cooperators will out-compete over groups of non-discriminative cooperators. In fact, cheaters select those individuals who are more able to detect cheating. Conversely, the hosts tolerate some amount of cheaters in their environment because they continuously challenge the hosts' own cognitive abilities, as parasites resident in our skin stimulate the immune system and act as a restraint against more virulent invaders because they keep the niche occupied (Preti & Miotto 2006).

Because cheaters are likely to take advantage of sexual partners, thus distributing their genes in the general population, some cheating mechanism with a genetic basis is likely to be fixed in the gene pool. Self-deception could be one of these mechanisms transmitted by cheaters to their victims, and it could help individuals detect cheating.

This is paradoxical, because self-deception is expected to favor cheating. However, both mirror-neuron theory (Rizzolatti & Craighero 2004) and the embodied cognition paradigm (Grafton 2009) posit that people are more able to detect those motor and emotional cues they have already experienced. In monkeys, "the different modes of presentation of events intrinsically different, as sounds, images or willed motor acts, are ... bound together within a simpler level of semantic reference, underpinned by the same network of audio–visual mirror neurons" (Gallese 2007, p. 660), that is, the animal quickly reacts to stimuli that are bound by experience to a predetermined outcome. In humans, "actions belonging to the motor repertoire of the observer (e.g., biting and speech-reading) or very closely related to it (e.g., monkey's lip-smacking) are mapped on the observer's motor system. Actions that do not belong to this repertoire (e.g., barking) are mapped and, henceforth, categorized on the basis of their visual properties" (Gallese 2007, p. 661).

Self-deception implies some kind of resetting of the subject's mental state, and this reflects in the fine motor disposition of the muscles involved in facial expression. People more likely to use a self-deception strategy will also be more able to detect these subtle cues, some kind of dissociation between the communicated content and the real, inner, secretly held content, operating at an unconscious level but withheld at the conscious one.

Therefore in the social war between cheaters and their victims, self-deception can both favor cheating and help detect cheating. It is not merely the advantage of being more convincing at deceiving the others that fixed self-deception as a strategy in our heritage, but also the contribution that the mental states related to self-deception give to cheating detection. This could be tested: People scoring higher on measures of self-deception should also be more able to detect cheating and deceptive attempts.

Aiming at self-deception: Deflationism, intentionalism, and biological purpose

doi:10.1017/S0140525X10002657

David Livingstone Smith

Department of Philosophy, University of New England, Biddeford, ME 04005.

dsmith@une.edu

http://realhumanature.com

Abstract: Deflationists about self-deception understand self-deception as the outcome of biased information processing, but in doing so, they

lose the normative distinction between self-deception and wishful thinking. Von Hippel & Trivers (VH&T) advocate a deflationist approach, but they also want preserve the purposive character of self-deception. A biologically realistic analysis of deception can eliminate the contradiction implicit in their position.

Self-deception is a form of motivated misbelief, but not every motivated misbelief counts as self-deception. Sometimes, motivated misbeliefs are accidental consequences of motivational states. We refer to these as "wishful thinking" rather than self-deception. Suppose *S* strongly desires *p* to be the case, and this desire distorts *S*'s assessment of the evidence so that *S* comes to unjustifiably believe *p*. In this scenario, neither *S* nor any process in *S* aims at producing misbelief. Because it has no aim, wishful thinking cannot succeed or fail. In the paradigmatic case of self-deception, *S* believes not-*p*, desires strongly *p*, and either *S* (or some mechanism in *S*) aims at causing *S* to misbelieve that *p*. In this case *S* (or some mechanism in *S*) succeeds in producing a misbelief if *S* ends up believing *p*, and fails if *S* ends up believing not-*p*. I will refer to this as the *normative distinction* between self-deception and wishful thinking.

Self-deception and wishful thinking both have interpersonal counterparts. Suppose you are driving through an unfamiliar town, and you ask a man for directions to the nearest gas station. He tells you to turn right at the next light, but unbeknownst to him, the gas station has closed down. He has not deceived you, because he did not intend to cause you to have a false belief, and therefore did not *succeed* in causing you to have a false belief. However if he had intended to lead you astray, he would have succeeded in deceiving you.

Modeling self-deception on interpersonal deception requires the assumption that the self-deceiver *intends* to cause himself to misbelieve. This way of looking at things requires that self-deceivers simultaneously believe that *p* and believe that not-*p*, and it also requires that self-deceivers conceal their self-deceptive intentions from themselves, both of which seem impossible. One response to these problems is to posit that self-deceivers have divided minds. The self-deceiver's mind contains a homunculus that intentionally causes him or her to misbelieve. There are numerous problems with this suggestion, not the least of which is its utter implausibility.

These sorts of problems motivate the *deflationist* approach to self-deception. Deflationists abjure modeling self-deception on interpersonal deception and argue that self-deception is an unintended consequence of information-processing biases. In Mele's (2001) formulation, these cognitive biases aim at avoiding costly errors, *not* at causing misbelief. Misbelief is, so to speak, a side effect. The problem is that this eliminates the normative distinction between self-deception and wishful thinking. The processes that cause *S* to misbelieve do not aim at doing this, so self-deception cannot succeed or fail, and no principled distinction can be drawn between self-deception and wishful thinking.

Von Hippel & Trivers (VH&T) make it clear that they are deflationists, but they also want to model self-deception on interpersonal deception and recognize (in the example of the husband coming home late from work) that intent is necessary for interpersonal deception. In short, they appear to be advocating two incompatible models of self-deception: deflationism, which denies that self-deception is intentional, and intentionalism, which requires it.

There is a reason VH&T have put themselves in this unenviable position. They recognize that the deflationist approach is empirically appealing, but it is important for their whole conception of self-deception that it is purposive. My guess is that VH&T are drawn to think about self-deception as similar to interpersonal deception because they want to capture its goal-directed character, but they do not notice that this plays havoc with their commitment to deflationism.

My diagnosis is that VH&T have made the erroneous assumption almost everyone working on self-deception makes: They mistakenly equate other-deception with *interpersonal* deception. This is a mistake because nonhuman organisms also engage in other-deception. A comprehensive theory of other-deception needs to encompass the whole spectrum, ranging from bee orchids' deception of their pollinators (which is clearly not intentional) to deliberate human lying (which clearly is).

The most promising formulation of a general theory of deception draws upon Millikan's (1984; 1993) analysis of biological purpose. Due to limitations of space, I will not set out the details of Millikan's theoretical apparatus but will present a somewhat simplified version of an analysis based on it.

> Deception $=_{df}$ For organisms O_1 and O_2, O_1 deceives O_2 iff O_2 possesses a character C with the purpose F of representing truly and O_1 possesses a character C^* with purpose F^* of causing C to misrepresent, and it is in virtue of performing F^* that C^* causes C to misrepresent.

This can be crudely paraphrased as follows. One organism deceives another just in case the deceiving organism has some characteristic that has the biological purpose of preventing the representational apparatus of the deceived organism from correctly representing its environment (i.e., fulfilling its biological purpose). The deceiving organism succeeds in causing the deceived organism's representational apparatus to fail. Using this analysis as a model for self-deception, we get:

> Self-deception $=_{df}$ O deceives itself iff O possesses character C with purpose F of representing truly, and character C^* with the purpose F^* of causing C to misrepresent, and it is in virtue of performing F^* that C^* causes C to misrepresent.

Roughly paraphrased, an organism deceives itself if and only if it has some characteristic with the biological purpose of causing the representational apparatus *in the same organism* from correctly representing its environment (fulfilling its biological purpose), and that characteristic succeeds in causing the representational apparatus to fail. This analysis of self-deception can then be made consistent with deflationism by adding that C^* fulfills its purpose by biasing the manner in which the organism processes information.

Evolution, lies, and foresight biases

doi:10.1017/S0140525X10002128

Thomas Suddendorf

School of Psychology, University of Queensland, Brisbane, Queensland 4072, Australia.

t.suddendorf@psy.uq.edu.au

http://www.psy.uq.edu.au/directory/index.html?id=39

Abstract: Humans are not the only animals to deceive, though we might be the only ones that lie. The arms race von Hippel & Trivers (VH&T) propose may have only started during hominin evolution. VH&T offer a powerful theory, and I suggest it can be expanded to explain why there are systematic biases in human foresight.

Von Hippel & Trivers (VH&T) argue that self-deception evolved not as a defense mechanism but as an offensive weapon in an evolutionary arms race of deception and deception-detection. Their proposal explains how the deceiver can also be the deceived and why evolution may have possibly selected for mechanisms that represent the world inaccurately. This is thus a powerful perspective that, as illustrated by the latter part of the target article, sheds new light on a range of phenomena. Here I will suggest that this approach might also offer a way out of the vexing problem of systematic biases in human foresight. First, however, I note that the authors fail to discuss when such an arms race may have got off the ground. They thus sidestep an important aspect of the evolution of self-deception.

Humans are clearly not the only creatures to deceive. Many animal signals are not honest but were selected to deceive

predators, prey, or competitors. This deception is not limited to simple mimicry but includes various curious behaviors and even counterdeception. Some primates, especially our closest living relatives, appear to engage in quite flexible "tactical" forms of deception (Whiten & Byrne 1988). Do they hence have the prerequisites for the purported arms race between deception and deception detection?

One important aspect of deception that nonhuman animals may not be capable of is lying. Parts of the target article seem to use the terms *deception* and *lying* interchangeably, perhaps reserving the latter to describe verbal deception. Yet to lie, one really must do more than declare something that is in fact not true. One must also know that it is not true; otherwise mistakes would be called lies. Furthermore, one must want the other to believe what one knows not to be true, to be true. Thus, lying implies intentionally implanting a false belief. In spite of persistent efforts, there is as yet no convincing evidence that nonhuman animals can represent false beliefs (Krachun et al. 2009; Penn et al. 2008). If this is correct, then they lack the very opportunity to deliberately manipulate such representations. Lies, and the self-deceptions they may have spawned, appear to have evolved only over the past five million years, after the split from the last common ancestor with chimpanzees.

A second purportedly unique skill that may have played an important role in the evolution of human deception and self-deception is *mental time travel* (Suddendorf & Corballis 1997). Humans can flexibly imagine a range of potential future episodes (and hence can plan complex deceptive ploys) and can also mentally reconstruct past events (and hence can uncover past deceptive ploys). These travels in both temporal directions are closely linked in mind and brain (e.g., Addis et al. 2007; Suddendorf & Corballis 2007). From Bartlett (1932) we know that recollecting past events is an active construction that may be biased. Selection for memory must be based on what fitness benefits it brings, not on how accurate it is per se. The same must be true for thinking about future events. Foresight is implied in various contexts in the target article (e.g., optimism, plans, and goal achievement), but was not addressed directly. Yet, I think (possibly because I deceive myself about the importance of something I have been working on for too long), that VH&T's theory can throw new light on the evolution of biases in foresight.

Our ability to imagine future scenarios has obvious adaptive benefits, allowing us to prepare in the present to secure future rewards or thwart future disaster. Why, then, is it that humans display systematic errors in anticipation? For example, various lines of research (see Gilbert 2006) have demonstrated that we tend to exaggerate the positive or negative emotional consequences of future events (e.g., of handing in one's PhD thesis; or of losing a leg). When the event occurs, we tend to feel not quite as happy as we had imagined, and we tend to cope much better with a negative event than we anticipated. We also systematically misjudge the likelihood of events. VH&T allude, for example, to the optimism bias where we generally tend to judge the likelihood of good things happening to our future self above that what is rational.

On the face of it, there are some clear benefits to these biases. One apparent benefit of exaggerating the hedonic value of future consequences, for instance, is that it may help motivate future directed action. One can only fully reap the benefits of anticipating future events if such thought can appropriately guide present action. The system that governs motivation, however, has long been based on present rewards. Evolution had to modulate this system rather than build a new one from scratch. An important problem foresight poses, then, is the need to motivate prudent action in the present when this is costly, or when it prevents more immediate hedonic rewards. To compete with current rewards and motivate future-directed action, it may thus make sense to exaggerate the future reward or punishment.

The logical problem, though, is that one would expect the system to learn with experience (and modulate decision making). People should learn to adjust their predictions and create a more accurate representation of future hedonic values. On some level, perhaps, we do appreciate the truth (the German vernacular, e.g., tells us that "Vorfreude ist die schönste Freude" [anticipated joy is the greatest joy]). Yet, we continue to display the same forecasting biases.

VH&T's theory explains how a system might have evolved that mis-represents the facts by the clever proposal of a social arms race between deception and deception detection. This approach could also offer a solution here. With language, a third purportedly uniquely human capacity, humans exchange their plans and coordinate them. Indeed, language may have evolved initially for the sharing of mental time travels (Suddendorf et al. 2009). In order to elicit cooperation on a project, one may benefit from exaggerating the likelihood and positive consequences of success (or negative consequences of failure). As in the argument of VH&T, I propose that one may be much better at doing this if one actually believes this exaggeration. Again, such belief may also reduce potential punishment if the future fails to bring what was promised. Thus, our biases in judging future hedonic values and likelihoods might be self-deceptions that have their origin in an evolutionary arms race between deception and deception detection mechanisms of a social other. VH&T's theory may potentially go a long way further in explaining the evolution of the complexities of the human mind than even they have anticipated.

Deception through self-deception: Take a look at somatoform disorders

doi:10.1017/S0140525X10002633

Alfonso Troisi

Department of Neuroscience, University of Rome Tor Vergata, 00161 Rome, Italy.

alfonso.troisi@uniroma2.it

Abstract: Patients with physical symptoms for which no organic cause can be found are distributed along a continuum of disease simulation that ranges from a sincere belief of having a serious disease to intentional presentation of false symptoms. The evolutionary hypothesis that self-deception improves the deception of others can explain such a combination of unconscious and intentional production of physical symptoms.

Arguing for their evolutionary approach to self-deception, von Hippel & Trivers (VH&T) complain that "to the best of our knowledge no one has examined whether self-deception is more likely when people attempt to deceive others. Thus, the theoretical possibility of self-deception in service of other deception remains just that" (sect. 6, para. 2). Indeed, convincing evidence for their evolutionary hypothesis that self-deception evolved to facilitate interpersonal deception comes from clinical studies of psychiatric patients with somatoform disorders.

Current psychiatric classification includes several different conditions (somatoform disorders, factitious disorders, and malingering) in which the clinical picture is dominated by physical symptoms for which no organic cause can be found. These disorders have in common several features, including "illness as a way of life," maladaptive use of medical care, refractoriness to palliative and symptomatic treatment, and the desire to seek those privileges afforded to the sick person by society (Krahn et al. 2008). This latter aspect points to the importance of deception of others in the psychological mechanisms underlying somatization and disease simulation. By simulating a disease and displaying care-eliciting behavior, the patient evokes predictable reactions from others (i.e., care of the sick) that are sanctioned in all human cultures and are probably related with invalid care behavior observed in several nonhuman species.

Even though these conditions share many common features, current diagnostic criteria separate somatoform disorders from factitious disorders and malingering. According to this diagnostic partitioning, the key difference is that patients with somatoform disorders actually believe they are ill, whereas patients with factitious disorders and malingerers feign illness and fake their physical symptoms. But the line between somatization (e.g., unconscious expression of emotional distress in physical terms) and deliberate simulation of physical symptoms is not easy to draw. Since the birth of modern psychiatry, clinical observations have repeatedly highlighted the difficulties in defining the patient's degree of voluntary control over symptom production.

In the nineteenth century, the French neurologist J.- M. Charcot was an international leader in the study of hysteria (a neurotic disorder currently classified under the rubric of somatoform disorders) (Goetz 2007). Speaking to his students on this topic, he provided a rare glimpse of his personal attitudes toward the creative spirit of hysterical simulators: "This leads me to say a word on simulation. You will meet with it at every point when dealing with the history of hysteria. One sometimes catches oneself admiring the amazing craft, sagacity, and perseverance which women, under the influence of this great *névrose*, will mobilize for the purpose of deception – especially when a physician is to be the victim." (Charcot 1889, p. 230). More recently, Cameron noted that "the difference between hysteria and malingering must finally rest upon the criterion of self-deception. . .. there are many cases in which pretense and self-deception are so intermingled as to make clear distinction impossible." (quoted in Ford 1983, p. 130). In line with these classical descriptions, contemporary clinicians acknowledge the striking mixture of conscious and unconscious control over symptom production that characterizes not only the different diagnostic subtypes of somatoform disorders, but also the temporal course of individual cases with the same diagnosis (Rogers 1988).

A recent study has demonstrated how intentional faking may evolve into a less conscious form of symptom reporting (Merckelbach et al. 2010). These authors conducted three experiments that addressed the residual effects of instructed feigning of symptoms. In experiment 1, undergraduates instructed to exaggerate symptoms on a malingering test continued to report more neurocognitive and psychiatric symptoms than did nonmalingering controls, when later asked to respond honestly to the same test. In experiment 2, students completed a symptom list of psychiatric complaints and then were asked to explain why they had endorsed two target symptoms that they did not, in actuality, endorse. A total of 57% of participants did not detect this mismatch between actual and manipulated symptom endorsement and even tended to adopt the manipulated symptoms when provided with an opportunity to do so. In experiment 3, it was found that self-deceptive enhancement is related to the tendency to continue to report neurocognitive and psychiatric symptoms that initially had been produced intentionally. Discussing the implications of their study for a better understanding of somatoform disorders, Merckelbach et al. concluded that "blindness" for the intentional aspect of symptom endorsement may explain the intrinsic overlap between feigning and somatoform complaints.

Twenty years ago, I collaborated on an article (Troisi & McGuire 1990) focusing on somatoform disorders from an evolutionary perspective. The core argument of the article was that the evolution of self-deception as a means to improve deception of others could explain why patients with somatoform disorders distribute along a continuum of disease simulation that ranges from a sincere belief of having a serious disease to intentional presentation of false symptoms. We argued against a diagnostic partitioning based on the distinction between unconscious somatization and intentional faking and emphasized the importance of assessing the relative contribution of deception and self-deception in each single case presenting with physical symptoms for which no organic cause can be found. Using the deception/self-deception dimensional model, we explained several clinical features of somatoform disorders. For example, the degree of self-deception explains the heterogeneity of the clinical picture. Some patients present with objective and dramatic manifestations (e.g., pseudoseizures), whereas others just complain of subjective symptoms (e.g., nausea). We hypothesized that patients who do not believe in their diseases tend to exaggerate the conspicuousness of disease manifestations in an effort to prove their validity and to convince others. By contrast, patients with an extreme degree of self-deception (that, in some cases, can reach a delusional intensity) limit their symptom production to vague complaints.

Like the original article by Trivers (1976/2006) on the evolution of self-deception, the Troisi & McGuire (1990) article on somatoform disorders has not been taken seriously in the psychiatric literature. I hope that VH&T's target article will convince researchers and clinicians that somatoform disorders are both a source of preliminary evidence for the interpersonal function of self-deception and a promising area to conduct experiments testing the evolutionary hypothesis.

Self-deception, lying, and the ability to deceive

doi:10.1017/S0140525X10002293

Aldert Vrij

Psychology Department, University of Portsmouth, Portsmouth PO1 2DY, United Kingdom.

Aldert.Vrij@port.ac.uk

http://www.port.ac.uk/departments/academic/psychology/staff/title,50475,en.html

Abstract: Von Hippel & Trivers (VH&T) argue that people become effective liars through self-deception. It can be said, however, that people who believe their own stories are not lying. VH&T also argue that people are quite good lie detectors, but they provide no evidence for this, and the available literature contradicts their claim. Their reasons to negate this evidence are unconvincing.

Von Hippel & Trivers (VH&T) consider self-deception as an offensive strategy evolved for deceiving others. Via self-deception, people can convince themselves that their lies are true, and consequently, they will no longer emit the cues of consciously mediated mendacity that could reveal their deceptive intent. Indeed, people who deceive themselves and do not display deception cues are difficult to catch, but it tells us nothing about lying skills because they are not lying. Lying is "a deliberate attempt to mislead others," and "falsehoods communicated by people who are mistaken or self-deceived are not lies" (DePaulo et al. 2003, p. 74).

VH&T describe different ways in which people can deceive themselves, including through "biased information search" (avoiding further information search, selective information search, and selective attention to available information), "biased interpretation of information," and "misremembering [information]." This means that they take into account only a few of the vast number of lies people can tell. How can a suspect who burgled a house last night and is interviewed by the police use such mechanisms? Or the man who has returned home late from work after talking to his female colleague and then is asked by his wife why he is late? (VH&T's example). They cannot.

VH&T report that "people are actually quite good at detecting deception." Research does not support this claim (Bond & DePaulo 2006; Vrij 2008). However, VH&T argue that lie detection research may have grossly underestimated people's ability to detect deception because it relies on conditions that "heavily advantage the deceiver." Those conditions include (1) the deceiver being unknown to the lie detector and (2) no opportunity to question the deceiver. Other conditions, however, give advantage

to *lie detector*, so that they are aware that they may be lied to. One important reason why lies in daily life remain undetected is that people tend to be credulous (DePaulo et al. 2003; Vrij 2008). It makes sense to be credulous, as people are more often confronted with truths than lies (DePaulo et al. 1996), but it hampers lie detection. The tendency to judge others as truthful becomes stronger as relationships become more intimate (Levine & McCornack 1992; McCornack & Parks 1986; Stiff et al. 1992). This could explain why research has shown that people are no better in detecting lies in friends or partners than in strangers (negating VH&T's claim). In fact, none of the studies where a direct comparison was made between the ability to detect truths and lies in strangers versus in friends or partners found a difference in accuracy rates (Anderson et al. 2002; Buller et al. 1991; Fleming et al. 1990; Millar & Millar 1995). One reason why no link between relationship closeness and accuracy at detecting deception seems to exist is that when close relationship partners attempt to detect deceit in each other, they bring to mind a great deal of information about each other. This information could be overwhelming, and the lie detector may deal with this by processing the information heuristically instead of carefully searching for genuine cues to deceit. Another explanation is that as relationships develop, people become more skilled at crafting communications uniquely designed to fool each other (Anderson et al. 1999).

There is no evidence either that the ability to interview *in itself* facilitates lie detection, as VH&T suggest (see Vrij 2008 for a review of those studies). It depends on how the interviews are conducted. In one experiment truth tellers went to a shop and bought an item that was hidden under a briefcase. Liars took money out of the briefcase. Therefore, both truth tellers' and liars' fingerprints were found on the briefcase. Swedish police detectives were given this fingerprint evidence and were requested to interrogate the suspect in the style of their choice. They obtained 56.1% accuracy. In contrast, police detectives who were taught an innovative interrogation technique aimed at using the piece of evidence strategically during the interrogation (by asking questions about the evidence without revealing it) obtained 85.4% accuracy (Hartwig et al. 2006).

Recent research revealed another way to detect lies via strategic interviewing. As VH&T correctly argue, liars experience more cognitive load than truth tellers. A lie catcher could exploit the differential levels of cognitive load to discriminate more effectively between them. Liars who require more cognitive resources than truth tellers will have fewer cognitive resources left over. If cognitive demand is further raised, which could be achieved by making additional requests, liars may have more difficulty than truth tellers in coping with these additional requests.

Ways to impose cognitive load include asking interviewees to tell their stories in reverse order or instructing them to maintain eye contact with the interviewer. In two experiments, half of the liars and truth tellers were requested to recall their stories in reverse order (Vrij et al. 2008) or to maintain eye contact with the interviewer (Vrij et al. 2010), whereas no instruction was given to the remaining participants. More cues to deceit emerged in the reverse order and maintaining eye contact conditions than in the control conditions. Observers who watched these videotaped interviews could distinguish between truths and lies better in the reverse order condition and maintaining eye contact conditions than in the control conditions. Vrij et al. (2010; in press) provided overviews of interviewing to detect deception research.

VH&T suggest future research examining how people detect lies in daily life. Such research has already been conducted (Park et al. 2002). Less than 2% of the lies were detected at the time the lie was told by relying exclusively on the liars' nonverbal behavior or speech content. Lies were mostly discovered via information from third parties (38%), physical evidence (23%), and confessions (14%).

In summary, VH&T's view that people who deceive themselves are lying can be challenged, and so can their view that people are quite good at detecting lies. People become better lie detectors by employing interview techniques aimed at strategically using the available evidence or by imposing cognitive load on the interviewees.

Authors' Response

Reflections on self-deception

doi:10.1017/S0140525X10003018

William von Hippel[a] and Robert Trivers[b]

[a]*School of Psychology, University of Queensland, St Lucia, QLD 4072, Australia;* [b]*Department of Anthropology, Rutgers University, New Brunswick, NJ 08901.*

billvh@psy.uq.edu.au
http://www.psy.uq.edu.au/directory/index.html?id=1159
trivers@rci.rutgers.edu
http://anthro.rutgers.edu/index.php?option=com_content&task=view&id=102&Itemid=136

Abstract: Commentators raised 10 major questions with regard to self-deception: Are dual representations necessary? Does self-deception serve intrapersonal goals? What forces shape self-deception? Are there cultural differences in self-deception? What is the self? Does self-deception have costs? How well do people detect deception? Are self-deceivers lying? Do cognitive processes account for seemingly motivational ones? And how is mental illness tied up with self-deception? We address these questions and conclude that none of them compel major modifications to our theory of self-deception, although many commentators provided helpful suggestions and observations.

R1. Dual representations are unnecessary

We begin our rejoinder with **Pinker**'s criticisms, because they cut to the heart of our proposal. Pinker's first point is that self-deception must involve dual representations, with truth and falsehood simultaneously stored. We disagree. An individual can self-deceive by failing to encode unwanted information in the first place. The research of Ditto and his colleagues (e.g., Ditto & Lopez 1992) provides the clearest example of this effect. By stopping their information search when the early returns were welcome (i.e., when the salivary test results suggested good health), participants in Ditto's experiments prevented themselves from ever encoding unwanted information that might have come along later. Thus, these individuals self-deceived without any representation of the truth.

Whereas **Pinker** proposes that the findings of Epley and Whitchurch (2008) cannot be taken as self-deception unless people can be shown to have accurate knowledge of their own appearance, we argue that this is unnecessary, even at an unconscious level. We further suspect that people's knowledge of their own biases is typically limited, and thus people might often be blissfully unaware – at any level – of the impact of their biased processing on their self-views. It should be noted, however,

that individuals who engage in biased encoding might occasionally have access to the possibility that unwanted information exists and that they avoided it. That is, they may have some representation of their own biased information gathering and its potential effect on the information they have in storage.

Smith agrees with **Pinker** that self-deception involves dual representations, but to Smith, this seems impossible. Smith then characterizes our mental dualism approach as constructing an imaginary internal homunculus, which he tells us has "numerous problems. . . not the least of which is its utter implausibility." Needless to say, constructing an internal homunculus was not part of our program, and what he thinks is impossible, we think are everyday events.

In his thoughts on deflationism, **Smith** confuses us with Mele (2001); he reintroduces a poorly defined distinction between wishful thinking and self-deception; and he ends by accusing us of believing that self-deception is both intentional and unintentional. If someone else managed to miss the point, let us state it clearly – we certainly believe that self-deception is intentional, in the sense that the organism itself intends to produce the bias, although the intention could be entirely unconscious. That is, humans have been favored by natural selection to self-deceive for its facilitative effect on the deception of others.

Bandura appears to agree with our argument that avoiding the truth is a form of self-deception and thus that the truth need not be harbored even in the unconscious mind. But he argues that the deceiving self must be aware of what the deceived self believes in order to concoct the self-deception. As the literature on biased processing shows, however, people can avoid unwanted information through a variety of biases that need only reflect the deceiving self's goals. As is the case with the orchid, natural selection does not require that these goals be available to conscious awareness. Thus, Bandura's dual representation of goals is also unnecessary for self-deception.

Finally, **Harnad** is also concerned about representation, but he appears to believe that our proposal includes the notion of self-deceivers as unconscious Darwinian robots. We are not sure what gave Harnad this idea. If he is arguing that a theory of self-deception would benefit from a better understanding of consciousness, we agree.

R2. Happiness, confidence, optimism, and guilt are interpersonal

Pinker's second point is that we have diluted the force of Trivers' (1976) original theorizing about self-deception by "applying it to phenomena such as happiness, optimism, confidence . . . in which the loop is strictly internal, with no outside party to be deceived." We disagree with his characterization of happiness, optimism, and confidence as strictly internal. Instead, we regard all three of these states as being of great *inter*personal importance, given that they signal positive qualities about the individual to others. Consider, for example, the effects of confidence on mating decisions. If you are considering entering into a relationship with a 25-year-old woman with low self-confidence, you may well reason (consciously or unconsciously) that she has known herself for 25 years and you have only known her for two weeks, so perhaps she is aware of important deficiencies that you have yet to discover. Similarly, an opponent's self-confidence is an important predictor of the outcome of an aggressive confrontation, and thus overconfidence can give a benefit to the degree that it induces self-doubt in the other or causes the other to retreat.

This possibility relates to **Suddendorf**'s analysis of the time course of the evolution of self-deception. Although he provides a compelling description of when self-deception to facilitate lying might have evolved, his analysis is limited to forms of self-deception that support deliberate and conscious efforts to manipulate the mental states of others. As noted earlier in this Response, self-deception should also underlie more general efforts to portray the self as better than it is, with overconfidence being the classic example. Because the benefits of overconfidence do not appear to be unique to human confrontations, coalition building, and mating efforts, it seems highly likely that self-deception is much older than our human lineage.

Despite these important disagreements with **Pinker**, it is worth returning to his central point about the importance of external parties. His eloquent dismissal of the logic of self-deception for purely internal purposes refutes the arguments proposed by **Egan** and **McKay, Mijović-Prelec, & Prelec** (**McKay et al.**) that self-deception could have evolved for its adaptive value in the absence of any interpersonal effects. But that does not mean that their suggestions cannot be resurrected and restated in interpersonal terms. For example, McKay et al. argue that self-deception is self-signaling intended to convince the self of its own morality and thereby eliminate the guilt associated with deception of others. We agree that this is an interpersonal benefit of self-deception, because inhibition of guilt appears to make people more successful in deceiving others (Karim et al. 2010).

Egan's motivational argument is similarly resurrected by **Suddendorf,** with the suggestion that foresight biases might have evolved because people often need to convince others to cooperate with their goals, and one way to persuade others is to self-deceive about the emotional impact of the outcome. This is an excellent suggestion that applies the interpersonal logic of Trivers' (1976) original idea regarding self-deception to the problem of self-motivation. If true, then affective forecasting errors should be greater when people must convince others to collaborate to pursue their goal than when the goal is a solitary pursuit. Finally, in a related vein, **Mercier** suggests that confirmation biases emerge not from self-deception but from the argumentative function of reasoning. But if reasoning evolved in part to serve an argumentative and thus persuasive function, we are brought back to our original proposal that people bias their information processing to more successfully convince others of a self-favorable view of the world.

R3. Self-deception is bounded by plausibility

We agree with **McKay et al.** that people will attempt to self-deceive but sometimes fail. We also agree that such failures will not stop people from trying again, but failures should guide their future efforts. In this sense, we are in

complete agreement with **Frey & Voland**'s proposal that self-deceptions are negotiated with the world; those that are challenged and shown to be false are no longer be believed by the self, and those that are accepted by others continue to be accepted by the self. This idea resonates with **Brooks & Swann**'s identity negotiation process, and it is a likely route by which self-deception might achieve the proper dosage that we alluded to in the target article. But it is important to note that by engaging in a modicum of self-deception *during* identity negotiation, people are likely to enhance their role in the relationship and the benefits they gain from it. The end result of such a negotiation is that people self-deceive to the degree that others believe.

Frey & Voland go on to argue for a model of negotiated self-deception that appears to be consistent with our proposal (see also **Buss). Lu & Chang** extend their arguments by suggesting that self-deception should be sensitive to probabilities and costs of detection. Thus, people might self-deceive more when they think they have a more difficult interpersonal deception to achieve.

If self-deception is sensitive to interpersonal opportunities and pitfalls, we are brought to **Gangestad**'s important point that this too is a co-evolutionary struggle, as people should be selected to resist the self-deception of others. Gangestad then asks whether self-deception might be most successful among those who have the most positive qualities and thus are the most believable when they self-enhance. In psychology we tend to think of individuals with secure high self-esteem as those who are comfortable with themselves, warts and all (Kohut 1977). But Gangestad's suggestion raises the possibility that secure high self-esteem might actually be a mixture of capability and self-deception. This perspective suggests that implicit self-esteem might correlate with self-enhancement in the Epley–Whitchurch paradigm because those who receive enough positive feedback to enable high implicit self-esteem may be best placed to convince others that they are better looking than they really are. That is, other aspects of their personality or capabilities (or perhaps even their physical attractiveness) might cause Epley and Whitchurch's self-enhancers to be highly regarded, and because they are highly regarded, other individuals are less likely to challenge their behavior when they act as if they are even more attractive than they are.

Frey & Voland attack this interpersonal story in their claim that costly signaling theory weakens the case for self-deceptive self-enhancement. Although we agree that costly signaling theory explains why perceivers place a premium on signals that are difficult to fake (e.g., honest signals of male quality such as physical size or symmetry), it does not follow from costly signaling theory that perceivers ignore signals that can sometimes be faked. Nor does it follow that people and other animals do not try to fake such signals when possible. One can thus infer that signals that can be faked – and are thereby viewed by receivers with a more jaundiced eye – will be more readily believed by others if they are also believed by the self. In this manner, self-deception can be accommodated within costly signaling theory (see also **Gangestad**).

If self-deception is negotiated with others, then one important outcome of this negotiation is that self-deceptions are likely to be plausible. This plausibility constraint has a number of implications, the first of which concerns **Brooks & Swann**'s point that although the benefits we ascribe to confidence may be accurate, that does not mean that they also apply to overconfidence. Although Brooks & Swann are certainly correct – in the sense that overconfidence has attendant costs that do not accompany confidence – it is also the case that so long as it is not dosed too liberally, overconfidence should be difficult to discriminate from confidence and thus should give people an advantage in addition to their justified confidence.

Plausibility constraints also address **Brooks & Swann**'s second argument that self-enhancement plays only a modest role in social interaction, in which they point to a meta-analysis that suggests that self-verification overrides self-enhancement. As a plausibility perspective makes apparent, this finding is not an argument against self-deception, but rather is consistent with the idea that the benefits of self-deception are dose-dependant. Reality is vitally important, and people ignore it at their peril. Thus, self-deception will be most effective when it represents small and believable deviations from reality. A well-tuned self-deceptive organism would likely be one that biases reality by, say, 20% in the favored direction (see Epley & Whitchurch 2008). Thus, self-verification strivings (i.e., an accuracy orientation) would account for 80% of the variance in people's understanding of the world and should thereby be at least as strong if not much stronger than self-enhancement strivings when strength is interpreted as variance accounted for. But if strength is interpreted as desire to know favorable information versus desire to know reality, then this desire should fluctuate as a function of current needs and opportunities.

Plausibility is also relevant to the suggestion made by **Brooks & Swann**, **Bandura**, and **Dunning** that infrequent self-deception and consequent deception of others may well be useful, but that excessive deception is likely to result in discovery and rejection. Discovery and rejection seem highly likely if people rely too regularly on deception as an interpersonal strategy; the threat of rejection from group members is one of the evolutionary pressures for telling the truth.

Dunning then conflates excessive boasting with self-deceptive self-enhancement, again overlooking plausibility constraints. This leads Dunning to conclude that self-deception might be more useful when we regularly interact with lots of strangers. Although this argument may hold for those who rely excessively on deception and self-enhancement, for those who practice deception in moderation, self-deception ought to have facilitated deception and self-enhancement even in (perhaps especially in) long-term, stable small groups. Thus, the fact that people evolved in small interdependent bands may be all the more reason for individuals to self-deceive on those occasions when they choose to or need to deceive others.

We see further perils of ignoring plausibility in the tacit assumption shared by some of our commentators that all deception must be accompanied by self-deception. For example, **Vrij** argues that the self-deceptive biases we describe in our proposal account for only a small portion of the vast number of lies that people tell in their everyday lives. Although that may well be true, there is an enormous ascertainment bias: We are much more aware of our

conscious deceptions than we are of our unconscious deceptions. More importantly, the existence of self-deception as an interpersonal strategy does not preclude deliberate deception. Plausibility constraints ensure that not all deception is capable of benefiting from self-deception.

In this manner, plausibility constraints also inform future research. For example, **Buss** describes a series of interesting hypotheses about self-deception within families and between the sexes regarding one's feelings. Deceptions about internal qualities such as feelings are difficult to disprove and thus seem likely candidates for self-deception. In contrast, the personal ad study that Buss describes, with men overestimating their height and women underestimating their weight, seems like a less plausible candidate for self-deception. Although men and women might believe their own exaggerations to some degree, it seems highly unlikely that they are self-deceived about the full extent of their claims, given the ready availability of contradictory evidence. These sorts of exaggerations are likely to be rapidly dismissed when people begin to negotiate an actual relationship, and thus initial deception of others in the absence of self-deception is likely to be more common in cases such as these.

Humphrey argues that an important cost to self-deception is loss of insight into the deceit of others. He suggests that when we deceive, we learn that others do the same, and if we self-deceive, we lose that learning opportunity. We agree that this is a cost but one that is limited by the percentage of our deception accompanied by self-deception. Plausibility constraints ensure that most of us tell plenty of deliberate lies on which we can later reflect. Humphrey goes on to note that Machiavellianism might be considered the flip-side of self-deception. This is an interesting suggestion and raises the testable hypothesis that the more one adopts either of these strategies, the less likely one is to adopt the other. This too is relevant to plausibility, given that different abilities and proclivities will make people differentially successful when using these potentially competing strategies.

Mercier notes that because lies are often detected by a lack of internal consistency, self-deception would facilitate lie detection rather than other deception, insofar as self-deceivers could no longer maintain internal consistency. But it is easy to imagine how plausibility constraints produce the necessary consistency in self-deception. Indeed, the lies that we tell others that are based on biased information processing may be just as internally consistent as the lies we tell knowingly, maybe even more so, because we do not need to worry about mixing truth and lies when we self-deceive. For example, if I bias my information gathering in the manner described by Ditto and Lopez (1992), then all the information at my disposal is internally consistent and supportive of the fact that I do not have a potential pancreatic disorder. Unbiased information gathering would have only put me in the potentially awkward position of learning that I do have the potential for the disorder, and then being forced to cover up that information.

Although the need to believe one's own self-deceptions increases the likelihood that they are internally consistent, it does not follow that we are our own fiercest critics, as **Fridland** suggests. That is, our ability to detect deception in others does not necessarily translate into detection of deception in ourselves. Furthermore, it does not follow that if we are better at detecting lies in close others, then we should be best in detecting them in ourselves. The flaw in Fridland's reasoning is that the motivation to detect deception is opposite in self versus other deception; we are strongly motivated to detect the latter but not the former. Indeed, the world is replete with individuals who are intolerant of behaviors in others that they excuse in themselves (Batson et al. 1997; Valdesolo & DeSteno 2008). By Fridland's logic, such hypocrisy should be impossible. Likewise, we do not see the self as simply the endpoint of a continuum of familiarity from strangers to close friends to the self – rather, the self is qualitatively different.

R4. Cultural differences are only skin deep

Heine disagrees with our claim that self-enhancement is pan-cultural, cites his meta-analyses that support his position that there is no self-enhancement in East Asian cultures, and dismisses the evidence and meta-analyses that are inconsistent with his perspective. Most of these arguments are a rehash of his debate with Sedikides, Brown, and others, and because these authors have already addressed the details of Heine's claims, we refer interested readers to their thorough rebuttals rather than devote the necessary ink here (e.g., see Brown 2010; Sedikides & Alicke, in press; Sedikides & Gregg 2008). It is important to note, however, that Heine has missed the bigger picture regarding the purpose of self-enhancement. From an evolutionary perspective, the critical issue is not that self-enhancement is intended to make the self feel better about the self, but rather that it is intended to convince others that the self is better than it really is. Self-enhancement is important because better selves receive more benefits than worse selves. Better selves are leaders, sexual partners, and winners, and worse selves are followers, loners, and losers, with all the social and material consequences of such statuses. Because East Asians also win friends, mates, and conflicts by being better than their rivals, we expect that they will gain an edge in coalition building, mating, and fighting by self-enhancing (just as Westerners do). But because different cultures have different rules about what it means to be moral and efficacious, as well as different rules about how best to communicate that information, it also follows that there should be cultural variation in self-enhancement.

By virtue of their collectivism, cultures in East Asia place a premium on harmony and fitting in with others, and thus modesty is an important virtue. As a consequence, East Asians are far less likely than individualist Westerners to claim to be great or to act in ways that suggest they believe they are great, given that immodesty itself provides direct evidence in East Asia that they are not so great after all. The importance of modesty in collectivist cultures raises the possibility that East Asians may self-enhance by appearing to self-denigrate – by exaggerating their modesty. That is, humble claims by East Asians could be made in service of self-enhancement. Consistent with this possibility, Cai et al. (2011) found that among Chinese participants, dispositional modesty was negatively correlated with explicit self-esteem but positively correlated with implicit self-esteem (measured via the IAT

[Implicit Association Test] and preference for one's own name). In contrast, among North American participants, dispositional modesty was negatively correlated with explicit self-esteem and uncorrelated with implicit self-esteem. Indeed, when Cai et al. (2011) instructed Chinese and North American participants to rate themselves in either a modest or immodest fashion, they found that immodest self-ratings reduced implicit self-esteem and modest self-ratings raised implicit self-esteem among Chinese participants but had no effect on North American participants. These results provide evidence that modesty can itself be self-enhancing, and just as important, that modesty demands will by necessity minimize explicit self-enhancement in East Asian cultures.

Nevertheless, these results do not provide clear evidence that East Asians believe they are better than they are (as we claim they should), because implicit measures are not well suited for demonstrating such a possibility. The Epley–Whitchurch (2008) paradigm is well suited for demonstrating that people believe they are better than they are. We expect that the Epley–Whitchurch paradigm would reveal self-enhancement just as clearly in East Asia as it does in the West. Furthermore, particularly strong evidence for our argument regarding the impact of modesty on claims versus beliefs would emerge if Easterners chose their actual or even an uglified self when asked to find their self in an array of their own real and morphed faces (Epley & Whitchurch, 2008, Study 1) but nevertheless found their attractive self more quickly in an array of other people's faces (Epley & Whitchurch, 2008, Study 2). Such a pattern of results would speak to the value of "claiming down" while "believing up" – or the co-occurrence of modesty and self-enhancement in collectivist cultures. Despite the allure of such a finding, it may be the case that the Epley–Whitchurch paradigm is too subtle for most people to use to demonstrate modesty, and thus East Asians may show self-enhancement in both explicit choices and reaction times. **Heine** would presumably predict that self-enhancement would not emerge in East Asia with either of these measures.

In a similar vein, **Egan** suggests that if self-deception serves deception of others, then people should be particularly likely to deceive themselves about moral issues, because getting others to believe one is more moral would cause others to "lower their guard." In apparent contrast to this prediction, Egan notes that Balcetis et al. (2008) found that collectivists self-enhance less in moral domains than do individualists. Unfortunately, Balcetis et al.'s study does not address Egan's point. Balcetis et al. asked people to make judgments regarding moral behaviors in which they might engage, and they found that collectivists were less likely than individualists to overestimate their tendency to engage in moral behaviors. But what does this finding mean with regard to self-deception? Are collectivists simply self-enhancing less on overt measures, as has been found many times in many domains (see **Heine**)? Or are collectivists more attuned to their own interpersonal behavior, and thereby more aware of their actual likelihood of engaging in a variety of interpersonal moral acts? It is unclear from the Balcetis et al. studies whether collectivism is truly associated with self-deceptive self-enhancement in moral domains, nor is it clear what such a finding would mean with regard to the evolution of self-deception.

R5. There is a central self and it's a big one

Guided by data from cognitive and social psychology that indicate that self-knowledge is too vast to load into working memory, **Kenrick & White** suggest that content domains determine which subselves are activated and guide information processing. This view leads Kenrick & White to redefine self-deception as *selectivity*. Although we agree that only certain aspects of the self are activated at any one time, we disagree with their proposed implication of this issue for self-deception. As Kenrick & White note, selectivity involves gathering, attending to, and remembering only that which is important to the active subself (or working self-concept; Markus & Wurf 1987). But people do not simply gather, label, and remember that which is selectively *important* to the active subself. Rather, people also gather, label, and remember information that is *biased* in favor of their current goals. Good news and bad news are equally important to the active subself, but self-deception selectively targets the good news – the better to persuade others that the state of the world is consistent with the goals of the self-deceiver.

In response to a similar set of concerns, **Kurzban** attempts to sidestep the problem of self-deception by deleting the concept of the "self" in favor of multiple, independent mental modules. In our opinion this is an error, as data from social psychology and cognitive neuroscience suggest otherwise. For example, the finding that brain regions involved in executive control can inhibit the activity of other brain regions (e.g., Anderson et al. 2004) suggests that there is a central self and that this central self has (limited) access to and control of information processing.

Kurzban also wants to avoid metaphors such as level of consciousness, and he argues instead that modularity provides an easy solution to the problem of self-deception. On the surface these arguments seem compelling, given that self-deception in a modular system seems simple, almost inevitable. Problems emerge, however, when we take the modularity metaphor seriously. If we accept the idea that the mind is composed of isolated modules, we are led to a question similar to that raised by **Bandura**: Which module directs behavior when two (or more) modules are equally relevant? Without a central self that controls, activates, and inhibits other systems, modules would continually be captured by external events in a manner that would disrupt sustained goal pursuit.

Perhaps more importantly, if a modular system shows *functional specificity* and *information encapsulation* (Kurzban & Aktipis 2006), why do systems that should logically be distinct leak into each other? For example, why does hand washing reduce cognitive dissonance (Lee & Schwarz 2010); why does touching a hard object make people more rigid in negotiations (Ackerman et al. 2010); why does holding a warm cup make someone else seem friendlier (Williams & Bargh 2008); and why do people who were excluded feel physically cold (Zhong & Leonardelli 2008)? Embodiment research demonstrates that modules, if they exist, are neither functionally distinct nor autonomous. Furthermore, the notion of levels of consciousness might be less metaphorical than the notion of modularity, given that research dating back to Ebbinghaus (1885) has shown that some information can be

consciously recollected, other information can be recognized as accurate even though it was not recollected, other information leaves a residue in consciousness even if the information itself is unconscious (e.g., the sense of familiarity – Jacoby 1991), and still other information is entirely unconscious but nevertheless drives behavior in a manner that is also outside of conscious awareness (Kolers 1976). Thus, the concept of modules allows us to escape neither the concept of the self nor levels of consciousness, leaving us to conclude that modularity does not provide an easy solution to the problem of self-deception after all.

Nevertheless, the existence of subselves can lead to competing goals at different levels of consciousness. This possibility leads **Huang & Bargh** to argue that unconscious pursuit of goals that are inconsistent with conscious desires is a form of self-deception. We agree that these dual systems of goal pursuit enable self-deception, and some of the examples they cite support such an interpretation. We disagree, however, that such deviations from conscious behavioral control are necessarily varieties of self-deception. Rather, these inconsistencies are evidence of competing goals that may or may not involve self-deception. For example, if a person previously held a certain attitude and then was persuaded to the opposite position, the original attitude might remain in an unconscious form and might continue to influence some types of goal-directed behavior (Jarvis 1998). This sort of slippage in the mental system facilitates self-deception, but it can emerge for non-motivational reasons as well, such as habit.

Although there is now substantial evidence for such dissociations between conscious and unconscious processes, **Bandura** argues that interconnections between brain structures suggest that mental dualisms of this sort are unlikely. It does not follow from rich neural interconnections, however, that people have conscious access to information that is processed outside of awareness. Conscious access is clearly limited, although as noted earlier, there is commerce between conscious and unconscious mind. Bandura goes on to question how the contradicted mind can produce coherent action. The Son Hing et al. (2008) experiment provides an answer to this question: This study shows that competing conscious and unconscious attitudes influence behavior under predictable circumstances (see also Hofmann et al. 2009). It should be kept in mind, however, that attitudinal ambivalence can also be reflected in consciousness, and thus people often knowingly behave in self-contradictory ways.

R6. Self-deception has costs

Funder raises two important issues. First, he asks why people would ever self-deceive in a downward direction. Despite the fact that most people self-enhance, some people self-diminish. If we accept the possibility that these individuals are as likely as self-enhancers to believe their self-diminishing claims, the question emerges whether self-diminishment may be favored by natural selection, and if so, why?

We believe that there are two major sources to "deceiving down." On the one hand, it is sometimes directly adaptive. In herring gulls and various other seabirds, offspring actively diminish their apparent size and degree of aggressiveness at fledging so they will be permitted to remain near their parents, thereby consuming more parental investment. In many species of fish, frogs, and insects, males diminish apparent size, color, and aggressiveness to resemble females and steal paternity over eggs (see Trivers 1985). These findings indicate that deceiving down can be a viable strategy in other species, and thus likely in humans as well, which should lead to self-deceptive self-diminishment. An important question would be to identify domains in which different types of people gain by self-diminishing.

On the other hand, people may also be socialized or otherwise taught that they are less capable, moral, or worthy than they really are. If acceptance of this negative message leads to self-diminishment biases, then perhaps such individuals' self-views represent an imposed variety of self-deception, whereby the individual self-deceives in a direction that benefits someone else, such as a parent, spouse, or dominant group (Trivers 2009). If so, this would appear to be an important cost to self-deception that may be borne by a substantial percentage of the population.

Related to this issue, **Funder**'s second point is that we make only passing reference to the costs of self-deception and focus almost exclusively on the associated gains. We plead guilty to this charge, because our goal was to establish why self-deception might have evolved, and that goal required attention to possible benefits. The costs of self-deception – particularly those regarding loss of information – seem apparent and real, and thus it is important to establish the benefits that select for self-deception in the first place. Of course, any mature evolutionary theory of the subject must be based on a cost/benefit analysis, and we acknowledge many of the costs that have been suggested.

For example, **Mercier** suggests that if we do not know that we lied, then we cannot apologize and make amends (see also **Khalil**). This is true, and it is a cost of self-deception. However, this cost must be weighed against the gain achieved by appearing less culpable when lying. Ignorance is typically a lesser crime than duplicity, but an apology may sometimes offset or even outweigh that benefit.

A different type of cost is raised by **Johansson, Hall, & Gärdenfors** (**Johansson et al.**), who suggest that people might accidentally convince themselves in their self-deceptive efforts to convince others, and thus they might not retain any representation of the truth. In support of this possibility, Johansson et al. discuss choice blindness, or the frequent inability of people to notice that their preferred option was switched after they made a choice. In their experiments on choice blindness, Johansson et al. found that people justified their non-chosen outcome just as vociferously as their chosen one, which suggests that they do not know why they chose as they did. Most intriguingly, Johansson et al. also describe evidence that these justified non-choices later become preferred options, suggesting that people have convinced themselves of their new preferences.

We agree that this is a possible cost and see it as another example of why simultaneous representation of truth and falsehood is not necessary for self-deception. Nevertheless, there may be some types of self-deception where the truth is lost and other types where the truth is retained and can later be accessed when the deceptive deed is done

(**Lu & Chang**). This is a provocative possibility, and there is evidence to suggest that such a system might operate at least some of the time. For example, Zaragoza and colleagues (e.g., McCloskey & Zaragoza 1985) have found that suggestive probes will cause people to recall false information in line with the suggested information. When individuals are later given a recognition task, however, they remain capable of recognizing the accurate information that they were originally presented. Thus, the possibility remains that memory might be distorted in favor of a self-deceptive goal and then revert back to reality when the deception ends, which would minimize the costs outlined by **Johansson et al**.

Frankish describes a cost similar to that of **Johansson et al.** in his argument that people often accept a claim as true for the purpose of exploring it or taking a certain perspective, despite knowing that it is false. Frankish proposes that this system has some slippage, as people can come to believe what they originally only accepted, perhaps by observing their own behavior. Acceptance appears to be another avenue by which people could end up self-deceiving in their efforts to deceive others. Particularly if it proves to be the case that people are more likely to adopt an accepting role when faced with congenial arguments, Frankish's suggestion would seem to provide another route for self-deception in service of offensive goals.

The notion of costs takes a different turn in the arguments put forward by **Khalil**, who tells us that in claiming that self-deception has evolved, we are implicitly assuming that it is "optimal." Unfortunately, he does not tell us what he means by this term or why we should be tagged with believing it. Apparently our argument that self-deception may be a tactic in deceiving others commits us to the notion that self-deception is optimal. We know of no possible meaning of "optimal" for which this is true, but we are not trained in economics. In biology and psychology, we do not believe that evolution generates optimality very often (if at all). Nor do we believe that self-deception is optimal, in the sense of not being able to be improved upon. Khalil argues that if he can show that self-deception is suboptimal, then our theory collapses, but he offers no logic to support this claim. Economics apparently produces theorems claiming that suboptimal solutions will be swept aside by economic forces, but we doubt the validity of such theorems even within economics, much less evolutionary biology and psychology.

We do agree that Adam Smith had some very insightful things to say about self-deception, that open confession and plea for forgiveness may often be preferable to self-deception, and that self-deception can sometimes be costly. But unlike **Khalil**, we believe that selection can favor individuals who fight their tendency to self-deceive in some circumstances while at the same time practicing it in others.

R7. It remains unclear how well people detect deception

Vrij's major point is that people are in fact poor lie detectors, and he claims that this conclusion is supported by the various experiments that he cites. We examine the details of this claim in the following paragraphs, but let us note at the outset that every one of the studies Vrij cites (as well as those cited by **Dunning** on this point) suffers from the same methodological limitations we discuss in our target article. Thus, none of the additional data that are raised address the criticisms made in the target article.

First, **Vrij** argues that experiments advantage detectors rather than deceivers because detectors know they might be lied to. Although this is true, most situations that motivate lying in everyday life also raise awareness that one might be lied to (as **Humphrey** notes). Second, Vrij argues that intimacy increases perceptions of truthfulness. Although this appears to be true much of the time, it does not follow that intimacy decreases accuracy in detection of deception. Rather, intimacy could lead to a stronger perceived likelihood of truthfulness (beta in signal detection terms) and simultaneously better discriminability (d′). The lack of evidence for better lie detection in friends over strangers described by Vrij speaks more to the methods used than to the answer to this question, because these studies suffer from the same limitations as the ones we reviewed. Third, the Hartwig et al. (2006) study raised by Vrij is interesting, but it does not speak to the questions that concern us about whether cross-examination helps people detect deception. Again, the lies told in Hartwig et al. were trivial, and thus cross-examination will not necessarily increase accuracy. Fourth, Vrij's own research on enhancing cognitive load is also interesting, but it is easy to see how his work supports our position that cross-examination helps detectors, because cross-examination also enhances cognitive load on the deceiver. The fact that Vrij can demonstrate effects of cognitive load with trivial lies provides further evidence that lie detection remains an important threat to liars. Fifth, Vrij raises the fact that Park et al.'s (2002) research shows that detection of lies in daily life relies almost entirely on third parties and physical evidence, and not on nonverbal behavior. As Vrij notes, however, 14% of the lies reported in this research were detected via confessions. Importantly, these were *solicited* confessions based on confrontations about suspicious behavior and the like. Additionally, the methodology in this study was based on recall, which is not only faulty but also influences the type of lies people may choose to disclose to the experimenters. Last, participants in the Park et al. study were simply asked about a recent lie they discovered, and thus as with most previous research, many of the lies they detected were likely to be trivial. Selection pressure should be heaviest on important lies, and thus we reiterate our claim that research is necessary to examine the detection of important lies in vivo.

Finally, **Dunning** raises a related possibility that people might do a better job deceiving if they do not care about the truth. This argument is plausible for deceptions that have few if any consequences for the deceiver (and we agree that "bullshitting" is a case in point), but the more important the deception is, the more people are forced to care about the truth because its discovery will cause them harm. Thus, this possibility seems unlikely for important deceptions.

R8. Self-deceivers are liars, whether they know it or not

Fridland and **Vrij** both argue that self-deceivers are not lying (although Fridland's claim is stronger, in that she

states that they are not even deceiving). This is true in the strictest sense of what it means to lie, but untrue once we understand that deception of others is the motivation for self-deception. For example, imagine I want to convince you that your spouse was not with my best friend while you were out of town. Imagine further that I have an acquaintance who mentions that he saw your spouse at 3:00 p.m. in the hair salon and at midnight in a bar. If I choose not to ask my acquaintance whom your spouse was with, or if I only ask my acquaintance whom she was with in the hair salon and avoid asking the more probative question of whom she was with in the bar, then I am lying when I later tell you that to the best of my knowledge she was not with my friend. Strictly speaking, what I am telling you is true. But the lie occurred when I initially gathered information in a biased manner that served my goal of convincing you of my friend's innocence regardless of what the truth might be.

R9. Motivation can drive cognition

Kramer & Bressan make the interesting suggestion that belief in God might be an unintended consequence of efficient cognitive processes rather than evidence of self-deceptive processes that impose order and a sense of control on the world. Kramer & Bressan suggest that people who have stronger schemas are more likely to have their attention captured by schema-violating events, with the end result that they attribute supernatural importance to what are in essence coincidences. But what if motivation drives cognition? What if people who have a stronger than average motivation to see order in their universe (perhaps because they feel insufficiently resourceful to cope in the absence of order and control) are more likely to establish strong schemas when they receive any support for those schemas from the environment? Such strong schemas then provide them with the order that they desire.

From **Kramer & Bressan**'s data we do not know why people have strong schemas or not – we only know that there are individual differences. If motivation influences the establishment of strong schemas, then Kramer & Bressan's data could be construed as support for our self-deception argument, because the people who crave order in the universe are more likely to see schema violations as meaningful events. The same argument holds for their description of the Amodio et al. (2007) study in which strong schemas were associated with the desire for strong government. Indeed, liberals have greater tolerance of ambiguity than conservatives (Jost et al. 2007), which again suggests that motivation might drive schema strength rather than the other way around. In contrast to Kramer & Bressan's claim, schema *strength* is not evidence of "efficient memory and attentional processes"; rather, it is the flexible use of schemas in information processing (i.e., assimilation *and* accommodation) that enables efficient memory and attention. Kramer & Bressan conclude by noting that illusions and conspiracy beliefs could also result from reduced motivation to accurately analyze the situation among those who are unlikely to gain control. In contrast to this possibility, people with a low sense of control typically engage in effortful (but often unsuccessful) struggles to regain control and comprehension (e.g., Bransford & Johnson 1973; Weary et al. 2010).

In contrast to the "accidental cognitive by-product" arguments of **Kramer & Bressan**, **Gorelik & Shackelford** suggest that religion and nationalism might be considered an imposed variety of self-deception. Their proposal extends the work of Kay et al. (2008) and Norris and Inglehart (2004) by considering various ways that self-deception might intertwine with religious and nationalistic practice.

Mercier makes a more general version of **Kramer & Bressan**'s argument by claiming that we need not invoke self-deception to explain self-enhancement; rather, it could be the result of an error management process that favors less costly errors. However, to support this argument he is forced to dismiss the cognitive load and self-affirmation findings, which he does by claiming that these manipulations reduce the tendency to engage in reasoning. Although cognitive load certainly disrupts reasoning, self-affirmation does not. For example, self-affirmation reduces some types of reasoning (i.e., self-image maintenance) while increasing other types (e.g., consideration of unwelcome information; for a review, see Sherman & Cohen 2006). Indeed, the different mechanisms but similar consequences of self-affirmation and cognitive load make up one of the reasons why they are so powerful in combination – one is motivational, and the other cognitive.

R10. Mental illness is associated with too little and too much self-deception

Preti & Miotto suggest that people with mental illness self-deceive less than mentally healthy people. We would agree that this is often the case (see Taylor & Brown 1988). Preti & Miotto conclude by noting that mirror systems might enable self-deceivers to detect deception better in others, at least when that deception is accompanied by self-deception. This is an intriguing prediction that also sets boundary conditions on **Humphrey**'s proposal that self-deception might inhibit detection of deception in others.

Troisi argues that somatoform disorders are often a form of self-deception, and he describes an experiment by Merckelbach et al. (2010) in which people were asked to explain why they had endorsed a symptom that they had not in fact endorsed. Over half of the participants did not notice the switch and showed evidence that they now believed they had the symptom. Similarly, people who had earlier feigned a symptom showed evidence that they now believed they had it. These findings seem similar to the choice blindness studies of **Johansson et al**. in which people also unintentionally deceived themselves. Such somatoform disorders could be regarded as a cost to a system that allows self-deception – if the mind evolved to allow people to be able to convince themselves of desired outcomes, then it might also allow people to convince themselves of unintended or unwanted beliefs. As Troisi notes, however, somatoform disorders can be regarded as an interpersonal strategy intended to elicit sympathy and care.

R11. Conclusions

Our goal has been to show that a comprehensive theory of self-deception can be built on evolutionary theory and

social psychology. Our commentators have raised areas where the data need to be strengthened, have noted alternative hypotheses, and have disagreed with us about the central tenets of our theory. They have also suggested refinements in logic and interpretation of the evidence and have made novel predictions. None of the empirical or conceptual challenges strikes us as fatal to the theory, and so it remains for future research to assess the merit of our ideas and how best to extend them. Call it self-deception if you will, but we think that our enterprise has passed the first test.

References

[The letters "a" and "r" before author's initials stand for target article and response references, respectively]

Ackerman, J. M., Becker, D. V., Mortensen, C. R., Sasaki, T., Neuberg, S. L. & Kenrick, D. T. (2009) A pox on the mind: Disjunction of attention and memory in the processing of physical disfigurement. *Journal of Experimental Social Psychology* 45:478–85. [DTK]

Ackerman, J. M., Nocera, C. C. & Bargh, J. A. (2010) Incidental haptic sensations influence social judgments. *Science* 328:1712–75. [rWvH]

Ackerman, J. M., Shapiro, J. R., Neuberg, S. L., Kenrick, D. T., Becker, D. V., Griskevicius, V., Maner, J. K. & Schaller, M. (2006) They all look the same to me (unless they are angry): From out-group homogeneity to out-group heterogeneity. *Psychological Science* 17:836–40. [DTK]

Addis, D. R., Wong, A. T. & Schacter, D. L. (2007) Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. *Neuropsychologia* 45:1363–77. [TS]

Albarracín, D. & Mitchell, A. L. (2004) The role of defensive confidence in preference for proattitudinal information: How believing that one is strong can sometimes be a defensive weakness. *Personality and Social Psychology Bulletin* 30:1565–84. [aWvH]

Alicke, M. D. & Sedikides, C. (2009) Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology* 20:1–48. [aWvH]

Amodio, D. M., Jost, J. T., Master, S. L. & Yee, C. M. (2007) Neurocognitive correlates of liberalism and conservatism. *Nature Neuroscience* 10:1246–47. [PK, rWvH]

Anderson, D. E., Ansfield, M. E. & DePaulo, B. M. (1999) Love's best habit: Deception in the context of relationships. In: *The social context of nonverbal behaviour*, ed. P. Philippot, R. S. Feldman & E. J. Coats, pp. 372–409. Cambridge University Press. [AV]

Anderson, D. E., DePaulo, B. M. & Ansfield, M. E. (2002) The development of deception detection skill: A longitudinal study of same-sex friends. *Personality and Social Psychology Bulletin* 28:536–45. [DD, AV, aWvH]

Anderson, M. C., Ochsner, K. N., Kuhl, B., Cooper, J., Robertson, E., Gabrieli, S. W., Glover, G. H. & Gabrieli, J. D. E. (2004) Neural systems underlying the suppression of unwanted memories. *Science* 303:232–35. [rWvH]

Ariely, D. & Norton, M. I. (2008) How actions create – not just reveal – preferences. *Trends in Cognitive Sciences* 12:13–16. [PJ]

Armitage, C. J., Harris, P. R., Hepton, G. & Napper L. (2008) Self-affirmation increases acceptance of health-risk information among UK adult smokers with low socioeconomic status. *Psychology of Addictive Behaviors* 22:88–95. [aWvH]

Armor, D. A. & Taylor, S. E. (1998) Situated optimism: Specific outcome expectations and self-regulation. In: *Advances in experimental social psychology, vol. 30*, ed. M. P. Zanna, pp. 309–79. Academic Press. [aWvH]

Aspinwall, L. G. & Richter, L. (1999) Optimism and self-mastery predict more rapid disengagement from unsolvable tasks in presence of alternatives. *Motivation and Emotion* 23:221–45. [LCE]

Assad, K. K., Donnellan, M. B. & Conger, R. D. (2007) Optimism: An enduring resource for romantic relationships. *Journal of Personality and Social Psychology* 93:285–97. [aWvH]

Babad, E. (1997) Wishful thinking among voters: Motivational and cognitive influences. *International Journal of Public Opinion Research* 9:105–25. [LCE]

Babcock, L. & Loewenstein, G. (1997) Explaining bargaining impasse: The role of self-serving biases. *Journal of Economic Perspectives* 11(1):109–26. [ELK]

Baker, R. R. & Bellis, M. A. (1993) Human sperm competition: Ejaculate adjustment by males and the function of masturbation. *Animal Behaviour* 46861–85. [aWvH]

Balcetis, E. (2008) Where the motivation resides and self-deception hides: How motivated cognition accomplishes self-deception. *Social and Personality Psychology Compass* 2(1):361–81. [HJL]

Balcetis, E. & Dunning, D. (2010) Wishful seeing: Desired objects are seen as closer. *Psychological Science* 21:147–52. [LCE, JYH]

Balcetis, E., Dunning, D. & Miller, R. L. (2008) Do collectivists know themselves better than individualists? Cross-cultural studies of the holier than thou phenomenon. *Journal of Personality and Social Psychology* 95:1252–67. [LCE, rWvH]

Bandura, A. (1982) Self-efficacy mechanism in human agency. *American Psychologist* 37:122–47. [MLB]

Bandura, A. (1986) *Social foundations of thought and action: A social cognitive theory.* Prentice-Hall. [AB]

Bandura, A. (1989) Human agency in social cognitive theory. *American Psychologist* 44:1175–84. [LCE]

Bandura, A. (1999) Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review* 3:193–209. [AB]

Bandura, A. (2008) The reconstrual of "free will" from the agentic perspective of social cognitive theory. In: *Are we free? Psychology and free will*, ed. J. Baer, J. C. Kaufman & R. F. Baumeister, pp. 86–127. Oxford University Press. [AB]

Bargh, J. A. (1994) The four horsemen of automaticity: Awareness, efficiency, intention, and control in social cognition. In: *Handbook of social cognition*, 2nd ed., ed. R. S. Wyer, Jr. & T. K. Srull, pp. 1–40. Erlbaum. [aWvH]

Bargh, J. A., Green, M. & Fitzsimons, G. (2008) The selfish goal: Unintended consequences of intended goal pursuits. *Social Cognition* 26:520–40. [JYH]

Bargh, J. A. & Huang, J. Y. (2009) The selfish goal. In: *The psychology of goals*, ed. G. B. Moskowitz & H. Grant, pp. 127–50. Guilford Press. [JYH]

Bargh, J. A. & Morsella, E. (2008) The unconscious mind. *Perspectives on Psychological Science* 3:73–79. [JYH]

Barrett, H. C. (2005) Enzymatic computation and cognitive modularity. *Mind and Language* 20:259–87. [RK]

Barrett, H. C. & Kurzban, R. (2006) Modularity in cognition: Framing the debate. *Psychological Review* 113:628–47. [RK]

Bartlett, F. C. (1932) *Remembering: A study in experimental and social psychology.* Cambridge University Press. [PK, TS]

Batson, C. D., Kobrynowicz, D., Dinnerstein, J. L., Kampf, H. C. & Wilson, A. D. (1997) In a very different voice: Unmasking moral hypocrisy. *Journal of Personality and Social Psychology* 72:1335–48. [rWvH]

Batson, C. D. & Thompson, E. R. (2001) Why don't moral people act morally? Motivational considerations. *Current Directions in Psychological Science* 10:54–57. [LCE]

Bem, D. J. (1967) Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review* 74:183–200. [PJ]

Bodner, R. & Prelec, D. (2003) Self-signaling in a neo-Calvinist model of everyday decision making. In: *Psychology of economic decisions, vol. I*, ed. I. Brocas & J. Carillo. Oxford University Press. [RM]

Boehm, J. K. & Lyubomirsky, S. (2008) Does happiness promote career success? *Journal of Career Assessment* 16:101–16. [aWvH]

Bok, S. (1980) The self deceived. *Social Science Information* 19:923–36. [AB]

Boles, T., Croson, R. & Murnighan, J. K. (2000) Deception and retribution in repeated ultimatum bargaining. *Organizational Behavior and Human Decision Processes* 83:235–59. [aWvH]

Bond, C. F., Jr. & DePaulo, B. M. (2006) Accuracy of deception judgments. *Personality and Social Psychology Review* 10:214–34. [DD, AV, aWvH]

Bond, C. F., Jr. & Fahey, W. E. (1987) False suspicion and the misperception of deceit. *British Journal of Social Psychology* 26:41–46. [aWvH]

Bosson, J. K., Swann, W. B. & Pennebaker, J. W. (2000) Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology* 79:631–43. [SJH]

Bower, G. H. (1991) Mood congruity of social judgments. In: *Emotion and social judgments*, ed. J. P. Forgas, pp. 31–54. Pergamon. [aWvH]

Boyd, R. & Richerson, P. J. (2009) Culture and the evolution of human cooperation, *Philosophical Transactions of the Royal Society (B)* 364:3281–88. [DD]

Bransford, J. D. & Johnson, M. K. (1973) Considerations of some problems of comprehension. In: *Visual information processing*, ed. W. G. Chase, pp. 383–438. Academic Press. [rWvH]

Bratman, M. E. (1992) Practical reasoning and acceptance in a context. *Mind* 101(401):1–15. [KF]

Bressan, P. (2002) The connection between random sequences, everyday coincidences, and belief in the paranormal. *Applied Cognitive Psychology* 16:17–34. [PK]

Bressan, P., Kramer, P. & Germani, M. (2008) Visual attentional capture predicts belief in a meaningful world. *Cortex* 44:1299–306. [PK]

Brewer, M. B. & Caporael, L. (2006) An evolutionary perspective on social identity: Revisiting groups. In: *Evolution and social psychology*, ed. M. Schaller, J. Simpson & D. Kenrick, pp. 143–61. Psychology Press. [DD]

Brissette, I., Scheier, M. F. & Carver, C. S. (2002) The role of optimism in social network development, coping, and psychological adjustment during a life transition. *Journal of Personality and Social Psychology* 82:102–11. [aWvH]

Brown, J. D. (2010) Across the (not so) Great Divide: Cultural similarities in self-evaluative processes. *Social and Personality Psychology Compass* 4:318–30. [rWvH]

Brown, S. P., Cron, W. L. & Slocum, J. W. (1998) Effects of trait competitiveness and perceived intraorganizational competition on salesperson goal setting and performance. *Journal of Marketing* 62:88–98. [aWvH]

Brugger, P., Regard, M., Landis, T. & Graves, R. E. (1995) The roots of meaningful coincidence. *Lancet* 345:1306–307. [PK]

Bruner, J. S. & Goodman, C. C. (1947) Value and need as organizing factors in perception. *Journal of Abnormal and Social Psychology* 42:33–44. [LCE]

Buhrmester, M. D., Blanton, H., & Swann, W. B. (in press) Implicit self-esteem: Nature, measurement, and a new way forward. *Journal of Personality and Social Psychology*. [SJH]

Buller, D. B., Strzyzewski, K. D. & Comstock, J. (1991) Interpersonal deception: I. Deceivers' reactions to receivers' suspicions and probing. *Communication Monographs* 58:1–24. [AV]

Buss, D. M. (1988) The evolution of human intrasexual competition: Tactics of mate attraction. *Journal of Personality and Social Psychology* 54:616–28. [aWvH]

Buss, D. M. (2003) *The evolution of desire: Strategies of human mating, rev. ed.* Free Press. [DMB]

Buss, D. M. (2009) The great struggles of life: Darwin and the emergence of evolutionary psychology. *American Psychologist* 64:140–48. [aWvH]

Buss, D. M. & Dedden, L. A. (1990) Derogation of competitors. *Journal of Social and Personal Relationships* 7:395–422. [aWvH]

Cai, H., Sedikides, C., Gaertner, L., Wang, C., Carvallo, M., Xu, Y., O'Mara, E. M. & Jackson, L. E. (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. [rWvH]

Carruthers, P. (2006) *The architecture of the mind: Massive modularity and the flexibility of thought.* Oxford University Press. [KF]

Carruthers, P. (2009a) An architecture for dual reasoning. In: *In two minds: Dual processes and beyond*, ed. J. St. B. T. Evans & K. Frankish, pp. 109–27. Oxford University Press. [KF]

Carruthers, P. (2009b) How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences* 32(2):121–82. [HM, UF]

Carver, C. S., Kus, L. A. & Scheier, M. F. (1994) Effects of good versus bad mood and optimistic versus pessimistic outlook on social acceptance versus rejection. *Journal of Social and Clinical Psychology* 13:138–51. [aWvH]

Carver, C. S. & Scheier, M. F. (2002) Optimism. In: *Handbook of positive psychology*, ed. C. R. Snyder & S. J. Lopez, pp. 231–43. Oxford University Press. [aWvH]

Cassidy, J. (1988) Child–mother attachment and the self in six-year olds. *Child Development* 59:121–34. [MLB]

Ceci, S. J., Loftus, E. F., Leichtman, M. D. & Bruck, M. (1994) The possible role of source misattributions in the creation of false beliefs among preschoolers. *The International Journal of Clinical and Experimental Hypnosis* 42:304–20. [aWvH]

Chaiken, S. & Trope, Y. (1999) *Dual-process theories in social psychology*. Guilford Press. [aWvH]

Chambers, J. R. & Windschitl, P. D. (2004) Biases in social comparative judgments: The role of nonmotivated factors in above-average and comparative-optimism effects. *Psychological Bulletin* 130:813–38. [MLB, SJH, aWvH]

Champlin, T. S. (1977) Self-deception: A reflexive dilemma. *Philosophy* 52:281–99. [AB]

Charcot, J. M. (1889) *Clinical lectures on diseases of the nervous system*. New Sydenham Society. [AT]

Chartrand, T. L., Dalton, A. N. & Fitzsimons, G. J. (2007) Nonconscious relationship reactance: When significant others prime opposing goals. *Journal of Experimental Social Psychology* 43:719–26. [aWvH]

Chartrand, T. L., Huber, J., Shiv, B. & Tanner, R. J. (2008) Nonconscious goals and consumer choice. *Journal of Consumer Research* 35:189–201. [aWvH]

Chrobak, Q. & Zaragoza, M. S. (2008) Inventing stories: Forcing witnesses to fabricate entire fictitious events leads to freely reported false memories. *Psychonomic Bulletin and Review* 15:1190–95. [aWvH]

Churchland, P. (1987) Epistemology in the age of neuroscience. *Journal of Philosophy* 84:544–53. [RK]

Clancy, S. A., Schacter, D. L., McNally, R. J. & Pitman, R. K. (2000) False recognition in women reporting recovered memories of sexual abuse. *Psychological Science* 11:26–31. [aWvH]

Coates, S. L., Butler, L. T. & Berry, D. C. (2006) Implicit memory and consumer choice: The mediating role of brand familiarity. *Applied Cognitive Psychology* 20:1101–16. [aWvH]

Cohen, G. L., Aronson, J. & Steele, C. M. (2000) When beliefs yield to evidence: Reducing biased evaluation by affirming the self. *Personality and Social Psychology Bulletin* 26:1151–64. [aWvH]

Cohen, L. J. (1992) *An essay on belief and acceptance.* Oxford University Press. [KF]

Cohen, S. (1986) *Behavior, health, and environmental stress*. Plenum Press. [aWvH]

Cohen, S., Alper, C. M., Doyle, W. J., Treanor, J. J. & Turner, R. B. (2006) Positive emotional style predicts resistance to illness after experimental exposure to rhinovirus or Influenza A virus. *Psychosomatic Medicine* 68:809–15. [aWvH]

Colvin, C. R. & Block, J. (1994) Do positive illusions foster mental health? An examination of the Taylor and Brown formulation. *Psychological Bulletin* 116:3–20. [aWvH]

Colvin, C. R., Block, J. & Funder, D. C. (1995) Overly positive self-evaluations and personality: Negative implications for mental health. *Journal of Personality and Social Psychology* 68:1152–62. [DCF, SJH, aWvH]

Coman, A., Manier, D. & Hirst, W. (2009) Forgetting the unforgettable through conversation: Socially shared retrieval-induced forgetting of September 11 memories. *Psychological Science* 20:627–33. [aWvH]

Conger, J. A. & Kanungo, R. N. (1987) Toward a behavioral theory of charismatic leadership in organizational settings. *Academy of Management Review* 12:637–47. [aWvH]

Conway, M. & Ross, M. (1984) Getting what you want by revising what you had. *Journal of Personality and Social Psychology* 47:738–48. [aWvH]

Corballis, M. C. (2007) The evolution of consciousness. In: *The Cambridge handbook of consciousness*, ed. P. D. Zelazo, M. Moscovitch & E. Thompson, pp. 571–95. Cambridge University Press. [JYH]

Correll, J., Spencer, S. J. & Zanna, M. P. (2004) An affirmed self and an open mind: Self-affirmation and sensitivity to argument strength. *Journal of Experimental Social Psychology* 40:350–56. [aWvH]

Crowne, D. P. & Marlowe, D. (1960) A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology* 24:1397–403. [AP]

Crowne, D. P. & Marlowe, D. (1964) *The approval motive.* Wiley. [AP]

Croyle, R. T., Loftus, E. F., Barger, S. D., Sun, Y.-C., Hart, M. & Gettig, J. (2006) How well do people recall risk factor test results? Accuracy and bias among cholesterol screening participants. *Health Psychology* 25:425–32. [aWvH]

Cuc, A., Koppel, J. & Hirst, W. (2007) Silence is not golden: A case for socially shared retrieval-induced forgetting. *Psychological Science* 18:727–33. [aWvH]

Cummins, D. D. (1999) Cheater detection is modified by social rank: The impact of dominance on the evolution of cognitive functions. *Evolution and Human Behavior* 20(4):229–48. [HJL]

D'Argembeau, A. & Van der Linden, M. (2008) Remembering pride and shame: Self-enhancement and the phenomenology of autobiographical memory. *Memory* 16:538–47. [aWvH]

Dawkins, R. (1976) *The selfish gene.* Oxford University Press. [JYH]

Dawkins, R. (1982) *The extended phenotype.* Freeman. [GG]

Dawkins, R. & Krebs, J. (1978) Animal signals: Information or manipulation? In: *Behavioural ecology: An evolutionary approach*, ed. J. Krebs & N. Davies, pp. 282–309. Blackwell Scientific. [RK]

Dawson, E., Gilovich, T. & Regan, D. T. (2002) Motivated reasoning and performance on the Wason Selection Task. *Personality and Social Psychology Bulletin* 28:1379–87. [aWvH]

Dawson, E., Savitsky, K. & Dunning, D. (2006) "Don't tell me, I don't want to know": Understanding people's reluctance to obtain medical diagnostic information. *Journal of Applied Social Psychology* 36:751–68. [aWvH]

de Jong, A., de Ruyter, K. & Wetzels, M. (2006) Linking employee confidence to performance: A study of self-managing service teams. *Journal of the Academy of Marketing Science* 34:576–87. [aWvH]

Dennett, D. (1981) *Brainstorms: Philosophical essays on mind and psychology.* MIT Press. [RK]

Dennett, D. C. (1991) *Consciousness explained.* Little, Brown and Co. [KF]

DePaulo, B. M. (1994) Spotting lies: Can humans learn to do better? *Current Directions in Psychological Science* 3:83–86. [aWvH]

DePaulo, B. M. (2004) The many faces of lies. In: *The social psychology of good and evil*, ed. A. G. Miller, pp. 303–26. Guilford. [aWvH]

DePaulo, B. M. & Kashy, D. A. (1998) Everyday lies in close and casual relationships. *Journal of Personality and Social Psychology* 74:63–79. [aWvH]

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M. & Epstein, J. A. (1996) Lying in everyday life. *Journal of Personality and Social Psychology* 70:979–95. [aWvH, AV]

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K. & Cooper, H. (2003) Cues to deception. *Psychological Bulletin* 129(1):74–118. [HM, aWvH]

DePaulo, B. M., Wetzel, C., Weylin Sternglanz, R. & Walker Wilson, M. J. (2003) Verbal and nonverbal dynamics of privacy, secrecy, and deceit. *Journal of Social Issues* 59:391–410. [AV]

Diener, E. & Diener, M. (1995) Cross-cultural correlates of life satisfaction. *Journal of Personality and Social Psychology* 68:653–63. [MLB]

Ditto, P. H. & Lopez, D. F. (1992) Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology* 63:568–84. [arWvH]

Ditto, P. H., Munro, G. D., Apanovitch, A. M., Scepansky, J. A. & Lockhart, L. K. (2003) Spontaneous skepticism: The interplay of motivation and expectation in responses to favorable and unfavorable medical diagnosis. *Personality and Social Psychology Bulletin* 29:1120–32. [aWvH]

Ditto, P. H., Scepansky, J. A., Munro, G. D., Apanovitch, A. M. & Lockhart, L. K. (1998) Motivated sensitivity to preference-inconsistent information. *Journal of Personality and Social Psychology* 75:53–69. [aWvH]

Donald, M. (1991) *Origins of the modern mind.* Harvard University Press. [JYH]

Drivdahl, S., Zaragoza, M. S. & Learned, D. (2009) The role of emotional elaboration in the creation of false memories. *Applied Cognitive Psychology* 23:13–35. [aWvH]

Ebbinghaus, H. (1885) *Memory: A contribution to experimental psychology*, trans. H. A. Ruger & C. E. Bussenius. Teachers College. [rWvH]

Effron, D., Cameron, J. S. & Monin, B. (2009) Endorsing Obama licenses favoring Whites. *Journal of Experimental Social Psychology* 45:590–93. [JYH]

Egan, L., Bloom, P. & Santos, L. R. (2010) Choice-induced preferences in the absence of choice: Evidence from a blind two choice paradigm with young children and capuchin monkeys. *Journal of Experimental Social Psychology* 46:204–207. [LCE]

Ekman, P. (1996) Why we don't catch liars. *Social Research* 63:801–17. [DD]

Elster, J. (1999) *Alchemies of the mind: Rationality and the emotions*. Cambridge University Press. [RM]

Epley, N. & Whitchurch, E. (2008) Mirror, mirror on the wall: Enhancement in self recognition. *Personality and Social Psychology Bulletin* 34:1159–70. [MLB, SP, arWvH]

Evans, J. St. B. T. (2007) *Hypothetical thinking: Dual processes in reasoning and judgement.* Psychology Press. [KF]

Evans, J. St. B. T. (2008) Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology* 59:255–78. [KF]

Evans, J. St. B. T. & Frankish, K., eds. (2009) *In two minds: Dual processes and beyond.* Oxford University Press. [KF]

Evans, J. St. B. T. & Over, D. E. (1996) *Rationality and reasoning.* Psychology Press. [KF]

Exline, J. J. & Lobel, M. (1999) The perils of outperformance: Sensitivity about being the target of a threatening upward comparison. *Psychological Bulletin* 125:307–37. [SJH]

Falk, C. F., Heine, S. J., Yuki, M. & Takemura, K. (2009) Why do Westerners self-enhance more than East Asians? *European Journal of Personality* 23:183–209. [SJH]

Fazio, R. H. & Olson, M. A. (2003) Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology* 54:297–327. [aWvH]

Fehr, E. & Gächter, S. (2002) Altruistic punishment in humans. *Nature* 415:137–40. [aWvH]

Fein, S. & Spencer, S. J. (1997) Prejudice as self-image maintenance: Affirming the self through derogating others. *Journal of Personality and Social Psychology* 73:31–44. [MLB, aWvH]

Ferguson, M. J. (2008) On becoming ready to pursue a goal you don't know you have: Effects of nonconscious goals on evaluative readiness. *Journal of Personality and Social Psychology* 95:1268–94. [JYH]

Festinger, L. (1954) A theory of social comparison processes. *Human Relations* 7:117–40. [aWvH]

Festinger, L. (1957) *A theory of cognitive dissonance.* Stanford University Press. [PJ]

Festinger, L. & Carlsmith, J. M. (1959) Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology* 58:203–11. [LCE, aWvH]

Fisher, M. L. (2004) Female intrasexual competition decreases female facial attractiveness. *Biology Letters* 271:283–85. [aWvH]

Fitzsimmons, G. & Anderson, J. (in press) Interdependent goals and relationship conflict. Chapter to appear In: *Social conflict and aggression*, ed. J. P. Forgas, A. Kruglanski & K. Williams. Psychology Press. [aWvH]

Fleming, J. M., Darley, J. H., Hilton, B. A. & Kojetin, B. A. (1990) Multiple audience problem: A strategic communication perspective on social perception. *Journal of Personality and Social Psychology* 58:593–609. [AV]

Flinn, M. V. (2007) Evolution of stress response to social threat. In: *The Oxford handbook of evolutionary psychology*, ed. R. I. M. Dunbar & L. Barrett, pp. 273–96. Oxford University Press. [UF]

Fodor, J. (1983) *The modularity of mind.* MIT Press. [RK]

Fodor, J. (2000) *The mind doesn't work that way.* Bradford Books/MIT. [RK]

Ford, C. V. (1983) *The somatizing disorders. Illness as a way of life*. Elsevier. [AT]

Förster, J., Liberman, N. & Higgins, E. T. (2005) Accessibility from active and fulfilled goals. *Journal of Experimental Social Psychology* 41(3):220–39. [JYH]

Frankfurt, H. G. (2005) *On bullshit*. Princeton University Press. [DD]

Frankish, K. (2004) *Mind and supermind.* Cambridge University Press. [KF]

Frankish, K. (2009) Systems and levels: Dual-system theories and the personal-subpersonal distinction. In: *In two minds: Dual processes and beyond*, ed. J. St. B. T. Evans & K. Frankish, pp. 89–107. Oxford University Press. [KF]

Frankish, K. (2010) Dual-process and dual-system theories of reasoning. *Philosophy Compass* 5(10):914–26. [KF]

Frankish, K. & Evans, J. St. B. T. (2009) The duality of mind: An historical perspective. In: *In two minds: Dual processes and beyond*, ed. J. St. B. T. Evans & K. Frankish, pp. 1–29. Oxford University Press. [KF]

Fredrickson, B. L. (1998) What good are positive emotions? *Review of General Psychology* 2: 300–19. [aWvH]

Fredrickson, B. L. (2001) The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American Psychologist* 56:218–26. [aWvH]

Fredrickson, B. L., Cohn, M. A., Coffey, K. A., Pek, J. & Finkel, S. M. (2008) Open hearts build lives: Positive emotions, induced through loving-kindness meditation, build consequential personal resources. *Journal of Personality and Social Psychology* 95:1045–62. [aWvH]

Frey, D. (1986) Recent research on selective exposure to information. In: *Advances in experimental social psychology, vol. 19*, ed. L. Berkowitz, pp. 41–80. Academic Press. [aWvH]

Frey, U. J. (2010) Modern illusions of humankind. In: *Homo Novus – A human without Illusions*, ed. U. J. Frey, C. Störmer & K. P. Willführ, pp. 263–88. Springer. [UF]

Frijda, N. H. & Mesquita, B. (1994) The social roles and functions of emotions. In: *Emotion and culture: Empirical studies of mutual influence*, ed. S. Kitayama & H. R. Markus, pp. 51–87. APA. [aWvH]

Funkhouser, E. (2005) Do the self-deceived get what they want? *Pacific Philosophical Quarterly* 86:295–312. [RM]

Gaertner, L., Sedikides, C. & Chang, K. (2008) On pancultural self-enhancement: Well-adjusted Taiwanese self-enhance on personally valued traits. *Journal of Cross-Cultural Psychology* 39:463–77. [aWvH]

Gallese, V. (2007) Before and below 'theory of mind': embodied simulation and the neural correlates of social cognition. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 362:659–69. [AP]

Gazzaniga, M. S. (1997) Why can't I control my brain? Aspects of conscious experience. In *Cognition, computation, and consciousness*, ed. M. Ito, Y. Miyashita & E. T. Rolls, pp. 69–79. Oxford University Press. [aWvH]

Gilbert, D. T. (2006) *Stumbling on happiness.* Knopf. [TS]

Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J. & Wheatley, T. (1998) Immune neglect: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology* 75:617–38. [aWvH]

Gilovich, T. (1991) *How we know what isn't so: The fallibility of human reason in everyday life.* Macmillan. [UF]

Gilovich, T., Savitsky, K. & Medvec, V. H. (1998) The illusion of transparency: Biased assessments of others' ability to read one's emotional states. *Journal of Personality and Social Psychology* 75:332–46. [aWvH]

Glass, D. C. & Singer, J. E. (1972) Behavioral after effects of unpredictable and uncontrollable events. *American Scientist* 60:457–65. [aWvH]

Godfrey, D. K., Jones, E. E. & Lord, C. G. (1986) Self-promotion is not ingratiating. *Journal of Personality and Social Psychology* 50:106–13. [SJH]

Goetz, C. G. (2007) J.-M. Charcot and simulated neurologic disease. Attitudes and diagnostic strategies. *Neurology* 69:103–109. [AT]

Gonsalves, B., Reber, P. J., Gitelman, D. R., Parrish, T. B., Mesulman, M. M. & Paller, K. A. (2004) Neural evidence that vivid imaging can lead to false remembering. *Psychological Science* 15:655–60. [aWvH]

Grafen, A. (1990) Biological signals as handicaps. *Journal of Theoretical Biology* 144:517–46. [RM]

Grafton, S. T. (2009) Embodied cognition and the simulation of action to understand others. *Annals of the New York Academy of Sciences* 1156:97–117. [AP]

Green, J. D., Sedikides, C. & Gregg, A. P. (2008) Forgotten but not gone: The recall and recognition of self-threatening memories. *Journal of Experimental Social Psychology* 44:547–61. [aWvH]

Greenwald, A. G. (1997) Self-knowledge and self-deception: Further consideration. In: *The mythomanias: The nature of deception and self-deception*, ed. M. S. Myslobodsky, pp. 51–72. Erlbaum. [HJL]

Greenwald, A. G. & Farnham, S. D. (2000) Using the implicit association test to measure self-esteem and self-concept. *Journal of Personality and Social Psychology* 79:1022–38. [SJH]

Greenwald, A. G., McGhee, D. E. & Schwartz, J. L. K. (1998) Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology* 74:1464–80. [aWvH]

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. & Banaji, M. R. (2009) Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology* 97:17–41. [aWvH]

Griskevicius, V., Cialdini, R. B. & Kenrick, D. T. (2006) Peacocks, Picasso, and parental investment: The effects of romantic motives on creativity. *Journal of Personality and Social Psychology* 91:63–76. [DTK]

Gur, R. C. & Sackeim, H. A. (1979) Self-deception: A concept in search of a phenomenon. *Journal of Personality and Social Psychology* 37:147–69. [aWvH]

Haight, M. R. (1980) *A study of self deception.* Humanities Press. [AB]

Haley, K. J. & Fessler, D. M. T. (2005) Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior* 26(3):245–56. [HJL]

Hall, L. & Johansson, P. (2008) Using choice blindness to study decision making and introspection. In: *Cognition – A smorgasbord*, ed. P. Gärdenfors & A. Wallin, pp. 267–83. Nya Doxa. [PJ]

Hall, L., Johansson, P. & Strandberg, T. (in preparation a) *Choice blindness and moral decision making.* [PJ]

Hall, L., Johansson, P., Tärning, B., Sikström, S. & Chater, N. (in preparation b) *Preference change through choice.* [PJ]

Hall, L., Johansson, P., Tärning, B., Sikström, S. & Deutgen, T. (2010) Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition* 117:54–61. [PJ]

Hamamura, T., Heine, S. J. & Takemoto, T. (2007) Why the better-than-average effect is a worse-than-average measure of self-enhancement. An investigation of conflicting findings from studies of East Asian self-evaluations. *Motivation and Emotion* 31:247–59. [SJH]

Harker, L. & Keltner, D. (2001) Expressions of positive emotion in women's college yearbook pictures and their relationship to personality and life outcomes across adulthood. *Journal of Personality and Social Psychology* 80:112–24. [aWvH]

Harnad, S. (1995) "Why and how we are not zombies. *Journal of Consciousness Studies* 1:164–67. Available at: http://cogprints.org/1601. [SH]

Harnad, S. (2000) Correlation vs. causality: How/why the mind/body problem is hard. *Journal of Consciousness Studies* 7(4):54–61. Available at: http://cogprints.org/1617/. [SH]

Harnad, S. (2002) Turing indistinguishability and the blind watchmaker. In: *Evolving consciousness*, ed. J. Fetzer, pp. 3–18. John Benjamins. Available at: http://cogprints.org/1615/. [SH]

Harnad, S. (2003) Can a machine be conscious? How? *Journal of Consciousness Studies* 10(4–5):69–75. Available at: http://eprints.ecs.soton.ac.uk/7718/. [SH]

Harnad, S. & Scherzer, P. (2008) First, scale up to the robotic Turing test, then worry about feeling. *Artificial Intelligence in Medicine* 44(2):83–89. Available at: http://eprints.ecs.soton.ac.uk/14430/. [SH]

Harris, P. R., Mayle, K., Mabbott, L. & Napper, L. (2007) Self-affirmation reduces smokers' defensiveness to graphic on-pack cigarette warning labels. *Health Psychology* 26:437–46. [aWvH]

Hartwig, M., Granhag, P. A., Strömwall, L. & Kronkvist, O. (2006) Strategic use of evidence during police interrogations: When training to detect deception works. *Law and Human Behavior* 30:603–19. [rWvH, AV]

Haselton, M. G., Buss, D. M., Oubaid, V. & Angleitner, A. (2005) Sex, lies, and strategic interference: The psychology of deception between the sexes. *Personality and Social Psychology Bulletin* 31:3–23. [DMB, aWvH]

Heine, S. J. (2005a) Constructing good selves in Japan and North America. In: *Culture and social behavior: The tenth Ontario symposium*, ed. R. M. Sorrentino, D. Cohen, J. M. Olson & M. P. Zanna, pp. 95–116. Erlbaum. [SJH]

Heine, S. J. (2005b) Where is the evidence for pancultural self-enhancement? A reply to Sedikides, Gaertner & Toguchi. *Journal of Personality and Social Psychology* 89:531–38. [SJH]

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

Heine, S. J., Kitayama, S. & Hamamura, T. (2007a) Inclusion of additional studies yields different conclusions: Comment on Sedikides, Gaertner & Vevea (2005), *Journal of Personality and Social Psychology. Asian Journal of Social Psychology* 10:49–58. [SJH]

Heine, S. J., Kitayama, S. & Hamamura, T. (2007b) Which studies test the question of pancultural self-enhancement? A reply to Sedikides, Gaertner & Vevea, 2007. *Asian Journal of Social Psychology* 10:198–200. [SJH]

Heine, S. J., Kitayama, S., Lehman, D. R., Takata, T., Ide, E., Leung, C. & Matsumoto, H. (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* 81:599–615. [SJH]

Heine, S. J., Lehman, D. R., Markus, H. R. & Kitayama, S. (1999) Is there a universal need for positive self-regard? *Psychological Review* 106:766–94. [SJH, aWvH]

Heine, S. J., Takata, T. & Lehman, D. R. (2000) Beyond self-presentation: Evidence for self-criticism among Japanese. *Personality and Social Psychology Bulletin* 26:71–78. [SJH]

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

Hertenstein, M. J., Hansel, C. A., Butts, A. M. & Hile, S. N. (2009) Smile intensity in photographs predicts divorce later in life. *Motivation and Emotion* 33:99–105. [aWvH]

Hofmann, W., Friese, M. & Strack, F. (2009) Impulse and self-control from a dual-systems perspective. *Perspectives on Psychological Science* 4:162–76. [rWvH]

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H. & Schmitt, M. (2005) A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin* 31:1369–85. [SJH, aWvH]

Horstmann G. (2002) Evidence for attentional capture by a surprising color singleton in visual search. *Psychological Science* 13:499–505. [PK]

Hrdy, S. B. (2009) *Mothers and others – The evolutionary origins of mutual understanding.* Harvard University Press. [UF]

Humphrey, N. (1978) Nature's psychologists. *New Scientist* 1109:900–904. [NH]

Humphrey, N. (1983) *Consciousness regained: Chapters in the development of mind.* Oxford University Press. [NH]

Humphrey, N. & Dennett, D. C. (1998) Speaking for our selves. In: *Brainchildren: Essays on designing minds*, ed. D. C. Dennet, pp. 31–58. Penguin Books. [RK]

Isaacowitz, D. M. (2006) Motivated gaze: The view from the gazer. *Current Directions in Psychological Science* 15:68–72. [aWvH]

Isaacowitz, D. M., Toner, K., Goren, D. & Wilson, H. R. (2008) Looking while unhappy mood-congruent gaze in young adults, positive gaze in older adults. *Psychological Science* 19:848–53. [aWvH]

Jacoby, L. L. (1991) A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language* 30:513–41. [rWvH]

Jang, S. A., Smith, S. W. Levine, T. R. (2002) To stay or to leave? The role of attachment styles in communication patterns and potential termination of romantic relationships following discovery of deception. *Communication Monographs* 69:236–52. [aWvH]

Jarvis, W. B. G. (1998) *Do attitudes really change*? Unpublished doctoral dissertation, Ohio State University. [rWvH]

Johansson, P., Hall, L., Sikstrom, S. & Olsson, A. (2005) Failure to detect mismatches between intention and outcome in a simple decision task. *Science* 310:116–19. [LCE, PJ]

Johansson, P., Hall, L., Sikström, S., Tärning, B. & Lind, A. (2006) How something can be said about telling more than we can know. *Consciousness and Cognition* 15(4):673–92. [PJ]

John, O. P. & Robins, R. W. (1994) Accuracy and bias in self-perception: Individual differences in self-enhancement and narcissism. *Journal of Personality and Social Psychology* 66:206–19. [DD, DCF]

Jordan, C., Spencer, S. J., Zanna, M. P., Hoshino-Browne, E. & Correll, J. (2003) Implicit self-esteem, explicit self-esteem and defensiveness. *Journal of Personality and Social Psychology* 85:969–78. [aWvH]

Josephs, R. A., Larrick, R. P., Steele, C. M. & Nisbett, R. E. (1992) Protecting the self from the negative consequences of risky decisions. *Journal of Personality and Social Psychology* 62:26–37. [aWvH]

Jost, J. T., Banaji, M. R. & Nosek, B. A. (2004) A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology* 25:881–919. [aWvH]

Jost, J. T. & Hunyady, O. (2005) Antecedents and consequences of system-justifying ideologies. *Current Directions in Psychological Science* 14:260–65. [aWvH]

Jost, J. T., Napier, J. L., Thorisdottir, H., Gosling, S. D., Palfai, T. P. & Ostafin, B. (2007) Are needs to manage uncertainty and threat associated with political conservatism or ideological extremity? *Personality and Social Psychology Bulletin* 33:989–1007. [rWvH]

Kahneman, D. & Frederick, S. (2002) Representativeness revisited: Attribute substitution in intuitive judgment. In: *Heuristics and biases: The psychology of intuitive judgment*, ed. T. Gilovich, D. Griffin & D. Kahneman, pp. 49–81. Cambridge University Press. [KF]

Karim, A. A., Schneider, M., Lotze, M., Veit, R., Sauseng, P., Braun, C. & Birbaumer, N. (2010) The truth about lying: Inhibition of the anterior prefrontal cortex improves deceptive behavior. *Cerebral Cortex* 20:205–13. [rWvH]

Kay, A. C., Gaucher, D., Napier, J. L., Callan, M. J. & Laurin, K. (2008) God and government: Testing a compensatory control explanation for the support of external systems. *Journal of Personality and Social Psychology* 95:18–35. [arWvH]

Keltner, D. & Kring, A. (1998) Emotion, social function, and psychopathology. *Review of General Psychology* 2:320–42. [aWvH]

Kenny, D. A. & Kashy, D. A. (1994) Enhanced co-orientation in the perception of friends: A social relations analysis. *Journal of Personality and Social Psychology* 67:1024–33. [aWvH]

Kenrick, D. T., Neuberg, S. L., Griskevicius, V., Becker, D. V., Schaller, M. (2010) Goal-driven cognition and functional behavior: The fundamental motives framework. *Current Directions in Psychological Science* 19:63–67. [DTK]

Khalil, E. L. (2009) Self-deceit and self-serving bias: Adam Smith on "general rules." *Journal of Institutional Economics* 5(2):251–58. [ELK]

Khalil, E. L. (2010) Adam Smith's concept of self-command as a solution to dynamic inconsistency and the commitment problem. *Economic Inquiry* 48(1):177–91. [ELK]

Khalil, E. L. (submitted) Making sense of self-deception. *Journal of Economic Psychology* [ELK]

Kim, P. H., Ferrin, D. L., Cooper, C. D. & Dirks, K. T. (2004) Removing the shadow of suspicion: The effects of apology versus denial for repairing

competence-versus integrity-based trust violations. *Journal of Applied Psychology* 89(1):104. [HM]

Klar, Y. & Giladi, E. E. (1997) No one in my group can be below the group's average: A robust positivity bias in favor of anonymous peers. *Journal of Personality and Social Psychology* 73(5):885–901. [MLB, SJH]

Klein, D. C., Fencil-Morse, E. & Seligman, M. E. P. (1976) Learned helplessness, depression, and the attribution of failure. *Journal of Personality and Social Psychology* 33:508–16. [aWvH]

Kohut, H. (1977) *The restoration of the self*. International Universities Press. [rWvH]

Kolers, P. A. (1976) Pattern-analyzing memory. *Science* 191:1280–81. [arWvH]

Korsgaard, C. (1989) Personal identity and the unity of agency: A Kantian response to Parfit. *Philosophy & Public Affairs* 18:101–32. [AB]

Krachun, C., Carpenter, M., Call, J. & Tomasello, M. (2009) A competitive nonverbal false belief task for children and apes. *Developmental Science* 12(4):521–35. [TS]

Krahn, L. E., Bostwick, J. M. & Stonnington, C. M. (2008) Looking toward DSM-V: Should factitious disorder become a subtype of somatoform disorder? *Psychosomatics* 49:277–82. [AT]

Krizan, Z. & Suls, J. (2008) Losing sight of oneself in the above-average effect: When egocentrism, focalism, and group diffuseness collide. *Journal of Experimental Social Psychology* 44:929–42. [SJH]

Krizan, Z. & Windschitl, P. D. (2009) Wishful thinking about the future: Does desire impact optimism? *Social and Personality Psychology Compass* 3:227–43. [LCE]

Kruger, J. (1999) Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology* 77:221–32. [SJH]

Kumashiro, M. & Sedikides, C. (2005) Taking on board liability-focused feedback: Close positive relationships as a self-bolstering resource. *Psychological Science* 16:732–39. [aWvH]

Kunda, Z. (1990) The case for motivated reasoning. *Psychological Bulletin* 108:480–98. [aWvH]

Kurland, J. A. & Gaulin, S. J. C. (2005) Cooperation and conflict among kin. In: *The handbook of evolutionary psychology*, ed. D. M. Buss, pp. 447–82. Wiley. [UF]

Kurzban, R. (in press) *Why everyone (else) is a hypocrite: Evolution and the modular mind.* Princeton University Press. [RK]

Kurzban, R. & Aktipis, C. A. (2006) Modular minds, multiple motives. In: *Evolution and social psychology*, ed. M. Schaller, J. Simpson & D. Kenrick, pp. 39–53. Psychology Press. [RK, rWvH]

Kurzban, R. & Aktipis, C. A. (2007) Modularity and the social mind: Are psychologists too self-ish? *Personality and Social Psychology Review* 11:131–49. [RK]

Kurzban, R. & Christner, J. (in press) Are supernatural beliefs commitment devices for intergroup conflict? In: *The psychology of social conflict and aggression (The Sydney Symposium of Social Psychology, vol. 13)*, ed. J. P. Forgas, A. Kruglanski & K. D. Willimas. [RK]

Kwang, T. & Swann, W. B., Jr. (2010) Do people embrace praise even when they don't feel worthy? A review of critical tests of self-enhancement versus self-verification. *Personality and Social Psychology Review* 14:263–80. [MLB]

Lahdenperä, M., Lummaa, V., Helle, S., Tremblay, M. & Russell, A. F. (2004) Fitness benefits of prolonged post-reproductive lifespan in women. *Nature* 428:178–81. [aWvH]

Lakin, J., Chartrand, T. L. & Arkin, R. (2008) I am too just like you: Nonconscious mimicry as an automatic behavioral response to social exclusion. *Psychological Science* 19:816–22. [aWvH]

Lane, R. D., Meringas, K. R., Schwartz, G. E., Huang, S. S. & Prusoff, B. A. (1990) Inverse relationship between defensiveness and lifetime prevalence of psychiatric disorder. *American Journal of Psychiatry* 147:573–78. [AP]

Lee, A. Y. (2002) Effects of implicit memory on memory-based versus stimulus-based brand choice. *Journal of Marketing Research* 39:440–54. [aWvH]

Lee, S. W. S. & Schwarz, N. (2010) Washing away postdecisional dissonance. *Science* 328:709. [rWvH]

Leeuwen, D. S. N. V. (2007) The spandrels of self-deception: Prospects for a biological theory of a mental phenomenon. *Philosophical Psychology* 20(3):329–48. [HJL]

Lerman, C., Croyle, R. T., Tercyak, K. P. & Hamann, H. (2002) Genetic testing: Psychological aspects and implications. *Journal of Consulting and Clinical Psychology* 70:784–97. [aWvH]

Levine, T. R. & McCornack, S. A. (1992) Linking love and lies: A formal test of the McCornack and Parks model of deception detection. *Journal of Social and Personal Relationships* 9:143–54. [AV]

Levine, T. R. & McCornack, S. A. (2001) Behavioral adaptation, confidence, and heuristic-based explanations of the probing effect. *Human Communication Research* 27:471–502. [aWvH]

Lichtenstein, S. & Slovic, P., eds. (2006) *The construction of preference.* Cambridge University Press. [PJ]

Lord, C. G., Ross, L. & Lepper, M. R. (1979) Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology* 37: 2098–109. [aWvH]

Lu, H. & Chang, L. (2010) Deceive yourself to deceive high but not necessarily low status others. Paper presented at the 22nd Annual Meeting of the Human Behavior and Evolution Society, Eugene, OR, June 2010. [HJL]

Lyubomirsky, S., King, L. & Diener, E. (2005) The benefits of frequent positive affect: Does happiness lead to success? *Psychological Bulletin* 131:803–55. [aWvH]

MacLeod, M. D. & Saunders, J. (2008) Retrieval inhibition and memory distortion: Negative consequences of an adaptive process. *Current Directions in Psychological Science* 17:26–30. [aWvH]

Maner, J. K., Kenrick, D. T. Becker, D. V., Delton, A. W., Hofer, B., Wilbur, C. J. & Neuberg, S. L. (2003) Sexually selective cognition: Beauty captures the mind of the beholder. *Journal of Personality and Social Psychology* 6:1107–20. [DTK]

Maner, J. K., Kenrick, D. T., Becker, D. V., Robertson, T. E., Hofer, B., Neuberg, S. L., Delton, A. W., Butner, J. & Schaller, M. (2005) Functional projection: How fundamental social motives can bias interpersonal perception. *Journal of Personality and Social Psychology* 88:63–78. [DTK]

Mann, S. & Vrij, A. (2006) Police officers' judgements of veracity, tenseness, cognitive load and attempted behavioural control in real life police interviews. *Psychology, Crime, & Law* 12:307–19. [aWvH]

Markus, H. & Wurf, E. (1987) The dynamic self-concept: A social psychological perspective. *Annual Review of Psychology* 38:299–337. [rWvH]

Markus, H. R. & Kitayama, S. (1991) Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review* 98:224–53. [SJH]

Marsland, A. L., Pressman, S. & Cohen, S. (2007) Positive affect and immune function. In: *Psychoneuroimmunology*, ed. R. Ader, pp. 261–79. Elsevier. [aWvH]

Martindale, C. (1980) Subselves. In *Review of Personality and Social Psychology*, ed. L. Wheeler, pp. 193–218. Sage. [DTK]

Mather, M., Canli, T., English, T., Whitfield, S., Wais, P., Ochsner, K., Gabrieli, J. D. E. & Carstensen, L. L. (2004) Amygdala responses to emotionally valenced stimuli in older and younger adults. *Psychological Science* 15:259–63. [aWvH]

Mather, M. & Carstensen, L. L. (2005) Aging and motivated cognition: The positivity effect in attention and memory. *Trends in Cognitive Science* 9:496–502. [aWvH]

Mayr, E. (1976) *Evolution and the diversity of life.* Harvard University Press. [JYH]

McCloskey, M. & Zaragoza, M. (1985) Misleading postevent information and memory for events: Arguments and evidence against memory impairment hypotheses. *Journal of Experimental Psychology: General* 114:1–16. Available at: http://csaweb113v.csa.com/ids70/view_record.php?id = 3&recnum = 29&log = from_res&SID = jqq1f6dbploe7phjilrkgf7gk4. [arWvH]

McCornack, S. A. & Parks, M. R. (1986) Detection deception and relational development: The other side of trust. In: *Communication yearbook, vol. 9*, ed. M. L. McLaughlin, pp. 377–89. Sage. [AV]

McKay, R. T. & Dennett, D. C. (2009) The evolution of misbelief. *Behavioral and Brain Sciences* 32(6):493–561. [RK, RM]

Mele, A. R. (1997) Real self-deception. *Behavioral and Brain Sciences* 20(1):91–136. [HJL, aWvH]

Mele, A. R. (2001) *Self-deception unmasked.* Princeton University Press. [DLS, rWvH]

Mercier, H. & Landemore, H. (in press) Reasoning is for arguing: Understanding the successes and failures of deliberation. *Political Psychology*. [HM]

Mercier, H. & Sperber, D. (in press) Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences.* [HM]

Merckelbach, H., Jelicic, M. & Pieters, M. (2010) The residual effect of feigning: How intentional faking may evolve into a less conscious form of symptom reporting. *Journal of Clinical and Experimental Neuropsychology* 2010 July 9:1–9. [Epub ahead of print, DOI: 10.1080/13803395.2010.495055]. [AT, rWvH]

Meyer, W.-U., Niepel, M., Rudolph, U. & Schützwohl, A. (1991) An experimental analysis of surprise. *Cognition and Emotion* 5:295–311. [PK]

Mezulis, A. H., Abramson, L. Y., Hyde, J. S. & Hankin, B. L. (2004) Is there a universal positive bias in attributions?: A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological Bulletin* 130:711–47. [SJH]

Mijović-Prelec, D. & Prelec, D. (2010) Self-deception as self-signaling: A model and experimental evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365(1538):227–40. [RM]

Millar, M. G. & Millar, K. U. (1995) Detection of deception in familiar and unfamiliar persons: The effects of information restriction. *Journal of Nonverbal Behavior* 19:69–84. [AV]

Miller, G. (2000) *The mating mind.* Penguin. [aWvH]

Millikan, R. G. (1984) *Language, thought and other biological categories.* MIT Press. [DLS]

Millikan, R. G. (1993) *White queen psychology and other essays for Alice.* MIT Press. [DLS]

Minsky, M. (1975) A framework for representing knowledge. In: *The psychology of computer vision*, ed. P. H. Winston, pp. 211–77. McGraw-Hill. [PK]

Miotto, P., De Coppi, M., Frezza, M., Rossi, M. & Preti, A. (2002) Social desirability and eating disorders. A community study of an Italian school-aged sample. *Acta Psychiatrica Scandinavica* 105:372–77. [AP]

Miotto, P. & Preti, A. (2008) Suicide ideation and social desirability among school-aged young people. *Journal of Adolescence* 31:519–33. [AP]

Monin, B. & Miller, D. T. (2001) Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology* 81:33–43. [JYH]

Motluk, A. (2001, December 22) It's a wonderful lie. *New Scientist* 2322:70–71. [NH]

Nairne, J. S., Pandeirada, J. N. S. & Thompson, S. R. (2008) Adaptive memory: The comparative value of survival processing. *Psychological Science* 19(2):176–80. [HJL]

Nardone, I. B., Ward, R., Fotopoulou, A. & Turnbull, O. H. (2007) Attention and emotion in anosognosia: Evidence of implicit awareness and repression? *Neurocase* 13:438–45. [aWvH]

Neely, J. H. (1977) Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General* 106:226–54. [aWvH]

Nickerson, R. S. (1998) Confirmation bias: A ubiquitous phenomena in many guises. *Review of General Psychology* 2:175–220. [HM]

Niepel, M., Rudolph, U., Schützwohl, A. & Meyer, W.-U. (1994) Temporal characteristics of the surprise reaction induced by schema-discrepant visual and auditory events. *Cognition and Emotion* 8:433–52. [PK]

Nisbett, R. E. & Wilson, T. D. (1977) Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84:231–59. [aWvH]

Nock, M. K., Park, J. L., Finn, C. T., Deliberto, T. L., Dour, H. J. & Banaji, M. R. (2010) Measuring the "suicidal mind:" Implicit cognition predicts suicidal behavior. *Psychological Science* 21:511–17. [aWvH]

Norris, P. & Inglehart. R. (2004) *Sacred and secular: Religion and politics worldwide.* Cambridge University Press. [arWvH]

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G. & Banaji, M. R. (2007) Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology* 18:36–88. [aWvH]

Nuttin, J. M. (1985) Narcissism beyond Gestalt and awareness: The name letter effect. *European Journal of Social Psychology* 15:353–61. [aWvH]

Oishi, S., & Diener, E. (2003) Culture and well-being: The cycle of action, evaluation, and decision. *Personality and Social Psychology Bulletin* 29:939–49. [SJH]

Olson, J. M. & Zanna, M. P. (1979) A new look at selective exposure. *Journal of Experimental Social Psychology* 15:1–15. [aWvH]

Park, H. S., Levine, T. R., McCornack, S. A., Morrisson, K. & Ferrara, M. (2002) How people really detect lies. *Communication Monographs* 69:144–57. [rWvH, AV]

Paulhus, D. L. (1991) Measurement and control of response bias. In: *Measures of personality and social psychological attitudes, vol. 1*, ed. J. P. Robinson, P. R. Shaver & L. S. Wrightsman, pp. 17–59. Academic Press. [AP]

Paulhus, D. L. (1998) Interpersonal and intrapsychic adaptiveness of trait self-enhancement: A mixed blessing? *Journal of Personality and Social Psychology* 74:1197–208. [DD, SJH, aWvH]

Paulhus, D. L. & John, O. P. (1998) Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality* 66(6):1025–60. [HJL]

Paulhus, D. L. & Reid, D. B. (1991) Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology* 60(2):307–17. [HJL]

Pears, D. (1985) The goals and strategies of self-deception. In: *The multiple self*, ed. J. Elster, pp. 59–77. Cambridge University Press. [RK]

Penn, D. C., Holyoak, K. J. & Povinelli, D. J. (2008) Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences* 31:109–78. [TS]

Penrod, S. D. & Cutler, B. L. (1995) Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, and Law* 1:817–45. [aWvH]

Peters, H. J. & Williams, J. M. (2006) Moving cultural background to the foreground: An investigation of self-talk, performance, and persistence following feedback. *Journal of Applied Sport Psychology* 18:240–53. [SJH]

Petty, R. E. & Cacioppo, J. T. (1986) The elaboration likelihood model of persuasion. In: *Advances in experimental social psychology, vol. 19*, ed. L. Berkowitz, pp. 123–205. Academic Press. [aWvH]

Pinker, S. (1997) *How the mind works.* Norton. [RK]

Plous, S. (1993) *The psychology of judgment and decision making.* McGraw-Hill. [UF]

Premack, D. & Woodruff, G. (1978) Does the chimpanzee have a theory of mind? *Behavioral & Brain Sciences* 1:515–26. [SH]

Preti, A. & Miotto, P. (2006) Mental disorders, evolution and inclusive fitness. *Behavioral and Brain Sciences* 29:419–20. [AP]

Preti, A., Vellante, M., Baron-Cohen, S., Zucca, G., Petretto, D. R. & Masala, C. (2010) The Empathy Quotient: A cross-cultural comparison of the Italian version. *Cognitive Neuropsychiatry* 2010 August 24:1–21 (DOI: 10.1080/13546801003790982). [AP]

Price, P. C. & Stone, E. R. (2004) Intuitive evaluation of likelihood judgment producers. *Journal of Behavioral Decision Making* 17:39–57. [aWvH]

Prinz, W. (2008) Mirrors for embodied communication. In: *Embodied communication in humans and machines*, ed. I. Wachsmuth, M. Lenzen & G. Knoblich, pp. 111–27. Oxford University Press. [UF]

Pyszczynski, T. & Greenberg, J. (1987) Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. *Advances in Experimental Social Psychology* 20:297–340. [aWvH]

Quattrone, G. & Tversky, A. (1984) Causal versus diagnostic contingencies: On self-deception and on the voter's illusion. *Journal of Personality and Social Psychology* 46:237–48. [RM]

Raghunathan, R. & Trope, Y. (2002) Walking the tightrope between feeling good and being accurate: Mood as a resource in processing persuasive messages. *Journal of Personality and Social Psychology* 83(3):510–25. [HM]

Ramachandran, V. S. (2009) Self-awareness: The last frontier. Available at: http://www.edge.org/3rd_culture/rama08/rama08_index.html. [aWvH]

Ramanaiah, N. V., Schill, T. & Leung, L. S. (1977) A test of the hypothesis about the two dimensional nature of the Marlowe-Crowne Social Desirability Scale. *Journal of Research in Personality* 11:251–59. [AP]

Reed, M. B. & Aspinwall, L. G. (1998) Self-affirmation reduces biased processing of health-risk information. *Motivation and Emotion* 22:99–132. [aWvH]

Rice, W. R. & Holland, B. (1997) The enemies within: Intragenomic conflict, interlocus contest evolution (ICE), and the intraspecific Red Queen. *Behavioral Ecology and Sociobiology* 41:1–10. [SWG]

Riskind, J. H., Moore, R. & Bowley, L. (1995) The looming of spiders: The fearful perceptual distortion of movement and menace. *Behaviour Research and Therapy* 33:171–78. [LCE]

Rizzolatti, G. & Craighero, L. (2004) The mirror-neuron system. *Annual Reviews of Neuroscience* 27:169–92. [AP]

Robins, R. W. & Beer, J. S. (2001) Positive illusions about the self: Short-term benefits and long-term costs. *Journal of Personality and Social Psychology* 80:340–52. [SJH]

Rogers, R. (1988) *Clinical assessment of malingering and deception.* Guilford Press. [AT]

Rorty, A. O. (1985) Self-deception, akrasia and irrationality. In: *The multiple self*, ed. J. Elster, pp. 115–32. Cambridge University Press. [RK]

Rosenhan, D. L. & Messick, S. (1966) Affect and expectation. *Journal of Personality and Social Psychology* 3:38–44. [ELK]

Rosenkranz, M. A., Jackson, D. C., Dalton, K. M., Dolski, I., Ryff, C. D., Singer, B. H., Muller, D., Kalin, N. H. & Davidson, R. J. (2003) Affective style and *in vivo* response: Neurobehavioral mechanisms. *Proceedings of the National Academy of Science* 100:11148–52. [aWvH]

Ross, M., Heine, S. J., Wilson, A. E. & Sugimori, S. (2005) Cross-cultural discrepancies in self-appraisals. *Personality and Social Psychology Bulletin* 31:1175–88. [SJH]

Rumelhart, D. E. (1984) Schemata and the cognitive system. In: *Handbook of social cognition, vol. 1*, ed. R. S. Wyer & T. K. Srull, pp. 161–88. Erlbaum. [PK]

Russell, B. (2009) *Unpopular Essays*. Routledge. [RM]

Sackeim, H. A. & Gur, R. C. (1979) Self-deception, other deception, and self-reported psychopathology. *Journal of Consulting and Clinical Psychology* 47:213–15. [aWvH]

Salmon, C. A. (2007) Parent–offspring conflict. In: *Family relations – An evolutionary perspective*, ed. C. A. Salmon & T. K. Shackelford, pp. 145–61. Oxford University Press. [UF]

Saucier, D. A., Miller, C. T. & Doucet, N. (2005) Differences in helping whites and blacks: A meta-analysis. *Personality and Social Psychology Review* 9:2–16. [aWvH]

Schmader, T. & Johns, M. (2003) Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology* 85:440–52. [aWvH]

Schmitt, D. P. & Buss, D. M. (1996) Mate attraction and competitor derogation: Context effects on perceived effectiveness. *Journal of Personality and Social Psychology* 70:1185–204. [aWvH]

Schmitt, D. P. & Buss, D. M. (2001) Human mate poaching: Tactics and temptations for infiltrating existing relationships. *Journal of Personality and Social Psychology* 80:894–917. [aWvH]

Schwarz, N. & Skurnik, I. (2003) Feeling and thinking: Implications for problem solving. In: *The nature of problem solving*, ed. J. Davidson & R. J. Sternberg, pp. 263–92. Cambridge University Press. [HM]

Schweitzer, M. E., Hershey, J. & Bradlow, E. (2006) Promises and lies: Restoring violated trust. *Organizational Behavior and Human Decision Processes* 101:1–19. [HM, aWvH]

Seamon, J. G., Williams, P. C., Crowley, M. J., Kim, I. J., Langer, S. A., Orne, P. J. & Wishengrad, D. L. (1995) The mere exposure effect is based on implicit memory: Effects of stimulus type, encoding conditions, and number of exposures on recognition and affect judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21:711–21. [aWvH]

Searcy, W. A. & Nowicki, S. (2005) *The evolution of animal communication: Reliability and deception in signaling systems.* Princeton University Press. [SWG]

Sedikides, C. & Alicke, M. D. (in press) Self-enhancement and self-protection motives. In: *Oxford handbook of motivation*, ed. R. Ryan. Oxford University Press. [rWvH]

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

Sedikides, C., Gaertner, L. & Vevea, J. L. (2005) Pancultural self-enhancement reloaded: A meta-analytic reply to Heine (2005). *Journal of Personality and Social Psychology* 89:539–51. [SJH, aWvH]

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* 10:59–67. [SJH]

Sedikides, C. & Gregg, A. P. (2008) Self-enhancement: Food for thought. *Perspectives on Psychological Science* 3:102–16. [rWvH]

Segerstrom, S. C. (2007) Optimism and resources: Effects on each other and on health over 10 years. *Journal of Research in Personality* 41:772–86. [aWvH]

Segerstrom, S. C. & Sephton, S. E. (2010) Optimistic expectancies and cell-mediated immunity: The role of positive affect. *Psychological Science* 21:448–55. [aWvH]

Shamir, B., House, R. J. & Arthur, M. B. (1993) The motivational effects of charismatic leadership: A self-concept based concept. *Organizational Science* 4:577–94. [aWvH]

Shank, R. & Abelson, R. (1977) *Scripts, plans, goals and understanding.* Erlbaum. [PK]

Shapiro, D. (1996) On the psychology of self-deception – truth-telling, lying and self-deception. *Social Research* 63:785–800. [RM]

Sherman, D. K. & Cohen, G. L. (2006) The psychology of self-defense: Self-affirmation theory. In: *Advances in experimental social psychology, vol. 38*, ed. M. P. Zanna, pp. 183–242. Elsevier Academic Press. [arWvH]

Sherman, D. K., Cohen, G. L., Nelson, L. D., Nussbaum, A. D., Bunyan, D. P. & Garcia, J. (2009) Affirmed yet unaware: Exploring the role of awareness in the process of self-affirmation. *Journal of Personality and Social Psychology* 97:745–64. [aWvH]

Sloman, S. A. (1996) The empirical case for two systems of reasoning. *Psychological Bulletin* 119(1):3–22. [KF]

Slovenko, R. (1999) Testifying with confidence. *Journal of the American Academy of Psychiatry and the Law* 27:127–31. [aWvH]

Slusher, M. P. & Anderson, C. A. (1987) When reality monitoring fails: The role of imagination in stereotype maintenance. *Journal of Personality and Social Psychology* 52:653–62. [aWvH]

Smith, A. (1759/1982) *The theory of moral sentiments*, ed. D. D. Raphael & A. L. Macfie. Liberty Fund. (Original work published in 1759.) [ELK]

Snyder, M. L., Kleck, R. E., Strenta, A. & Mentzer, S. J. (1979) Avoidance of the handicapped: An attributional ambiguity analysis. *Journal of Personality and Social Psychology* 37:2297–306. [aWvH]

Solberg Nes, L. S. & Segerstrom, S. C. (2006) Dispositional optimism and coping: A meta-analytic review. *Personality and Social Psychology Review* 10:235–51. [aWvH]

Son Hing, L. S., Chung-Yan, G. A., Hamilton, L. K. & Zanna, M. P. (2008) A two-dimensional model that employs explicit and implicit attitudes to characterize prejudice. *Journal of Personality and Social Psychology* 94:971–87. [arWvH]

Spalding, L. R. & Hardin, C. D. (1999) Unconscious unease and self-handicapping: Behavioral consequences of individual differences in implicit and explicit self-esteem. *Psychological Science* 10:535–39. [aWvH]

Spencer, S. J., Fein, S., Wolfe, C. T., Fong, C. & Dunn, M. A. (1998) Automatic activation of stereotypes: The role of self-image threat. *Personality and Social Psychology Bulletin* 24:1139–52. [aWvH]

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G. & Wilson, D. (2010) Epistemic vigilance. *Mind & Language* 25(4):359–93. [HM]

Sroufe, L. A. (1989) Relationships, self, and individual adaptation. In: *Relationship disturbances in early childhood: A developmental approach*, ed. A. J. Sameroff & R. N. Emde, pp. 70–94. Basic. [MLB]

Stanovich, K. E. (1999) *Who is rational?: Studies of individual differences in reasoning.* Erlbaum. [KF]

Stanovich, K. E. (2004) *The robot's rebellion: Finding meaning in the age of Darwin.* University of Chicago Press. [KF]

Steele, C. M. (1988) The psychology of self-affirmation: Sustaining the integrity of the self. In: *Advances in experimental social psychology, vol. 21*, ed. L. Berkowitz, pp. 261–302. Academic Press. [aWvH]

Steele, C. M. & Liu, T. J. (1983) Dissonance processes as self-affirmation. *Journal of Personality and Social Psychology* 45:5–19. [LCE, aWvH]

Steinel, W. & De Dreu, C. K. W. (2004) Social motives and strategic misrepresentation in social decision making. *Journal of Personality & Social Psychology* 86:419–34. [aWvH]

Sternberg, R. J. & Kolligan, J., eds. (1990) *Competence considered.* Yale University Press. [LCE]

Stevens, J. R. & Hauser, M. D. (2004) Why be nice? Psychological constraints on the evolution of cooperation. *Trends in Cognitive Sciences* 8:60–65. [AP]

Stich, S. (1983) *From folk psychology to cognitive science: The case against belief.* Bradford. [RK]

Stich, S. (1990) *The fragmentation of reason.* MIT Press. [AP]

Stiff, J. B., Kim, H. J. & Ramesh, C. N. (1992) Truth biases and aroused suspicion in relational deception. *Communication Research* 19:326–45. [AV]

Stouten, J., De Cremer, D. & van Dijk, E. (2006) Violating equality in social dilemmas: Emotional and retributive reactions as a function of trust, attribution, and honesty. *Personality and Social Psychology Bulletin* 32:894–906. [aWvH]

Su, J. C. & Oishi, S. (2010) *Culture and self-enhancement. A social relation analysis.* Unpublished manuscript. [SJH]

Suddendorf, T., Addis, D. R. & Corballis, M. C. (2009) Mental time travel and the shaping of the human mind. *Philosophical Transactions of the Royal Society B-Biological Sciences* 364(1521):1317–24. [TS]

Suddendorf, T. & Corballis, M. C. (1997) Mental time travel and the evolution of the human mind. *Genetic Social and General Psychology Monographs* 123:133–67. [TS]

Suddendorf, T. & Corballis, M. C. (2007) The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences* 30(3):299–351. [TS]

Sugiyama, L. S., Tooby, J., Cosmides, L. (2002) Cross-cultural evidence of cognitive adaptations for social exchange among the Shiwiar of Ecuadorian Amazonia. *Proceedings of the National Academy of Sciences* 99:11537–45. Available at: http://www.uoregon.edu/~sugiyama/docs/cogadaptsocex.pdf. [LCE]

Suh, E., Diener, E., Oishi, S. & Triandis, H. C. (1998) The shifting basis of life satisfaction judgments across cultures: Emotions versus norms. *Journal of Personality and Social Psychology* 74:482–93. [SJH]

Sundie, J. M., Kenrick, D. T., Griskevicius, V., Tybur, J., Vohs, K. & Beal, D. J. (in press) Peacocks, Porsches, and Thorsten Veblen: Conspicuous consumption as a sexual signaling system. *Journal of Personality and Social Psychology.* [DTK]

Surbey, M. K. (2004) Self-deception: Helping and hindering personal and public decision making. In: *Evolutionary psychology, public policy, and personal decisions*, ed. C. B. Crawford & C. A. Salmon, pp. 117–44. Erlbaum. [NH]

Swann, W. B., Jr. (1983) Self-verification: Bringing social reality into harmony with the self. In: *Social psychological perspectives on the self, vol. 2*, ed. J. Suls & A. G. Greenwald, pp. 33–66. Erlbaum. [MLB]

Swann, W. B., Jr. (1987) Identity negotiation: Where two roads meet. *Journal of Personality and Social Psychology* 53:1038–51. [MLB]

Swann, W. B., Jr. (in press) Self-verification theory. In: *Handbook of theories of social psychology*, ed. P. Van Lang, A. Kruglanski & E. T. Higgins. Sage. [MLB, aWvH]

Swann, W. B., Jr. & Bosson, J. (2008) Identity negotiation: A theory of self and social interaction. In: *Handbook of personality psychology: Theory and research*, ed. O. John, R. Robins & L. Pervin, pp. 448–71. Guilford Press. [MLB]

Swann, W. B., Jr., Chang-Schneider, C. & McClarty, K. (2007) Do our self-views matter? Self-concept and self-esteem in everyday life. *American Psychologist* 62:84–94. [MLB]

Taylor, S. E. (1989) *Positive illusions: Creative self-deception and the healthy mind.* Basic Books. [RK]

Taylor, S. E. (1989/1995) *Mit Zuversicht: Warum positive Illusionen für uns so wichtig sind.* Rowohlt. (Original work published in 1989.) [UF]

Taylor, S. E. & Armor, D. A. (1996) Positive illusions and coping with adversity. *Journal of Personality* 64:873–98. [LCE, SJH]

Taylor, S. E. & Brown, J. D. (1988) Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin* 103:193–210. [LCE, DCF, SJH, arWvH]

Taylor, S. E. & Brown, J. D. (1994) Positive illusions and well being revisited: Separating fact from fiction. *Psychological Bulletin* 116:21–27. [DCF, aWvH]

Thornhill, R. & Gangestad, S. W. (2009) *The evolutionary biology of human female sexuality*. Oxford University Press. [aWvH]

Tice, D. M., Butler, J. L., Muraven, M. B. & Stillwell, A. M. (1995) When modesty prevails: Differential favorability of self-presentation to friends and strangers. *Journal of Personality and Social Psychology* 69:1120–38. [SJH]

Tomaka, J., Blascovich, J. & Kelsey, R. M. (1992) Effects of self-deception, social desirability, and repressive coping on psychophysiological reactivity to stress. *Personality and Social Psychology Bulletin* 18:616–24. [aWvH]

Tooby, J. & Cosmides, L. (1992) The psychological foundations of culture. In: *The adapted mind: Evolutionary psychology and the generation of culture*, ed. J. H. Barkow, L. Cosmides & J. Tooby, pp. 19–136. Oxford University Press. [JYH, RK]

Trivers, R. (1974) Parent–offspring conflict. *American Zoologist* 14: 49–64. [UF]

Trivers, R. (1976/2006) Foreword. In: *The selfish gene*, R. Dawkins, pp. 19–20. Oxford University Press. (Original work published in 1976.) [EF, HJL, AP, SP, AT, arWvH]

Trivers, R. (1985) Deceit and self-deception. In: *Social evolution*, pp. 395–420. Benjamin/Cummings. [DD, SWG, HJL, arWvH]

Trivers, R. (1991) Deceit and self-deception: The relationship between communication and consciousness. In: *Man and beast revisited*, ed. M. Robinson & T. L. Tiger, pp. 175–91. Smithsonian Press. [DD]

Trivers, R. (2000) The elements of a scientific theory of self-deception. *Annals of the New York Academy of Sciences* 907:114–31. [AP, HJL, aWvH]

Trivers, R. (2009) Deceit and self-deception. In: *Mind the gap*, ed. P. Kappeler & J. Silk, pp. 373–93. Springer-Verlag. [arWvH]

Troisi, A. & McGuire, M. T. (1990) Deception in somatizing disorders. In: *Psychiatry: A world perspective, vol. 3*, ed. C. N. Stefanis, pp. 973–78. Elsevier. [AT]

Trope, Y. & Neter, E. (1994) Reconciling competing motives in self-evaluation: The role of self-control in feedback seeking. *Journal of Personality and Social Psychology* 66:646–57. [aWvH]

Valdesolo, P. & DeSteno, D. A. (2007) Moral hypocrisy: Social groups and the flexibility of virtue. *Psychological Science* 18:689–90. [LCE]

Valdesolo, P. & DeSteno, D. A. (2008) The duality of virtue: Deconstructing the moral hypocrite. *Journal of Experimental Social Psychology* 44:1334–38. [arWvH]

Veltkamp, M., Aarts, H. & Custers, R. (2008) Perception in the service of goal pursuit: Motivation to attain goals enhances the perceived size of goal-instrumental objects. *Social Cognition* 26:720–36. [JYH]

Vohs, K. D. & Heatherton, T. F. (2001) Self-esteem and threats to self: Implications for self-construals and interpersonal perceptions. *Journal of Personality and Social Psychology* 81:1103–18. [SJH]

Vohs, K. D. & Schooler, J. W. (2007) The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science* 19:49–54. [aWvH]

Voland, E. (2007) We recognize ourselves as being similar to others: Implications of the "social brain hypothesis" for the biological evolution of the intuition of freedom. *Evolutionary Psychology* 5:442–52. [UF]

von Hippel, W. & Gonsalkorale, K. (2005) "That is bloody revolting!" Inhibitory control of thoughts better left unsaid. *Psychological Science* 16:497–500. [aWvH]

von Hippel, W., Lakin, J. L. & Shakarchi, R. J. (2005) Individual differences in motivated social cognition: The case of self-serving information processing. *Personality and Social Psychology Bulletin* 31:1347–57. [aWvH]

Vrij, A. (2000) *Detecting lies and deceit*. Wiley. [aWvH]

Vrij, A. (2004) Why professionals fail to catch liars and how they can improve. *Legal Criminology Psychology* 9:159–81. [aWvH]

Vrij, A. (2008) *Detecting lies and deceit: Pitfalls and opportunities, 2nd ed.* Wiley. [AV]

Vrij, A., Fisher, R., Mann, S. & Leal, S. (2006) Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences* 10:141–42. [aWvH]

Vrij, A., Granhag, P. A., Mann, S. & Leal, S. (in press) Outsmarting the liars: Towards a cognitive lie detection approach. *Current Directions in Psychological Science.* [AV]

Vrij, A. & Mann, S. (2005) Police use of nonverbal behavior as indicators of deception. In: *Applications of nonverbal communication*, ed. R. E. Riggio & R. S. Feldman, pp. 63–94. Erlbaum. [aWvH]

Vrij, A., Mann, S., Fisher, R., Leal, S., Milne, B. & Bull, R. (2008) Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. *Law and Human Behavior* 32: 253–65. [AV]

Vrij, A., Mann, S., Leal, S. & Fisher, R. (2010) "Look Into My Eyes": Can an instruction to maintain eye contact facilitate lie detection? *Psychology, Crime, & Law* 16:327–48. [AV]

Weary, G., Tobin, S. J. & Edwards, J. E. (2010) The causal uncertainty model revisited. In: *Handbook of the uncertain self*, ed. R. M. Arkin, K. C. Oleson & P. J. Carroll, pp. 78–100. Psychology Press. [rWvH]

Wegner, D. M. (2005) Who is the controller of controlled processes? In: *The new unconscious*, ed. R. Hassin, J. S. Uleman & J. A. Bargh, pp. 19–36. Oxford University Press. [RK]

Weinstein, N. D. (1980) Unrealistic optimism about future life events. *Journal of Personality and Social Psychology* 39:806–20. [aWvH]

Westbrook, R. A. (1980) Consumer satisfaction as a function of personal competence/efficacy. *Journal of the Academy of Marketing Science* 8:427–37. [aWvH]

Westen, D., Blagov, P. S., Harenski, K., Kilts, C. & Hamann, S. (2006) Neural bases of motivated reasoning: An fMRI study of emotional constraints on partisan political judgment in the 2004 U.S. presidential election. *Journal of Cognitive Neuroscience* 18:1947–58. [aWvH]

Whiten, A. & Byrne, R. W. (1988) Tactical deception in primates. *Behavioral and Brain Sciences* 11:233–73. [TS]

Whitson, J. A. & Galinsky, A. D. (2008) Lacking control increases illusory pattern perception. *Science* 322:115. [aWvH]

Williams, L. E. & Bargh, J. A. (2008) Experiencing physical warmth promotes interpersonal warmth. *Science* 322:606–607. [rWvH]

Wills, T. A. (1981) Downward comparison principles in social psychology. *Psychological Bulletin* 90:245–71. [aWvH]

Wilson, T. D. & Gilbert, D. (2003) Affective forecasting. *Advances in Experimental Social Psychology* 35:345–411. [aWvH]

Wilson, T. D., Lindsey, S. & Schooler, T. (2000) A model of dual attitudes. *Psychological Review* 107:101–26. [aWvH]

Wilson, T. D., Wheatley, T. P., Kurtz, J. L., Dunn, E. W. & Gilbert, D. T. (2004) When to fire: Anticipatory versus post-event reconstrual of uncontrollable events. *Personality and Social Psychology Bulletin* 30:340–51. [aWvH]

World Health Organization (2009) WHO Statistical Information System (WHOSIS). Available at: http://www.who.int/whosis. [aWvH]

Wrangham, R. (1999) Is military incompetence adaptive? *Evolution and Human Behavior* 20:3–17. [DMB]

Zahavi, A. & Zahavi, A. (1997) *The handicap principle: A missing piece of Darwin's puzzle.* Oxford University Press. [UF]

Zaragoza, M. S. & Mitchell, J. (1996) Repeated exposure to suggestion and the creation of false memories. *Psychological Science* 7:294–300. [aWvH]

Zaragoza, M. S., Payment, K. E., Ackil, J. K., Drivdahl, S. B. & Beck, M. (2001) Interviewing witnesses: Forced confabulation and confirmatory feedback increase false memories. *Psychological Science* 12:473–77. [aWvH]

Zarnoth, P. & Sniezek, J. A. (1997) The social influence of confidence in group decision making. *Journal of Experimental Social Psychology* 33:345–66. [aWvH]

Zhong, C. B. & Leonardelli, G. J. (2008) Cold and lonely: Does social exclusion literally feel cold? *Psychological Science* 19:838–42. [rWvH]

Zuckerman, M., Koestner, R. & Alton, A. O. (1984) Learning to detect deception. *Journal of Personality and Social Psychology* 46:519–28. [aWvH]