Forthcoming in: *Collected Essays in Psychology and Economics*,
I. Brocas and J. Carillo (eds), Oxford University Press, 2002,
available at www.ecare.ulb.ac.be/ecare/Juan/book.htm.

Self-signaling and diagnostic utility in everyday decision making[1]

Ronit Bodner

Drazen Prelec

May, 2001

(final version)

Abstract

A self-signaling action is an action chosen partly to secure good news about one's traits or abilities, even when the action has no causal impact on these traits and abilities. We discuss some of the odd things that happen when self-signaling is introduced into an otherwise rational conception of action. We employ a signaling game perspective in which the diagnostic signals are an endogenous part of the equilibrium choice. We are interested (1) in pure self-signaling, separate from any desire to be regarded well by others, and (2) purely diagnostic motivation, that is, caring about what an action might reveal about a trait even when that action has no causal impact on it. When diagnostic motivation is strong, the person's actions exhibit a rigidity characteristic of personal rules. Our model also predicts that a boost in self-image positively affects actions even though it leaves true preferences unchanged — we call this a "moral placebo effect."

---

[1] The chapter draws on (co-authored) Chapter 3 of Bodner's doctoral dissertation (Bodner, 1995) and an unpublished MIT working paper (Bodner and Prelec, 1997). The authors thank Bodner's dissertation advisors France Leclerc and Richard Thaler, workshop discussants Thomas Schelling, Russell Winer, and Mathias Dewatripont, and George Ainslie, Michael Bratman, Juan Carillo, Itzakh Gilboa, George Loewenstein, Al Mela, Matthew Rabin, Duncan Simester and Florian Zettelmeyer for comments on these ideas (with the usual disclaimer). We are grateful to Birger Wernerfelt for drawing attention to Bernheim's work on social conformity. Author addresses: Bodner – Director, Learning Innovations, 13\4 Shimshon St., Jerusalem, 93501, Israel, learning@netvision.net.il; Prelec — E56-320, MIT, Sloan School, 38 Memorial Drive, Cambridge, MA 02139, dprelec@mit.edu.

1 Psychological evidence

When we make a choice we reveal something of our inner traits or dispositions, not only to others, but also to ourselves. After the fact, this can be a source of pleasure or pain, depending on whether we were impressed or disappointed by our actions. Before the fact, the anticipation of future pride or remorse can influence what we choose to do.

In a previous paper (Bodner and Prelec, 1997), we described how the model of a utility maximizing individual could be expanded to include *diagnostic utility* as a separate motive for action. We review the basic elements of that proposal here. The inspiration comes directly from signaling games in which actions of one person provide an informative signal to others, which in turn affects esteem (Bernheim, 1994). Here, however, actions provide a signal to ourselves, that is, actions are *self-signaling*. For example, a person who takes the daily jog in spite of the rain may see that as a gratifying signal of willpower, dedication, or future well being. For someone uncertain about where he or she stands with respect to these dispositions, each new choice can provide a bit of good or bad "news." We incorporate the value of such "news" into the person's utility function.

The notion that a person may draw inferences from an action he enacted partially in order to gain that inference has been posed as a philosophical paradox (e.g. Campbell and Sawden, 1985; Elster, 1985, 1989). A key problem is the following: Suppose that the disposition in question is altruism, and a person interprets a 25¢ donation to a panhandler as evidence of altruism. If the boost in self-esteem makes it worth giving the quarter even when there is no concern for the poor, than clearly, such a donation is not valid evidence of altruism. Logically, giving is valid evidence of high altruism only if a person with low altruism would not have given the quarter. This reasoning motivates our equilibrium approach, in which inferences from actions are an *endogenous* part of the equilibrium choice.

As an empirical matter several studies have demonstrated that diagnostic considerations do indeed affect behavior (Quattrone and Tversky, 1984; Shafir and Tversky, 1992; Bodner, 1995). An elegant experiment by Quattrone and Tversky (1984) both defines the self-signaling phenomenon and demonstrates its existence. Quattrone and Tversky first asked each subject to take a cold pressor pain test in which the subject's arm is submerged in a container of cold water until the subject can no longer tolerate the pain. Subsequently the subject was told that recent medical studies had discovered a certain inborn heart condition, and that people with this condition are “frequently ill, prone to heart-disease, and have shorter-than-average life expectancy.” Subjects were also told that this type could be identified by the effect of exercise on the cold pressor test. Subjects were randomly assigned to one of two conditions

in which they were told that the bad type of heart was associated with either increases or with decreases in tolerance to the cold water after exercise. Subjects then repeated the cold pressor test, after riding an Exercycle for one minute. As predicted, the vast majority of subjects showed changes in tolerance on the second cold pressor trial in the direction correlated of "good news"—if told that decreased tolerance is diagnostic of a bad heart they endured the near-freezing water longer (and vice versa). The result shows that people are willing to bear painful consequences for a behavior that is a signal, though not a cause, of a medical diagnosis.

An experiment by Shafir and Tversky (1992) on "Newcomb's paradox" reinforces the same point. In the philosophical version of the paradox, a person is (hypothetically) presented with two boxes, A and B. Box A contains either nothing or some large amount of money deposited by an "omniscient being." Box B contains a small amount of money for sure. The decision-maker doesn't know what Box A contains choice, and has to choose whether to take the contents of that box (A) or of both boxes (A+B). What makes the problem a paradox is that the person is asked to believe that the omniscient being has already predicted her choice, and on that basis has already either "punished" a greedy choice of (A+B) with no deposit in A or "rewarded" a choice of (A) with a large deposit. The dominance principle argues in favor of choosing both boxes, because the deposits are fixed at the moment of choice.

This is the philosophical statement of the problem. In the actual experiment, Shafir and Tversky presented a variant of Newcomb's problem at the end of another, longer experiment, in which subjects repeatedly played a Prisoner's Dilemma game against (virtual) opponents via computer terminals. After finishing these games, a final "bonus" problem appeared, with the two Newcomb boxes, and subjects had to choose whether to take money from one box or from both boxes. The experimental cover story did not mention an omniscient being but instead informed the subjects that "a program developed at MIT recently was applied during the entire session [of Prisoner's Dilemma choices] to analyze the pattern of your preference." Ostensibly, this mighty program could predict choices, one or two boxes, with 85% accuracy, and, of course, if the program predicted a choice of both boxes it would then put nothing in Box A. Although it was evident that the money amounts were already set at the moment of choice, most experimental subjects opted for the single box. It is "as if" they believed that by declining to take the money in Box B, they could change the amount of money already deposited in box A.

Although these are relatively recent experiments, their results are consistent with a long stream of psychological research, going back at least to the James-Lange theory of emotions which claimed that people infer their own states from behavior (e.g., they feel afraid if they see themselves running). The notion that people adopt the perspective of an outside observer when interpreting their own actions has been extensively explored in the research on self-perception (Bem, 1972). In a similar vein, there is an

extensive literature confirming the existence of "self-handicapping" strategies, where a person might get too little sleep or under-prepare for an examination. In such a case, a successful performance could be attributed to ability while unsuccessful performance could be externalized as due to the lack of proper preparation (e.g. Berglas and Jones, 1978; Berglas and Baumeister, 1993). This broader context of psychological research suggests that we should view the results of Quattrone and Tversky, and Shafir and Tversky not as mere curiosities, applying to only contrived experimental situations, but instead as evidence of a general motivational "short circuit." Motivation does not require causality, even when the lack of causality is utterly transparent. If anything, these experiments probably underestimate the impact of diagnosticity in realistic decisions, where the absence of causal links between actions and dispositions is less evident.

Formally, our model distinguishes between *outcome utility* — the utility of the anticipated causal consequences of choice — and *diagnostic utility* — the value of the adjusted estimate of one's disposition, adjusted in light of the choice. Individuals act so as to maximize some combination of the two sources of utility, and (in one version of the model) make correct inferences about what their choices imply about their dispositions. When diagnostic utility is sufficiently important, the individual chooses the same action independent of disposition. We interpret this as a personal rule. We describe other ways in which the behavior of self-signaling individuals is qualitatively different from that of standard economic agents. First, a self-signaling person will be more likely to reveal discrepancies between resolutions and actions when resolutions pertain to actions that are contingent or delayed. Thus she might honestly commit to do some worthy action if the circumstances requiring the action were remote (temporally or probabilistically), but would in fact regret the commitment if those circumstances obtained. Second, self-signaling gives rise to moral placebo effects, where a change in mere beliefs about one's traits or abilities may affect actions even though the new beliefs leave one's actual disposition unchanged.

2 Definition of self-signaling

We begin with a general definition (Bodner and Prelec, 1997). A self-signaling person is characterized by: (1) a parameterized *outcome utility function*, $u(x,\theta)$, where x is the outcome and θ the unknown parameter; (2) a *self-image distribution* $f(\theta)$; (3) a *meta-utility function* $V(\theta)$. The outcome utility function represents the expected value of the causal consequences of choosing x. The θ-parameter is an index of the unknown, momentary disposition (a "type"). In the experiment by Quattrone and Tversky, it would be an index of true, momentary tolerance for cold water following exercise. The self-

image captures what the person knows about the disposition before making the choice. The meta-utility function represents the person's preferences over dispositions, which is to say, "preferences-over-preferences," given that dispositions are a preference parameter. In the Quattrone-Tversky experiment, subjects would naturally prefer to have tolerance levels θ that are associated with healthy hearts.

The total utility created by choosing outcome x from a set, C={x,y,z..}, equals the sum of the outcome utility of x, and the diagnostic utility of choosing x over all other outcomes y,z.. etc.. The outcome utility is simply the value $u(x,\theta)$, computed with the actual θ. The diagnostic utility is the expectation of $V(\theta)$ calculated with respect to an *interpretation function*, $f(\theta|x,C)$, which revises the self-image in light of the choice x. The total utility created by choosing x is:

Utility of choosing x from C = Outcome utility of x + Diagnostic utility of choosing x from C

$$U(x,C,\theta) = u(x,\theta) + \alpha_{\Pi} f(\theta|x,C)V(\theta). \quad (1)$$

Why do we not take an expectation of $u(x,\theta)$ over the unknown θ in equation 1? The question brings to the surface the distinction between a standard unknown parameter and a parameter like θ which has the property of being "known" by the decision-making mechanism but which cannot be introspected before the choice.[2] The gut knows θ, but the mind does not. A θ-parameter, such as a person's cold water tolerance, willpower, disposition to alcoholism, altruism, exerts influence at the moment of choice, but cannot be deduced by merely imagining what one might do in a given situation. Precisely, it represents that part of uncertainty that can only be resolved by a <u>real</u>, as opposed to a hypothetical choice.

Equation 1 defines the self-signaling decision problem. To complete it, one has to specify the revised distribution, $f(\theta|x,C)$. There are several ways to do that, depending on how much rationality one wishes to ascribe to the decision-maker's interpretation of his choices:

<u>Exogenous interpretations</u>

The most direct way to complete the model is to postulate an exogenous interpretation function $f(\theta|x,C)$. Psychologically, the function could be internalized through socialization or persuasion (e.g., as in the original formulation of Newcomb's paradox). On this zero-level model, the self-signaling implications of an action are independent of motives. The fact that feelings of shame or guilt can arise

---

[2] The assumption that people only remember choices, but not the motivational and informational states which led up to those choices, is invoked (in very different settings) by Ariely et al. (2000), Benabou and Tirole (2000), Hirschleifer and Welch (1998), and Koszegi (1999). Koszegi in particular postulates an "ego-utility" function, whose argument is the expectation of θ, conditional on past actions, i.e., Koszegi's agent evaluates $V(E\{\theta|x,C\})$ instead of $E\{V(\theta)|x,C\}$ as in our model. We discuss the work of Benabou and Tirole in more detail in the concluding section of this paper.

even in situations where a person is not "at fault" suggests that such interpretations have psychological reality.

Face-value interpretations

One level up, we have "face-value" interpretations, which derive $f(\theta|x,C)$ endogenously from the assumption that an action reveals θ that maximizes outcome-utility component of total utility. A person whose actions were made innocent of their diagnostic implications would be justified in maintaining face-value inferences. Diagnostic utility would be experienced as an unintentional byproduct of choice, not something that consciously affected choice.

True interpretations

A third possibility is that the revised distribution $f(\theta|x,C)$ is rational in a stronger sense, namely, that it is consistent with maximization of both components of (1) and with Bayes' rule. A choice reveals θ for which that choice is optimal. The only fine point is how to interpret actions that cannot be construed as optimal for any θ. To ensure that all actions have interpretations, we define "plausibility of θ given choice x from C" as the difference between the total utility of action x (assuming that θ is the actual disposition) and the highest total utility attainable from C (again, assuming θ):

Plausibility of θ given choice x from C = utility of x for θ - maximum utility θ can gain from C

$$P(\theta,x,C) \;=\; U(x,C,\theta) - \max_{y \in C} U(y,C,\theta). \qquad (2)$$

Interpretations are true if $f(\theta|x,C)$ places positive probability only on those θ for which x maximizes total utility, and if no such θ exists, on the most plausible θ, according to (2).[3] Bayes' rule resolves "ties," where more than one θ is maximally plausible. We call these interpretations 'true' because they correctly discount the signaling value of an action for the fact that the action is partly motivated by self-signaling. They are 'true-to-reality,' so to speak.

We can summarize the unbiased, fully rational self-signaling model with two assumptions:

Assumption 1 (Optimization) I choose so as to maximize total utility (equation 1), taking into account what my action might reveal about my true preferences.

[3] For face-value interpretations, the criterion in (2) would have the same form but with outcome utilities $u(x,\theta)$ replacing total utilities $U(x,C,\theta)$.

Assumption 2 (True interpretations) What my choice reveals about my true preferences is precisely that they are the preferences for which that choice maximizes total utility, or comes closest to maximizing total utility (equation 2).

Readers familiar with economic theory will recognize here the concepts of signaling games (see Bodner and Prelec, 1997a, for a reduction of this problem to a degenerate signaling game).[4] Those not familiar with economic signaling models may be struck by the circularity in the assumptions: In order to know how to solve the optimization problem in Assumption 1 one needs to have ready an interpretation of each possible action, but the interoperations of actions in Assumption 2 presupposes knowledge of which action is optimal for each θ. The circularity is intentional and tailor-made for game-theoretic analysis. The solutions are 'equilibria' in which actions are optimal in light of how they will be interpreted, and interpretations are true given the optimality of actions.

Some examples

Here are seven examples where self-signaling might play a role:

Example 1 x is '$ left at the casino" and θ 'disposition to gambling.'

Example 2 x is '$ donation to the United Way' and θ 'the level of true concern about the activities supported by United Way.'

Example 3 x is '$ spent on a wine bottle (no special occasion)" and θ is 'financial responsibility.'

Example 4 x is 'taking or not taking a drink before noon' and θ is disposition to alcoholism.

Example 5 x is 'embarking or not embarking on a dangerous mountain-climbing expedition,' and θ is 'perseverance.'

Example 6 x is 'quitting or not quitting your regular job' and θ is 'acting ability.'

Example 7 x is 'voting or not voting' and θ is 'dedication to the candidate.'

Remark 1 In many, if not most of these examples the act in question has significant causal consequences as well. Taking the drink in Example 4 changes body chemistry, increasing physiological

[4] The game is between a Sender that executes choices and a Receiver whose only function is to set the diagnostic utility term equal to the expectation of $V(\theta)$, as in (1). Could we interpret the Receiver as a sort of Conscience or Superego struggling to control the Sender's baser impulses? We resist such an interpretation because any genuine model of a "higher Self" should ascribe to it preferences over outcomes, presumably different from the preferences of the "lower Self." For example, in Thaler and Shefrin's Planner-Doer model (1981), the conflict between the Planner and the Doer arises because both sides have preferences over consumption streams, but the Planner's preferences exhibit a smaller discount rate. The Receiver here has no preferences over actions or over outcomes — it is simply a Bayesian calculator, enforcing a kind of objectivity on interpretations of actions.

dependence on alcohol. To the extent that his effect is recognized by the decision-maker, it would be absorbed in u(x,θ), along with any other causal consequences. However, in this example, and in many others like it, the purely causal consequences of the deviant action are slight, and a person's concern cannot be plausibly ascribed to it. Our model hopes to explain is why seemingly trivial actions can have great subjective significance. In the self-signaling model, the diagnostic value of an action may be all out of proportion to its scale — a small gesture can reveal character quite effectively (e.g., stealing 25¢ still counts as theft).

Remark 2 In many, if not most of these examples, the person may not be intrinsically concerned about having the particular disposition; rather she is concerned about specific consequences that might result from having this disposition. For instance, a person may not care about endurance and perseverance per se, but only the career benefits that she expects to flow from this trait. One can distinguish therefore between intrinsic and instrumental self-signaling.[5] There is no difficulty in dealing with instrumental self-signaling provided that the future consequences of a given disposition do not involve any additional future decisions. The calculation of V(θ) would simply expand to include not just intrinsic concern about θ but also about all future consequences that derive from θ, adjusted by probability and time delay. This would be the natural way to model 'heart condition' in the Quattrone and Tversky experiment. If, however, these future consequences are predicated on additional decisions by that same person, then the problem would have to be analyzed as a dynamic game between a succession of multiple-selves, each of which is endowed with a self-signaling utility structure. The same applies to the final example, voting. The level of personal dedication to a candidate is not likely to have great intrinsic importance. However, to the extent that such dedication levels are correlated across the population, my dedication predicts the dedication of others, and hence their inclination to vote. Assuming other voters self-signal as well, in equilibrium my vote may be valid evidence for whether others will vote.

Remark 3 Remark 2 notwithstanding, the distinction between intrinsic and instrumental self-signaling may not be as clear cut psychologically as it is logically. Consider an ostensibly instrumental disposition like willpower, and imagine that you have just scored in the top decile on some accurate psychological test of this trait. In order to feel happy by this news you don't need to elaborate in detail exactly how willpower might prove useful in the future. You have a rough sense that willpower will help in all kinds of situations, so the news is good. The decile rank in willpower functions exactly like the heart condition

[5] Similarly, Koszegi (1999) distinguishes between "pure self-image" and "anxiety or worry about the future."

in the Quattrone and Tversky experiment — it is a general predictor of future quality of life on some important dimensions. Psychologically, therefore, the intrinsic self-signaling model may extend to forms of instrumental self-signaling, at least in situations where the underlying traits have diffuse potential benefits, contingent on unforeseeable opportunities and choices.

Remark 4 Example 3 raises issues of generalization, as discussed by Gilboa and Gilboa- Schechtman (2001) in this volume. Is the action diagnostic of "general financial irresponsibility," or "overindulgence on wines," or "overindulgence on Burgundies after a hard day at the office?" There is nothing in the model that picks out one or another level of generalization. Formally, the level of generalization could be made endogenous by treating it as a cognitive decision variable, subject to similar diagnostic motivation as the action itself (i.e., narrow generalizations would facilitate excuses). Generalization involves psychological notions of similarity and grouping, which fall outside of standard economic modeling (Prelec, 1991; Gilboa and Schmeidler, 1995)..

Remark 5 To the extent that a disposition is revealed through a series of choices, wouldn't a lifetime of behavioral evidence overwhelm the information content of a single action? Note, first, that as far as feelings are concerned, we often do seem to ignore the long-run track record and give excess weight to the most recent experience. This psychological bias would enhance diagnostic motivation. Second, and more important, the problem of drawing inferences from actions is not as simple as we had made it out to be. One can often choose how much to blame oneself and how much to blame outside circumstances.[6] A more realistic model would write the momentary disposition at time t as the sum of a stable component and an unobservable temptation, $\theta(t) = \theta + \tau(t)$, and then consider how the person might extract the signal, θ, from the noisy behavioral record. Moreover, if θ is subject to drift, e.g., as in: $\theta(t) = \theta(t-1) + \tau(t)$, then the mere passage of time would create uncertainty about θ and restore diagnostic motivation. A person would periodically need to "check" that $\theta(t)$ is still in the good range.

Remark 6 Finally, doesn't the true interpretation model require a weird combination of self-ignorance (i.e., of one's own dispositions) and self-awareness (i.e., of one's propensity to self-signal)? Well, as a matter of sheer cognitive skill, we are certainly able to discount behavioral signals *of other people* when we suspect ulterior motives on their part. What is less clear, however, is whether in fact we apply to our own actions the same rigorous interpretive standards that we apply to the actions of others. To the extent that we fail to do so, the face-value model of interpretations will be more correct.

---

[6] Benabou and Tirole (2000) analyze the attribution-of-blame problem when memory of past actions and past temptation levels is imperfect. See also Bodner and Prelec (1997; Section 5).

3 Face-value interpretations and excessive self-esteem

Looking at equation 1, one might think that the optimal choice would reflect a compromise between outcome and diagnostic utility, so that actions would diverge from the natural, outcome-utility maximizing levels in the direction diagnostic of preferred dispositions. This is how things turn out with 'face-value' interpretations, but not with 'true' interpretations. In the true-interpretations case, the signaling value of good actions is discounted for the diagnostic motive, which creates an escalating pressure for behavioral perfection. The generic result (described in Section 4) is that either diagnostic utility wins, and the person does the same 'perfect' thing irrespective of disposition, or natural impulses win, and the person simply ignores the diagnostic component of the utility structure.

Let's now consider the case where the disposition pertains to a nasty vice, such as gambling, and could be at one of three levels: *Low*, *Moderate*, and *High,* all equally likely. The level, x, of the problematic activity could be zero or any positive amount up to some maximum level. The meta-utility function $V(\theta)$ is decreasing in θ, so that lower dispositions are intrinsically preferred. Outcome utility, $u(x,\theta)$, is twice continuously differentiable, strictly concave, with $\partial u/\partial x$ increasing in θ, which indicates that higher θ implies more appetite for the activity. These assumptions ensure the existence of a unique equilibrium. We will refer to the consumption levels that maximize just outcome utility as the *natural consumption levels*, and, for the sake of interest, assume that they are positive for all θ, e.g., even a person with the best *Low* disposition would prefer to gamble a small amount.

With these assumptions, and with face-value interpretations, we might observe an equilibrium such as described in Table 1 below. The three natural levels are labeled as "Light," "Moderate," and "Heavy," in the middle column. The first set of arrows indicates optimal actions for each disposition, and the second set of arrows the corresponding face-value interpretations. Solid arrows indicate interpretations of actions selected with positive probability in equilibrium, and dashed arrows actions never selected in equilibrium. In equilibrium, the person cuts back gambling by one step from the natural level. Face-value interpretations generously ignore the diagnostic motive for the reduction, and induce a self-image that is too good by one Δ-step, except for the best disposition, which is correctly diagnosed.[7]

---

[7] Other models that give rise to an excessively positive self-image are presented by Carrillo and Mariotti (2000), Brocas and Carillo (1999, 2000), Benabou and Tirole (1999), and Koszegi (1999).

f(□)	Disposition to vice	Consumption	Face-value interpretation	True interpretation
		Abstain		
.33	*Low*	Light	*Low*	*{Low or Moderate}*
.33	*Moderate*	Moderate	*Moderate*	*High*
.33	*High*	Heavy	*High*	*High*

Table 1 An equilibrium that can arise only with face-value interpretations, and which produces on average an overly positive self-image (a 2/3 chance of *Low* and 1/3 chance of *Moderate*).

If interpretations are true, then this obviously will not work. The rightmost column in the table gives the true interpretations in the postulated equilibrium. The Moderate level of consumption is diagnostic of the *High* rather than the *Moderate* disposition, and Light consumption of either *Low* or *Moderate*, rather than *Low* for sure.

As an empirical hypothesis, face-value interpretations may closer to psychological reality. There is much evidence that self-assessments are excessively positive (e.g., Taylor and Brown, 1988). Furthermore, in the specific context of the Quattrone-Tversky experiment, which is our empirical cornerstone, most of the participants did not acknowledge ex-post any conscious efforts to influence the results of the cold-pressor test. The small minority of subjects who did confess to trying to bias the results were also relatively pessimistic about their own chances of having a good heart. The subject population apparently divided into a self-satisfied, face-value interpretations majority, and a pessimistic, true-interpretations minority.[8]

4 True interpretations and rule-governed action

What consumption levels would then be consistent with true interpretations? There are only three possibilities, identified in Tables 2, 3, and 4. When diagnostic utility has no weight whatsoever, consumption is just at the natural levels, shown in Table 2. The right column gives the obvious interpretations. In this case, you do as you please and what you do makes transparent who you are.

[8] To explain this kind of self-deception, it would be sufficient to hypothesize that the plausibility function in equation 3 places less weight on diagnostic utility than does the optimal action rule in equation (1).

f(□)	Disposition to vice		Consumption		True Interpretation
			Abstain	⇢	
.33	*Low*	→	Light	→	*Low*
.33	*Moderate*	→	Moderate	→	*Moderate*
.33	*High*	→	Heavy	→	*High*

Table 2 A separating equilibrium, when diagnostic utility is weak.

What happens when the weight of diagnostic utility is increased slightly from zero? Initially, nothing — the optimal actions remain the same. A person with a *High* disposition would perhaps feel some inclination to reduce consumption, but could only generate a positive signal by reducing consumption to the moderate level, set by the *Moderate* disposition. If diagnostic utility weak, it will not justify the discrete reduction in consumption. As diagnostic utility becomes stronger, a different, partially separating equilibrium emerges, in which consumption falls to zero for the two better dispositions, and remains Heavy for *High* (Table 2). Notice that zero consumption does not maximize outcome utility for any disposition. The fact that it emerges as on optimal choice with self-signaling is an example of "excessive virtue," sustained by the harsh interpretation of any positive level of consumption. The psychological intuition might go as follows: A person who is concerned about his inclinations to gambling, and who has, as a result, never ventured into a casino, would treat even one lapse as evidence of a strong gambling urge. The person might say — "given how much I care not to discover that I have a taste for gambling, then I must indeed have a strong taste for it if I succumb on this occasion!" Moderation is not an option.[9]

[9]A separate benefit of abstention is that it denies the person the opportunity to learn how much she really likes a potentially addictive substance. Carrillo (1998) shows how by "enforcing ignorance" abstention can become the optimal second-best strategy in a situation where fully-informed hyperbolic-discounting agents cannot precommit to moderation.

f(□)	Disposition to vice	Consumption	True Interpretation
.33	*Low*	Abstain	*{Low* or *Moderate}*
.33	*Moderate*	Light	
.33	*High*	Moderate	*High*
		Heavy	

Table 3 A partially separating equilibrium.

Finally, when diagnostic utility is completely dominant, a person with a *High* disposition will also abstain (Table 3). It is natural to interpret this pooling equilibrium as a rule, inasmuch the same action is taken independent of disposition. In this situation, even though abstention is certain, this fact does not provide any reassurance about the underlying disposition — it is still equally likely to be any of the three types. Indeed, in a repeated choice setting, perfect compliance over many trials would still not reveal the underlying disposition, and it is this very fact that sustains the diagnostic motive. If one could infer that one had a perfect disposition after a certain number of abstentions, then one could afford 'to relax' and gamble occasionally. *Here, the rule remains in force precisely because following the rule is not informative*.

f(□)	Disposition to vice	Consumption	True Interpretation
.33	*Low*	Abstain	*{Low*, *Moderate* or *High}*
.33	*Moderate*	Light	
.33	*High*	Moderate	*High*
		Heavy	

Table 4 Strong diagnostic utility promotes complete pooling — an "abstention rule."

Tables 2, 3, and 4 are the only possibilities that the model allows, if interpretations are true. Interestingly, one cannot have an equilibrium that pools on the Light or Moderate level. If, for example, pooling-on-Light is entertained as an equilibrium, then the best disposition (*Low*) would be motivated to reveal itself by reducing consumption just a tiny bit.

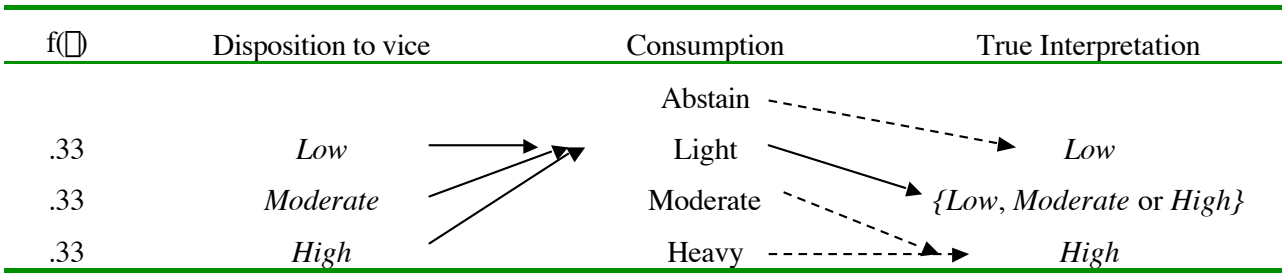

Table 5 An impossible equilibrium, with pooling on Light consumption. Interpretations are true but actions are not optimal. Any consumption level below Light signals the best disposition, so there is an incentive to reduce consumption by a small amount, from "Light" to "Light-θ". Note how the impossibility result depends on x being a continuous variable.

Looking over the three equilibrium cases (Tables 1, 2, 3), we see that the motivational struggle between outcome utility and diagnostic utility can be resolved in one of only two ways: Either outcome utility is stronger, and the person does as she pleases, or diagnostic utility is stronger, and the person chooses to abstain from consumption altogether.

Would these conclusions survive under different dimensional assumptions about x and θ? With continuous dispositions, we find separating equilibria with all dispositions consuming less than their natural, outcome-utility-maximizing level (Bodner and Prelec, 1997; Section 4; see also Bernheim, 1994). With a discrete rather than continuous action set, a single utility structure (equation 1) may yield multiple equilibria, including equilibria with pooling on positive consumption levels (e.g., something like Table 5). Informally, the model seems more sensitive to dimensional assumptions about the action space than about dispositional space.

5 Resolutions are more consistent with the ideal self

True interpretations allow a short menu of three generic equilibria, and which one obtains will depend on the relative weight of outcome and diagnostic utility. One way in which this balance can be tilted is if mere "intentions" or "resolutions to act" replace actions. Now, by a resolution we usually mean one of two things. The first sense is that there is something that one intends to do at some time in the future. A second sense is that there is something that one intends to do if called upon to do so (e.g., "I resolve to abstain if the tempting occasion presents itself"). The first sense has to do with delay, and the second sense combines uncertainty and delay. In either case, we can compare resolutions, where consequences are uncertain and delayed, with actions having the same utility structure but where the

outcomes are certain and immediate. In both cases, we imagine that the resolution is binding (i.e., there is no possibility of escaping the commitment ex-post).

Looking first at the case of a contingent resolution, the ex-ante decision problem is whether to commit to consume a given amount (or none at all) if the opportunity arises. Because the decision is binding, the self-signaling utility structure in equation 1 is changed in only one respect—there is a probability, p, that the person will actually have to implement his resolution, and probability $(1-p)$ that he will not. Hence, the only change is that outcome utilities are reduced by a factor of p, while the diagnostic part of the equation remains the same. The reduction in relative outcome weight promotes choices diagnostic of good dispositions, in equilibrium (Bodner and Prelec, 1997).

Bodner's experiments (Bodner, 1995) on contingent charitable pledges exhibit exactly this phenomenon. Contingent pledges are promises to give in the event that one is called upon to do so. Bodner found that the pledges of subjects who regarded themselves as insufficiently altruistic were relatively more sensitive to the stated probability of being called upon to give: they were relatively more generous when that probability was small. In effect, such subjects were purchasing self-esteem "on the cheap," by pledging more when the likelihood of actual sacrifice was low.

Self-esteem can also be purchased on the cheap when the decisional consequences are far away in time. Outcome utility is now temporally, rather than probabilistically discounted, again making the good action a more likely choice. The chronic discrepancy between future plans and actual behavior may therefore be explained by two facts: (1) the diagnostic utility of a resolution is immediate while the outcome cost of that resolution is discounted, and (2) people fail to anticipate the reversal in preference, i.e., that they are naïve in the sense of O'Donoghue and Rabin (1999).

This account provides a single explanation of what are otherwise different categories of dynamic inconsistency. Anything that selectively lowers the weight of outcome utility (low physical salience, uncertainty, time distance, and so forth) will result in choices being more driven by meta-utility. When the weight of outcome utility is restored (by making outcomes salient, certain, imminent, etc..), the person may regret his earlier choice.

6 Moral placebos

Consider the following problem: Imagine that you have a friend concerned about an underlying disposition to vice. He has abstained so far but is not sure whether he will be able to maintain this policy in the future. What beliefs would be most effective in maintaining his abstention policy, holding constant the combined utility structure, u(x,□) and V(□)? Your only manner of influence is to shift f(□) in some particular direction by reasoning and persuasion, i.e., you cannot change his actual disposition.

Informally, one may identify two approaches here. The first approach, "from fear," would say that it is good to increase subjective probabilities of bad dispositions. The second approach, "from self-worth," would say that boosting the probability of good dispositions increases the subjective stake in abstaining, making abstention more likely.

It turns out that the second approach is more effective if interpretations are true. Let us compare the plausibility of a pooling or "abstain" equilibrium under optimistic and pessimistic prior beliefs. When the self-image is optimistic, then abstention is interpreted as a high likelihood (90%) of the best disposition, while any consumption triggers the most undesirable interpretation. Hence the diagnostic "stakes" are very high, increasing the chance that this equilibrium will be maintained (Table 6 below).

f(□)	Disposition to vice	Consumption	True Interpretation
		Abstain	
.90	*Low*	Light	{Probably (90%) *Low*}
.09	*Moderate*	Moderate	
.01	*High*	Heavy	*High*

Table 6 An optimistic self-image increases the likelihood of an Abstaining equilibrium.

When the self-image is pessimistic (Table 7), then the diagnostic benefits of abstaining are slight. A person who abstains still faces a subjective 90% probability of having a bad disposition, hence the diagnostic penalty for indulging is small. The likelihood that this pooling equilibrium will arise with pessimistic beliefs is low. This constitutes an argument in favor of "positive thinking" insofar more favorable opinions about one's disposition provide a larger stake for abstaining.[10]

[10] Benabou and Tirole (2000) derive a similar result (Proposition 1) in context of an intertemporal, multiple-selves model.

f(θ)	Disposition to vice	Consumption	True Interpretation
.01	*Low*	Abstain / Light	{Probably (90%) *High*}
.09	*Moderate*	Moderate	
.90	*High*	Heavy	*High*

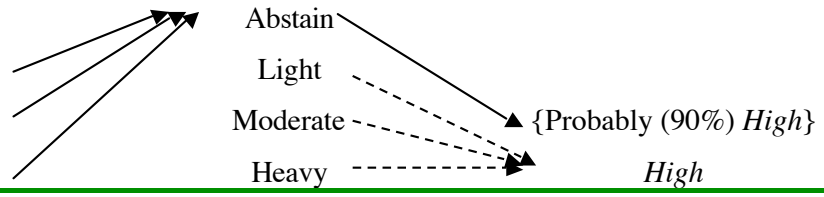

Table 7 A pessimistic self-image is not likely to produce an Abstaining equilibrium. The diagnostic benefits of abstaining are smaller than in Table 6.

We can call the effect shown across Tables 6 and 7 a *moral placebo*, inasmuch the ability to abstain is supported by a positive self-image that need not have any basis in reality. Consider a person having a bad *High* disposition, but who is not sure of that. At low levels of temptation (low weight on outcome utility), the pooling equilibrium in Table 3 obtains and the person abstains from consumption. At high levels of temptation (high weight on outcome utility), the separating equilibrium in Table 1 or 2 obtains, and he consumes heavily. At intermediate levels of temptation, however, whether one or the other equilibrium obtains depends on his self-image f(θ). By replacing the probabilities in Table 7 with those in Table 6 (i.e., consuming the "moral placebo"), the person's ability to withstand temptation increases.

Moral placebo effects open the door to other nonstandard influences on actions. If dispositions are correlated across individuals in a particular group, then observing the actions of others in the group provides information about one's own disposition. For instance, if Persons *A* and *B* believe that their dispositions are drawn from a common prior, $f(\theta^A,\theta^B)$, and if *A* sees *B* abstaining from consumption, that may make A's self-image more favorable, thereby increasing the likelihood that she will abstain. In other words, the information that someone else has resisted temptation promotes my own abstention, even if there are no causal connections between us and my action is entirely private

Could a moral placebo effect be created by information provided by one's own earlier choices, that is, could past actions — decisional precedents — influence present choices purely by changing self-beliefs? The general answer is Yes, but the particulars will depend on how the intertemporal choice problem is set up, and on whether the intertemporal dependencies are themselves taken into account in earlier choice (i.e., whether the individuals are 'naïve' or 'sophisticated' in the sense of O'Donoghue and Rabin (1999)). The extent that past choices can provide information about the current disposition depends also on whether a person believes in a fixed disposition ("Calvinism") or whether she thinks that dispositions are variable. Under the Calvinist variant, once a bad θ is revealed no further action can repair the damage. If, however, θ is a stochastic process, then one's own earlier choice provides

information in exactly the same way as do choices of other persons. If the information is positive, i.e., if one has a good track record, the chances of abstaining in the future will tend to go up. When dispositions are not fixed, therefore, an action diagnostic of favored dispositions has a multiplier effect, increasing the probability that such an action will be repeated the second time that a similar choice opportunity arises. It is as if each moral success contributes an increment of "moral capital," making future successes more likely,

7 One Self or many Selves?

The conjecture that that people can learn from their own actions has figured prominently in the psychological writings of George Ainslie (1992), and has, more recently, been given a different theoretical justification by Benabou and Tirole (2000, 2001). Because the objectives of Benabou and Tirole have some overlap with our own, we now briefly outline and compare their model.

Benabou and Tirole (2000) work within an intertemporal, multiple-selves context, where each temporal self is strategically sophisticated with respect to its future incarnations. They distinguish between 'private' information, which one cannot communicate to future selves except through actions, and 'public' information, available to all selves (there are also intermediate cases, where information has some chance of being incorrectly recalled). The critical feature of this model relative to our own is that it upholds the rationality of the individual decision making process at a given point in time. A key issue for any such model is to reconcile two seemingly contradictory requirements:

(1) Self-transparency: Every Current Self knows it's own preferences (including "willpower").
(2) Self-signaling: The Current Self is influenced by actions (i.e., "signals") of Past Selves, even though those actions do not affect her preferences *per se*.

(1) is an economic model-building principle, namely, that incomplete self-knowledge can only arise intertemporally, when you either forget your past actions or motives, or when you cannot forecast your future preferences. (2) is the desired result. The challenge resides is making knowledge of past actions informationally useful to the current Self, notwithstanding its 'self-transparent' nature. Benabou and Tirole respond by dividing a stylized self-control problem into two subdecisions, a "morning" decision whether to indulge all day or embark on a work project, and a second, "afternoon" decision whether to finish the work or stop short of the goal. This scenario repeats on the second day (and could in principle be extended to subsequent days). Hyperbolic discounting (or "weakness of will") makes the

morning Self curious about what the afternoon Self will do: Specifically, the morning Self prefers to start working only if she believes the afternoon Self will follow through and finish the job. Even though the morning Self knows her willpower exactly, the morning decision is different from the afternoon decision, so the morning Self has to look backward to what happened "under similar conditions" yesterday afternoon to estimate the chances that willpower *this* afternoon will be up to snuff. The imperfect correlation of willpower across time periods endows earlier decisions with informational value, even to a self-transparent Self. The morning Self conditions her actions on her expectations of what she will do later that afternoon and her action yesterday afternoon provides the best evidence for that.

Benabou and Tirole (2000) demonstrate that self-transparency and strategic sophistication are logically compatible with many of the self-reputation phenomena that have been discussed in the psychological literature, most notably by Ainslie (1975, 1986, 1992). The model is close in spirit if not in detail to Ainslie's story, in the sense that Ainslie also develops a rich psychology of self-control from hyperbolic discounting alone, without requiring any conflict between different personae *within* a given temporal Self (see especially the discussion of Freudian mechanisms in Ainslie, 1982).

Our model is based on somewhat different set of psychological intuitions. Looking at the psychological evidence, we are impressed with the seamless quality of self-signaling and self-deception. When individuals manipulate their 'medical test' results (Quattrone and Tversky, 1984), personality self-reports (Sanitioso et al., 1990; Kunda, 1990; Dunning et al., 1995), or problem solving strategies in a desired direction (Ginossar and Trope, 1987), it is hard to discern two agents, the one who *signals* and the one who is *signaled to*.[11] For this reason, we favor the hypothesis that all acts of volition are *at the outset* biased by diagnostic motivation. This is a doctrine of volitional "original sin" insofar the desire for good news compromises thoughts and actions as they are formed and enter into consciousness. A person may be rationally aware of this biasing force, as in the "true interpretations" version, or completely unaware of it, as in the "face-value" version, but in either case there is no moment in time when he is truly self-transparent. Without self-transparency, the notion of self-signaling becomes more literal, involving one Self (at most).

---

[11] To be fair, nothing in the intertemporal approach requires each Self to be fully conscious and to occupy an extended time period. Indeed, Ainslie (1992) has conjectured that the sensation of pain is produced by yielding to the temptation of an extremely brief attentional pleasure, a pleasure so brief that it cannot be detected consciously yet sufficiently strong to 'lock' attention on the pain-producing stimulus.

REFERENCES

Ainslie, G., (1975). “Specious reward: A behavioral theory of impulsiveness and impulse control,” *Psychological Bulletin,* 82, 463-509.

Ainslie, G., (1982). “A behavioral economic approach to the defense mechanisms: Freud's energy theory revisited,” *Social Science Information*, 21, 735-779.

Ainslie, G., (1986) "Beyond microeconomics: Conflict among interests in a multiple self as a determinant of value." In *The multiple Self,* Elster, J. (ed.), Cambridge: Cambridge University Press.

Ainslie, G. (1992) *Picoeconomics: The Strategic Interaction of Successive Motivational States within the Person*. New York Cambridge University Press.

Ariely, D., Loewenstein, G., and D. Prelec. (2000). “Coherent arbitrariness: Duration sensitive pricing around an arbitrary anchor.” MIT mimeo.

Bem, D. J. (1972), “Self-perception theory,“ In *Advances in experimental social psychology* (Vol. 6). L. Berkowitz (Ed.), New York: Academic Press.

Benabou, R., and J. Tirole (1999). “Self-confidence: Intrapersonal strategies,” IDEI mimeo, June.

Benabou, R., and J. Tirole (2000). “Willpower and personal rules,” Princeton mimeo, June.

Benabou, R., and J. Tirole (2001). “Self-knowledge and self-regulation: An economic approach” (this volume).

Berglas, S. and E. E. Jones, (1978) "Drug choice as a self-handicapping strategy in response to noncontingent success," *Journal of Personality and Social Psychology,* Vol. 36, 4, 405-417.

Berglas, S. and Baumeister, R. (1993). *Your Own Worst Enemy: Understanding the Paradox of Self-Defeating Behavior*. BasicBooks: New York.

Bernheim, D. B. (1994), “A theory of conformity,” *Journal of Political Economy*, 102, 5, 841-877.

Bodner, R. (1995). “Self knowledge and the diagnostic value of actions: The case of donating to a charitable cause.” unpublished Ph.D. dissertation, MIT, Sloan School of Management.

Bodner, R. and D. Prelec. (1997). “The diagnostic value of actions in a self-signaling model, MIT mimeo, January.

Bratman, M. E., (1995). “Planning and temptation” in Friedman and Clark, eds., *Mind and Morals,* Bradford/MIT press.

Brocas, I., and Carrillo, J. (1999). "Entry mistakes, entrepreneurial boldness and optimism," ULB-ECARES mimeo, June.

Brocas, I., and Carrillo, J. (2000). "Information and self-control," this volume.

Campbell, R. and Sowden, Lanning eds. (1985) *Paradoxes of Rationality and Cooperation* Vancouver: University.

Carrillo, J. (1998). "Self control, moderate consumption, and craving," CEPR D.P. 2017, November.

Carrillo, J., and T. Mariotti (2000). "Strategic ignorance as a self-disciplining device," *Review of Economic Studies*, 76(3), 529-544.

Elster, J. (1985). "Weakness of will and the free-rider problem." *Economics and Philosophy. 1*. 231-265.

Dunning, D., Leuenberger, A., and Sherman, D. A. (1995). A new look at motivated inference: Are self-serving theories of success a product of motivational forces? *Journal of Personality and Social Psychology*, 69, 58-68.

Gilboa, I. and E. Gilboa-Schechtman, "Mental accounting and the absentminded driver," this volume.

Gilboa, I. and D. Schmeidler (1995), "Case-based decision theory," Quarterly Journal of Economics 110: 605-639.

Ginossar, Z., and Trope, Y. (1987). Problem solving in judgment under uncertainty. *Journal of Personality and Social Psychology*, *52*, 464-474l.

Hirschleifer, D., and Welch, I. (1998). "A rational economic approach to the psychology of shange: Amnesia, inertia, and impulsiveness." mimeo, November.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480-498.

Koszegi, B. (1999). "Self-image and economic behavior," MIT mimeo, October.

Laibson, D. I. (1997). "Golden eggs and hyperbolic discounting," *Quarterly Journal of Economics*, *112*: 443-478.

Nozick, R. (1969) "Newcomb's problem and two principles of choice," in Nicholas Rescher et. al., eds., in *Essays in Honor of Carl G. Hempel.* Dordrecht: Reidel.

O'Donoghue, T., and Rabin, M. (1999). "Doing it now or later," *American Economic Review*, *89(1),* 103-124.

Prelec, D. "Values and principles: Some limitations on traditional economic analysis," in *Socioeconomics: Toward a New Synthesis,* A. Etzioni and P. Lawrence (Eds.), New York: M.E. Sharpe, 1991.

Quattrone, G. A., and A. Tversky, (1984) "Causal versus diagnostic contingencies: On self-deception and on the voter's illusion," *Journal of Personality and Social Psychology*, 46, 2, 237-248.

Sanitioso, R., Kunda, Z., and Fong, G. T. (1990). Motivated recruitment of autobiographical memory. *Journal of Personality and Social Psychology*, *59*, 229-241.

Shafir, E. and A. Tversky, (1992). "Thinking through uncertainty: Nonconsequential reasoning and choice." *Cognitive Psychology*, 24, 449-474.

Taylor, S. E. and Brown, J. D. (1988). "Illusion and well-being: A social psychological perspective on mental health," *Psychological Bulletin*, *103*, 193-210.

Thaler, Richard and H. M. Shefrin, (1981). "An economic theory of self-control, " *Journal of Political Economy*, 89, 393-410.