From: Time and Decision, G. Loewenstein, D. Read, & R.F.
Baumeister (eds.) Russell Sage Press, New York, 2003.

# Self-signaling and self-control

Drazen Prelec and Ronit Bodner

Sloan School, MIT

# 1 Motivation without causality

Self-control is a hallmark virtue of human character. To lack self-control is to be governed by momentary pleasures even when these pleasures place larger values at risk. Willing the tired body to exercise or the tired mind to another hour of work are but two examples of active self-control — the tolerance of pain in return for a larger but more remote and uncertain gain. Turning down a chocolate dessert or an attractive sexual encounter are examples of passive self-control — avoiding immediate gratification in order to preserve broader personal objectives or self-esteem. The importance of self-control for individual happiness and social welfare is not controversial (Baumeister, Heatherton and Tice, 1994).

Theoretical approaches to self-control have largely focused on the temporal aspect of the problem, the conflict between the near and the far. This conflict, however, does not capture one key aspect of the self-control decision. Let us take a workhorse example — dieting. The dieter faces a series of temptations each of which involves a choice between (a) yielding and eating the tempting morsel, or (b) resisting temptation and gaining… gaining what precisely? At the level of a single action, the benefits of self-control are obscure. The caloric impact of one act of indulgence is negligible and will not affect the waistline, let alone any larger objectives. As described by Herman and Polivy in their chapter on dieting (this volume):

> “She [the dieter] can resist that tempting plate of cookies, but there’s no certainty that she will become slim as a result. The stubborn fact of the matter is that weight loss is only vaguely connected to food intake… Finally, it may be that the reward that the dieter is after is something more than slimness. For some dieters, slimness may be an end in itself; but for many other dieters, slimness is simply a means to an end. Being slim is how one

> becomes attractive; being slim is how one becomes healthy; being slim is how one displays one's virtue."

The gap between a single act of dietary self-sacrifice and its benefits — slimness, leading to beauty, health and virtue — is not only temporal but also *causal*. A single low-calorie salad does not cause slimness, beauty, health or virtue. Rather, the benefits or costs are properties of a longrun policy or lifestyle, and an integral policy or lifestyle cannot be obtained with a single decision, except in extreme cases, like joining the Foreign Legion. In an earlier paper (Prelec, 1991), one of us referred to this as a mismatch of "scale", where the effects of an action only become visible if the action is repeated many times, across different contexts, and in conjunction with other supporting activities (e.g., exercise in the case of dieting).

Our ability to persevere in spite of the tenuous connections between individual actions and broad lifestyle objectives is both essential to normal functioning and a major challenge for psychology and economics. As several authors have pointed out (e.g., Elster, 1989), the theoretical problem resembles the paradox of voting in political science. Voters go through some trouble to cast their vote, though their individual effort cannot affect the outcome (no major election has been, or is likely to be decided by exactly one vote). Here, again, we have an instance of motivation without clear causality. Whatever motivates the voter, it cannot be only the ceremonial or participatory pleasures of the exercise. Few would stand for hours to vote at a broken polling booth. People want their vote to count, and they will vote only if they think there is some chance their vote *will count*, i.e., affect the final outcome.

The feeling that larger objectives are put at risk even by small-stakes choices seems essential to self-control, to persistence and endurance over the long haul. Intuitively, success in a small matter is a *signal* of success in larger matters; it allows one to enjoy for a moment the expectations of attaining the larger good.[1] As such, the signal — whatever it's intrinsic nature — is an independent source of immediate pleasure, as nicely described by Herman and Polivy (this volume):

> "Indeed, the very hunger pangs that signal abstention, and which are allegedly so difficult to bear, provide a clear signal of success. Going to bed hungry is usually regarded as an aversive experience; for the dieter, however, it may provide a sense of accomplishment. This hedonic dialectic—in which pain confers pleasure, and dieting is an exercise in masochism—makes the application of standard behavioral economic calculations questionable."

In a recent paper (Bodner and Prelec, 1997, 2001) we have proposed a "self-signaling" model of diagnostic motivation, which we now extend to self-control. The model rests on a distinction between two types of reward (or utility): reward that flows directly from the causal consequences of choice, whether these consequences are immediate or delayed, and diagnostic reward, which is the pleasure or pain derived from learning something positive or negative about one's own internal state, disposition, ability, or future prospects. People are presumed to be chronically uncertain about where they stand with respect to these broad attributes, which in turn makes their choices diagnostic. For example, taking a drink before noon is diagnostic of alcoholism; hard exercise is diagnostic of health, willpower, perhaps even financial success, and so on. Anticipation of such diagnostic reward, or fear of diagnostic pain promotes self-control and inhibits self-indulgence.

[1] In Ainslie's terms, you "stake the expectation" of longrun success on successful performance in the here and now (Ainslie, 1992; 2001). As with voting, the logic follows the Kantian categorical imperative: If I fail now, then I will fail on all subsequent occasions of this type (Gilboa and Gilboa-Schechtman, 2001).

The model is an exercise in behavioral economics in that the person's choices are governed by maximization of utility or pleasure, albeit with the diagnostic component added to the equation. More importantly, economic analysis provides determinate predictions about the amount of good or bad news that any given action might yield — it provides, in other words, a precise theory of what actions *mean*. The core idea is that once you know that your choices are governed by maximization of total utility, then, even if you are somewhat in doubt about your true wishes and desires, you can assess in advance what any given choice will reveal about them.

## 2 Self-signaling in the lab and in everyday life

A self-signaling action is an action taken in order to obtain good news about one's underlying disposition or future prospects, even when this action has no causal impact on the disposition or prospects. The defnition immediately raises philosophical problems — if I take an action to get good news, doesn't make invalidate the "good news" (Campbell and Sawden, 1985; Elster, 1985)?

Whatever its logical status, self-signaling is a psychological reality, as several studies have shown (Dunning et al., 1995; Quattrone and Tversky, 1984; Sanitioso et al., 1990; Shafir and Tversky, 1992). The "cold-water test" experiment by Quattrone and Tversky (1984) is especially elegant, both as a definition of the self-signaling phenomenon and a demonstration of its existence. Quattrone and Tversky began by asking subjects to keep their hand submerged in a container of cold water until they could no longer tolerate the pain. Subjects were then told (via a "scientific lecture") that a certain inborn heart condition, leading to a shorter-than-average life expectancy, could be diagnosed by the effect of exercise on cold tolerance. One half of the subjects were told that having a bad heart would increase cold tolerance, the other half that it would decrease tolerance following exercise. Having absorbed this interesting piece of information, subjects then rode a treadmill vigorously for about a minute and repeated the cold water test. The vast majority showed changes in tolerance on the second cold trial in the direction correlated with "good news," thus demonstrating the influence of diagnostic motivation on behavior. Most subjects were not aware of any attempt to bias their test results.

Although the set-up in this experiment is somewhat unusual, the results fit well with a gread deal of other psychological research. We know that people manipulate personality self-reports (Sanitioso et al., 1990; Kunda, 1990; Dunning et al., 1995), problem solving strategies (Ginossar and Trope, 1987), and charitable pledges (Bodner, 1996) in a diagnostically favorable direction. From the literature on "self-handicapping," we know that a person might get too little sleep or under-prepare for an examination in order to create a situation where 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). The general notion that people adopt the perspective of an outside observer when interpreting their own actions is the core hypothesis in Bem's influential (1972) self-perception theory. It can also be traced back at least to the James-Lange theory of emotions (James, 1896/1981), which claimed that people infer their own emotions from behavior (e.g., they feel afraid if they see themselves running). This broader context of psychological research suggests that the phenomenon reperted by Quattrone and Tversky is not an isolated laboratory curiosity. If anything, the experimental results should underestimate the impact of diagnosticity in realistic decisions, where the absence of causal links between actions and dispositions is less transparent.

A key difference between the laboratory and the real-world setting is that in the real world it is hard to identify decisions with no causal consequences whatsoever. Here, for example, are five situations where part of the motive for engaging or not engaging in an activity is diagnostic. Each example identifies an action that might provide evidence of an underlying state:

(i) Donating time or money to a charitable cause as evidence of true concern for the cause.

(ii) Taking a drink before noon as evidence of a drinking problem.

(iii) Jogging during a 'heat wave' as evidence of willpower.

(iv) Purchasing an expensive item on credit as evidence of financial irresponsibility.

(v) Voting as evidence that one is dedicated to the candidate, and hence that other people like me will also take the trouble to vote.

In many of the examples — certainly in (i), (ii), (iii) and (iv) — the action in question also has causal implications. Taking the jog (example (ii)) does lead to an incremental improvent in fitness. What we propose, however, is that this small physical benefit does not exhaust the motivational forces compelling the person to exercise, that the larger part of the satisfaction may be due to the signal that the action provides. With diagnostic motivation the significance of an action need not bear any relation to the actual stakes – a small-scale action can be just as diagnostic as a large one. If you take a dollar from collection plate when no one is looking, that reveals that you are thief just as much as if you had taken the whole plate.[2]

[2] A second issue that arises in realworld examples is to what extent the signaling is to the 'self' and what extent it is signaling to others. Certainly, the motivational force of many the examples on the

Looking over the five examples, we can discern at least three distinct motives for self-signaling, which may be labelled *intrinsic*, *instrumental*, and *magical.* Self-signaling is intrinsic when a person cares about about an underlying trait or disposition independent of behaviors that might flow from this. For example, one might wish to believe that one has a good heart or soul, loves one's spouse and family, cares about those in need, appreciates good wine, has a particular sexual orientation, etc.. With instrumental self-signaling, by contrast, a person is concerned only with the specific consequences that a particular disposition promotes. For instance, a person may not care about endurance and perseverance per se, but only the career benefits that she expects to derive from this trait.[3] Finally, magical self-signaling applies to cases like example (v), where one cares about an underlying disposition because it correlates across the population, and hence predicts how other will behave. "Magical" refers to the feeling that by taking an action one is causing other people to behave in the same way. We shall not discuss the magical case further here (for experimental evidence, see Quattrone and Tversky, 1984; Shafir and Tversky, 1992).

## 3 The self-signaling model

In economic theory, a person is defined by their desires (as formalized with utilities) and beliefs (as formalized with probability distributions). Individual differences, in traits, dispositions, character, resolve into differences in desires or

list would be enhanced if others were there to observe the choice. That said, we claim that even if no one is watching, there is still some diagnostic pleasure or pain, in these examples.

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

beliefs. To say that a person is not sure of some underlying trait or disposition is to say that they are not sure of their true utility function (they might not be sure of their beliefs as well, but we will avoid this complication here).

To capture this notion formally, we let □ be the index of the unknown underlying disposition, x a possible choice outcome, and u(x,□) the *outcome-utility* generated by x *if there was no choice in the matter*. For example, a compulsory charitable contribution, e.g., via a tax, will make a generous person (high □ on the generosity dimension) feel better than a selfish person, other things being equal. In the model, a person' current beliefs about □, or "self-image," are defined by a probability distribution, f(□). The value of this self-image is, in turn, determined by a second utility function, V(□), which indicates how much pleasure or pain a person would feel if he or she found out their □ for sure, e.g., by taking some kind of infallible psychological test. A person's initial endowment of self-esteem, before making a choice, equals the expectation of V(□) or $\square_{\square}$V(□)f(□).

The range of internal characteristics that fall under the notion of a disposition is broad. Stable traits, such as endurance, perseverance, or intelligence, which establish expectations of future success, would be one category. A second category would be inclinations to deviant behavior, addictions, sexual proclivities, and such. A third category would be moral dispositions, altruism, virtue, etc.. All of these dispositions contribute to the self-image, and hence to self-esteem. The necessary property of a disposition is that it can influence behavior. This leads to an apparent paradox: How can an unknown disposition have any affect on choice? The question brings to the surface the distinction between conscious and implicit knowledge of □. A dispositional parameter, in our view, has the property of being implicitly "known" by the decision-making mechanism even though it cannot be introspected before the choice.[4] The “gut” knows □ but the conscious mind does not. Willpower, inclination to alcoholism, altruism, or cold water tolerance exert influence at the moment of choice, but cannot be deduced by merely imagining what one might do in a given situation.

By choosing one outcome over others, therefore, a person reveals something about his or her dispositions. Hence, the choice leads to an updating of the self-image, from $f(\square)$ to $f(\square|x)$. The change in self-image marked by the choice generates a separate form of utility — *diagnostic utility* — equal to the gain or loss in self-esteem: $\square_{\square}V(\square)f(\square|x)$ - $\square_{\square}V(\square)f(\square)$, when the prior beliefs $f(\square)$ are updated to $f(\square|x)$. Diagnostic utility registers the extent one is ‘morally’ impressed or disappointed by one’s own choices.

Our first assumption is that diagnostic utility is fully taken into account before a choice is made. The “hand” that chooses maximizes total utility, which is the sum of outcome <u>and</u> diagnostic utility:

$$\text{Total utility} = \text{Outcome utility} + \text{"Diagnostic utility,"}$$

$$= \quad u(x,\square) \quad + \square\ \square_{\square}\, V(\square)\, (f(\square\,|x) - f(\square)).$$

This looks straightforward, until one considers the updated inferences, $f(\square|x)$. Where do these distributions come from? The simplest assumption, though perhaps not the most accurate one psychologically, is that interpretations are *true*, which means that the revised distribution $f(\square|x)$ is consistent with maximization of <u>both</u>

[4] 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).

components of total utility. What does this mean psychologically? It means that the decision maker is fully aware that he or she is partly motivated by the desire to get good news. Whatever action he or she chooses will thereby be revealed as the best action all things considered, including in "all things" the desire to find out something good about themselves. More precisely, by choosing x over other options y or z, the person reveals him or herself to be exactly the kind of person for whom x produces more total utility than either y or z. This assumption precludes unrealistic, self-serving inferences. For example, suppose that the disposition in question is altruism, and a person interprets a 25¢ donation as evidence of altruism. This would not be a true interpretation, because the generous portion of diagnostic utility would make it worth giving the quarter even when there was no real concern for the poor. In other words, the diagnostic value of an action is properly discounted for the presence of diagnostic motivation.

The assumption of true interpretations 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 behavior is the same irrespective of whether the inferences pertain to someone else's or our own states. Just as we might discount someone else's good behavior as being due only to a desire to impress, so too we could discount our own behavior for ulterior motives, according to the true interpretation assumption.

## 4 Self-control with intrinsic and instrumental self-signaling

We now look at how self-signaling promotes self-control in a simple example. Imagine a person facing a decision whether to *Indulge* or *Abstain* in some problematic activity, let us say, an opportunistic sexual encounter. In the situation we have in mind the person faces a temptation, but does not know — before deciding — how strong the temptation really is, nor whether the temptation arises because he or she is “loose” (which is a permanent disposition or trait) or because there is exceptional attraction or “chemistry” with this partner on this unique occasion. Whatever happens, the decision will be diagnostic and provide some clues about all of these things. We imagine also that it is the first time such a situation has arisen.

The effective temptation to *Indulge* is an example of a unknown disposition, □, as discussed in the previous section. It combines two parameters, also unknown, a stable inclination toward this kind of thing, □*, and a “pull of the moment” or chemistry variable, □:

Temptation = inclination + chemistry on this occasion,

□ = □* + □.

A person experiences a temptation, but does not know how much of this is due to inlination (□*) or to chemistry (□). To simplify calculations, we will assume that □* and □range from 0 to 1 with all values equally likely. Temptation therefore ranges from 0 to +2 (and conforms to a triangular distribution). A level of temptation close to +2 means that both appetite and chemistry are high on this occasion; a level close to zero means that both are low; intermediate levels of temptation could be due to a combination of high appetite and low chemistry, to low appetite and high chemistry, or to moderate levels of both appetite and chemistry.

In this situation, the gain in outcome utility of *Indulging* rather than *Abstaining* is proportional to temptation, or, more precisely, u(*Indulge*,□) – u(*Abstain*,□) = □ - T.

T is non-diagnostic cost of  having such an encounter, capturing the risks and other costs associated with a single episode.  A person with no diagnostic motivation would indulge only on those occasions when temptation exceeded the cost, i.e.,. when □ > T. In this example, we assume that T is small (T < 1) so that most people, most of the time would succumb to temptation if self-signaling was not a factor in the decision.

Diagnostic motivation now enters into the picture in one of two ways, exemplified by two individuals, Tom and Harry. Tom is concerned about intrinsic appetite, which he regards as an important and undesirable character trait. He just doesn't want to be the kind of person who enjoys this activity, perhaps because it would reveal a coarse hedonistic nature, or a weakness of character inconsistent with his moral and religious ideals. Harry, on the other hand, is more pragmatic. He is not interested in intrinsic appetite per se, but does care about appetite insofar it conveys information about the likelihood of indulging in such escapades in the future.  For him, too, the action is a signal, but it is a signal of his *behavioral* tendency to indulge. For prudential reasons, Harry prefers to believe that such consummated encounters will be few and far between.

Basically, Tom wants to have his true desires — his identity — aligned with his ideals, while Harry wants to believe that he will not engage in sex with strangers, or at least not often.  Their V-functions are essentially the same, but with respect to different arguments:

For Tom: V(□*) = 1 — □*,

For Harry: V(□*) = 1 — (Longrun Probability of Indulging given □*).

Our motivating question, then, is whether diagnostic concerns will affect the likelihood of abstaining, for either Tom or Harry. Precisely, we would like to know whether the *threshold level* of temptation — the level at which temptation is just sufficient to tip the decision in favor of indulging — is higher than the natural level, T, which obtains when diagnostic concerns do not arise. Both Tom and Harry believe that their intrinsic appetite is fixed and unaffected by what they do, i.e., self-control is not a 'muscle' which gains with exercise, in the sense discussed by Baumeister and Vohs (this volume).

Tom's inferential problem Let us consider Tom first. Suppose that Tom passes the test on this occasion and abstains. What can Tom infer about his intrinsic appetite from this happy outcome? To assess the implications of that single decision, Tom needs to know his threshold level of temptation for situations of this kind. He needs to know whether abstaining is a surprising decision, as it would be if the threshold was low, or the expected one, as it would if the threshold was high.

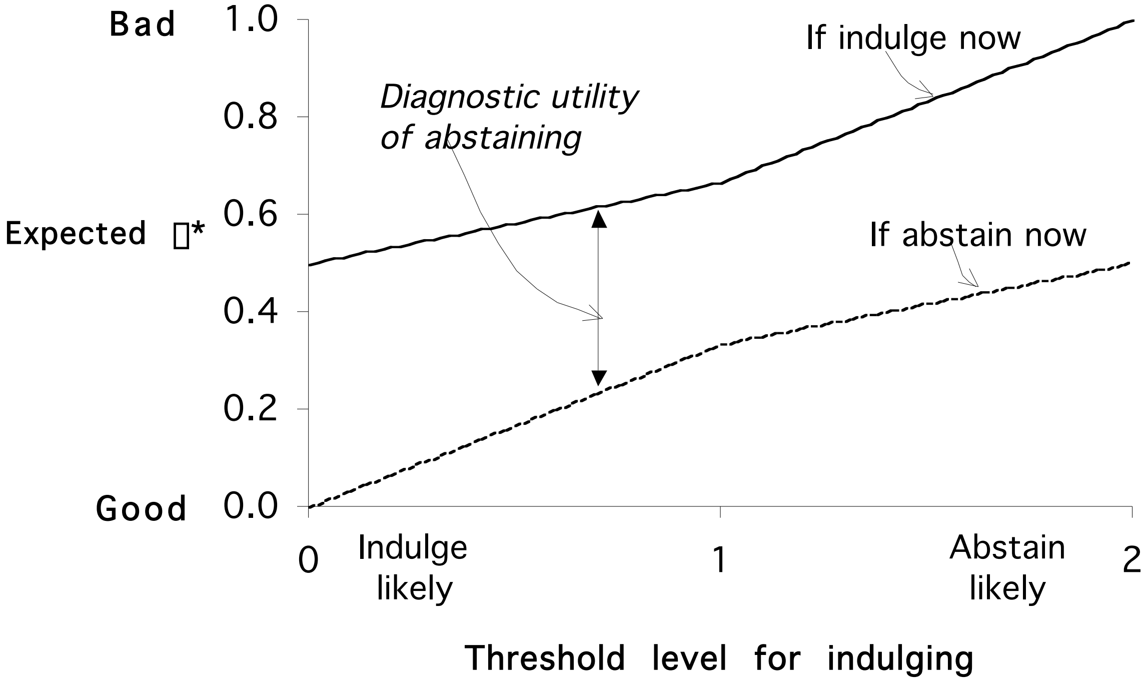


Figure 1

Figure 1 shows how inferences about intrinsic appetite change as a function of the action chosen and the presumed threshold, given on the x-axis. When the threshold is low and indulging likely, then indulging provides little new information, and the estimate of □* which follows from indulging is close to the initial estimate of 0.5 (recall that □* ranges from zero to one, with all values equally likely). Abstaining, however, is a strong signal of low □*, i.e., little intrinsic appetite for the activity.

The opposite inferences obtain at the right-hand side of the graph, when the threshold is high and abstention nearly certain. Now the single decision to Abstain provides little new information, leaving the best guess of □* close to 0.5. Indulging is a big surprise, leading to a sharp negative revision in Tom's self-image. We have here the ingredients for ‘compulsive’ adherence to a norm, where the virtuous action is taken not because it leads to increased self-esteem but because failure to adhere to the norm would trigger a harsh negative inference. Indeed, as we can see from the  upward sloping lines in Figure 1, as the threshold goes up and abstention becomes more certain, the interpretation of either action — abstaining or indulging — becomes more pessimistic.  “Positive expectations about behavior depress the interpretations about dispositions” is the paradoxical conclusion here.

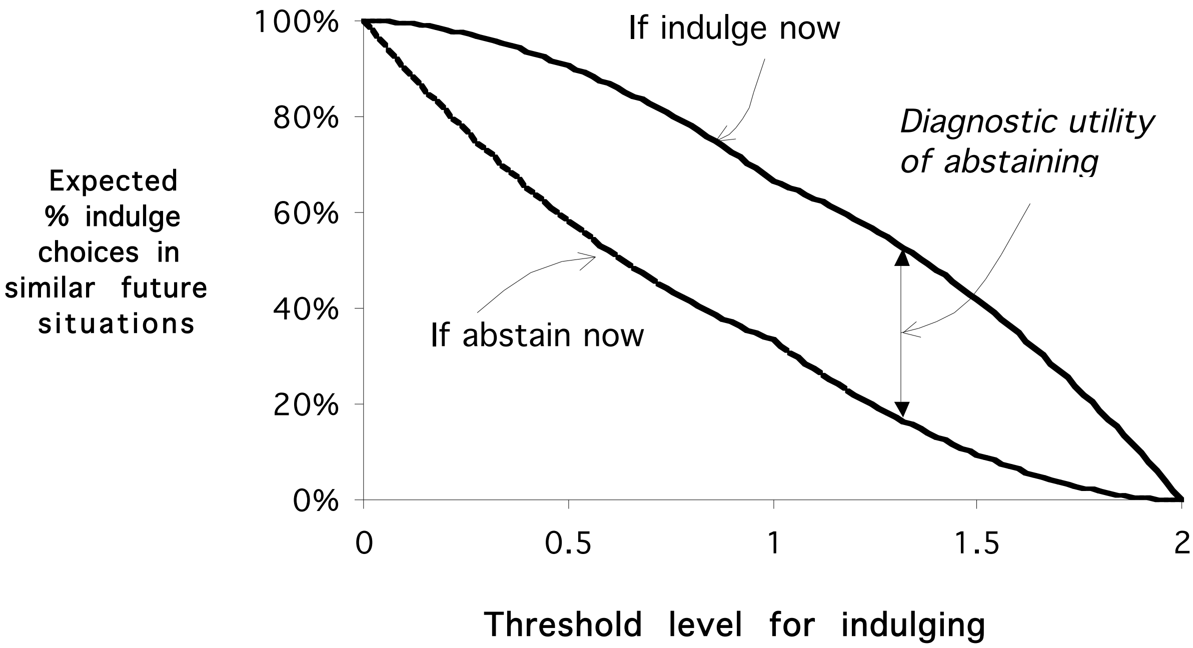


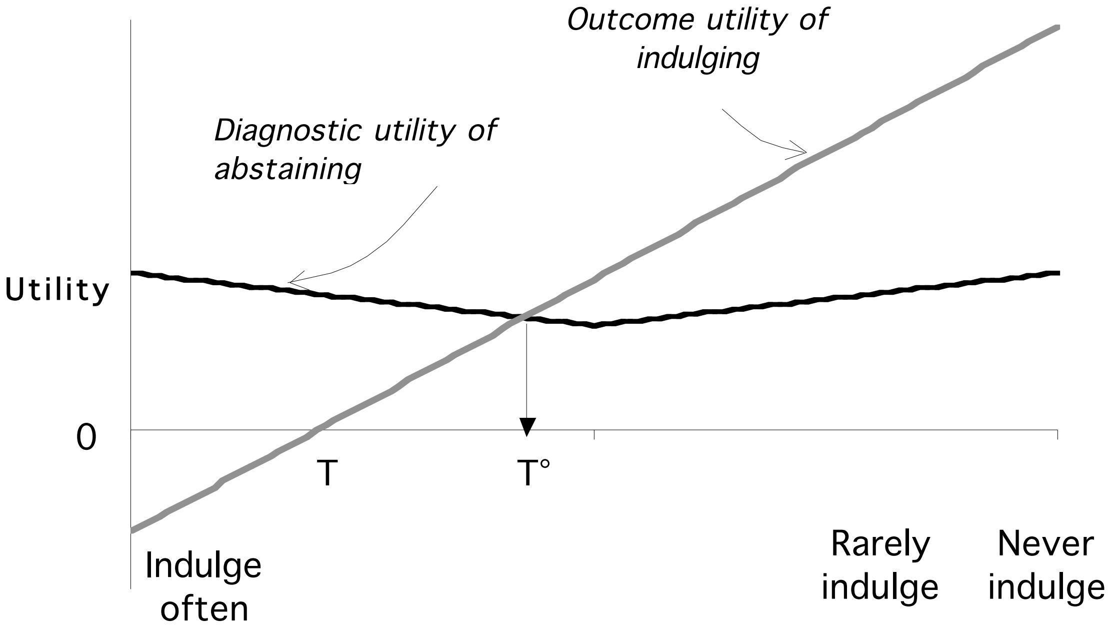
Figure 2

Harry’s inferential problem   The situation is entirely different for Harry, who cares only for the longrun behavioral predictions that flow from the current action. First, the inferential problem invoves one extra step. Not only does he have to infer something about his intrinsic appetite but he also has to assess what this particular level of appetite implies about future behavior. Harry begins, therefore, with the analysis displayed in Figure 1. Assuming a particular threshold, indulging or abstaining leads to revised self-image distributions, from the uniform distribution f(□*) to f(□*|*Indulge*) or f(□*|*Abstain*). With these distributions in hand, he is able to assess how likely it is that temptation will exceed threshold and lead to indulging on a representative future occasion. These computed probabilities of Indulging and Abstaining are displayed in Figure 2.[5] We see that as the threshold goes up and abstention becomes more certain, the interpretation of either action — abstaining or indulging — now becomes more *optimistic*.

Figures 1 and 2 summarize what either action means, as a sourrce of information about intrinsic appetite (Figure 1) or behavioral propensity (Figure 2). Because meta-utility is linear in either appetite (for Tom) or behavioral propensity (for Harry), the diagnostic utility of doing "the right thing" and abstaining is proportional to the difference between the two graphs in either figure. For Tom, diagnostic utility does not vary much with threshold — the gap between the two lines in Figure 1 remains roughly constant. For Harry, diagnostic utility diminishes as the threshold moves toward zero or two, and one or the other action becomes almost certain. If the threshold is near two, for example, then Harry will indulge only if both his intrinsic appetite and the pull of the moment are very high. Under these conditions, an otherwise surprising "lapse" might reveal high intrinsic appetite, without shaking Harry's belief that such lapses will remain relatively rare.

[5] We are assuming, also, that there is no learning over successive choices, as if memory of earlier choices is erased when the next choice come up. If we allow for memory then actions would at some point lose most of their diagnosticity — a lifetime of behavioral evidence would surely overwhelm the information content of the next action. This is psychologically implausible (and we are not aware of any evidence that diagnostic concerns diminish over the lifecycle). As anyone who has given a bad lecture knows, we often do ignore the long-run record and give excess weight to the most recent performance.

Balancing diagnostic and outcome utilities   Whether Tom or Harry actually indulges on any given occasion depends on the balance between outcome utility (which is temptation net of cost) and diagnostic disutility.  Figures 3 and 4 compare these two components of total utility, first for Tom (Figure 3) and then for Harry (Figure 4). Both figures are set up to graphically compute the threshold level of temptation, which is the level where the outcome utility of indulging exactly matches the diagnostic utility of abstaining. The x-axis in both figures represents levels of temptation as candidates for such a threshold. The y-axis then displays two graphs, a graph of outcome utility  as function of temptation, which is an increasing straight line, and a graph of diagnostic utility as function of candidate threshold temptation level, which is copied directly from either Figure 1 (for Tom) or from Figure 2 (for Harry).

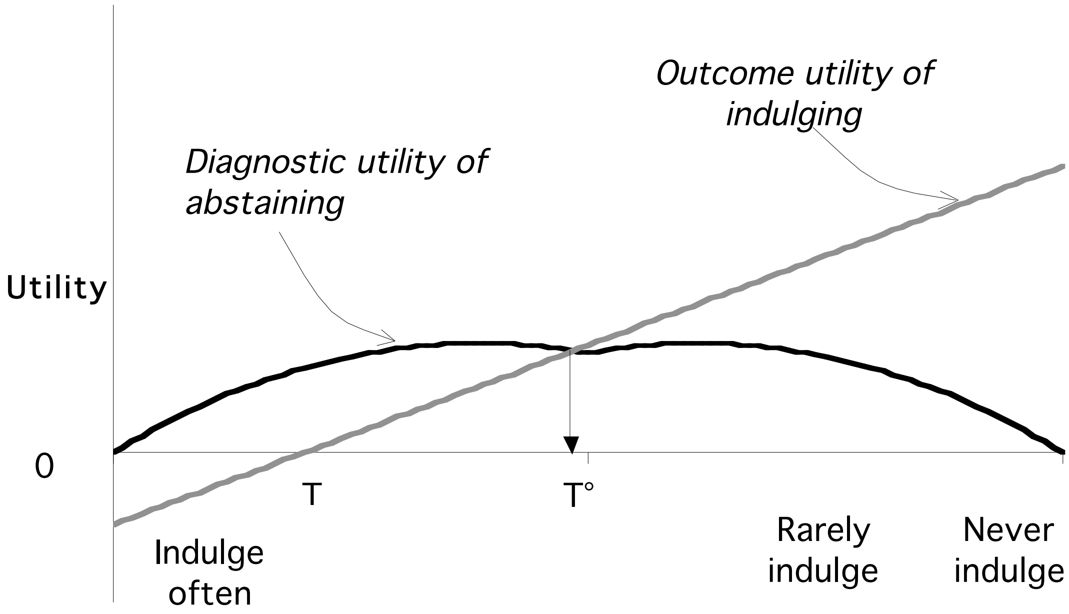


T   =  "Natural" threshold, ignoring diagnostic utility
T°     =  Threshold  (with true interpretations)

Figure 3

A useful benchmark here is T, the level of temptation that would be sufficient to motivate indulgence in the absence of diagnostic motivation. This, of course, is the just  the level where the outcome utility graph intersects the horizontal axis. When diagnostic  utility is introduced into the picture, then this level of temptation is no longer sufficient.

Looking at Tom's problem in Figure 3 first, the point where outcome utility of indulging matches diagnostic utility of abstaining is given by T°. If temptation is greater than T°, then outcome utility exceeds diagnostic utility by a positive amount and a person will indulge.  Conversely, if temptation falls below T°, then diagnostic utility of abstaining  is greater and Tom will abstain.  A similar situation obtains for Harry, whose dilemma is captured by the pair of graphs in Figure 4. Again, the threshold level of temptation is given by T°,  which is greater than the diagnosticity-free level of T.



T = Natural threshold, ignoring diagnostic utility
T° = Threshold with true interpretations

Figure 4

Two important points emerge here. First, *self-signaling promotes self-control*, irrespective of whether a person is intrinsically (Tom) or instrumentally (Harry) concerned about underlying dispositions. Second, this conclusion follows even though there is full awareness of diagnostic motivation. The model, therefore, instantiates Ainslie's view that "doing good for its diagnostic value may not invalidate that diagnostic value" (Ainslie, 1992, p. 203).

# 5 Salience and self-control

That self-control is not just a matter of time discounting and time preference was apparent ever since the classical delay-of-gratification experiments by Mischel and his associates (Mischel, 1974; Mischel et al., this volume). In Michel's paradigm, children were placed in a room by themselves and taught that they could summon the experimenter by ringing a bell. The children would then be shown a superior and inferior prize and told that they would receive the superior prize if they could wait for the experimenter to return. Children found it harder to wait for the delayed reward if made to wait in the physical presence of either one of the reward objects, which presumably induced craving. Time discounting alone cannot explain this, nor can it explain why why self-control breaks downs especially under the influence of strong physical drives or emotions (Loewenstein, 1996) or when a person is fatigued (Muraven and Baumeister, 2000).

The negative effect of salience on self-control emerges naturally with the self-signaling framework, for the simple reason that operations that change salience affect the two parts of the total utility equation asymmetrically. If salience is reduced, for example, by making the tempting reward delayed in time, or uncertain, or physically less available, this changes the outcome utility part of the equation while leaving the diagnostic utility part unchanged. Hence, the balance of power will shift in favor of diagnostic considerations.

A conceptually simple demonstration of this arises in context of 'contingent resolutions,' which are binding decisions that only take effect if a certain contingency is realized. Suppose, suppose to stay with the earlier example, that Tom or Harry have to decide what to do before knowing for sure whether the opportunity is really there. Intuition suggests that it will be easier to abstain in that case; the rewards, after all, are somewhat hypothetical.[6]

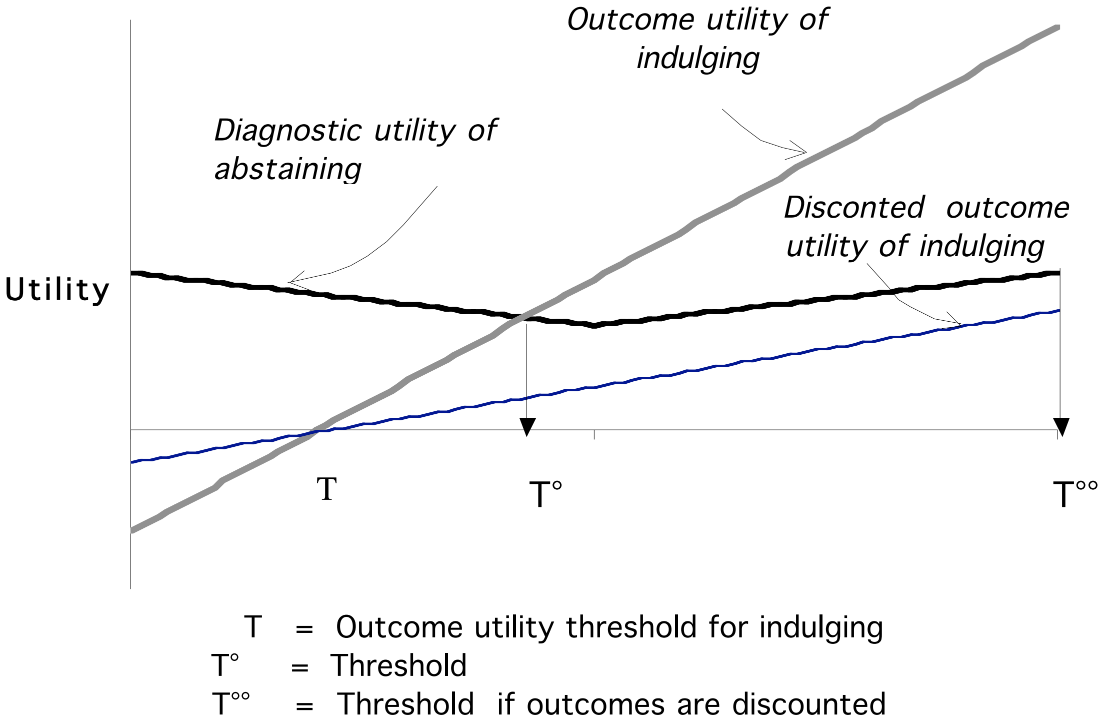


Figure 5

The model expresses this by reducing the outcome utility of indulging, from u(*Indulge*,□) to: *p*u(*Indulge*,□), where *p* is the probability that the tempting

[6] Bodner‘s experiments (Bodner, 1995) on contingent charitable pledges revealed a version of 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.

opportunity will present itself. The ‘discounted’ line with smaller slope in Figure 5 gives the reduced levels of outcome utility. Because the outcome utility of indulging is now everywhere lower than the diagnostic utility of abstaining, the person in this case would resolve to abstain irrespective of disposition. However, if the resolution wasn’t binding, then the person might also be inconsistent, resolving to abstain but then reversing his decision and indulging if the opportunity to do so actually came up.

The same argument applies to a situation where consequences are removed in time. Outcome utility is now *temporally* discounted, again making *Abstain* a more likely choice. The chronic discrepancy between future plans and actual behavior may therefore be explained by the fact that the diagnostic utility of a resolution is immediate while the cost of the resolution in outcome utility is delayed, and the fact that people fail to anticipate reversals in preference, i.e., that they are naïve in the sense of O’Donoghue and Rabin (1999, this volume).

This account provides a common 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 metautility. When the weight of outcome utility is restored, by making outcomes salient, certain, imminent, etc.., the person may regret the earlier choice.

## 6 Self-signaling without awareness

The reader may be wondering at this point whether the model that has been sketched out rests on an odd combination of self-ignorance (i.e., of one’s own dispositions) and self-insight (i.e., of one’s propensity to self-signal). While we are certainly able to discount the behavioral signals of other people when we suspect ulterior motives on their part, it is not at all self-evident that we will apply to our own actions the same rigorous standards that we apply to the actions of others.

As a purely theoretical matter, one can formulate a variant self-signaling model in which the person is presumed ignorant of the self-signaling motive, and accepts the evidence of his or her actions at *face-value*. The updated inferences, f(□|x), are in that case based on the assumption that an action reveals the disposition that maximizes only the outcome-utility component of total utility, ignoring the diagnostic component. Here, there would be no discounting for diagnostic motivation. Diagnostic utility would be experienced as an unintentional byproduct of choice, not something that consciously affected choice.

How does this affect the impact of diagnostic motivation on self-control? In broad terms, the impact is still there,

Third — and this is perhaps less apparent from the Figures — self-signaling will also promote self-control for individuals in Case IV, who are not aware of diagnostic motivation. Such individuals will experience diagnostic utility consistent with a threshold of T (because they believe the threshold is T), but will have an actual threshold at some higher level, T' (which could be above or below T°) where this diagnostic utility exactly matches outcome utility. This discrepancy between the true threshold, which is relatively high, and the presumed threshold, which is relatively low, means that they actions will tend to be "better-than-expected." Hence, individuals with face-value interpretations will leave the choice situation with an excessively positive self-image, on average.[7]

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

The two ways of interpreting actions generate four self-signaling "psychologies," outlined in the Table below. The columns in the Table indicate whether diagnostic motivation is present (right column) or absent (left column). The rows in the Table indicate whether inferences from actions are discounted for diagnostic motivation (bottom row) or not discounted (top row). The diagonal entries are "rational" in the sense that interpretations and motivations are consistent with each other.

We will not dwell on left side of the Table, where preferences are free from diagnostic utility. Case I is the rational economic model, where a person does as she pleases, and her choices hence directly reveal what she likes. Case IV describes a logical possibility of dubious realism, except perhaps as a model of certain forms of psychosis. Here, a person is not diagnostically motivated, but interprets his actions as if he were so motivated. The result is a kind of paranoid self-scrutiny, a search for ulterior motives that are not really there.

The interesting contrast is between Cases II and III. In Case II, diagnostic motivation makes 'good' actions more likely, yet this biasing affect is ignored in making inferences. A person generously gives himself full credit for doing the good thing, even when part of the motive was precisely to get the credit. The result is an excessively positive self-image. As an empirical hypothesis, Case II may be quite close to the truth. There is much evidence that self-assessments are excessively positive (e.g., Taylor and Brown, 1988). What is distinctive about this particular explanation of excessive self-esteem, is that it doesn't postulate any direct self-deception; the incorrect self-image is a byproduct of a cognitive "blind spot," a lack of awareness that good actions were motivated by diagnostic concerns.

The salient feature of Case III, the fully rational self-signaling model, is that it promotes escalation of virtuous conduct. Because good behavior is discounted for diagnostic motives, being reasonably good may no longer be "good enough." As inferior dispositions mimic behavior diagnostic of superior ones, the superior ones need to do even better in order to differentiate themselves. The process naturally tends towards behavioral perfection, where no further improvements on the relevant dimension are possible. At the verbal level, a person may characterize his own behavior in terms of rules that either proscribe an activity altogether (in the case of vices) or insist that an activity be performed without exception (in the case of virtues).

Perfection, however, doesn't secure a positive inference: With Case III, a person, sadly, can never improve their self-image (on average). The futility of the signaling effort does not diminish its motivational force, however. How is this possible? There is a logic of compulsion at work here, with motivation sustained not by positive diagnostic benefits but by fear of negative inferences that would be triggered by even a single lapse on the dimension of concern. To a compulsive, washing hands does not guarantee that they are clean, but a failure to wash does confirm that they are dirty. Like Baron von Munchausen's horse, the rational self-signaler drinks and drinks but the thirst cannot be quenched.

|  | *Preferences free from diagnostic utility*<br>*□ = 0* | *Preferences subject to diagnostic utility*<br>*□ > 0* |
|---|---|---|
| *Face-value interpretations of actions*<br>*(as if □ = 0)* | **I. Standard economic model**<br>• you do as you please<br>• actions reveal who you are | **II. Normal self-deception**<br>• you bias behavior toward actions diagnostic of good dispositions<br>• improve future prospects<br>• create overly positive intrinsic self-image |
| *Interpretations discounted for diagnostic motivation*<br>*(as if □ > 0)* | **IV. Paranoid self-scrutiny**<br>• you do as you please<br>• second-guess actions for nonexistent motives<br>• overly negative intrinsic self-image and excessive pessimism about future prospects | **III. Rational self-signaling**<br>• you seek behavioral perfection<br>• tend toward “always” “never” rules<br>• improve future prospects<br>• fail to improve intrinsic self-image (on average) |

## 7 Summary

We have described here a model that accounts for some of the diagnostic implications of choice. Is this a theory of self-control, however? In an important sense, it is not. One can certainly imagine situations where self-control is needed to *overcome* diagnostic motivation. Consider a case of a person debating whether to enroll in a dating service. Signing up for the service has long-term causal benefits — more partners to choose from, and so on. However, it also has diagnostic costs: he has to confront a certain kind of personal failure, namely, that unlike so many others he was not able to find a partner in the usual "romantic" way. This diagnostic pain is realized at the moment he signs up, before any benefits are realized. In this case, willpower must be applied to overcome the diagnostic hurt associated with an otherwise sensible action. Diagnosticity is a general source of motivation, which often supports long-term objectives but which can sometimes work against them, as in this example.

In our model, diagnostic utility enters the equation as just another source of pleasure, to be balanced against other pleasures. Nothing in the model refers to the problem of effort and willpower, which is intimately involved with self-control (Muraven and Baumeister, 2000). A person with strong diagnostic motivation will not feel any conflict — he or she will do the "right thing" smoothly, without strain. The model, therefore, explains why we get out bed, brush our teeth, and go about our daily business without much fuss. At least in present form, it does not shed light on intrapersonal conflict *per se*.

## References

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

Ainslie, G. (2001) *Breakdown of will*. Cambridge: Cambridge University Press.

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

Baumeister, R.F., Heatherton, T.F., Tice, D.M. (1994). *Losing control: How and why people fail at self-regulation*. San Diego, CA: Academic Press.

Baumeister, R.F.,and K. D. Vohs. “Willpower,” in G. Loewenstein, D. Read, & R.F. Baumeister (Eds.) *Time and Decision*. New York: Russell Sage Foundation.

Bem, D. J. (1972), Self-perception Theory, In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 6). 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.

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.

Bodner, R. (1995) "Self Knowledge and the Diagnostic Value of Actions: The Case of Donating to a Charitable Cause", 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.

Bodner, R. and D. Prelec. (2001). "Self-signaling in a neo-Calvinist model of everyday decision making," in *Essays in Psychology and Economics*, J. Carrillo and Isabelle Brocas (eds.), Oxford University 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," in *Essays in Psychology and Economics*, J. Carrillo and Isabelle Brocas (eds.), Oxford University Press.

Burkhardt, F. and F. Bowers, eds. *The Work of William James: The Principles of Psychology, Volume I,* Harvard University Press.

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

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

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.

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

Elster, J. (1989). *The Cement of Society: A Study of Social Order*. Cambridge University Press, Cambridge, UK.

Gilboa, I. and E. Gilboa-Schechtman, "Mental accounting and the absentminded driver," in *Essays in Psychology and Economics*, J. Carrillo and Isabelle Brocas (eds.), Oxford University Press.

Ginossar, Z., and Trope, Y. (1987). Problem solving in judgment under uncertainty. *Journal of Personality and Social Psychology*, *52*, 464-4741.

Herman, C.P. and Polivy, J. (2002). “Dieting as an Exercise in Behavioral Economics.” in G. Loewenstein, D. Read, & R.F. Baumeister (Eds.) *Time and Decision*. New York: Russell Sage Foundation.

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

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

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

Loewenstein, G. F. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, 65, 272-292.

Mischel, W. (1974). Processes in Delay of Gratification. in D. Berkowitz, *Advances in Experimental Social Psychology*, 7, pp. 249—292.

Mischel, W., Ayduk, O., and Mendoza-Denton, R. (2002) “Sustaining delay of gratification over time: A hot/cool systems perspective.” in G. Loewenstein, D. Read, & R.F. Baumeister (Eds.) *Time and Decision*. New York: Russell Sage Foundation.

Muraven, M., and Baumeister, R.F. (2000) Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin*. 126: 247-259.

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

O’Donoghue, T., and Rabin, M. “Self-awareness and self-control,” in G. Loewenstein, D. Read, & R.F. Baumeister (Eds.) *Time and Decision*. New York: Russell Sage Foundation.

Prelec, D. (1991) 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.

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.