Protesting too much: Self-deception and self-signaling

Ryan McKay[1], Danica Mijović-Prelec[2] and Dražen Prelec[2,3]

[1] Department of Psychology, Royal Holloway, University of London
[2] Sloan School and Neuroeconomics Center, MIT
[3] Department of Economics, and Department of Brain & Cognitive Sciences, MIT

Abstract

von Hippel and Trivers 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, citing Labouchere on Gladstone, 'Unpopular Essays'*

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 and Trivers (hereafter 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 sub-agents 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, e.g., reproductive quality (Grafen, 1990). The sender plays 'offense' by emitting signals that exaggerate his qualities; 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 takes 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* (Quattrone & Tversky, 1984; Bodner & Prelec, 2003). 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, e.g., about ones' self-esteem, sexuality, future prospects, family relationships, etc., 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 sub-selves, 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.

Acknowledgements: 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.

References

Bodner, R. & Prelec, D. (2003). Self-signaling in a neo-Calvinist model of everyday decision making. In *Psychology of economic decisions*, Vol. I. (eds I. Brocas & J. Carillo). London, UK: Oxford University Press; **Elster, J**. (1999) *Alchemies of the mind: Rationality and the emotions*. Cambridge, UK: Cambridge University Press. **Funkouser, E**. (2005) Do the self-deceived get what they want? *Pacific Phil. Q., 86*, 295–312. **Grafen, A.** (1990). Biological signals as handicaps. *Journal of Theoretical Biology, 144*, 517-546; **McKay, R. & Dennett, D.** (2009). The evolution of misbelief. *Behavioral and Brain Sciences, 32*(6), 493-561; **Mijović-Prelec, D. & Prelec, D.** (2010). Self-deception as self-signaling: A model andexperimental evidence. *Philos Trans R Soc Lond B Biol Sci., 365*(1538), 227-40; **Quattrone, G. & Tversky, A.** (1984). Causal versus diagnostic contingencies: On self-deception and on the voter's illusion. *J Pers Soc Psychol, 46*, 237–248. **Shapiro, D**. (1996). On the psychology of self-deception—truth-telling, lying and self-deception. *Social Res., 63*, 785–800.