Rational Self-deception

Eli Spiegelman

August, 2012
Abstract:

This note presents a simple criterion for identifying cases in which the optimal beliefs are the same as the objectively true beliefs. In other cases, self-deception is compatible with expected utility maximization. The criterion is that payoff-relevant uncertainty must be resolved at the moment that utility is perceived. The context is a three-period model, in which the agent is faced with a state-contingent optimal action, in which one state yields a higher payoff. In period 0 she observes the objective prior probability that each state will occur, but may alter her beliefs about these probabilities (self-deceive). The beliefs she chooses in period 0 determine her action in period 1 as a standard maximization procedure. In period 2, a signal yields information about the state of the world. It is shown that the objective prior is optimal if and only if the signal in period 2 is perfectly revealing.

Keywords: self-deception, beliefs, uncertainty, rationality, savoring, procrastination
JEL Codes: D81, D83
Introduction

Jack recently saw a movie about the production of industrial chicken. The movie showed, in graphic and (to Jack) horrifying detail, the process by which agro-industrial firms maximize the flow of product. Jack is recalling the movie now, in the grocery store, while deliberating between a conventional chicken ($7.99) and an organic, free range chicken ($21.99). Jack is faced with a common but peculiar quandary. His preferences are such that, if the process by which the industrial chicken was made was distasteful enough, he would prefer to spend the extra $14 on the organic chicken. However, he is not now, and likely will never be, in possession of the information required to determine whether in fact this is the case; after all, the movie is not a perfect indicator, and particularly not for the specific chickens on the shelf before him. Still, he must make a decision, or go hungry. The point of this article is that in this case Jack has an incentive to believe that the conventional chicken is not really so bad as a balanced appraisal of the movie might suggest.

As mentioned, Jack’s position is peculiar but not uncommon. Whenever preferences contain some component that goes beyond materially observable facts, the incompleteness of the information affects not just the decision process but also the evaluation of the outcome itself. We therefore enter the world of belief-dependent utility, psychological games (Battigalli and Dufwenberg, 2007, 2009) and psychological decision-making (Caplin and Leahy, 2001). The central proposition below is that this change also yields an incentive to manipulate beliefs, which has been so far under-studied in the literature. By beliefs I refer to the standard concept of a subjective probability distribution over the conceivable states of the world, where these might include other people’s strategies or beliefs. Beliefs are traditionally considered to be the information we use to guide us from actions to outcomes. In this case, it is relatively immediate that we “prefer” to have beliefs that are consistent with some objective truth. If I get $5 for correctly calling a coin toss, and lose $1 for missing it, then in general I would prefer my beliefs
to be accurate, in the sense of corresponding to the true probabilities involved. By contrast, consider flipping the coin \( n \) times, and my beliefs about flip number \( n + 1 \). In this case, since I will never know the true value, the incentives change. One could imagine a “sour grapes effect”, in which I would prefer to believe I would have missed that toss. Alternatively, one could imagine that I might prefer to believe it would have saved me.

To take another large field of relevance, there are many “social preference” parameters that we may never exactly observe. For example, consider the fairness or morality of a certain choice, or the honesty or intentions of those with whom we interact. Or consider a decision that may have some reflection on our character or abilities, or have some future effects, which depend on an as-yet unrealized state of the world in which we have some kind of stake. In all these cases, the evaluation of different action choices depends on beliefs, rather than on outcomes or actions themselves. Thus, as we argue below, in all these cases, correct beliefs may not be optimal.

Theory

The basic decision problem is relatively straightforward. An agent \( X \) must make a choice, called an action \( a \) from a finite set of actions \( A = \{a_1, a_2, \ldots, a_N\} \). In addition, there is a finite set of possible states of the world\(^1\), \( \Omega = \{\omega_1, \omega_2, \ldots, \omega_M\} \). Payoffs \( \pi \) map the product of actions and states to the real line. I make the further assumptions:

A1: For any state of the world, there is a single action that gives the highest payoff.

A2: Any action is payoff-maximizing for at most one state.

A3: States can be ordered such that the payoff earned by the maximizing action is a decreasing function of the order of the states.

---

\(^1\) I assume that the action space is constant across states because I will later allow \( X \) to assign positive probability to all states. If some states allowed different actions, then she could make inferences about the state by observing the possible actions. In effect, \( \Omega \) refers to all the possible states given the actions available – all those that cannot be ruled out. Compare the discussion in Yildiz (2004).
A1 and A2 imply that the structure of the decision is a coordination “game,” played against nature. A3 says that all states give different maximal payoffs, which makes the coordination game Pareto ranked.

The decision that X has to make has three steps. First (t = 0) she develops prior beliefs about the state of the world. These beliefs, as mentioned above, take the form of a probability distribution over the states of the world. I suppose there is some kind of psychological commitment to these priors, so that they maintain relevance in the following steps. In the second phase (t = 1), based on these priors, X makes a decision that maximizes expected payoffs given these beliefs. In the third step (t = 2), she gets more information about the true state, and updates her beliefs.

There is nothing very unusual in this set-up, unless it’s the fact of breaking the decision down into so many parts. At this point, I introduce two new assumptions. The first is the problem that I am interested in investigating: self-deception. In phase t = 0, the agent observes exogenous information that leads to non-degenerate prior beliefs p. However, she may costlessly adjust these beliefs to some ρ, which will then become her “committed” beliefs. Thus, I make the assumption that

A4: For any state ω ∈ Ω, ρω may not be equal to pω, although \[ \sum_{ω=1}^{M} ρ_ω = 1 \]

The second assumption provides the solution I offer as a necessary and sufficient condition to make that self-deception rational. In phase t = 2, the additional information that X receives may not be fully revealing. In particular, X observes a signal q from some set Θ that maps the state space to probability distributions over the states. This mapping is simply X’s interpretation of the signal as suggesting a particular state. Specifically, it is known that the signal q(ω) is most likely to occur in state ω for any ω in Ω. For simplicity, suppose that there is a “better than average chance” of getting the “right” signal, and an equal, “worse than average” chance of getting any signal other than the right one.
A5: For any \( x \in \Omega \), \( \Pr[q = q(x)|\omega = x] = \theta,\ 1 \geq \theta > \frac{1}{M}, \Pr[q = q(x)|\omega \neq x] = \epsilon = \frac{1 - \theta}{M - 1} \).

Having observed this signal, \( X \) updates her beliefs by Bayesian reasoning, arriving at final beliefs \( \mu \), defined as follows:

\[
\forall x, y \in \Omega, \mu_x = \Pr[\omega = x | q(y)] = \begin{cases} \frac{\rho_x \theta}{\rho'q(y)} & \text{if } x = y \\ \frac{\rho_x \epsilon}{\rho'q(y)} & \text{otherwise} \end{cases}
\]  

where the denominator is a vector multiplication. Given these assumptions, \( X \)'s decision problem is to choose \( \rho \) in \( t = 0 \) so as to maximize expected utility in \( t = 2 \). Crucially, in \( t = 0 \), \( X \) knows that the choice of \( \rho \) will determine her action in \( t = 1 \), but cannot have any effect on the true state of the world or the likelihood of seeing different signals. The question becomes: When will \( X \) choose \( \rho \neq p \)? And the answer is the main proposition of this paper:

**Proposition:** \( \rho = p \) is optimal if and only if \( \theta = 1 \).

Before demonstrating this proposition, note that for any action \( a \), the expected utility that \( X \) will perceive in \( t = 2 \), from the perspective of \( t = 0 \), is

\[
EV(a | \rho) = \sum_{s \in \Omega} p_s (\theta \mu_s \pi(a, s) + (1 - \theta) \sum_{s', \omega} \mu_{s', \omega} \pi(a, x))
\]

where \( \mu \) is as defined in (1). Next, the assumptions above imply that, if the perceived likelihood of any state is high enough, \( X \) will choose the action that maximizes payoff in that state. Call this action \( a^*(\omega) \). Now we proceed with the proof. IF: Suppose \( \theta = 1 \). In this case the bottom branch of (1), and indeed the second term inside (2) both disappear. Further, the top branch of (1) is identically equal to 1. This means that expected utility does not depend directly on \( \rho \) in \( t = 2 \). The
expected utility for any action is simply \( \sum_{s=1}^{M_0} p_s \pi(a, s) \), which – given the assumed structure of the problem – has a unique solution. However, the action taken at \( t = 1 \) does depend on \( \rho \). There may be many values of \( \rho \) which lead to the optimal solution, but \( \rho = p \) is clearly one of them.

**ONLY IF:** Now suppose \( \theta < 1 \). Focus on the actions that maximize utility in some state. Expression (2) implies that the expected utility of each of those actions will be monotonically increasing in the prior likelihood laid on the state in which it maximizes utility. This in turn implies that the maximum expected utility must be a corner solution; the optimum beliefs are degenerate. Since the exogenous priors \( p \) were non-degenerate, this implies that at the optimal solution, \( \rho \neq p \). QED

**An example**

Agent X lives for 3 periods \( t = 0, 1, 2 \). There are two states of the world, \( L \) or \( H \). In period 0 her non-degenerate, objective priors, denoted \( p_L \) and \( (1 - p_L) \), are determined by some exogenous process. At this point, by “some mechanism” she costlessly adjusts these priors to \( \rho_L \) and \( (1 - \rho_L) \). She may, but need not, choose \( \rho_L = p_L \). If she does not, she self-deceives. In period 1 she must take an action U or D to maximize her expected period-2 utility, based on the values in Table 1, below\(^2\), and \( \rho \).

<table>
<thead>
<tr>
<th>Action</th>
<th>State</th>
<th>L</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Payoffs by state

In period 2, the terminal stage, X observes a signal \( \sigma \in \{l, h\} \), which is correlated with the state of the world. Suppose for simplicity that \( \Pr[h \mid H] = \Pr[l \mid L] = \theta \), with \( 1 \geq \theta > 0.5 \). I will refer to \( \theta \) as the informativeness of the signal, and distinguish cases where \( \theta = 1 \) from those

\(^2\)The values are without loss of generality. The key for the model is just that the optimal choice is different in the different states, and that some states have preferred outcomes to others. As long as this condition holds, indeed, the model generalizes immediately to \( N \) states of the world.
where $\theta < 1$. The latter corresponds, for instance, to the “interim” utility in Benabou and Tirole (henceforth BT) (Bénabou and Tirole 2010). Also like BT, I allow utility to be based on a (non-degenerate) probabilistic assessment of the situation. Define $X$’s beliefs at the end of period 2 as $\mu_m (1 - \mu_m)$ that the state of the world is $L$ and $H$, respectively, based on the signal $\sigma = h, l$. The decision tree for this scenario is shown in Figure 1, below.

To clarify the decision context, at $t = 0$ $X$ chooses beliefs $\rho$, knowing that while these will determine her choice at $t = 1$, they can have no impact on the true state of the world, or the probability of seeing signal $h$ or $l$. Thus in period 0, $X$ will use the objective prior, $p$, and the informativeness of the signal, $\theta$, to determine what beliefs, $\rho$, will yield the greatest expected utility in period 2. Given the values in Table 1, $X$ knows in period 0 that in period 1 she will choose action $D$ if and only if her beliefs are $\rho < 2/3$. The final beliefs, which determine utility, are as follows. Upon seeing a low signal, the belief that the state is $L$ is

![Decision Tree](image-url)
\[
\mu_i = \Pr[L | I] = \frac{\rho \theta}{\rho \theta + (1 - \rho)(1 - \theta)}
\]

On the other hand, if the signal \( h \) appears, \( X \) will believe she is in state \( L \) with probability

\[
\mu_h = \Pr[L | h] = \frac{\rho (1 - \theta)}{\rho (1 - \theta) + (1 - \rho) \theta}
\]

In this example it is clear that if \( \theta = 1 \) then (3) equals 1 and (4) equals zero, for any \( \rho \). This breaks the information sets A, B, C and D in Figure 1, and implies that the expected value of \( \mu \) is \( p \), regardless of \( \rho \). Suppose without loss of generality that \( p < 2/3 \), so it is rational for \( X \) to choose \( D \). Would it be rational for \( X \) to choose to adjust her beliefs? Any belief \( \rho < 2/3 \) will result in the same action as \( p \). Hence, from the pre-adjustment point of view, there would be no incentive to choose any \( \rho < 2/3 \) different from \( p \). An adjustment to \( \rho > 2/3 \) would result in choosing \( U \) instead of \( D \) at \( t = 1 \). Given \( p \), this would result in a lower expected payoff, so \( X \) does strictly better when \( \rho = p \) than for any \( \rho > 2/3 \). Thus when \( q = 1 \), the exogenous beliefs are optimal.

On the other hand, if \( q < 1 \), then the expected utility for each available action \( A = U, D \) is

\[
EV(A | \rho) = \Pr(l)[\mu_i U(A, L) + (1 - \mu_i) U(A, H)] + \Pr(h)[\mu_h U(A, L) + (1 - \mu_h) U(A, H)]
\]

For any action, and any value of \( \theta \) in \([0.5, 1)\), this function is monotone in \( \rho \). Therefore, the optimal beliefs will be corner solutions. Since by assumption the priors are non-degenerate, this means that the optimal beliefs are different than, and in particular, more extreme than the objective priors for all values of \( \theta \) in \([0.5, 1)\). Furthermore, the expected utility will be higher if \( X \) chooses to believe the state is \( H \) and choose action \( D \) than in the reverse case. Figure 2, below, shows the value functions for each action when \( \theta = 0.9 \) and \( p = 0.5 \). As \( \theta \) goes to 1, the functions for actions \( U \) and \( D \) converge to horizontals at 0.5 and 1, respectively. However, for any value of
$\theta < 1$, the “envelope function,” defined as the expected value of the action which does best, has a global maximum at $\rho = 0$.

![Figure 2: Value functions of actions $U$ and $D$ by $\rho$ when $\theta = 0.9$, $p = 0.5$](image)

**Discussion**

In cases of self-deception a person manages by some mechanism to engage in “motivated reasoning” and thereby manipulate her beliefs in some way. Roughly speaking, the self-deceiver chooses to believe $x$ (the “target belief”), rather than any other $y$ (each of which is logically a not-$x$), and specifically rather than some $y$ which is at least as well supported by observed data as $x$ is. Such behavior raises two important questions: how, and why?

This note addresses the second question. The first – how do people selectively integrate the available information into their beliefs – is an important empirical problem. However, granting that they can does not explain why they would choose to do so. Beliefs matter because,  

---

3 (Szabados 1973) argues that self-deception requires $y$ is strictly better supported than $x$ by the data; otherwise it is wishful thinking. (Mele 1997) Suggests the 4-part sufficient conditions for $S$ to be self-deceived (1) the belief $x$ that $S$ acquires is false; (2) $S$ treats data relevant to the truth of $x$ in a motivationally biased way; (3) the biased treatment is a nondeviant cause of $S$’s acquiring the belief $x$; (4) the body of data $S$ possesses provides greater warrant for $\neg x$ than for $x$ (p. 95).
in general, the preferred action will change as the state of the world changes, and beliefs summarize information about the state of the world. Acting on beliefs different from those the data support seems destined to lead to sub-optimal outcomes. Yet self-deception is a very intuitive phenomenon.

The model presented above makes plain a sufficient and necessary condition for self-deception to operate as a motivation in the context of Bayesian rationality, that is, subjective expected utility maximization. The condition is an uncertainty postulate, in that it requires that beliefs about the true state of the world remain non-degenerate at the moment when utility is experienced. Relaxing the standard assumption that states are revealed at the end of the game gives beliefs a quite radically different role. They no longer simply guide the decision-maker towards the optimal outcome. Rather, they take on affective benefit, which supplies the motivation to manipulate them.

In particular, we have found that in the case where (a) the optimal action depends on the state of the world, and (b) some states of the world are “better” than others, it is necessary and sufficient that final utilities remain uncertain for self-deception to be a motivating force. Not only that, but in these cases agents do best when they can establish “unshakeable beliefs” that their preferred state of the world is correct. The intuition behind the result is that, when uncertainty is resolved, the manipulated beliefs may still affect the action choice at $t = 1$, but can no longer have any impact on beliefs at $t = 2$. The final beliefs therefore become irrelevant for the decision made at $t = 0$. By contrast, when $\theta < 1$, $\rho$ has a direct impact on utility through the beliefs in (3) and (4).

It will perhaps be noticed that an alternate formulation for the decision problem above would have $X$’s utility depend directly on the final belief. In this case, expression (5) is interpreted as the utility function, which is maximized by choosing $\rho$, which then nails down $\mu$, and thus also $A$. This modification has no effect on the proposition above. It makes the model more familiar, in that the “outcomes” (including the belief) are known with certainty at the end of
the decision problem. However, it obscures the mechanism by which the result is obtained. The reason why \( \rho = p \) is optimal if and only if \( \theta = 1 \) is that in the contrary case the state of the world is both relevant to the actor’s choice problem and also unknown. The interpretation of (5) as a utility function also makes a more complex theoretical structure, at least graphically. Sacrificing both “praxeological realism” and theoretical simplicity to make the model conform to decision problems that use beliefs in a fundamentally different way seems unwarranted.

While the condition of persistent uncertainty is unfamiliar in the literature, moreover, in the background it does characterize all of the theoretical papers I am aware of, and I have not been able to construct an example that convincingly contradicts it. A similar trope occurs in the empirical literature, where a robust finding is that self-serving biases are more prominent in treatments that preserve ambiguity or uncertainty about the outcomes (Felson 1981; Dunning, Meyerowitz et al. 1989; Dana, Weber et al. 2007; Haisley and Weber 2008; Valdesolo and DeSteno 2008; Sloman, Fernbach et al. 2010).

An interesting implication of this unresolved uncertainty is that it implies that the agent must have preferences that go beyond her own material well-being. Material well-being does not exist probabilistically, and hence no material benefit can possibly be perceived as long as uncertainty about its magnitude remains. *Homo economicus*, whose preferences are identical to material well-being, is necessarily impervious to such temptation. Perhaps unfortunately for the rest of us, “extra-material” incentives yield abundant opportunities. As illustrated in the introductory paragraph, one area of particular interest may be the demand for environmental goods. This is interesting because the self-deception may, in general, push people in several different directions. Those who, for instance, buy hybrid cars will downplay the environmental costs of battery disposal. Those who buy conventional chicken will downplay the benefits of ecologically sensitive production methods. There are also several examples from the previous literature. In BT, agents want both to maximize their lifetime income (material incentive) and to
have a good impression of that income along the way (psychological incentive). Lord, Ross et al. (Lord, Ross et al. 1979), found that people want both to have correct beliefs (on the appropriateness of capital punishment, in this example), and also that those beliefs agree with their previous ideas or general ideological framework. In studies such as (Quattrone and Tversky 1984; Ditto, Pizarro et al. 2009) subjects would like to know whether they have a proclivity to some disease, but also would like the result to be negative. Knowing their true proclivity would potentially allow people to take preventative – or at least preparative – measures, reducing the material cost the disease imposes. However, that would cost them the belief that they were healthy in the interim. Without the interim incentive, there would be no reason to manipulate beliefs in the way those researchers interpret their subjects as doing.

Conclusion
The condition of persistent uncertainty answers the why of self-deception. Self-deception occurs because when uncertainty is unresolved, prior beliefs no longer have solely instrumental value, pushing the actor towards the best possible outcome. They also have direct benefit, since they influence the interpretation of later information. This principle seems to unify much fascinating work in economics. Savoring, procrastination, self-image management, fairness and reciprocity are all (non-exhaustive) examples of cases where an agent may have an interest in maintaining certain beliefs about choices. They all also share the characteristic that the object of the beliefs is not known with certainty at the time utility is perceived.

The model also suggests answers for the first question of how beliefs are manipulated. Essentially, X “reverse engineers” her prior beliefs to bring observed data into line with what she would prefer to be true. Knowing that the action D will give her a higher utility in state H than the action U will in state L, she manages to believe that H is likely enough to justify that choice by conditioning the observed information with on a distorted (e.g., selectively attended) base understanding, or prior. However, the essential message of the model will hold with other kinds
of manipulation. For instance, $X$ could equivalently manipulate the signal $\sigma$, for instance by interpreting it as implying a different $\theta$. Which model is more accurate is an empirical question, and may well depend on the case considered. It remains a promising area for future research.
Works cited


