Representations and Decision Rules in the Theory of Self-Deception

<Commentary on von Hippel & Trivers [BBS 34(1) 2011] – revised by CE>

<Running head: *Commentary/*The evolution and psychology of self-deception>

<CT>Representations and decision rules in the theory of self-deception

<CA>Steven Pinker. Department of Psychology, Harvard University.

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

<C-AB>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.

<C-Text begins>

> "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), 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*.

<C-text ends>

Trivers, R. 1976. Foreword. In R. Dawkins, *The selfish gene*. New York: Oxford University Press.

Epley, N., & Whitchurch, E. (2008). Mirror, mirror on the wall: Enhancement in self

recognition. *Personality and Social Psychology Bulletin*, *34*, 1159–1170.