ISSN 1471-0498

DEPARTMENT OF ECONOMICS
DISCUSSION PAPER SERIES

SELF-KNOWLEDGE AND SELF-DECEPTION

H. Peyton Young

Number 383
January 2008

Manor Road Building, Oxford OX1 3UQ

Self-Knowledge and Self-Deception

H. Peyton Young

University of Oxford
Johns Hopkins University
The Brookings Institution

This version: December 31, 2007

I am indebted to Roland Benabou, Chris Carroll, Joshua Epstein, Alan Gibbard, Joe Harrington, Edi Karni, Ali Khan, Kislaya Prasad, Marszena Rostek, Brian Skyrms, and Ennio Stacchetti for helpful comments on an earlier draft.

Abstract

A person is concerned about *self-image* if his utility function depends, not only on his actions, but also on his beliefs about what sort of person he is. This dual motivation problem makes it difficult, and in some cases impossible, for someone to learn who he really is based solely on his revealed behavior. Indeed, there are very simple situations, involving just two actions and two possible identities, such that, if there is any initial uncertainty about one's true identity, it will never be resolved even when one has an infinite number of opportunities to act.

> *Sometimes it happens that with the sharpest self-examination we can find nothing beside the moral principle of duty which could have been powerful enough to move us to this or that action…yet we cannot from this infer with certainty that it was not really some secret impulse of self-love…that was the actual determining cause of the will.* Immanuel Kant, 1785

1. Homo ambiguus

Suppose that you come upon a motorist who is stranded by the highway. Your first thought is to stop and help, which is the compassionate thing to do. On second thought, you realize that your motive might be merely to *appear to yourself* as being compassionate. In this case the act is pointless, because you will fail to 'fool yourself' into thinking you are more compassionate than you really are. Therefore you should not stop to help. But this would establish *beyond a doubt* that you are not compassionate, so maybe it would be a good idea to stop after all.

We can model this situation as a game between your possible identities and yourself as impartial judge. In effect, you want to deduce your true identity from your actions, that is, you want to know what sort of person you are based on your revealed preferences.[1] The difficulty is that you may have dual motives for acting: you derive utility from the action itself, and also from the signal the action conveys about who you are. Kant drew attention to this dual motivation problem in *Fundamental Principles of the Metaphysics of Morals* (1785), and argued

[1] This approach is similar in spirit to Bem (1967, 1972) , who proposed that self-perception is based on the empirical observation of one's actions.

that it prevents one from knowing *for sure* whether a given act was done for the "right" reason, or merely to satisfy one's vanity.

In this paper I shall examine this question from a game-theoretic perspective. In particular, I shall demonstrate variants of the Samaritan game in which: i) you get to observe yourself infinitely often and you have perfect recall; ii) all of your possible 'identities' are perfectly rational; iii) the resulting game has only mixed equilibria, which are separating in the sense that your different identities behave differently in equilibrium; iv) if you start with any amount of uncertainty about your true identity, then with probability one the uncertainty will never be resolved.

The analytical framework employs the concept of a psychological game first introduced by Geanakoplos, Pearce, and Stachetti (1989), and developed by Benabou and Tirole (2001, 2004, 2006, 2007) and Bodner and Prelec (1997, 2003) among others. The present set-up differs from these papers in several key respects however. First, unlike Geanakoplos, Pearce, and Stachetti (1989) players have no scope for choosing their beliefs; beliefs evolve objectively by Bayesian updating. Second, unlike the models of Bodner and Prelec and Benabou and Tirole, different 'identities' can have different preferences for who they want to be. For example, one type of person may want to think he is altruistic while another type may want to think he is selfish and 'tough-minded'. (Virtually all of the prior literature on self-signaling and dual motivation assumes that different identities agree on what identity it would be desirable to have.) [2] Third, I shall

[2] There is another literature on identity and the multiple self that is more distantly related to the approach taken here; see in particular Ainslie (1986, 1992), Elster (1986), and Akerlof and Kranton (2000, 2002, 2005).

not assume any 'irrational factors' or 'behavioral anomalies' such as time inconsistency, imperfect recall, selective memory, and wishful thinking, which are often invoked in the behavioral economics strand of the literature. The aim is to keep the number of moving parts to a minimum in order to focus exclusively on the problem of *rationally deducing who you are from what you do*, and to show that this problem can be very severe even in elementary situations involving just two possible identities. Moreover, even in these situations, the *structure* of the equilibrium is not so elementary, which makes the problem interesting from an analytical standpoint.

2. The model

The model is premised on the notion advanced by Adam Smith that we judge ourselves as we imagine an impartial observer would judge us: "When I endeavour to examine my own conduct, when I endeavour to pass sentence upon it, and either to approve or condemn it, it is evident that, in all such cases, I divide myself, as it were, into two persons; and that I, the examiner and judge, represent a different character from that other I, the person whose conduct is examined into and judged of." (*The Theory of the Moral Sentiments*, 1759, part III, chapter I).

Specifically, let us suppose that each person consists of an observing "Self," who has beliefs but not preferences, and a "self" that has preferences and takes actions. The self is the familiar *homo economicus* who maximizes expected utility, whereas the Self is a separate aspect of the person: an 'impartial spectator' who

tries to deduce who the self really is based on the evidence. The complication is that *the self's utility function contains the Self's opinion as one of its arguments.*[3]

To make this idea precise, let I be a finite set of n possible *selves*, which we shall also refer to as *identities* or *types*. A *self* is defined by its utility function, which has two arguments: *actions*, and *beliefs* of the Self about who the self is or might be. Let A denote the (finite) set of actions available to the self. Let p_{ia} be the probability that self i plays action $a \in A$, and let p_i denote the corresponding probability distribution over A. The *Self's beliefs* consist of a probability distribution $\vec{\theta} = (\theta_1, ..., \theta_n)$ on I, where θ_i is the probability that the Self currently assigns to the true self being i. Assume for the moment that *i* takes the belief vector as given, and that $i's$ utility function takes the additively separable form

$$u_i(p_i, \vec{\theta}) = \sum_{a \in A} u_{ia} p_{ia} + \sum_{j \in I} v_{ij} \theta_j \ . \qquad (1)$$

Here u_{ia} is the "intrinsic" utility to i from playing action a, and v_{ij} is the utility from *self-image*, that is, the utility that i gets from the Self's belief that "i" is actually "j". In the context of the Samaritan game, for example, "i" might be an uncompassionate person who gets negative utility from stopping to help, but gets positive utility from the Self's belief that he is compassionate. (Bodner and Prelec (2003) refer to this aspect of the utility function as *diagnostic utility*.)

[3] The idea that beliefs may have consumption value has been proposed in various contexts; see in particular Akerlof and Dickens (1982), Schelling (1986), and Geanakoplos, Pearce, and Stacchetti (1989).

Unlike much of the previous literature on self-signalling and dual motivation, I shall assume that the values v_{ij} are completely arbitrary. For example, an i might want to have the image of being a j, whereas a k might find such an image repugnant. Note that the self-image value that i places on j includes all aspects of j's "personality", *including j's desire to be known as someone else*. This definition is not circular; it merely says that the desirability of a given image takes into account all aspects of the utility function defining that image. For example, you (the self) might wish to have the image of someone who is compassionate and is not concerned with self-image (compassionate and selfless). Then again, you might find such an image unnatural, and prefer to view yourself as someone who is compassionate and proud of it.

The Self's beliefs change over time as the Self observes more data, which consist of the self taking particular actions. In choosing an action, therefore, the self must consider the intrinsic utility from the action itself, as well as the change in the Self's beliefs that the action will induce -- its implications for future self-image. First we shall examine the implications of this situation when the self gets to act only once, then we shall extend the analysis to infinitely repeated games and forward-looking agents.

Assume that the Self's prior belief is given by the vector $\vec{\theta}$. Let θ_{ja} denote the Self's posterior belief that the self is j given that action a occurs. Imagine that you are, in fact, self i. Since you are rational, you choose actions to maximize your expected utility. This would suggest choosing $\vec{p}_i$ to maximize

$$u_i(\vec{p}_i, \vec{\theta}) = \sum_{a \in A} p_{ia}[u_{ia} + \sum_{j \in I} v_{ij}\theta_{ja}]. \qquad (2)$$

This is misleading, however, because the posterior values θ_{ja} depend not only on $\vec{p}_i$, but also on what every *possible self* would do *if it were the true self*. Thus, instead of (2), we need to write

$$U_i(\vec{p},\vec{\theta}) = \sum_{a\in A} p_{ia}[u_{ia} + \sum_{j\in I} v_{ij}\theta_{ja}(\vec{p},\vec{\theta})], \tag{3}$$

where $\vec{p} = (\vec{p}_1, \vec{p}_2, ..., \vec{p}_n)$ and each $\vec{p}_i$ maximizes $U_i(\vec{p},\vec{\theta})$ given the strategies of the other possible selves.

The numbers $\theta_{ja}(\vec{p},\vec{\theta})$ are computed as follows: given that a occurs, the Self's posterior beliefs are

$$\theta_{ja} = f_j(a,\vec{p},\vec{\theta}) = \theta_j p_{ja} / \sum_k \theta_k p_{ka} \text{ if } \sum_k \theta_k p_{ka} > 0.^{4} \tag{4}$$

Denote the posterior probability distribution by $\vec{\theta}_a = f(a,\vec{p},\vec{\theta})$. The payoff functions $U_i(\vec{p},\vec{\theta})$ defined by (3) and (4) constitute an n-person game among the various possible selves, which we shall call a *game of self-knowledge*. In effect, it is a one-person game of incomplete information where the n possible selves are the

[4] If $\sum_k \theta_k p_{ka} = 0$ the posterior can be chosen arbitrarily. For the sake of concreteness, I shall assume that $\theta_{ja} = 0$ if $p_{ja} = 0$ and $p_{ka} > 0$ for some $k \neq j$, that is, if j could not have played a but someone else could have done so; otherwise let $\theta_{ja} = \theta_j$. With this choice each payoff function $U_i(\vec{p},\vec{\theta})$ is continuous in i's own strategy $\vec{p}_i$.

player "types."[5] The selves wish to establish "reputations" with the Self, but unlike the usual situation where players seek to establish reputations with other players, here there is no other player. The Self takes no actions and has no payoffs. It grinds out a posterior belief mechanically, and this belief has consumption value for the sole actual player, which is the self.

An equilibrium of such a game may be defined as follows. Let $N(\vec{\theta})$ be a function that selects a Nash equilibrium of $G_{\vec{\theta}}$ for each prior $\vec{\theta}$ (assuming that an equilibrium exists). This selection is made by the Self and can be thought of as an adjunct of the Self's beliefs. Specifically, it is the strategy-tuple $\vec{p} = N(\vec{\theta})$ that the Self believes will be played when the Self's priors are $\vec{\theta}$. (In the examples discussed below the equilibrium is unique so the selection issue does not arise.)

We shall say that the belief vector $\vec{\theta}$ is *stable* if the posterior equals the prior for all actions a that have positive probability under $\vec{\theta}$ and $N(\vec{\theta})$. In other words, $\vec{\theta}$ is *stable* if for all $a \in A$,

$$\sum_{i \in I} p_{ia}\theta_{ia} > 0 \Rightarrow \vec{\theta} = f(a, N(\vec{\theta}), \vec{\theta})), \qquad (5)$$

where f is defined as in (4).

[5] Unlike standard games of incomplete information, however, the prior probability distribution over types need not be objectively given: the prior beliefs may be purely subjective (though we shall assume they are governed by the rules of Bayesian inference).

3. Learning dynamics

We are now in a position to examine the dynamics of such a process. Fix a mapping $\vec{p} = N(\vec{\theta})$ from beliefs to equilibrium strategies. (Recall that $N(\cdot)$ is part of the Self's belief system.) Consider a series of discrete time periods $t = 0,1,2...$, and suppose that the Self's initial belief vector is $\vec{\theta}^0$. In period 1 an action is generated according to the probability distribution $\vec{p}_i = N_i(\vec{\theta}^0)$. Given that action a occurs, the posterior is $\vec{\theta}_a^1 = f(a, N(\vec{\theta}^0), \vec{\theta}^0))$. Hence the vector $\vec{\theta}_a^1$ occurs with probability $p_{ia} = (N_i(\vec{\theta}^0))_a$. The process is then repeated next period, with $\vec{\theta}_a^1$ playing the role of $\vec{\theta}^0$. More generally, if at the end of period t the realized posterior beliefs are $\vec{\theta}^t$, then the posterior beliefs at the end of period $t+1$ are

$$\vec{\theta}^{t+1} = f_i(a, N(\vec{\theta}^t), \vec{\theta}^t)) \text{ with probability } p_{ia}^t = (N_i(\vec{\theta}^t))_a. \qquad (6)$$

Let Θ be the set of all probability distributions $\vec{\theta}$ on the n possible selves. Expression (6) defines a Markov process P_i on Θ, namely, *the actual process generating the posterior beliefs when i is the true self*. There are n distinct Markov processes associated with a given game of self-knowledge, one for each possible self. Obviously, if the Self begins with a prior that puts probability zero on the true self, self-knowledge will never be attained. A less obvious result (established in section 6) is that *there are games in which self-knowledge will never be attained if there is any uncertainty about the true self to begin with*. Before proving this we turn to some preliminary considerations.

4. Games of self-knowledge with two actions and two types

A game of self-knowledge is *2 x 2* if there are two possible selves or identities $I=\{1,2\}$ and two possible actions $A=\{a,b\}$. (The terms *self*, *identity*, and *type* will be used interchangeably.) In a *2 x 2* game, the Self's beliefs can be represented by a scalar $\theta\in(0,1)$, which is the probability that the Self assigns (at a given point in time) to the true self or type being $i=1$. Let type i put probability p_i on action a and probability $1-p_i$ on action b. Given a prior θ, and strategies p_1 and p_2, denote the posterior by θ_a if a occurs and θ_b if b occurs. Then we have

$$\theta_a=\theta p_1/\left(\theta p_1+(1-\theta)p_2\right) \quad \text{and} \quad \theta_b=\theta(1-p_1)/\left(\theta(1-p_1)+(1-\theta)(1-p_2)\right), \qquad (7)$$

assuming the respective denominators are positive.[6] The payoff function to type 1 can be written as follows:

$$p_1\left[u_{1a}+\theta_a v_{11}+(1-\theta_a)v_{12}\right]+(1-p_1)\left[u_{1b}+\theta_b v_{11}+(1-\theta_b)v_{12}\right]. \qquad (8)$$

This can be simplified by setting $\alpha_1=u_{1a}-u_{1b}$, $\gamma_1=v_{11}-v_{12}$, and dropping the constant terms. Recalling that θ_a and θ_b are functions of p_1, p_2, and θ, we can write 1's payoff function as follows:

$$U_1(p_1,p_2,\theta)=\alpha_1 p_1+\gamma_1 p_1\theta_a+\gamma_1(1-p_1)\theta_b. \qquad (9)$$

[6] If $\theta p_1+(1-\theta)p_2=0$ and $p_1>0$ then $p_2=0$ and $\theta_a=1$, whereas if $p_2>0$ then $\theta_a=0$. If $p_1=p_2=0$, then by assumption $\theta_a=\theta$.

Similarly, setting $\alpha_2 = u_{2a} - u_{2b}$ and $\gamma_2 = v_{21} - v_{22}$, we have

$$U_2(p_1, p_2, \theta) = \alpha_2 p_2 + \gamma_2 p_2 \theta_a + \gamma_2 (1 - p_2)\theta_b, \tag{10}$$

Notice that the second term of (9) goes to zero as $p_1 \to 0$; likewise the third term goes to zero as $p_1 \to 1$. Hence $U_1(p_1, p_2, \theta)$ is continuous in p_1 for every $\theta \in (0,1)$; similarly $U_2(p_1, p_2, \theta)$ is continuous in p_2.

5. The Samaritan game: first version

In this section we consider the Samaritan game under the assumption that one type of person (type 1) is compassionate and does not care about his reputation (*the good Samaritan*), whereas the other (type 2) is not compassionate but wants to have the reputation with the Self of being compassionate (*the fake Samaritan*). Action a is 'stop and help', whereas action b is 'don't stop'. Without any real loss of generality let $\alpha_1 = 1, \gamma_1 = 0$ and $\alpha_2 = -1, \gamma_2 = \gamma > 0$. In effect, γ measures the payoff from self-image relative to the payoff from acting (see table 1).

Payoffs from acting		
	a	*b*
type 1	1	0
type 2	−1	0

Payoffs from self-image		
	type 1	*type 2*
type 1	0	0
type 2	γ	0

Table 1. Payoffs from acting and from self-image.

Given the prior θ and the strategies p_1, p_2, the probabilities of the various outcomes are shown in Table 2.

	a	*b*
type 1	θp_1	$\theta(1-p_1)$
type 2	$(1-\theta)p_2$	$(1-\theta)(1-p_2)$

Table 2. Probabilities of the four outcomes conditional on θ and p_1, p_2.

It follows that the players' utility functions are

$$\begin{aligned} U_1(p_1, p_2, \theta) &= p_1 \\ U_2(p_1, p_2, \theta) &= -p_2 + \gamma\theta p_1 p_2 / \left(\theta p_1 + (1-\theta)p_2\right). \end{aligned} \tag{11}$$

It is a dominant strategy for type 1 to play $p_1 = 1$. Hence (11) simplifies to $U_2(p_1, p_2, \theta) = -p_2 + \gamma\theta p_2 / \left(\theta + (1-\theta)p_2\right)$. Notice that when $0 < \theta < 1$ this is a strictly concave function of p_2. The equilibrium may be pure or mixed depending on the values of γ and θ, namely,

$$\begin{aligned} \gamma > 1 \text{ and } 0 < \theta < \gamma^{-1/2} &\Rightarrow p_2 = (\gamma^{1/2} - 1)\theta/(1-\theta) \\ \gamma > 1 \text{ and } \gamma^{-1/2} \le \theta \le 1 &\Rightarrow p_2 = 1 \\ \gamma \le 1 &\Rightarrow p_2 = 0. \end{aligned} \tag{12}$$

To summarize: the second type imitates the first type $\left(p_1 = p_2 = 1\right)$ if the payoff from her self-image is high enough $(\gamma > 1)$ *and* the Self puts a sufficiently high

prior probability ($\theta \geq \gamma^{-1/2}$) on the true type being 1. If the payoff from self-image is not high enough ($\gamma \leq 1$), or if $\gamma > 1$ and $0 < \theta < \gamma^{-1/2}$, type 2 is content to play her preferred action b regardless of the Self's beliefs, which of course immediately reveals her true identity.

The most interesting situation arises when $\gamma > 1$ and $0 < \theta < \gamma^{-1/2}$. In this case type 2 plays a mixed strategy. Since type 1 always plays a, type 2 immediately reveals her true identity if she happens to plays b. In this case the posterior is $\theta_b = 0$. However, if the outcome of 2's mixed strategy is a, the Self's posterior belief is

$$\theta_a = \theta / \left(\theta + (1-\theta)\, p_2 \right) = \gamma^{-1/2}. \tag{13}$$

This means that the next time the game is played (assuming there is a next time), type 2's strategy will be to imitate type 1 and play a for sure ($p_2 = 1$). The reason is that the beliefs next period will be $\theta = \gamma^{-1/2}$, which is covered by the middle case in (12). Hence in the next period, and in all subsequent periods, the Self's beliefs remain unchanged at $\theta = \gamma^{-1/2}$: the beliefs are stable.

We can now describe the solution to the Samaritan game in terms of the equilibrium Markov processes P_1, P_2. In the process P_1, which corresponds to the true type being 1, action a is played for sure. Hence the posterior θ' (which plays the role of prior next period) bears the following relationship to the current prior θ:

$$
P_1: \qquad \begin{array}{c} \theta = 0 \;\Rightarrow\; \theta' = 0 \\ 0 < \theta < \gamma^{-1/2} \;\Rightarrow\; \theta' = \gamma^{-1/2} \\ \gamma^{-1/2} \le \theta \le 1 \;\Rightarrow\; \theta' = \theta \end{array} \tag{14}
$$

Consider now the process P_2 , which corresponds to the true type being 2. Assume that $\gamma > 1$ (the other case is trivial). If $0 < \theta < \gamma^{-1/2}$, then in the first period, action a is played with the probability p_2 given in (12). Hence the posterior $\theta' = \gamma^{-1/2}$ occurs with probability p_2, and the posterior $\theta' = 0$ occurs with probability $1 - p_2$. The one-period transition probabilities for P_2 are summarized below:

$$
P_2: \qquad \begin{array}{c} \theta = 0 \;\Rightarrow\; \theta' = 0\,, \\ \gamma^{-1/2} \le \theta \le 1 \;\Rightarrow\; \theta' = \theta\,, \\ 0 < \theta < \gamma^{-1/2} \;\Rightarrow\; \begin{array}{c} P\left[\theta' = \gamma^{-1/2}\right] = (\gamma^{1/2} - 1)\theta/(1-\theta) \\ P\left[\theta' = 0\right] = (1 - \gamma^{1/2}\theta)/(1-\theta) \end{array} \end{array} \tag{15}
$$

The processes P_1 and P_2 are illustrated in Figure 1 below. Notice that for a wide range of beliefs θ the two types behave differently, so this is not (initially) a pooling equilibrium; rather, the behaviors pool eventually with positive probability. (In the next section we shall analyze a variant of the game in which the behaviors eventually pool with probability one for any nondegenerate choice of prior belief.)

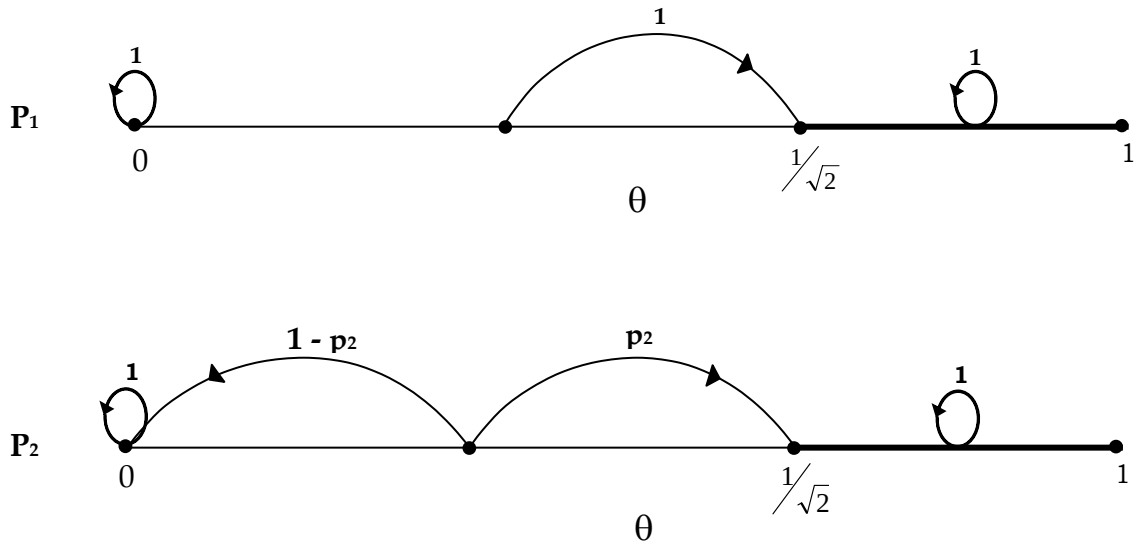

Figure 1. The equilibrium Markov processes on the belief space $\theta \in [0,1]$ for the Samaritan game (first version) with $\gamma = 2$. The thick segments consist of fixed points.

Assume now that the game is played infinitely often and that the types are myopic -- they do not look ahead. (The case of non-myopic types will be considered in section 7.) Suppose that the Self starts with the prior $\theta = \theta^0$. Each possible self i gives rise to a stochastic process P_i, which generates a random sequence of posteriors $\theta^0, \theta_i^1, \theta_i^2, ..., \theta_i^t, ...$. The Self is *initially uncertain* if $0 < \theta^0 < 1$. The *uncertainty is unresolved* for a given realization $\theta^0, \theta_i^1, \theta_i^2, ..., \theta_i^t, ...$ if the sequence $\{\theta_i^t\}$ is bounded away from both 0 and 1. In this case self-knowledge is never achieved.

Proposition 1. *Consider the infinitely repeated Samaritan game (first version) with myopic selves. If the fake Samaritan has a sufficiently high payoff from self-image, and there is any initial uncertainty about the true self* $0 < \theta^0 < 1$, *then under both* P_1 and P_2 *there is a positive probability that self-knowledge is never achieved*.

Let us interpret this result in the context of the motorist who comes upon the person stranded by the highway. The motorist's Self deliberates about her situation (she is sure of her gender). She reaches the conclusion that if she were a fake Samaritan (FS), she would employ a randomization device to decide whether or not to stop, whereas if she were a good Samaritan (GS) she would stop for sure. Thus if she observes that she is not stopping, she can be sure that she is an FS. However, there is a positive probability that no matter who she is (FS or GS) she will in fact stop, that is, stopping occurs with positive probability in both states of the world. In this case she updates her beliefs about who she is, but she will still not know *for sure* who she is. Moreover (this is the crux of the matter) these updated beliefs have the property that *on all future occasions she will always stop*, so the uncertainty will never be resolved. These results continue to hold when the selves are forward-looking, as will be shown in section 7.

6. The Samaritan game: second version.

We now complicate the situation by supposing that there are two types who differ in their preferred actions and *each type prefers to be the other type*. For example, suppose that in the Samaritan game one type is compassionate "deep down" but wants to think of himself as hard-nosed (the *macho Samaritan*), while the other type is uncompassionate but wants to have the image of being a macho Samaritan. This latter type is not the same as the fake Samaritan of the previous example, who was content to see herself as a plain-vanilla good Samaritan. This is a more sophisticated fake -- a *faux Samaritan* – who is uncompassionate by nature but treasures the image of being a macho Samaritan, that is, of being compassionate "deep down" and not wanting to show it.

It is easy to conjure up other examples with a similar structure. For instance, some may people crave the limelight but prefer to be known as homebodies (politicians who claim they want to 'spend more time with the family'), while others are essentially homebodies but would like to think of themselves as celebrities (Walter Mitty). In the economic sphere there are undoubtedly Veblenesque characters who want to show off their wealth even though they have frugal tastes (conspicuous consumption), while others may be natural spendthrifts who want to avoid the appearance of profligacy (inconspicuous consumption). The essential element in all of these examples is that there are two types who differ in their intrinsic preferences for acting, and want to have the image of being the opposite type. In the symmetric case, the payoffs from self-image take the form

	type 1	*type 2*
type 1	0	γ
type 2	γ	0

We can therefore write the utility functions in the form:

$$U_1(p_1, p_2, \theta) = \alpha_1 p_1 + \gamma_1 p_1 \theta_a + \gamma_1 (1 - p_1)\theta_b$$
$$U_2(p_1, p_2, \theta) = \alpha_2 p_2 + \gamma_2 p_2 \theta_a + \gamma_2 (1 - p_2)\theta_b,$$

where the posteriors (conditional on a or b being played) are

$$\theta_a = \theta p_1 / \left(\theta p_1 + (1-\theta) p_2\right),\ \theta_b = \theta(1 - p_1) / \left(\theta(1 - p_1) + (1-\theta)(1 - p_2)\right). \quad (16)$$

In the present situation let us assume that type 1 prefers to play a while type 2 prefers to play b. Without loss of generality let $\alpha_1 = 1,\ \alpha_2 = -1$. Since θ_a and θ_b are the probabilities the Self places on being type 1 (conditional on a or b), we also have $\gamma_1 < 0$ and $\gamma_2 > 0$. To simplify the exposition we shall assume that the game is *symmetric* in the sense that $\gamma_2 = -\gamma_1 = \gamma > 0$; the solutions are qualitatively similar when $|\gamma_1| \neq |\gamma_2|$. Thus we can write

$$U_1(p_1, p_2, \theta) = p_1 - \gamma[p_1\theta_a + (1-p_1)\theta_b],$$
$$U_2(p_1, p_2, \theta) = -p_2 + \gamma[p_2\theta_a + \gamma_2(1-p_2)\theta_b]. \qquad (17)$$

The larger γ is, the more the selves care about their self-image relative to the intrinsic payoff from taking their preferred action. Suppose first that $0 < \gamma < 1$. Since the selves are myopic, they would rather take their preferred action than try to manipulate their self-image. Hence in equilibrium we have:

$$0 < \gamma < 1 \Rightarrow \bar{p}_1 = 1,\ \bar{p}_2 = 0 \text{ for all } \theta. \qquad (18)$$

Next let us consider the case $\gamma > 1$. We begin by establishing the following.

Claim 1. The utility functions $U_i(p_i, \cdot)$ are concave in p_i, and

$$\partial U_1 / \partial p_1 = 1 + \gamma[(1-\theta_a)^2 - (1-\theta_b)^2] \qquad (19)$$
$$\partial U_2 / \partial p_2 = -1 + \gamma[(\theta_a^2 - \theta_b^2]. \qquad (20)$$

Proof. It is a straightforward (albeit tedious) exercise to verify (19) and (20) by differentiating (17) and collecting terms. Given this, it follows readily that the $U_i(p_i,\cdot)$ are concave. In particular, we know that the conditional posterior θ_a is nondecreasing in p_1 while θ_b is nonincreasing in p_1 (see (16)). It follows from this and (19) that $\partial U_1/\partial p_1$ is nonincreasing in p_1 . Similarly, $\partial U_2/\partial p_2$ is nonincreasing in p_2. Hence the functions $U_i(p_i,\cdot)$ are concave. □

Claim 2. *Assume that* $\gamma > 1$. *If* θ *is sufficiently close to 1, then in equilibrium both players choose* a, *whereas if* θ *is sufficiently close to 0 both players choose* b*:*

$$1/2+1/2\gamma \le \theta \le 1 \;\Rightarrow\; \overline{p}_1 = \overline{p}_2 = 1, \tag{21}$$

$$0 \le \theta \le 1/2-1/2\gamma \;\Rightarrow\; \overline{p}_1 = \overline{p}_2 = 0. \tag{22}$$

In both cases the posterior equals θ, *that is,* θ *is stable*.

The logic is the following: when the Self attaches sufficiently high probability θ to the true type being 1, the return to the *first type* from trying to deceive the Self is small, and is outweighed by the cost of playing the less preferred action, b. Meanwhile the *second type* is happy to go along with the Self's mistaken belief. Hence they both play action a for sure. But this means that the Self has no basis for revising his belief, so the posterior is θ. A similar logic applies when θ is sufficiently close to zero. The fact that the cutoff points are $\theta = 1/2 \pm 1/2\gamma$ can be verified by direct calculation.

Claim 3. *Assume that* $\gamma > 1$. *If* $1/2 - 1/2\gamma < \theta < 1/2 + 1/2\gamma$ *the unique equilibrium is mixed:*

$$\overline{p}_1 = \frac{(\gamma+1)(2\gamma\theta - \gamma + 1)}{4\gamma\theta} \text{ and } \overline{p}_2 = \frac{(\gamma-1)(2\gamma\theta - \gamma - 1)}{4\gamma(1-\theta)}. \tag{23}$$

Moreover the conditional posteriors are

$$\theta_a = 1/2 + 1/2\gamma\,,\ \theta_b = 1/2 - 1/2\gamma\,. \tag{24}$$

Proof. Necessary conditions for an interior (mixed) Nash equilibrium are $\partial U_1 / \partial p_1 = 0$ and $\partial U_2 / \partial p_2 = 0$. Since the U_i are concave, these conditions are also sufficient. From this and (19) and (20) we obtain

$$\theta_a{}^2 - \theta_b{}^2 = 1/\gamma \tag{25}$$

$$(1-\theta_a)^2 - (1-\theta_b)^2 = -1/\gamma \tag{26}$$

The unique solution is given by expression (24). The computation of the equilibrium $(\overline{p}_1, \overline{p}_2)$ is found by substituting these values of θ_a and θ_b into the formula for the conditional posteriors (see expression (16)). □

A key point to notice is that, in equilibrium, the conditional posteriors are *independent* of the prior θ so long as θ lies in the interval $(1/2 - 1/2\gamma,\ 1/2 + 1/2\gamma)$. This does not imply, however, that the equilibrium *strategies* $\overline{p}_1$ and $\overline{p}_2$ are independent of θ, as can be seen immediately from (23).

The results in claims 1-3 can be summarized in the following diagram of the Markov process P_1 that specifies how the posteriors $\theta \in [0,1]$ evolve when the true type is 1. (The process P_2 is similar.)

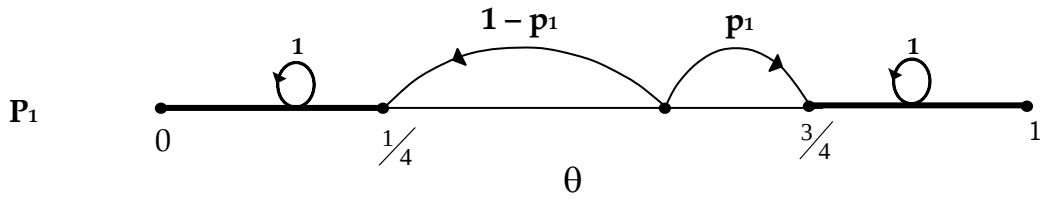

Figure 2. The Markov process for type 1 in the Samaritan game (second version) with $\gamma = 2$.

An interesting property of this process is that it is either in an absorbing state, or moves in one period to an absorbing state. The reason is that the conditional posteriors that result from a mixed equilibrium (see (24)) satisfy the conditions for a pure equilibrium in which the selves act alike (see (21) and (22)). Hence after one period there is no further movement in θ. It follows that, if there is any uncertainty initially, the learning process may change the degree of uncertainty but will never eliminate it. The striking feature of this result is that it holds for *any* prior beliefs that are not exactly correct, not just some beliefs.

Proposition 2. *Consider the infinitely repeated Samaritan game (second version) with symmetric payoffs and myopic selves. If the payoff from self-image is sufficiently high, and there is any initial uncertainty about the true self* $0 < \theta^0 < 1$, *then under both* P_1 and P_2 *it is certain that self-knowledge will never be achieved.*

7. Forward-looking selves

The preceding results are not an artifact of the assumption that the selves are myopic: the same kind of behavior can result when they are forward-looking and they play a subgame perfect equilibrium of the repeated game. This point can be illustrated with the first version of the Samaritan game. Suppose that each possible self (or type) i has a discount factor $0 < \delta_i < 1$. Since type 1 does not care about self-image, he chooses his preferred action (a) in each period irrespective of 2's behavior. Let us therefore focus on type 2. If at any time t action b is played, the Self knows that 2 is the true type, hence in this case $\theta_i^{t'} = 0$ for all $t' \geq t$. Once this happens, 2's optimal strategy is to play action b for all $t' \geq t$.

Hence we may describe 2's strategy as follows: let p_2^t be 2's probability of playing a in period t, conditional on a having always been played up until now. Given any other history, 2 plays b for sure from period t on. Given a string of a's through period t, the Self's posterior at time t is:

$$\theta^t = \frac{\theta^0}{\theta^0 + (\prod_{1 \leq s \leq t} p_s)(1-\theta^0)} \ . \tag{27}$$

Otherwise $\theta^t = 0$. Type 2's expected payoff in period t is $p_2^t(\gamma\theta^t - 1) + 1$, so 2's expected discounted payoff can be written

$$\sum_{t \geq 1} \delta_2^{t-1} (\prod_{1 \leq s \leq t} p_2^s)(\gamma\theta^t - 1) + 1/(1-\delta_2) \, . \tag{28}$$

Assume that $0<\theta^0<1$. Letting $r^0=(1-\theta^0)/\theta^0$ and $x^t=\prod_{1\le s\le t} p_2^s$, we can rewrite the expected payoff in the following simple form

$$\sum_{t\ge 1}\delta_2^{t-1}(x^t)/(1+r^0x^t)+1/(1-\delta_2)\,. \tag{29}$$

In period t the constrained optimum x^t, subject to $x^t\le 1$, is

$$\bar{x}^t=\ \min\ \left\{1,(\gamma^{1/2}-1)/r^0\right\}. \tag{30}$$

Now observe that type 2 can choose the probabilities $\bar{p}_2^t$ so that the constrained optimum $\bar{x}^t$ is achieved in *every* period t, namely,

$$\bar{p}^1=\ \min\ \left\{1,(\gamma^{1/2}-1)/r^0\right\} \text{ and } \bar{p}^t=1 \text{ for all } t\ge 2\,. \tag{31}$$

A little algebra shows that this is the same solution as in the myopic case. On the one hand, if $0<\theta^0<\gamma^{-1/2}$, type 2 randomizes in the first period and with probability $\bar{p}^1$ the posteriors are constant thereafter: $\theta^t=\gamma^{-1/2}$ for all $t\ge 1$. On the other hand, if $\gamma^{-1/2}<\ \theta^0<1$, type 2 plays action a for sure and the posteriors are $\theta^t=\theta^0$ for all $t\ge 1$. Thus Proposition 1 holds in the non-myopic case.

The second version of the Samaritan game is considerably more complicated to analyze when the player-types are forward-looking. I shall not attempt a complete characterization of the equilibria in this case. However, there certainly *exist* equilibria of this game in which initial uncertainty is never fully resolved. In particular, the equilibrium in the myopic case can also be sustained as a

subgame perfect equilibrium when the types are forward-looking: it suffices to assume that the myopic equilibrium will be played after any deviation.

8. Concluding remarks

This paper has been concerned with the problem of learning one's true identity when the intrinsic payoffs from acting are mixed with payoffs from self-image. The model is admittedly rather spare, and lacks a number of elements that one might wish to include in a more 'realistic' analysis. For example, we might want to relax the assumption that the selves are strictly rational, and introduce such features as wishful thinking, forgetfulness, confirmation bias and so forth. It seems likely, however, that these factors will only increase the opportunities for self-deception and make the quest for self-knowledge more difficult.

A second way in which the model could be made more realistic is to enlarge the space of available actions. Moreover, by allowing a wider range of choices, one might hope to make the quest for self-knowledge easier. The reason is that a large action space allows you to 'put yourself to the test', that is, to choose a costly action that only a very desirable sort of person (very brave, very generous, etc.) would be willing to bear. The trouble with this argument is that there seems to be little justification for increasing the range of possible actions while keeping fixed the range of possible types. On the contrary, for every costly action there might be a very desirable type of person who is willing to bear the cost, as well as an image-conscious type of person who is willing to bear the cost *because he wants to be seen as the very desirable type*. What matters is the trade-off rate between self-image and action, not the range of images and actions *per se*. The Samaritan game shows that, when people care enough about their self-image

relative to their actions, any uncertainty about who they are may never be resolved.

References

Ainslie, George (1986), "Beyond microeconomics: conflict among interests in a multiple self as a determinant of value," in Jon Elster, ed., *The Multiple Self*, Cambridge and New York: Cambridge University Press.

Ainslie, George (1992), *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person (Studies in Rationality and Social Change)*. Cambridge and New York: Cambridge University Press.

Akerlof, George, and William Dickens (1982), "The economic consequences of cognitive dissonance," *American Economic Revie*w, 72, 307-319.

Akerlof, George A., and Rachel Kranton (2000), "Economics and identity," *Quarterly Journal of Economics*, 115, 715-733.

Akerlof, George, and Rachel Kranton (2002), "Identity and schooling: some lessons for the economics of education," *Journal of Economic Literature*, 40, 1167-1201.

Akerlof, George, and Rachel Kranton (2005), " Identity and the economics of organizations," *Journal of Economic Perspectives*, 19, 9-32.

Bem, Daryl (1967), "Self-perception: an alternative interpretation of cognitive dissonance phenomena," *Psychological Review*, 74, 183-200.

Bem, Daryl (1972), "Self-perception theory," in L. Berkowitz, ed., *Advances in Experimental Social Psychology*, vol. 6, 1-62. New York: Academic Press.

Benabou, Roland, and Jean Tirole (2001), "Self-knowledge and self-regulation: an economic approach," Working Paper, Princeton University and Université de Toulouse.

Benabou, Roland, and Jean Tirole (2004), "Willpower and personal rules," *Journal of Political Economy*, 112, 848-886.

Benabou, Roland, and Jean Tirole (2006a), " Incentives and prosocial behavior," *American Economic Review*, 96, 1652-1678.

Benabou, Roland, and Jean Tirole (2006b), "Identity, dignity, and taboos: beliefs as assets," Working Paper, Princeton University and Université de Toulouse.

Bodner, Ronit, and Drazen Prelec (1997), "The diagnostic value of actions in a self-signaling model," Working Paper, Massachusetts Institute of Technology.

Bodner, Ronit, and Drazen Prelec (2003), "Self-signaling and diagnostic utility in everyday decision making," in I. Brocas and J. Carillo, eds, *The Psychology of Economic Decisions. Vol 1: Rationality and Well-Being*. Oxford University Press.

Elster, Jon (1986), ed., *The Multiple Self*, Cambridge and New York: Cambridge University Press.

Geanakoplos, John, David Pearce, and Emilio Stacchetti (1989), "Psychological games and sequential rationality," *Games and Economic Behavior*, 1, 60-79.

Kant, Immanuel (1785), *Fundamental Principles of the Metaphysics of Morals.* Allen W. Wood, ed., *Basic Writings of Kant*. New York: Modern Library [2001, p. 165].

Schelling, Thomas C. (1986), "The mind as a consuming organ," in Jon Elster, ed., *The Multiple Self*, Cambridge and New York: Cambridge University Press.

Smith, Adam (1759), *The Theory of the Moral Sentiments*. Oxford: Clarendon Press [1975].