PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2010) **365**, 227–240
doi:10.1098/rstb.2009.0218

Self-deception as self-signalling: a model and experimental evidence

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

[1]*Sloan School of Management and Neuroeconomics Center,* [2]*Department of Economics, and* [3]*Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, USA*

Self-deception has long been the subject of speculation and controversy in psychology, evolutionary biology and philosophy. According to an influential 'deflationary' view, the concept is an over-interpretation of what is in reality an instance of motivationally biased judgement. The opposite view takes the interpersonal deception analogy seriously, and holds that some part of the self actively manipulates information so as to mislead the other part. Building on an earlier self-signalling model of Bodner and Prelec, we present a game-theoretic model of self-deception. We propose that two distinct mechanisms collaborate to produce overt expressions of belief: a mechanism responsible for action selection (including verbal statements) and an interpretive mechanism that draws inferences from actions and generates emotional responses consistent with the inferences. The model distinguishes between two modes of self-deception, depending on whether the self-deceived individual regards his own statements as fully credible. The paper concludes with a new experimental study showing that self-deceptive judgements can be reliably and repeatedly elicited with financial incentives in a categorization task, and that the degree of self-deception varies with incentives. The study also finds evidence of the two forms of self-deception. The psychological benefits of self-deception, as measured by confidence, peak at moderate levels.

Keywords: self-deception; self-signalling; over-optimism; motivated reasoning; behavioural economics; multiple selves

1. INTRODUCTION

Any definition of self-deception is likely to be controversial, so we start with an actual incident, witnessed by one of us a number of years ago.

It was sherry hour, a casual gathering of a few doctoral students, all good friends. A veteran student had just finished a lengthy disquisition on her recent scholarly progress and post-graduation aspirations. Warming to the topic, she asserted that she would complete her dissertation within the year. 'Are you kidding, you're never going to finish it,' remarked another with a smile, his guard down on account of the drink. The comment was not unjust; the student had nothing to show for some half dozen years in the programme. Yet, it hit the mark a bit too well, and in an instant its author found himself wiping the contents of a full glass of sherry from his face and shirt.

Like many true events, this one allows multiple interpretations. Two are relevant here, as picking out two modes of self-deception. To begin with, one could take the student's claim at face value: she is convinced that the dissertation will be completed on schedule, all evidence to the contrary. In the construction of this conviction, periodic extravagant affirmations played a key role, substituting for the absence of actual progress. Words became evidence, following the logic—'if it wasn't true, then why would I say it?' (and if true, how perverse to deny it?).

This would be one interpretation. On a second interpretation, the student understood very well that her scholarly prospects were dim. Yet, almost as a matter of personal ritual, she felt compelled to state a contrary belief, and perhaps for the moment she did entertain it. However, the belief was fragile, easily punctured by the offhand remark. She expressed conviction, but did not experience conviction, not in an authentic way. Tossing the sherry was a way of saying—'Don't treat me like a fool, I have an idea how things stand, but why must you spell it out'.

Regardless of which reading is more faithful to the actual event, each refers to a genuine psychological possibility, requiring explanation. Here we present a formal theory of self-deception that relies on a single psychological mechanism—self-signalling—to generate self-deception in both of these alternative modes. The theory distinguishes among three levels of belief: *deep belief*, *stated belief* and *experienced belief*. Deep belief drives action, including overt statements of belief; experienced belief determines the emotional state following the statement. When stated belief does not match deep belief, we have attempted self-deception. The attempt succeeds if experienced belief matches stated belief. It misfires to the extent that the person discounts her own statement, with emotional response falling short of what might be expected on the basis of the words alone.

* Author for correspondence (mijovic@mit.edu).

One contribution of 12 to a Theme Issue 'Rationality and emotions'.

Deep beliefs are presumed to be largely inaccessible. This psychological opacity endows statements with self-signalling value, and creates a motive for self-deception. The formal model casts these assumptions into a signalling game, leading to predictions about how incentives and self-knowledge jointly determine whether self-deception is attempted and whether it succeeds. The two modes of self-deception arise as consequences of different levels of psychological awareness about the self-deception mechanism. According to the model, awareness should reduce the credibility of stated beliefs, as one might expect, but it need not eliminate the gap between stated and deep beliefs. In the full-awareness case, a person may be compelled to utter self-deceptive statements even though they have no effect on experienced belief (Shapiro 1996). This would correspond to the ritualistic interpretation of the earlier incident.

We introduce the model in §§3 and 4. It extends the self-signalling model developed by R. Bodner & D. Prelec (Bodner & Prelec 1995, 2003), and is also broadly related to recent economic models of intrapersonal psychological interactions (Benabou & Tirole 2002, 2004; Bernheim & Thomadsen 2005; Brocas & Carrillo 2008). This is followed by a new experimental test, presented in §5. In the study, subjects are asked to provide repeated assessments of their own performance in a competitive decision task. Self-assessments cannot affect actual performance, but can affect the subjects' expectations of winning the contest, leading potentially to self-deception. Consistent with the model, we find that financial incentives influence the degree of self-deception, and that the benefits of self-deception, as measured by confidence ratings, accrue to subjects exhibiting an intermediate level of self-deception, who are presumably unaware of their self-deception.

2. BACKGROUND

Self-deception is an ancient subject. Classical philosophers, beginning with Aristotle and St Augustine, treated it at length, focusing especially on the connections between self-deception, morality and the emotions (Elster 1999). Two thousand years of speculation and commentary have failed to exhaust the topic or forge a consensus interpretation. The notion of self-deception remains integral to Western understanding of human character, as shown by religious moralistic literature, drama and fiction, and by secular world-views such as Marxism, psychoanalysis and atheism, which promise to strip the scales from our self-deceiving eyes.

The modern scholarly literature on self-deception is similarly large, and rife with controversy. According to Gur & Sackeim's (1979) influential formulation, a self-deceived individual (i) holds two contradictory beliefs, *p* and *not-p*, (ii) holds them simultaneously, (iii) is unaware of holding one of the beliefs, and (iv) is motivated to remain unaware of that belief. There is an analogy here to inter-personal deception, where one party (the deceiver) knows or believes something and has a reason for inducing opposite beliefs in another party (the deceived). The interpersonal analogy highlights the distinction between those false beliefs that are arrived at by chance or through error, and those for which some intentional agency is responsible.

Moving from the inter-personal to the intra-personal level, the definition raises two paradoxes (Mele 1997). The static paradox concerns the state of mind of the self-deceived individual: how can he hold two incompatible beliefs, *p* and *not-p*? The dynamic paradox concerns the process of becoming self-deceived: how can a person intentionally acquire a belief or remain unaware of a belief? Recognition that one is generating or suppressing beliefs would seem to destroy the effectiveness of the effort itself.

An influential 'deflationary' response to these two paradoxes has been to deny both, and to assimilate self-deception to the general category of motivationally biased judgements (Mele 1997, 1998). On this view, the interpersonal metaphor is misguided, and most if not all self-deception is not intentional. The opposite view takes the interpersonal analogy seriously, and holds that some part of the self actively manipulates information so as to mislead the other part. The psychoanalytic tradition falls squarely in this camp.

Some manifestations of self-deception lend themselves naturally to deflationary interpretations. Consider the finding that most people rate themselves as superior on virtually any desirable characteristic (Brown & Dutton 1995; Dunning & Hayes 1996). For example, 94 per cent of university professors rate themselves as above average in professional accomplishment relative to their peers (Gilovich 1991). Such findings may only show that most people give special weight to criteria favouring their own case. Once the self-serving bias is in place, the better-than-average conclusion can emerge even if specific pieces of evidence are evaluated in an impartial way. At no moment is it necessary for the individual to believe both *p* and *not-p*. Indeed, even rational inference can give rise to the better-than-average effect in some circumstances (J.-P. Benoit & J. Dubra 2009, unpublished data).

Self-serving beliefs can also be generated ad hoc through contrived cover stories, as shown by Kunda in a series of elegant demonstrations (Kunda 1990). In one case, subjects were asked to evaluate the credibility of a (fake) scientific study linking coffee consumption and breast cancer. Female subjects who also happened to be heavy coffee drinkers were especially critical of the study, and the least persuaded by the presented evidence. This is only a sample of the literature documenting how evidence consistent with the favoured hypothesis receives preferential treatment (Ditto & Lopez 1992; Dawson *et al*. 2002; Norton *et al*. 2004; Balcetis & Dunning 2006). Moreover, this phenomenon occurs largely outside of awareness (Kunda 1987; Pyszczynski & Greenberg 1987; Pronin *et al*. 2004). No one questions the reality of motivated reasoning or perception. The critical issue is whether motivational biases are sufficient to explain self-deception.

From the perspective of the 'real self-deception' side, motivated reasoning explanations seem to ignore three critical aspects of self-deception. First, they do not account for the strong emotions generated

when self-deceptive beliefs are challenged. What prevents the self-deceived from enjoying their false beliefs with smug complacency? There is no explanation for the brittle quality of self-deception (Audi 1985; Bach 1997),[1] and the defensiveness associated with a self-deceptive personality.

Second, the motivated reasoning view denies the special significance of mistaken beliefs about the *self*. Yet, the concept of self-deception and the most salient examples of self-deception have historically been restricted to beliefs about the self (Holton 2000). To reinforce this intuition, let us suppose that the student in our story had not been talking about the prospects for her dissertation but about some impersonal issue. Let us say that she believes that the 1969 Apollo moon landing is a gigantic hoax, and that she derived these views from a highly motivated interpretation of the evidence. In that case, we might call her biased, but it would be odd to accuse her of self-deception.

Finally, and perhaps most tellingly, under the motivated reasoning view it is hard to make sense of the notion of failed self-deception, a point made by Funkhouser in his provocatively entitled article, 'Do the self-deceived get what they want?' (Funkouser 2005). If self-deception is merely the manifestation of a bias, then the self-deceived will by definition get what they want. A bias that misfires, i.e. one that leaves beliefs unchanged, is no bias at all.

In their original study of self-deception, Gur and Sackeim attempted to demonstrate the coexistence of two incompatible beliefs by exploiting the fact that people dislike the recorded sound of their own voice. In their experiment, subjects heard fragments of speech and were asked to identify the speaker (Gur & Sackeim 1979). Non-recognition of own voice was often accompanied by physiological indications (galvanic skin response) suggestive of detection. Hence, the verbal assessment—'this is not my own voice'—was in conflict with the physiologically based assessment—'this is indeed my own voice'.

This interpretation has been criticized on grounds that physiological signs do not necessarily rise to the level of belief (Mele 1997). Similar objections were raised by Mele against arguments from blindsight cases (the phenomenon where a patient claims blindness but is able to detect visual stimuli above chance; Weisenkrantz 1986). An ideal demonstration would be one where a *single* voluntary response conveys two incompatible propositions. A neuropsychological case study indicates how this may be done in principle (Mijovic-Prelec *et al*. 1994). The patient in question suffered from unilateral visual neglect following a right hemisphere stroke, and to all appearances was unaware of details in the left visual space. However, under experimentally controlled conditions, when asked to judge the presence or absence of a randomly placed target, his verbal denial of left-side targets was suspiciously fast, much faster than his tentative response to null trials when no target was present—the two response time distributions were essentially non-overlapping. The speed of response matched the speed of detection of right-side targets, showing that the left-side target was noticed and that the patient realized the futility of searching for it elsewhere. A single response thus conveyed two contradictory propositions simultaneously: one voluntary response dimension (search time) conveyed *p*, while the other, equally voluntary, semantic dimension conveyed *not-p*.[2]

Among studies with normal human subjects, an experiment by Quattrone & Tversky (1984) provides perhaps the cleanest challenge to deflationary accounts. Their experiment took place at a medical facility, adding credibility to the unusual cover story. Subjects were first asked to keep their hand submerged in a container of cold water until they could no longer tolerate the pain. This was followed by a debriefing, which explained that a certain inborn heart condition could be diagnosed by the effect of exercise on cold tolerance. The consequences of this condition included a shorter lifespan and reduced quality of life. Some subjects were told that having a bad heart would increase cold tolerance, while the others were told the opposite. Backing this up were charts showing different lifespan distributions associated with the two types of heart. Having absorbed this information, subjects were put on an exercycle for a minute after which they repeated the same cold water tolerance test. The majority showed changes in tolerance on the second cold trial in the direction correlated with 'good news'. In effect, they were cheating on their own diagnosis.

Apart from the Quattrone–Tversky experiment, several other studies provide support for self-signalling. For example, respondents adjust answers to personality questionnaires so as to obtain a profile diagnostic of a good outcome (Kunda 1990; Sanitioso *et al*. 1990; Dunning *et al*. 1995); they also adjust problem solving strategies (Ginossar & Trope 1987), and charitable pledges in a diagnostically favourable direction (Bodner 1995). In a recent paper, Dhar & Wertenbroch assess self-signalling directly in the context of consumer choices between goods that could be perceived as virtues (apples, organic pasta) or vices (cookies, steak) (R. Dhar & K. Wertenbroch 2007, unpublished data). They manipulate whether the choice set is homogeneous (containing only vice or only virtues) or mixed, the idea being that selections from mixed sets are diagnostic for self-control, whereas selections from a homogeneous set are not diagnostic. Consistent with the self-signalling hypothesis, they find that consumers are willing to pay relatively more for a virtuous good in a mixed set, when its selection would also generate positive diagnostic utility, but relatively more for a vice good in a homogeneous set, when its selection would avoid negative diagnostic utility.

3. SELF-DECEPTION AS SELF-SIGNALLING

One can attempt to provide a motivated reasoning interpretation of self-signalling. Thus, for example, Mele (1997) states that

> One can hold (i) that sincere deniers (in the Quattrone–Tversky experiment), due to a desire to live a long, healthy life, were motivated to believe that they had a healthy heart; (ii) that this motivation (in conjunction with a belief that an upward/downward shift in tolerance would constitute evidence for the favoured

> proposition) led them to try to shift their tolerance; and (iii) that this motivation also led them to believe that they were not purposely shifting their tolerance . . .

According to this view, the trying and the false belief that one is not trying are both motivated by the desire for good news, but it does not follow that either the trying or the belief is intentional. However, to assimilate the results of Quattrone and Tversky to this deflationary point of view, one has to expand the powers ascribed to the concept of motivation. The mechanism responsible for trying to shift tolerance must register the difference between the natural tolerance level, corresponding to an absence of trying, and the shifted tolerance level obtained as a result of the trying. In other words, it must register both the true and the fake tolerance. It must not only be able to bias the interpretation of evidence, it must also be able to manufacture the evidence itself.

There is clearly a need to explain how a person can simultaneously try to do something and to be unaware of so trying. We will shortly provide an interpretation of self-deception that treats it as a special case of a self-signalling. Because the model draws on Bayesian game theory, we first say a few words about this modelling technology.

The basic building block is a rational agent, defined by preferences (utility function), beliefs (subjective probabilities), and an action or choice set. Faced with alternative actions, the agent is presumed to select the one that maximizes expected utility. New information is incorporated into his beliefs according to Bayes' rule. Strategic interactions among agents are modelled with Bayesian game theory. The standard solution concept here is the Nash equilibrium, which characterizes mutual consistency among different players' strategies. Briefly, strategies are in equilibrium if every player is maximizing expected utility, on the assumption that other players are following strategies specified by the equilibrium.

With these tools one can model self-deception in roughly three ways. The first is to adjust the Bayesian model of belief formation. For example, in a model by G. Mayraz (2009, unpublished data) subjective probabilities of outcomes are inflated or reduced in direct proportion to their utilities. In effect, the valuation of an uncertain outcome is treated as if it were an additional piece of information bearing on the likelihood of the outcome. The second is to treat the individual as a series of temporal selves, with earlier selves manipulating the beliefs of the later selves, e.g. by suppressing information directly or by exploiting future selves' recall of earlier actions but not of the motives that gave rise to those actions (Caplin & Leahy 2001; Benabou & Tirole 2002, 2004; Bernheim & Thomadsen 2005; Koszegi 2006*a*,*b*; Gottlieb 2009). The third approach is to add psychological structure by partitioning the decisionmaker into several simultaneously interacting entities, which could be called selves or modules depending on how much true agency and self-awareness they have (Thaler & Shefrin 1981; Bodner & Prelec 2003; Brocas & Carrillo 2008; Fudenberg & Levine 2008).

The self-signalling model takes the behaviour revealed in the Quattrone and Tversky experiment as prototypical for self-deception. It was introduced by Bodner & Prelec (1995),[3] as a formal decision model for non-causal motivation, that is, motivation to generate actions that are diagnostic of good outcomes but that have no causal ability to affect those outcomes. With respect to our threefold classification, it is a psychological structure model, partitioning the decision maker into two collaborative entities, one responsible for action selection and the other responsible for action interpretation. We first provide a short summary of the original model and then discuss how it accounts for self-deception as a byproduct of the self-signalling process.

Self-signalling presumes the existence of an underlying characteristic that is (i) personally important, (ii) introspectively inaccessible, and (iii) potentially revealed through actions. We let the parameter θ represent this characteristic, with θ° indicating its actual value, x a possible outcome, and $u(x, \theta)$ the utility (reward or satisfaction) generated by the outcome x in the absence of any choice (i.e. a forced receipt of x). Uncertainty about θ is defined by a probability distribution, $p(\theta)$, which may be taken as the current self-image with respect to this characteristic. The value of the self-image is, in turn, determined by a second function, $v(\theta)$, which indicates how much pleasure or pain a person would feel from discovering true θ.

By intentionally choosing one outcome over others, a person learns something about his or her inaccessible characteristics. Hence, an action leads to an updating of the self-image, from $p(\theta)$ to $p(\theta|x)$. The change in self-image generates a second form of utility, called *diagnostic utility*: $\Sigma_\theta v(\theta)p(\theta|x) - \Sigma_\theta v(\theta)p(\theta)$, produced by replacing $p(\theta)$ with the updated $p(\theta|x)$. Diagnostic utility captures the extent to which one's own choice provides good or bad news about θ.

In the context of the Quattrone–Tversky experiment, θ would correspond to cold sensitivity, $u(x, \theta)$ to the (dis)pleasure associated with x seconds of exposure to cold water in context of the experimental instructions, and $v(\theta)$ to relief or anxiety associated with discovering one's cold sensitivity level. The total utility of choosing to hold one's hand in cold water for x seconds would then be the sum of outcome and diagnostic utility:

$$\left.\begin{aligned}&\text{Total utility} = \text{outcome utility} + \text{diagnostic utility},\\ &V(x, \theta^{\circ}) = u(x, \theta^{\circ}) + \lambda\Sigma_\theta v(\theta)p(\theta|x),\end{aligned}\right\} \tag{3.1}$$

where λ represents the weight of diagnostic utility. For notational simplicity we omit the constant term $-\Sigma_\theta v(\theta)p(\theta)$.

This is the model as stated in Bodner & Prelec (2003). However, in a self-deception scenario, what is at stake is a desired deep belief, e.g. that one's spouse is not having an affair. A husband may recognize certain problematic pieces of evidence but remain unsure about his own reading of them. Self-signalling is extended to such cases by treating one's

interpretation of evidence as the relevant inaccessible characteristic. Formally, θ_S is the probability of event S, and $u(x,\theta)$ an expectation over these events: $u(x, \theta) = \Sigma_S \theta_S U(x, S)$, where $U(x, S)$ is the utility of x if the event S occurs. The self-signalling equation then becomes[4]

$$V(x, \theta^{\circ}) = u(x, \theta^{\circ}) + \lambda \Sigma_{\theta} u(x, \theta) p(\theta|x). \quad (3.2)$$

In §5, we will apply this equation to the explicit financial incentives that are set up by our experiment. But first we need to complete the model by specifying $p(\theta|x)$.

4. TWO MODES OF SELF-DECEPTION

Previously, we had referred to the static and dynamic paradoxes of self-deception as central to the debate on the subject. The present model addresses the static paradox, on the coexistence of different beliefs, by postulating three levels of belief. Deep belief is associated with the inaccessible characteristic, whose actual value is θ^0. Stated belief is associated with the signalling action x, which either directly or indirectly expresses belief. Experienced belief is associated with the self-inference that follows the statement, $p(\theta|x)$.

Regarding the second, dynamic paradox, the model allows resolution in one of two ways, both of which have psychological plausibility. Observe that to complete the model we need to specify how $p(\theta|x)$ is derived from the choice and from $p(\theta)$. There are two endogenous rules for computing this distribution (Prelec & Bodner 2003), that is, rules that require no new parameters beyond the ones already given: $u(x, \theta)$ and $p(\theta)$. These rules generate the two variants of self-signalling.

The first, *face-value* rule assumes that the inferential mechanism operates without awareness of diagnostic motivation. The updated inferences, $p(\theta|x)$, are then based on the assumption that an action reveals the characteristic that maximizes only the outcome-utility component of total utility, ignoring the diagnostic component. Formally, this corresponds to the requirement that: $p(\theta|x) > 0$ implies: $u(x, \theta) \geq u(y, \theta)$, for any other choice y. That is, by choosing x I demonstrate deep beliefs such that x maximizes standard expected utility given these deep beliefs (with ties resolved by Bayes' rule). There is no discounting for diagnostic motivation. Diagnostic utility would be experienced as an unintentional byproduct of choice, not something that consciously affected choice.

The second *rational* rule, assumes full awareness about the self-signalling motive expressed in equation (3.1). $p(\theta|x)$ must then fully reflect the fact that actions are motivated by the anticipated inferences that flow from them. The signalling value of an ostensibly virtuous action is thereby reduced, or 'discounted' for diagnostic motivation. Formally, this corresponds to the requirement that: $p(\theta|x) > 0$ implies: $V(x, \theta) \geq V(y, \theta)$, for any other choice y. This carries to a logical conclusion the basic idea in self-perception theory (Bem 1972), namely, that the process of inferring underlying beliefs and desires from external behaviour is the same irrespective of whether the inferences pertain to someone else or to ourselves. Just as we might discount someone else's good behaviour as being due only to a desire to impress, so too we could discount our own behaviour for ulterior motives, according to the true interpretation assumption.[5]

Now we can indicate how the model resolves the dynamic paradox of self-deception. Recall that the paradox centres on the question whether the attempt to self-deceive destroys the credibility of the resulting belief. The paradox disappears if there is consistency between choice of x as a function of θ, and inference about θ as a function of observed x. This is what the equilibrium requires: the experienced beliefs $p(\theta|x)$ place positive probability only on those characteristics θ that maximize utility in light of $p(\theta|x)$—total utility for the rational variant, or outcome utility for the face-value variant. Regardless of which inferential rule is used, the beliefs experienced following the self-deceptive action will be consistent with the level of insight one has into one's tendency to self-deceive.

Self-deception is attempted whenever a person selects an action that does not maximize $u(x, \theta)$; however, the attempt is successful to the extent that $p(\theta|x)$ changes relative to $p(\theta)$. Which situation obtains depends crucially on awareness. With face-value interpretations, self-deception if attempted always succeeds. There is no discounting for self-deceptive motivation. In contrast, rational interpretations lead to a discounting of self-deceptive actions and statements. The crucial point, however, is that discounting does not eliminate the motive to self-signal, even in the extreme case where the self-deceptive statement has no self-credibility. Intuitively, this is because discounting affects positive and negative statements asymmetrically. Self-serving statements and predictions may be weakly believed or not believed at all, while the pessimistic may remain totally credible. For example: 'I will finish my dissertation on schedule,' may provide little reassurance that the dissertation will be finished. However, the opposite statement, that 'I will not finish my dissertation on schedule,' is clear evidence that the dissertation will indeed not be finished. In that case, a positive statement becomes mandatory not because it will be believed, but because of fear of the all-too-credible power of a negative statement. The function of the positive statement is not to convince but merely to preserve uncertainty about deep beliefs.

The self-signalling model allows, therefore, for two modes of self-deception. In the first mode, self-deceptive statements lead to changes in experienced belief, which is consistent with the traditional understanding of self-deception. In the second mode, self-deceptive statements have a ritualistic quality, leaving little or no trace on experienced belief. One might call this is an ideological or 'personal-correctness' mode, by analogy with political-correctness in the social domain.[6] A 'correctness regime'—whether personal or social/political—is characterized by rigid standards of expression and an intolerance of minor deviations from 'official belief'. But in neither case is public conformity solid evidence of underlying support or conviction. This residual uncertainty about deep belief may be the source of the defensiveness and touchiness associated with self-deception.

5. A SELF-DECEPTION EXPERIMENT

Much of the lay interest in self-deception derives from its alleged destructive consequences, from the feeling that people engage in self-deception in spite of the evident harm. Yet, the issue of cost is rarely addressed in experiments on self-deception, or in experiments on motivated reasoning (for an exception in the context of negotiations, see Babcock & Loewenstein 1997). It is generally considered sufficient to show that a particular manipulation biases judgements away from the truth. Subjects generally do not suffer any loss as a result of their experimentally induced self-deception.

A second unresolved issue is the link between awareness and self-deception. Indeed, the conceptual distinction between attempted and successful self-deception is not always observed. The impact of awareness is shown by an intriguing subsidiary result reported by Quattrone and Tversky. In the debriefing to the main experiment, they found that a significant minority of subjects acknowledged trying to influence the test after the fact, and were pessimistic about their heart condition. These subjects were evidently trying to self-deceive, but were not successful in the attempt.

These two issues motivate the study that we now describe. The specific experimental setting also hopes to capture some of the characteristics present in the dissertation incident. If one were to abstract from the details, these characteristics could be expressed as follows:

(i) There is a remote, important goal, such as the success of a research programme or dissertation.
(ii) Interim signs of progress arrive regularly. They are ambiguous and require explicit assessment.
(iii) There are costs to providing over-optimistic assessments, but these costs will only be revealed at the end of the enterprise.
(iv) While optimistic assessments of interim progress may provide momentary psychological relief, they do nothing to increase the chances that the goal will actually be achieved. There are no benefits of the 'self-fulfilling prophecy' kind.

Self-signalling implies that if the desire for good news is strong enough, it will bias interim assessments even if such biasing reduces overall chances of achieving the long-run goal. Moreover, we should observe the bias even if the judgemental task is novel, and incentives purely financial, i.e. unrelated to any chronic self-esteem concerns that subjects might bring to the laboratory. In other words, we should be able to generate self-deception repeatedly, reliably, and with arbitrary stimuli and incentives.

(a) *Procedure*

The subjects were 85 students at Princeton University, recruited through PLESS, the Princeton Experimental Economics Lab. The experiment involved many repetitions of a difficult categorization and prediction task; the 'large remote goal' was a chance of winning a $40 bonus if their overall performance was exceptionally good, according to criteria described below.

The experiment had two phases. In the first phase, they saw a series of 100 Korean characters on the computer screen and, following the presentation of each character, they were asked to classify it as more 'male-like' or 'female-like' in appearance. Individuals who had some familiarity with Korean characters were excluded from the study. The subjects therefore could only view the characters as abstract figures. They were given no special instructions about how to make this judgement, except to try to use their intuition and to take into account the entire configuration of the sign. Following each classification, they also rated their confidence on a five point scale.

To create incentives for careful responding, they were told (truthfully) that there is a correct answer for each sign, determined by the majority opinion of a group of previously tested subjects. They were told nothing about the composition, size, or incentives of this group, except that it was given the same instructions to use intuition and take into account the entire configuration of the sign.

Having been informed about the consensus-based answer key, subjects were told that they would receive $0.02 for each correct binary gender classification, correctness defined according to this key (there were no separate incentives for confidence ratings). In economic terms, the incentives corresponded to a 'beauty contest' game, where the winning answer is the one that matches majority opinion. Importantly, subjects never received any feedback on the accuracy of their classifications. While deliberately ambiguous, these instructions nevertheless generated considerable agreement in classifications (60–65% on average). The sorting largely conformed to conventional stereotypes; for example, 'female-like' signs were more likely to contain circles or numerous smaller diagonal strokes. Examples of signs eliciting high consensus are shown in figure 1.

The sole purpose of the initial classification in phase 1 was to create a subjective answer key, one for each participant, capturing that participant's best guess of how the peer group will assign gender. These answers could then be compared against subsequent classifications under incentive conditions designed to promote self-deception.

In phase II subjects encountered the same set of signs, in a different order, and were again asked to classify them according to gender (and rating confidence on the same five point scale). However, at the beginning of each trial, before the sign was displayed, subjects were asked to anticipate (by pressing the M or F key) whether the next sign would be more male-like or female-like. Because the signs arrived in random order, the gender of the next sign was unpredictable, and the subjects were forced to purely guess. As in phase I, each correct response (anticipation and classification) was credited with $0.02, with the total only revealed at the end of the experiment. In summary, a subject who somehow managed to respond with perfect accuracy would receive $2 in phase I, and $4 in phase II ($2 for the 100 perfect anticipations and $2 for the perfect classifications).

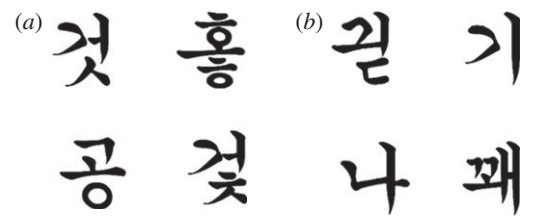

Figure 1. Examples of four signs judged to be more female-like (*a*) or more male-like (*b*) by a clear majority of respondents. There was no significant bias towards one or the other gender category in C1 or C2 classifications. However, there was a slight bias towards male anticipations: 51.9% for male, versus 48.1% for female, ($p < 0.001$ by χ^2-test). Subject's gender (41% female, 59% male) had no impact on classifications or anticipations.

This incentive structure was set up to generate a potential motive for self-deception. Suppose, for example, that a participant anticipated that the next sign would be 'male'. If the next sign had a more female-like shape, then the participant would face a dilemma, namely, whether to acknowledge the anticipation error or to reinterpret the sign as in fact looking more male.

To modulate the strength of the self-deception motive, we added to these piece-rate accuracy incentives an additional bonus of $40, which depended on overall performance relative to other subjects. The criteria for assigning the bonus differed across the two treatment groups. In the classification bonus group, the bonus was reserved for the top three subjects according to ex-post classification accuracy in phase II. In the anticipation bonus group, it was reserved for the top three subjects according to anticipation accuracy. As a result, the motive for self-deceptive, i.e. anticipation-confirming classifications was relatively weaker in the classification bonus condition and relatively stronger in the anticipation bonus condition.

We refer to this as a self-deception 'motive' rather than 'incentive' because the experiment in fact provides no financial incentives for self-deception. A subject that indulges in self-deceptive classifications will not thereby increase the actual accuracy of his or her anticipations, but will probably decrease the accuracy of classifications. Hence, a self-deceptive response pattern in the anticipation bonus condition purchases spurious psychological benefits (the feeling that one has a higher chance of winning the $40 bonus) with real financial costs.

(b) *Predictions*

These benefits would not appear in the analysis if one applied standard decision theory to the second classification decision. However, they do figure in the self-signalling equation. Suppose that the subject has anticipated that the next sign would be Male. Upon observing the sign, she has to decide whether to classify it as Male (m) or Female (f). The financial rewards of either response depend on actual gender, whether the sign (S) is male ($S = M$) or female ($S = F$), and are shown in the decision matrix below, where a is the reward for correct anticipation and c the reward for correct classification (table 1).

Table 1. The payoff matrix for the classification response, following an anticipation that the sign will be male. The reward for correct classification is c, while the reward for correctly having anticipated that the sign will be male is a. The terms a and c include *both* the $0.02 piece-rate payment for accuracy and any subjective impact on the expectation of winning the $40 bonus. In the classification bonus condition, the bonus increases the value of c, while in the anticipation bonus condition, it increases the value of a.

anticipation = male	sign is actually male ($S = M$)	sign is actually female ($S = F$)
confirming classification $x = m$	$a + c$	*0*
disconfirming classification $x = f$	a	c

The subject gets credit for classifying correctly, but also wishes to believe that the stimulus is male, to validate the correctness of the preceding anticipation. Given these incentives, the self-signalling equation (3.2) derives the utilities for the two responses

$$V(x = m, \theta^{\circ}) = (a + c)\theta^{\circ}_{\mathrm{M}} + \lambda(a + c)E(\theta_{\mathrm{M}}|x = m)),$$
$$V(x = f, \theta^{\circ}) = a\theta^{\circ}_{\mathrm{M}} + c\theta^{\circ}_{\mathrm{F}} + \lambda(aE(\theta_{\mathrm{M}}|x = f) + cE(\theta_{\mathrm{F}}|x = f)),$$

where $\theta^{\circ}_{\mathrm{M}}$ is actual deep belief, introspectively inaccessible, while $E(\theta_{\mathrm{M}}|x = m)$ is the expectation of this belief inferred from classifying the sign as male. If categorization is symmetric, a reasonable simplification, $E(\theta_{\mathrm{M}}|x = m) = E(\theta_{\mathrm{F}}|x = f)$, indicates that the subject will categorize the stimulus as male if

$$V(x = m, \theta^{\circ}) - V(x = f, \theta^{\circ}) = c(\theta^{\circ}_{\mathrm{M}} - \theta^{\circ}_{\mathrm{F}}) + \lambda a(E(\theta_{\mathrm{M}}|x = m) - E(\theta_{\mathrm{M}}|x = f)) > 0$$

In the absence of self-signalling ($\lambda = 0$), the subject will categorize the sign as male if, and only if the probability of male is greater than $\frac{1}{2}$ i.e. if $\theta^{\circ}_{\mathrm{M}} > \theta^{\circ}_{\mathrm{F}}$. With self-signalling, one has to factor in the diagnostic utility of selecting male, which is proportional to $E(\theta_{\mathrm{M}}|x = m) - E(\theta_{\mathrm{M}}|x = f)$.

Previously, we mentioned two rules for specifying the inferences that a person might draw from her own actions. With face-value interpretations, the subject falsely believes that she is not affected by diagnostic considerations, and therefore assumes that if she classified the stimulus as male, this must mean that she indeed believes deep down that $\theta_{\mathrm{M}} > \theta_{\mathrm{F}}$, which is to say that $\theta_{\mathrm{M}} > 0.5$. This implies that $E(\theta_{\mathrm{M}}|x = m) = E(\theta_{\mathrm{M}}|\theta_{\mathrm{M}} > 0.5) > E(\theta_{\mathrm{M}}|\theta_{\mathrm{M}} < 0.5) = E(\theta_{\mathrm{M}}|x = f)$.

With rational interpretations, the situation is more complex, because awareness of diagnostic motivation discounts the signal; the subject appreciates that there is now a lower bar $\theta^* < 0.5$ for classifying the sign as male, and consequently that $E(\theta_{\mathrm{M}}|x = m = E(\theta_{\mathrm{M}}|\theta_{\mathrm{M}} > \theta^*)$. However, discounting preserves the basic directional implication, namely, that a male classification provides positive information that

Table 2. Distribution of trial patterns for the two different treatment groups. The labelling MFF, for example, refers to an initial classification of male in phase I, and an anticipation of female followed by a classification as female in phase II.

	confirming trials (C2 = A)		disconfirming trials (C2 ≠ A)			
	consistent MMM or FFF	self-deceptive MFF or FMM	inconsistent MMF or FFM	honest MFM or FMF	SD-Inc	(SD-Inc)/Inc
classification bonus (%)	42.2	18.3	11.7	27.8	+6.6	+55
anticipation bonus (%)	43.8	23.4	8.6	24.3	+14.8	+173

the sign was in fact male, i.e. $E(\theta_M|x=m) = E(\theta_M|\theta_M > \theta^*) > E(\theta_M|\theta_M < \theta^*) = E(\theta_M|x=f)$.

If λ is large enough, anticipations will be confirmed irrespective of deep belief, that is, even if $\theta^{\circ}_M = 0$. This would imply that $E(\theta_M|x=m) = E(\theta_M) = 0.5$ (assuming symmetry), and $E(\theta_M|\ x=f) = 0$. In other words, confirming the anticipation conveys no information and simply preserves the prior 50–50 odds while disconfirming the anticipation—a counterfactual response that never occurs—would prove that the anticipation was incorrect. In the extreme case where $\theta^{\circ}_M = 0$, the optimal classification will be male provided that,

$$V(x=m,\theta^{\circ}) - V(x=f,\theta^{\circ}) = c(0-1) + \lambda a(0.5-0) > 0.$$

Therefore, under rational interpretations if the weight of diagnostic utility exceeds the threshold: $\lambda > 2c/a$, the only possible response is the confirming one, even though this response has no impact on experienced beliefs.

With either face-value or rational interpretations, the diagnostic utility of a male categorization, following a male anticipation, should be positive. The model thus predicts that anticipation-confirming classifications will increase with anticipation incentives (a) and decrease with classification incentives (c).

(c) *Results*

To summarize, participants made five responses in connection with each sign: an initial classification in phase I (C1) followed by a confidence rating (R1), and in phase II a blind anticipation (A) followed by a second classification (C2) and confidence rating (R2). The responses can be mapped onto the theoretical variables in the following way. C2 corresponds to x. If we let θ_C denote the probability that a classification is correct, then R1 is an ordinal indicator of the prior expectation that C1 is correct, $E(\theta_{C1}) = \Sigma_\theta \theta_{C1} p(\theta_{C1})$, and R2 is an indicator of posterior expectation $E(\theta_{C2}|\mathrm{C2}) = \Sigma_\theta \theta_{C2} p(\theta_{C2}|\mathrm{C2})$. Therefore, the difference R2 − R1 will be our proxy measure of diagnostic utility.

Collapsing across the male/female categories and ignoring the confidence ratings, trials can be sorted into one of four types. A consistent trial corresponds to the pattern C2 = A = C1, where all three responses coincide. An honest pattern corresponds to C2 ≠ A and C2 = C1, that is, the subject acknowledges that the preceding anticipation was incorrect, and confirms the original gender classification in phase I. A self-deceptive pattern corresponds to: C2 = A ≠ C1, that is, the sign changes gender so as to make the preceding anticipation seem correct. An inconsistent pattern corresponds to C2 ≠ A = C1, that is, the subject changes mind about the gender even though the anticipation was consistent with his original classification. The frequency of inconsistent patterns provides a baseline for assessing whether there is statistically significant self-deception, or whether the trials labelled as self-deceptive reflect simple variability in classifications.

Table 2 presents the breakdown of trial patterns, by treatment group. Two results stand out: first, the proportion of self-deceptive patterns is greater than the proportion of inconsistent patterns, which define the error baseline. Hence, the second classification judgement is influenced by the preceding anticipations at the aggregate level.[7] Second, this impact of anticipations is greater in the anticipation bonus condition, relative to the classification bonus condition. The table provides two measures of impact, as the absolute or relative per cent increase in self-deceptive patterns, over the inconsistent baseline. Depending on which measure one adopts, the gap between the self-deceptive and inconsistent shares is between two to three times greater in the anticipation bonus condition. This confirms that the impact of anticipations on subsequent classifications is controlled in large measure by the financial incentives.

Figure 2 displays the self-deceptive and inconsistent pattern percentages for all 85 subjects, indicating treatment by colour. The impact of treatment is evident here as well. This can be confirmed statistically by counting the number of subjects with significant self-deception at the individual level, and then comparing between groups. A logistic regression of C2 against C1 and A simultaneously provides a sensitive individual-level test (the inclusion of C1 in the regression controls for bias towards one or the other gender classification, as well as for chance correlation between C1 and A). In the absence of self-deception, the coefficient on A should be non-significant. In the classification bonus treatment, 53 per cent of subjects are significantly self-deceptive at the 0.05 level, and 27 per cent at the 0.001 level; these percentages rise to 73 and 45 per cent, respectively, in the anticipation bonus condition. Comparing treatments, the difference in proportions is significant ($\chi^2 = 5.93$, $p < 0.02$ for $p = 0.05$ cutoff, $\chi^2 = 3.13$, $p < 0.08$ for $p = 0.001$ cutoff).

In what follows, we will refer to subjects with self-deception at the 0.001 level as the high self-deception (SD) group ($N = 30$), and those with self-deception at only 0.05 level as the moderate SD group ($N = 20$).[8]

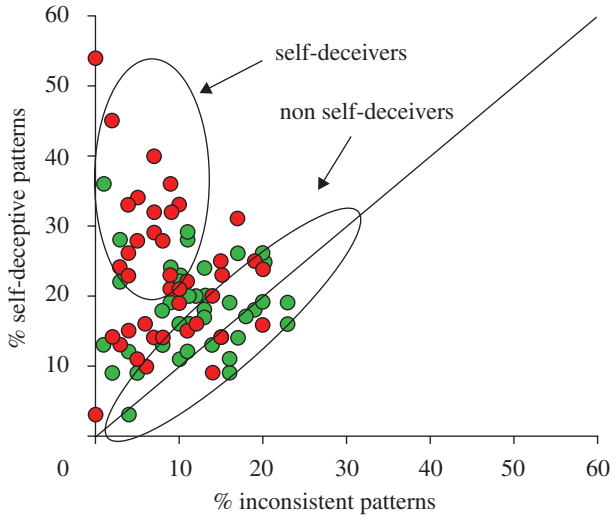

Figure 2. The impact of incentive condition on self-deception. Per cent of inconsistent patterns gives the baseline for assessing self-deception. The majority of subjects with strong-deception come from the anticipation bonus condition. The ovals are approximate (green circles, subjects with \$40 classification bonus; red circles, subjects with \$40 anticipation bonus).

There are no indications that self-deception is associated with lower effort; if anything, the relationship runs the other way. The average accuracy at C1 (according to the peer group answer key) increases from 61.5 to 63.2 and 66.2 per cent for the non-, moderate and high SD groups (the difference between high and none is significant, $t(63) = 2.10$, $p < 0.05$, as is the difference between high and the rest, $t(83) = 2.00$, $p < 0.05$). However, this difference disappears at C2, where the average accuracies are 62.3, 63.8 and 62.8 per cent. The change in accuracy is significant for the high SD group only (matched-pair t-test, $t(2973) = 3.38$, $p < 0.0005$). It appears that the high SD subjects exhibit greater motivation and engagement with the task initially, but their advantage disappears in the second phase, as result of self-deception.

(d) *What psychological benefits are obtained for the reduction in objective accuracy?*

According to self-signalling theory there is a diagnostic utility benefit, which we cannot measure directly but which should be revealed through the confidence ratings that follow each classification response. The benefit is modulated by awareness: it should be higher with face-value interpretations, and lower or nonexistent with rational interpretations. A plausible proxy for awareness is the overall rate of anticipation-confirming responses.[9] These rates vary in a predictable manner across the groups: 53 per cent (non-SD), 63 per cent (moderate SD), and 76 per cent (high SD).[10] High confirmation rates ought to raise doubts about the integrity of the confirming response. The subjects presumably understand that their anticipations are random guesses, and that being correct three times out of four is simply not sustainable.

The average confidence ratings (1–5 scale) are not significantly different for the three groups, at 3.08, 3.32 and 2.92, respectively, but are directionally consistent with the hypothesis that the benefits of self-deception peak at moderate levels. Moreover, among subjects with statistical self-deception (pooling moderate and high groups), the correlation between confidence and confirmation rate is significantly negative ($r = -0.40$, $t(48) = -3.05$, $p < 0.005$).

A more appropriate indicator of diagnostic utility is the difference between the second and the first confidence ratings, R2 − R1. This removes variation in intrinsic confidence that subjects might have with respect to the classification task, as well as variation in how they use the rating scale. On normative grounds, one would expect confidence to increase following C2 = C1, suggestive of a less ambiguous sign, and no change in confidence following C2 = A, because the anticipation has no information value. What one observes, instead, is that confirming responses (C2 = A) increase and disconfirming responses decrease confidence (matched-pairs, $t(82) = +1.66$ for C2 = A, $p < 0.05$; $t(82) = -1.96$, $p < 0.03$ for C2 ≠ A). In contrast, classification confirming responses (C2 = C1) have no impact on confidence.

Looking at the three groups separately, the moderate SD group experiences an increase in confidence following confirmation ($t(19) = +2.11$, $p < 0.05$), the high SD group experiences a marginally significant *decrease* in confidence following disconfirmation ($t(27) = -1.76$, $p < 0.05$ one-tailed), and the non-SD group does not register significant changes in confidence following either type of response. Hence, one could say that the moderate SD group is motivated by the benefits of confirmation, and the high SD group by the costs of disconfirmation, which is consistent with discounting of the confirming judgements as predicted by the model.

The net benefits of confirmation are highest at moderate rates, as shown in figure 3, which displays quadratic regression of change in confidence on confirmation rate. As expected, the quadratic term is significant, but only following a confirming response. According to the estimated fit, the boost in confidence reaches a maximum at about 65 per cent confirmation rate, which is presumably high enough to have impact but not so high to raise suspicion. This relationship is driven by the changes in confidence experienced after a confirmation, and specifically among subjects in the anticipation bonus condition.

Response time data provide additional evidence of different processing at high self-deception levels. Figure 4 displays C2 response time as percentage of C1 response time, by trial pattern and level of self-deception. This nets out differences in response time between subjects, and also nets out stimulus-specific differences in response time, due to differential difficulty of classifying different stimuli.

Subjects without statistical self-deception show no difference in C2 response times as a function of trial pattern. Moderate self-deception is associated with longer C2 response times on honest and inconsistent trials. The pattern that clearly emerges with high self-deception subjects is fast confirming response times, that is, whenever C2 = A.

To better understand the significance of this, we computed individual subject correlations between

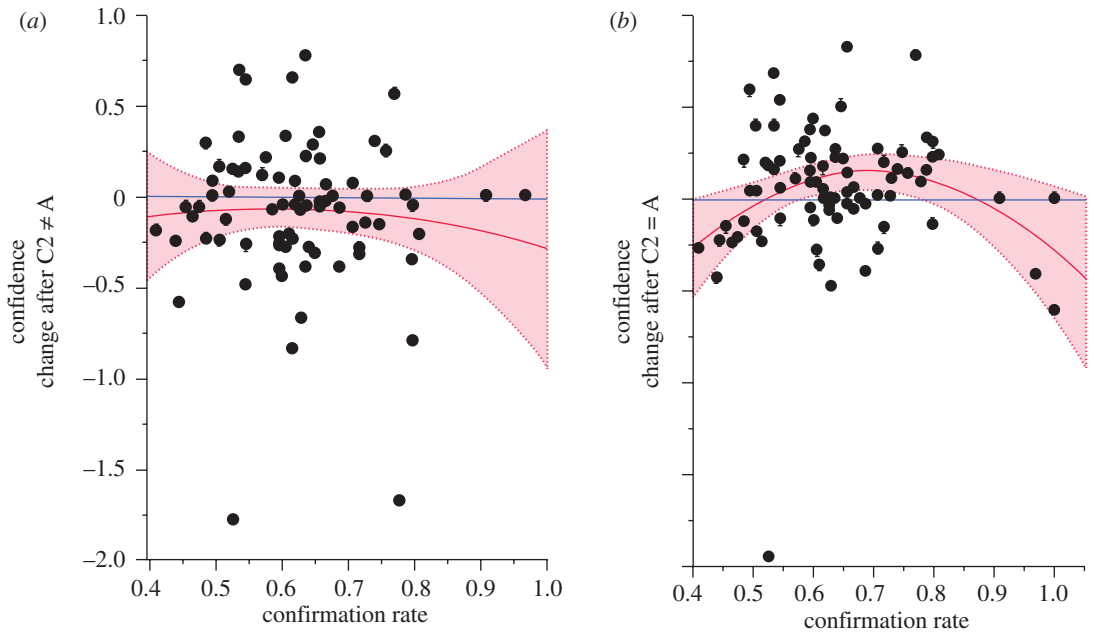

Figure 3. Average change in the 1–5 confidence rating (R2−R1) following a disconfirming (C2 ≠ A, *a*) or confirming response (C2 = A, *b*) plotted by subject against subject's confirmation rate. The solid line is best fitting quadratic, with shaded 95% confidence interval. The linear term is not significant in either (*a*) or (*b*); the negative quadratic term is significant in (*b*) ($p < 0.002$). If the analysis is conducted on the difference between (*a*) and (*b*) (which corresponds to the diagnostic utility of confirming rather than disconfirming), the negative quadratic remains highly significant ($t = -3.47$, $p < 0.001$), and linear becomes positively significant ($t = +2.52$, $p < 0.02$).

change in log-response time and change in confidence. Normally, one would interpret response time as an (inverse) indicator of response confidence. However, a fast response time could also indicate higher motivation without confidence, or a desire to move away quickly from a problematic stimulus to the next task.

The fraction of subjects showing an anomalous positive coefficient, implying lower confidence for faster response times, is higher (non-significantly) in the high SD group (27%, compared with 20% and 14% for the moderate SD and non-SD). The difference in correlation coefficients between the high SD group and the remaining subjects is significant ($t(83) = 2.40$, $p < 0.02$), as is the correlation between the correlation coefficients and individual confirmation rates (Spearman $\rho = 0.21$, $p < 0.07$). The standard relationship between fast response time and confidence appears to deteriorate at high confirmation rates. Subjects with high SD have a higher propensity to confirm and they confirm more quickly, but these faster response times are no longer a reliable indicator of confidence.

This suggests that the self-deception we observe here is probably not a biased sifting of perceptual evidence. A sifting of evidence would presumably prolong response time on self-deceptive trials, which is opposite to the pattern we observe in figure 4. Rapid response times associated with self-deception suggest suppression of evidence, rather than a second-look at the evidence.

(e) *Discussion*

Two main findings emerge from the experiment. First, it is possible to induce costly self-deception in a repeated decision task, by presenting subjects with the prospect of a large and essentially non-contingent financial bonus. Actions that provide favourable news about the chances of winning the bonus are selected more often than they should be. The extent of this self-deception is in turn sensitive to the financial parameters, as predicted by the self-signalling model. A majority of subjects exhibit some statistical self-deception, but some avoid it altogether, even with high incentives.

Second, among subjects with self-deception there is great variation in the extent of statistical bias towards the diagnostically favourable response. Subjects with moderate levels of bias appear to derive some psychological benefit from self-deception, reflected in their higher confidence ratings. In contrast, subjects with the most severe bias show no improvement in confidence relative to subjects without any bias at all.

These results strongly point to the conclusion that differences in levels of bias are associated with differences in awareness that one is biased. While we do not measure awareness directly, common sense suggests that a self-assessed success rate of 60 per cent (rather than the unbiased 50%) can sneak by under-the-radar, like small rates of cheating (von Hippel *et al.* 2005; Mazar *et al.* 2008); however, a rate of, say, 80 per cent is definitely too good to be true. Confirming responses will deliver the psychological benefit in confidence only if the overall bias in confirmation rate does not stray outside of some reasonable margin.

Granting this, one still needs to explain the extravagant bias observed in so many subjects. As a group, these subjects are certainly not careless, as shown by their greater accuracy in the initial phase of the experiment, before the bonus opportunity is revealed.

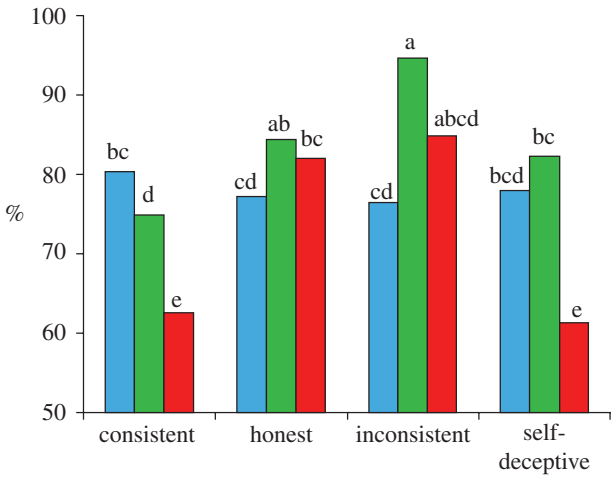

Figure 4. C2 response times expressed as per cent of C1 response times, plotted separately for subjects with high, moderate and no self-deception, and for different types of trials produced by the C2 response. Levels connected by the same letter are not significantly different ($p < 0.05$, t-test; blue bar, no self-deception ($n = 35$); green bar, moderate SD $0.001 < p < 0.05$ ($n = 20$); red bar, high SD $p < 0.001$ ($n = 20$).

If they are strongly motivated to win, they will also appreciate that they have to do exceptionally well to have any chance; being right a little more than half the time is not enough. In the absence of feedback, a high self-estimated success rate, while not necessarily credible, preserves some possibility of winning the bonus, while a more candid, average rate would subjectively rule it out. This would explain the briskness of the confirming responses, and the lack of any boost in confidence following them.[11]

An interesting question is whether subjects 'see' the characters differently, as a result of their anticipations. This would be consistent with the findings on motivated perception of ambiguous letters and animal drawings by Balcetis & Dunning (2006), and with the Berns *et al.* (2005) fMRI replication of Asch's classic experiment on conformity. A potentially important difference between our paradigm and that of Balcetis and Dunning is that in our case the desired response category was changing rapidly from one trial to the next. In that sense, the task frustrated the development of a stable bias towards one or the other category. (We also find little evidence that more ambiguous signs are more vulnerable to self-deceptive reclassification, whether ambiguity is measured by initial confidence rating, initial response time, or concordance across subjects.) Therefore, while we cannot address the perceptual question conclusively, it seems that some other mechanism must be responsible for the very highest rates of confirmation observed in our study.

6. CONCLUDING REMARKS

We have proposed here a theory of genuine self-deception, that is, self-deception conceived strictly on the interpersonal model. The equations of the model could apply equally well to the interaction of two individuals, each with distinct beliefs, actions, and objectives, with one individual attempting to deceive the other. From the equations, one cannot tell whether this is a model of self-deception or just plain deception. In the presentation of the theory, we have not emphasized the interpersonal interpretation, because the postulated personae or 'selves' are necessarily speculative. In this concluding section we will comment on the interpersonal interpretation in more detail. This will clarify the psychological architecture implicit in the model and allow us to comment briefly on the potential evolutionary benefits of this architecture.

At the formal level, the self-signalling model represents the interaction between two agents or 'interests' (Ainslie 1986), whom we may identify as actor and interpreter. The actor has private motives that are hidden from the interpreter. He makes a choice, potentially revealing something about these motives. The interpreter observes the choice, infers the underlying motive, and then grades the motive according to a preset formula. The grade matters to the actor, it enters into his utility function as the diagnostic utility component. It does not matter to the interpreter, he does not care what grade he gives as long as it is the correct grade. So, there is conflict but it is not a conflict over ultimate objectives or ongoing behaviour, only a conflict over interpretations of actions and underlying motives. The actor wishes to extract a good grade, if possible a better grade than he deserves, while the interpreter strives for objectivity.

Returning to the psychological, intrapersonal level, the same model now refers to the interaction of two optimizing modules, one responsible for behaviour selection and the other for accurate online behaviour evaluation. The interpreting module has some characteristics of a conscience or superego in that it scrutinizes behaviour impartially for underlying motivation. However, it falls short of being a conscience in lacking intrinsic values or preferences.

What might be the reason for this psychological architecture? Why split the mind into two elements and render one element ignorant? Trivers developed a provocative evolutionary rationale for self-ignorance in his theory of self-deception (Trivers 1985). According to him, we are unaware of our true motives so as to be better able to deceive others. The sincere deceiver is presumed to have advantage in not having to pretend, to hold two distinct attitudes in mind at the same time. This would be especially true of emotions, which are notoriously difficult to disguise. Complete unawareness of one's true motives would make deception of others effortless.

Trivers has in mind an unconcerned ignorance of motive. In contrast, the model developed here deals with concerned ignorance: The person is unsure about his characteristics and this precisely is the source of worry. Uncertainty makes actions informative, and sustains diagnostic motivation. This leads to a different rationalization of self-ignorance, as means of enhancing the motivational significance of actions.

It is notoriously hard to assess the significance of an additional day's progress, whether for a dissertation or some other remote goal. Assessed coldly, the impact of

even a very good day may be negligible. Yet, success requires persistence, and persistence must be rewarded before the final outcome is known. Such rewards cannot come from the outside but only from the organism itself. If the organism acquires the ability to self-reward, then it must also acquire an objective, external attitude towards it's own actions. This argues for the structural separation of modules responsible for action selection and those responsible for interpreting and rewarding those actions. It also argues for denying internal information to the interpretational mechanisms. As custodian of self-reward, the interpreter should take into account the external evidence that actions would provide to an outside observer, and not the internal, corruptible evidence of feelings and intentions. The less the interpreter knows about internal parameters, the greater the chances that it will enforce objective criteria for delivering self-reward.

On this view, genuine self-deception, as opposed to mere bias, is a byproduct of this specific modular architecture. Like ordinary deception, it is an external, public activity, involving overt statements or actions directed towards an audience, whether real or imagined. Modelling this process as a signalling game, as we have done in this paper, provides benefits that we hope will be exploited further in future work. First, the formal theory raises conceptual possibilities that might otherwise be overlooked. In particular, it draws attention to the possibility of a stable state of inauthentic belief, characterized by a chronic mismatch between what a person says and what they truly believe and experience. Second, the theory motivates experimental studies, such as the one presented here. Finally, it guides search for brain mechanisms that might in principle carry out the computations required by the model.

We are grateful to Ravi Dhar, Guy Mayraz, Trey Hedden, Stephanie Carpenter, and Arnaldo Pereira-Diaz for extensive comments on the manuscript; to Dan Ariely, Jiwoong Shin, and Andreja Bubic for experimental help and advice; to the Institute for Advanced Study for hospitality and financial support; to the Psychology Department of Zagreb University for hosting a pilot study; and to Tom Palfrey and the Princeton Laboratory for Experimental Social Science for hosting the experiment reported here. We also wish to acknowledge numerous discussions with our MIT colleagues and collaborators, John Gabrieli, Richard Holton, Nina Wickens, Kristina Fanucci, Paymon Hosseini and Alexander Huang, as well as comments by seminar participants at the Robinson College (University of Cambridge) Workshop on Rationality and Emotions, the Institute for Advanced Study, GREQAM-Marseille, Sorbonne, Zurich Institute for Research in Experimental Economics, Toulouse School of Economics and Brown University, among others.

ENDNOTES

[1]Bach (1997) expresses this nicely: 'For example, what makes the betrayed husband count as self deceived is not merely that his belief that his wife is faithful is sustained by a motivationally biased treatment of his evidence. He could believe this even if he had no tendency to think about the subject ever again. He counts as a self-deceiver only because sustaining his belief that his wife is faithful requires an active effort to avoid thinking that she is not. In self-deception, unlike blindness or denial, the truth is dangerously close at hand. His would not be a case of self-deception if it hardly ever occurred to him that his wife might be playing around and if he did not appreciate the weight of the evidence, at least to some extent. If self-deception were just a matter of belief, then once the self-deceptive belief was formed, the issue would be settled for him; but in self-deception it is not. The self-deceiver is disposed to think the very thing he is motivated to avoid thinking, and this is the disposition he resists'.

[2]See also Levy's (2008) arguments about anosognosia for hemplegia (denial of paralysis) as a real case of self-deception.

[3]As coauthored chapter of Bodner's (1995) doctoral dissertation.

[4]It is important not to confuse self-signalling with evidential decision theory (EDT; Gibbard & Harper 1978). The decision criterion in EDT is $\Sigma_S U(x, S)p(S|x)$, which resembles the second part of (2), $\Sigma_\theta u(x, \theta)p(\theta|x)$. The key difference is that actual deep beliefs θ° do not appear in the EDT formula. From a formal standpoint, closest to the present approach is the memory-anticipation model of Bernheim & Thomadsen (2005). In their model, at time-zero the individual selects an action affecting outcomes at time-two, in light of information that she knows will be later forgotten. At interim time-one the person tries to retroactively infer this information from actions already taken, leading to anticipatory emotions about outcomes at time-two. The individual at time-zero then has a reason to take actions supportive of positive interim emotions, knowing full that these emotions will be disappointed later. In the philosophical debate, Mele (1997) mentions this type of scenario and allows it to be a genuine, albeit rare example of intentional self-deception; for Audi (1997) it is a distinct phenomenon, more properly termed 'self-caused deception'.

[5]We are sidestepping important details, namely: (i) What inferences follow from an action that is suboptimal for any θ and thus, strictly speaking, should not occur (this is the problem of beliefs 'off-the-equilibrium-path')? (ii) When does an equilibrium exist, and when is it unique? See Cho & Sobel (1990) for a general treatment of these issues.

[6]For an analogous treatment of social conformity see Bernheim (1994).

[7]Note that a disconfirming response strategy ($C2 \neq A$) would guarantee that one of the two responses is always correct. This would provide a hedging benefit for subjects who are risk averse at the level of a single trial. We find no evidence of hedging in the data.

[8]It is interesting that the high SD group includes some subjects from the classification bonus treatment. These subjects may have been motivated by the \$0.02 reward for anticipations. Alternatively, this may reflect intrinsic motivation associated with self-confirming responses (or, equivalently, to a disinclination to acknowledge error, even if the financial consequences are minor).

[9]If $\text{Prob}(A = C1) = 0.5$, then the confirmation rate equals to the combined frequency of consistent and self-deceptive trials. However, $\text{Prob}(A = C1)$ could deviate from 0.5 through chance, or if a subject favours one category. To compensate for unequal base rates of $A = C1$ and $A \neq C1$ we work with the corrected rate: $(\text{Prob}(C2 = A|A = C1) + \text{Prob}(C2 = A|A \neq C1))/2$. The correlation between this index and the raw frequency of consistent and self-deceptive trials is $+0.97$, so for practical purposes we can regard them as the same.

[10]They also vary between treatments: 58.5 per cent versus 66.5 per cent for classification and anticipation bonus groups, respectively, $t(83) = 3.33$, $p < 0.002$.

[11]The notion that moderate levels of self-deception are beneficial for self-esteem and mental health has been debated extensively (Lockhard & Paulhus 1988; Taylor & Brown 1988).

REFERENCES

Ainslie, G. 1986 Beyond microeconomics: conflict among interests in a multiple self as a determinant of value. In *The multiple self* (ed. J. Elster), pp. 133–175. Cambridge, UK: Cambridge University Press.

Audi, R. 1985 Self-deception and rationality. In *Self-deception and self-understanding* (ed. M. W. Martin), pp. 169–194. Lawrence, KS: University of Kansas.

Audi, R. 1997 Self-deception vs. self-caused deception: a comment on Professor Mele. *Behav. Brain Sci.* **20**, 104. (doi:10.1017/S0140525X97230037)

Babcock, L. & Loewenstein, G. 1997 Explaining bargaining impasse: the role of self-serving biases. *J. Econ. Perspect.* **11**, 109–126.

Bach, K. 1997 Thinking and believing in self-deception. *Behav. Brain Sci.* **20**, 105. (doi:10.1017/S0140525X97240033)

Balcetis, E. & Dunning, D. 2006 See what you want to see: motivational influences on visual perception. *J. Pers. Soc. Psychol.* **91**, 612–625. (doi:10.1037/0022-3514.91.4.612)

Bem, D. 1972 Self-perception theory. In *Advances in experimental social psychology* (ed. L. Berkowitz). New York, NY: Academic Press.

Benabou, R. & Tirole, J. 2002 Self-confidence and personal motivation. *Q. J. Econ.* **117**, 871–915. (doi:10.1162/003355302760193913)

Benabou, R. & Tirole, J. 2004 Willpower and personal rules. *J. Polit. Econ.* **112**, 848–887. (doi:10.1086/421167)

Bernheim, B. 1994 A theory of conformity. *J. Polit. Econ.* **102**, 841–877. (doi:10.1086/261957)

Bernheim, B. & Thomadsen, R. 2005 Memory and anticipation. *Econ. J.* **115**, 271–304. (doi:10.1111/j.1468-0297.2005.00989.x)

Berns, G., Chappelow, J., Zink, C., Pagnoni, G., Martin-Skurski, M. & Richards, J. 2005 Neurobiological correlates of social conformity and independence during mental rotation. *Biol. Psychiatry* **58**, 245–253. (doi:10.1016/j.biopsych.2005.04.012)

Bodner, R. 1995 Self-knowledge and the diagnostic value of actions: the case of donating to a charitable cause. Doctoral dissertation, Sloan School, Massachusetts Institute of Technology, Cambridge, MA.

Bodner, R. & Prelec, D. 1995 The diagnostic value of actions and the emergence of personal rules in a self-signaling model. In *Self-knowledge and the diagnostic value of one's actions* (ed. R. Bodner), ch. 2, pp. 53–67. Doctoral dissertation, Sloan School, Massachusetts Institute of Technology, Cambridge, MA, USA.

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), pp. 105–126. London, UK: Oxford University Press.

Brocas, I. & Carrillo, J. D. 2008 The brain as a hierarchical organization. *Am. Econ. Rev.* **98**, 1312–1346. (doi:10.1257/aer.98.4.1312)

Brown, J. D. & Dutton, K. A. 1995 Truth and consequences—the costs and benefits of accurate self-knowledge. *Pers. Soc. Psychol. Bull.* **21**, 1288–1296. (doi:10.1177/01461672952112006)

Caplin, A. & Leahy, J. 2001 Psychological expected utility theory and anticipatory feelings. *Q. J. Econ.* **116**, 55–79. (doi:10.1162/003355301556347)

Cho, I. & Sobel, J. 1990 Strategic stability and uniqueness in signaling games. *J. Econ. Theory* **50**, 381–413. (doi:10.1016/0022-0531(90)90009-9)

Dawson, E., Gilovich, T. & Regan, D. T. 2002 Motivated reasoning and performance on the Wason selection task. *Pers. Soc. Psychol. Bull.* **28**, 1379–1387. (doi:10.1177/014616702236869)

Ditto, P. H. & Lopez, D. F. 1992 Motivated skepticism—use of differential decision criteria for preferred and nonpreferred conclusions. *J. Pers. Soc. Psychol.* **63**, 568–584. (doi:10.1037/0022-3514.63.4.568)

Dunning, D. & Hayes, A. 1996 Evidence for egocentric comparison in social judgment. *J. Pers. Soc. Psychol.* **71**, 213–229. (doi:10.1037/0022-3514.71.2.213)

Dunning, D., Leuenberger, A. & Sherman, D. 1995 A new look at motivated inference—are self-serving theories of success a product of motivational forces. *J. Pers. Soc. Psychol.* **69**, 58–68. (doi:10.1037/0022-3514.69.1.58)

Elster, J. 1999 *Alchemies of the mind: rationality and the emotions*. Cambridge, UK: Cambridge University Press.

Fudenberg, D. & Levine, D. 2008 A dual self model of impulse control. *Am. Econ. Rev.* **96**, 1449–1476. (doi:10.1257/aer.96.5.1449)

Funkouser, E. 2005 Do the self-deceived get what they want? *Pacific Phil. Q.* **86**, 295–312.

Gibbard, A. & Harper, W. 1978 Counterfactuals and two kinds of expected utility. In *Foundations and applications of decision theory* (eds C. A. Hooker, J. J. Leach & E. F. McClennen), pp. 125–162. Dordrecht, the Netherlands: Reidel.

Gilovich, T. 1991 *How we know what isn't so: fallibility of human reason in everyday life*. New York, NY: Free Press.

Ginossar, Z. & Trope, Y. 1987 Problem solving in judgment under uncertainty. *J. Pers. Soc. Psychol.* **52**, 464–474. (doi:10.1037/0022-3514.52.3.464)

Gottlieb, D. 2009 Imperfect memory and choice under risk. Doctoral dissertation, Department of Economics, Massachusetts Institute of Technology.

Gur, R. C. & Sackeim, H. A. 1979 Self-deception: a concept in search of a phenomenon. *J. Pers. Soc. Psychol.* **37**, 147–169. (doi:10.1037/0022-3514.37.2.147)

Holton, R. 2000 What is the role of the self in self-deception? *Proc. Aristotelian Soc.* **101**, 53–69.

Koszegi, B. 2006*a* Ego utility, overconfidence, and task choice. *J. Eur. Econ. Assoc.* **4**, 673–707.

Koszegi, B. 2006*b* Emotional agency. *Q. J. Econ.* **121**, 121–156.

Kunda, Z. 1987 Motivated inference: self-serving generation and evaluation of evidence. *J. Pers. Soc. Psychol.* **53**, 636–647. (doi:10.1037/0022-3514.53.4.636)

Kunda, Z. 1990 The case for motivated reasoning. *Psychol. Bull.* **108**, 480–498. (doi:10.1037/0033-2909.108.3.480)

Levy, N. 2008 Self-deception without thought experiments. In *Delusions and self deception: affective and motivational influences on belief-formation* (eds T. Bayne & J. Fernández), pp. 227–242. Hove: Psychology Press.

Lockhard, J. & Paulhus, D. 1988 *Self-deception: an adaptive mechanism?* Englewood Cliffs, NJ: Prentice-Hall.

Mazar, N., Amir, O. & Ariely, D. 2008 The dishonesty of honest people. *J. Market. Res.* **45**, 633–644. (doi:10.1509/jmkr.45.6.633)

Mele, A. R. 1997 Real self-deception. *Behav. Brain Sci.* **20**, 91–136.

Mele, A. R. 1998 Motivated belief and agency. *Phil. Psychol.* **11**, 353–369. (doi:10.1080/09515089808573266)

Mijovic-Prelec, D., Shin, L. M., Chabris, C. F. & Kosslyn, S. M. 1994 When does 'no' really mean 'yes'? A case study in unilateral visual neglect. *Neuropsychologia* **32**, 151–158. (doi:10.1016/0028-3932(94)90002-7)

Norton, M. I., Vandello, J. A. & Darley, J. M. 2004 Casuistry and social category bias. *J. Pers. Soc. Psychol.* **87**, 817–831. (doi:10.1037/0022-3514.87.6.817)

Prelec, D. & Bodner, R. 2003 Self-signaling and self-control. In *Time and decision* (eds G. Loewenstein, D. Read & R. F. Baumeister), pp. 277–300. New York, NY: Russell Sage Press.

Pronin, E., Gilovich, T. & Ross, L. 2004 Objectivity in the eye of the beholder: divergent perceptions of bias in self versus other. *Psychol. Rev.* **111**, 781–799. (doi:10.1037/0033-295X.111.3.781)

Pyszczynski, T. & Greenberg, J. 1987 Toward an integration of cognitive and motivational perspectives on social

inference—a biased hypothesis testing model. *Adv. Exp. Soc. Psychol.* **20**, 297–340. (doi:10.1016/S0065-2601(08)60417-7)

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. (doi:10.1037/0022-3514.46.2.237)

Sanitioso, R., Kunda, Z. & Fong, G. T. 1990 Motivated recruitment of autobiographical memory. *J. Pers. Soc. Psychol.* **59**, 229–241. (doi:10.1037/0022-3514.59.2.229)

Shapiro, D. 1996 On the psychology of self-deception—truth-telling, lying and self-deception. *Soc. Res.* **63**, 785–800.

Taylor, S. & Brown, J. 1988 Illusion and well-being: a social psychological perspective on mental health. *Psychol. Bull.* **103**, 193–210. (doi:10.1037/0033-2909.103.2.193)

Thaler, R. & Shefrin, H. M. 1981 An economic theory of self-control. *J. Polit. Econ.* **39**, 392–406.

Trivers, R. 1985 *Social evolution*. Menlo Park, CA: Benjamin/Cummings Pub. Co.

von Hippel, W., Lakin, J. L. & Shakarchi, R. J. 2005 Individual differences in motivated social cognition: the case of self-serving information processing. *Pers. Soc. Psychol. Bull.* **31**, 1347–1357. (doi:10.1177/0146167205274899)

Weisenkrantz, L. 1986 *Blindsight: a case study and implications*. Oxford, UK: Oxford University Press.