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