

CHRISTOPHER L. STEPHENS

WHEN IS IT SELECTIVELY ADVANTAGEOUS TO HAVE  
TRUE BELIEFS? SANDWICHING THE BETTER SAFE THAN  
SORRY ARGUMENT

(Received in revised form 25 November 2000)

**ABSTRACT.** Several philosophers have argued that natural selection will favor reliable belief formation; others have been more skeptical. These traditional approaches to the evolution of rationality have been either too sketchy or else have assumed that phenotypic plasticity can be equated with having a mind. Here I develop a new model to explore the functional utility of belief and desire formation mechanisms, and defend the claim that natural selection favors reliable inference methods in a broad, but not universal, range of circumstances.

INTRODUCTION

What sorts of belief and desire formation policies should we expect to evolve by natural selection, and under what conditions? Several authors have suggested that reliability will be favored by natural selection.<sup>1</sup> Quine (1969), for example, states that “creatures invertebrates wrong in their inductions have a pathetic but praiseworthy tendency to die out before reproducing their kind” (p. 126). Daniel Dennett (1987) writes that “[n]atural selection guarantees that *most* of an organism’s beliefs will be true, *most* of its strategies rational” (p. 75). And before recanting his fondness for evolutionary theories of content, Jerry Fodor (1981) claimed that “Darwinian selection guarantees that organisms either know the elements of logic or become posthumous” (p. 121).

In contrast, other philosophers have been more skeptical.<sup>2</sup> Stephen Stich (1990, p. 60), for instance, states that “[t]he thesis at hand is that natural selection prefers reliable inferential systems – those that do a good job at producing truths and avoiding falsehoods – to unreliable ones. And it is my contention that this thesis is false”. While the possible advantage of reliable belief formation devices is



*Philosophical Studies* **105**: 161–189, 2001.

© 2001 Kluwer Academic Publishers. Printed in the Netherlands.

intuitively clear, the possible advantage of *unreliable* belief formation devices is more puzzling. Stich elaborates his argument as follows:

To complete the argument, it remains to note that a very cautious, risk-averse inferential strategy – one that leaps to the conclusion that danger is present on very slight evidence – will typically lead to false beliefs more often, and true one's less often, than a less hair-trigger one that waits for more evidence before rendering a judgment. Nonetheless, the unreliable, error-prone, risk-averse strategy may well be favored by natural selection. For natural selection does not care about truth; it cares only about reproductive success. And from the point of view of reproductive success, it is often better to be safe (and wrong) than sorry (Stich 1990, p. 62).

I call this the *better-safe-than-sorry argument*. Suppose there is only a small chance that a tiger is nearby. It is better to be safe and believe that *a tiger is nearby* rather than believe that *no tiger is nearby* and be sorry if it turns out you're wrong. The better-safe-than-sorry argument figures prominently in many attempts to show that rules or mechanisms that lead to many false beliefs can be adaptive.<sup>3</sup>

Neither the proponents nor the skeptics, however, have spent much time examining detailed models exploring the conditions under which evolution by natural selection would favor various kinds of belief and desire formation policies. Philosophers have typically answered the question about whether natural selection favors rational or true beliefs in a *yes-or-no* fashion. My contention is that this debate has been pursued at too abstract a level of analysis. With respect to those few who *have* developed models, they have attempted to analyze belief formation policies independently from desire formation policies and decision rules. I argue below that this is a mistake. After a brief examination of the existing approaches, I develop a new model to characterize the functional utility of belief and desire formation mechanisms. The-better-safe-than-sorry argument, it turns out, has limited scope. Certain kinds of conditions exist in which we should expect organisms' beliefs and desires to be reliable. But there is no monolithic answer to our question about the evolutionary value of believing the true and desiring the good. Rather than yes-or-no, the answer to our central question about whether natural selection favors reliable belief and desire formation

rules is: *it depends*. *On what it depends* is something that I'll attempt to assess.

#### GODFREY-SMITH AND SOBER'S MODELS

Peter Godfrey-Smith (1991, 1996) and Elliott Sober (1994) have developed models about when an organism should be flexible and pay attention to information in its environment, and when it should be inflexible and follow a certain strategy regardless of the information available in the environment. These models answer the question of when it pays an organism to be inflexible or flexible in response to various environmental cues that may be more or less reliable indicators of the state of the world. Godfrey-Smith (1991, 1996) in particular, uses signal detection theory to help determine when an environmental cue is good enough so that it is worth paying attention to.<sup>4</sup>

Both Sober and Godfrey-Smith reach similar conclusions: flexibility is better than inflexibility if the error characteristics of the learning device are small enough (and whatever evolutionary cost there is to flexibility does not outweigh the benefits). One consequence of their approach is that even though, for any given action, a true belief may be better than a false one, it is still the case that rules which lead to more false beliefs can be selected for. The *better-safe-than-sorry argument* is an instance of this phenomenon. Rules that lead to a high percentage of false beliefs can be selected for in situations where the organism does not have access to highly reliable cues and there is a high cost to a false negative (*e.g.*, thinking there is no tiger present when there is).

Although Sober and Godfrey-Smith's models can be applied to policies for forming beliefs, they are in fact more general; they are models of phenotypic plasticity. The model to be developed here, on the other hand, is explicitly *mental*. In addition to representing both beliefs and desires, my model uses two distinct decision tables in order to represent both what is going on in the organism's mind and what is going on in the world.<sup>5</sup> Finally, it is important to note that my model is about *when* we should expect an organism to have more or less reliable ways of finding out about the environment – independent of the *manner* in which the organism does so. A more

complex model could embed a signal detection framework *within* the model I provide below. In so doing, one could find out not only when we should expect an organism to have more reliable ways of finding out about the world, but how it should do so (*e.g.*, whether by innate prejudice or by learning in response to some sorts of cues).

#### A NEW MODEL

The following model compares both reliable belief and desire formation rules with rules that deviate from maximum reliability. Although the concept of a reliable belief formation rule is familiar in epistemology, the idea of a reliable desire formation rule is less standard. Desires, after all, are not true or false, so how could a desire formation rule be reliable? My answer is that a desire formation rule is reliable to the extent that it generates desires that, if satisfied, enhance the organism's fitness.<sup>6</sup> Desires for what is fitness enhancing, and for things that are coextensive with what is fitness enhancing, are what I call desires for the good.<sup>7</sup> This is clarified below.

A central aspect of the model concerns the relationship between what the organism *thinks* is true and what *is* the case, and what it *wants* and what will actually *maximize fitness*. Consider the decision problem in Table I. Two actions  $A_1$  and  $A_2$ , and two possible states of the world, described by the propositions  $S_1$  and  $S_2$ , with objective probabilities  $p$  and  $(1 - p)$ , respectively:<sup>8</sup>

TABLE I. Objective Decision Matrix

	$S_1$	$S_2$
	$p$	$1 - p$
$A_1$	$w_{11}$	$w_{12}$
$A_2$	$w_{21}$	$w_{22}$

Here  $w_{11}$ ,  $w_{12}$ ,  $w_{21}$ ,  $w_{22}$  represent the fitness outcomes of the alternative actions, conditional on the state of the world. Their actual values are important only in so far as they establish an

interval ranking of outcomes. For  $A_1$  to be favored over  $A_2$  (objectively),  $pw_{11} + (1 - p)w_{12} > pw_{21} + (1 - p)w_{22}$ , which reduces to:

( $\alpha$ ) *Objective Criterion for  $A_1$  to be the Best Action:*

$$\frac{p}{(1 - p)} > \frac{(w_{22} - w_{12})}{(w_{11} - w_{21})}$$

The *objective criterion for best action* states the objective conditions under which  $A_1$  will be favored by natural selection over  $A_2$ .<sup>9</sup> In order to better understand this criterion, some terminology will be useful. The *importance* of a proposition is the fitness difference between performing  $A_1$  and  $A_2$  if the proposition is true.<sup>10</sup> The proposition that a predator is nearby has a large importance because, if true, it makes a big difference what the organism does. In contrast, the proposition that no predator is nearby is relatively unimportant, because it doesn't matter as much what the organism does, if it is true. So the criterion says that the ratio of the probability of  $S_1$  to the probability of  $S_2$  must be greater than the ratio of the importance of  $S_2$  to the importance of  $S_1$ . More precisely, the importance of  $S_1$  is defined as  $(w_{11} - w_{21})$  and the importance of  $S_2$  is  $(w_{22} - w_{12})$ .<sup>11</sup>

*Criterion ( $\alpha$ )* is a representation of what will maximize fitness. It is therefore *objective* in the sense that what the better action is depends on the facts about what is really fitness enhancing, and not on what the organism happens to believe or value. However, since multiple combinations of beliefs, desires and decision rules can lead to the same action, we are interested in the range of possible decision procedures that could lead the organism to satisfy this criterion.

Now let's turn to the organism's subjective decision procedure, in which it decides what to do by consulting its subjective probabilities and utilities. Let  $q$  = the value of the organism's subjective probability that  $S_1$  is true, and  $(1 - q)$  the organism's subjective probability that  $S_2$  is true.  $U_{11}, \dots, U_{22}$  represent the organism's subjective utilities as shown in Table II.

TABLE II. Subjective Decision Matrix

	$S_1$	$S_2$
	$q$	$1 - q$
$A_1$	$U_{11}$	$U_{12}$
$A_2$	$U_{21}$	$U_{22}$

Then, for  $A_1$  to be favored over  $A_2$ , we have:

( $\beta$ ) *Subjective Criterion for  $A_1$  to be the Best Action:*

$$\frac{q}{(1 - q)} > \frac{(U_{22} - U_{12})}{(U_{11} - U_{21})}$$

The *subjective criterion for best action* states the conditions under which the organism will *decide* to do  $A_1$  over  $A_2$ . In order for the organism to ‘do the right thing’ from the evolutionary point of view, it needs a decision procedure that assigns some combination of values to  $q$  and  $U_{ij}$  so that it will perform the action that the criterion ( $\alpha$ ) says is the objectively best action.

The following optimality criterion describes how the subjective and objective probabilities and utilities must be related for the organism to perform the action that is fitter. The optimal organism must assign values so that:

*Optimality Criterion:  $\alpha$  if and only if  $\beta$ .*

There are several ways for an organism to satisfy this criterion. One way is to *believe the true* and *desire the good*. That is:

$$(I) \quad p = q \text{ and } w_{ij} = U_{ij} \text{ for all } i, j.$$

Since  $p$  is the objective probability that  $S_1$  is true, if the organism assigns its subjective degree of belief  $q = p$ , it *believes the true*. Similarly, since the  $w_{ij}$ ’s are the objective fitness values for the various outcomes, if each of the organism’s subjective utilities (its  $U_{ij}$ ’s) is equal to its corresponding  $w_{ij}$ ’s, the organism *desires the good*.

It is important to be clear about what believing the true and desiring the good means. If an organism believes the true, for example, it does not mean that the organism must have a true belief about the proposition in question in any given decision problem in which that proposition is used. Rather, it means that the organism's subjective probability estimate in a given proposition matches its objective frequency. For instance, if the frequency of poisonous mushrooms in the organism's environment is 0.1, then an organism believes the true if it has a degree of confidence of 0.1 in the belief that mushrooms are poisonous.<sup>12</sup> However, when picking a particular mushroom, an organism that believes the true in my sense may have a mistaken belief about whether it is poisonous. It is not *specific* states of the world that determine objective utilities; rather, it is an average across (evolutionary) time that determines fitness.

It is also important not to misinterpret what it means to desire the good. Here it is also important to keep in mind that a perfect matching of subjective and objective utilities does not make sense from a decision theoretic point of view. It is the interval scales that matter (how much you prefer one outcome to another and what the probabilities of the outcomes are). Positive linear transformations preserve linear scales (see Resnik 1987, p. 83).

Recall that the importance of a proposition is determined by the outcome difference between performing one act as opposed to another if that proposition is true. The *expected importance* is the probability of a proposition multiplied by its importance. We now need to introduce a pair of related concepts:

The *objective* expected importance of proposition  $S_1$  is  $p(w_{11} - w_{12})$ ; for  $S_2$ , it is  $p(w_{22} - w_{21})$ .

The *subjective* expected importance of proposition  $S_1$  is  $q(U_{11} - U_{12})$ ; for  $S_2$ , it is  $q(U_{22} - U_{21})$ .

So another way to meet the optimality criterion is for the *subjective expected importances* to match the *objective expected importances*.

$$(II) \quad p(w_{11} - w_{21}) = q(U_{11} - U_{21}), \text{ and } (1 - p)(w_{12} - w_{22}) = (1 - q)(U_{12} - U_{22})$$

Condition (II) is weaker than Condition (I) because (I) entails (II) but (II) can be satisfied without satisfying (I). To satisfy condition (II) there is no need for the organism to believe the true and desire

the good. For example, the organism can assign a value to  $q$  much lower than the objective frequency  $p$  provided it also assigns a value to the utility difference ( $U_{11} - U_{21}$ ) that is much greater than ( $w_{11} - w_{21}$ ).

Although condition (II) is sufficient to meet the optimality condition, it is not necessary. A *necessary and sufficient* condition for an organism to meet the optimality criterion if the fitness value of  $A_1$  is favored over that of  $A_2$  is:

$$(III) \quad p(w_{11} - w_{21}) > (1 - p)(w_{22} - w_{12}) \text{ if and only if } q(U_{11} - U_{21}) > (1 - q)(U_{22} - U_{12})$$

Condition (III) provides a necessary and sufficient condition since it is algebraically equivalent to the optimality criterion. Condition (III) is thus weaker than condition (I) and condition (II). To satisfy condition (III), there is no need for the organism's subjective expected importances to match the objective expected importances, much less for the organism to believe the true and desire the good. What are the situations under which natural selection would favor organisms that obey condition (I) and not just (II) or (III)?

In situations where one action *dominates* another, it does not matter what probabilities the organism assigns to  $q$  and  $(1 - q)$ , as long as the organism's subjective utility *ordering* is the same as the objective fitness ordering. A dominant act is one in which an organism is better off performing a certain action *regardless* of what the state of the world is. For example, suppose the organism is trying to decide whether to eat the fruit in front of it. It does not know how nutritious the fruit is – in this environment, assume that half the fruit of this kind is very nutritious, and half is only mildly nutritious. Suppose the (objective) payoffs are as in Table III:

TABLE III. Objective Fitness Outcomes for Fruit Eating

	Very nutritious 0.5	Mildly nutritious 0.5	← objective frequencies
$A_1$ : Eat	10	5	
$A_2$ : Don't eat	0	0	

In this situation, as long as the organism assigns higher subjective utilities to eating ( $U_{11} > U_{21}$  and  $U_{12} > U_{22}$ ), it does not matter what subjective probabilities it assigns to the states. All such probability assignments will result in the organism's choosing to eat, which is the fitness enhancing decision. In this kind of problem, any deviation from the reliabilist ideal can be selected for. To satisfy condition (III), there is no need to satisfy condition (II), much less condition (I). Also, although a more general case exists in which it does not matter what subjective probabilities or utilities are assigned (because it does not matter which action is performed), no case exists in which it does not matter what you desire but does matter what you believe. The utilities always matter, if anything does.

Now suppose  $S_1$  is the proposition that the mushrooms in front of the organism are nutritious, and  $S_2$  the proposition that the mushrooms are poisonous, and that  $A_1$  is the action of eating, and  $A_2$  the action of not eating, the mushrooms. Suppose further that the frequency of poisonous mushrooms is 50%, and suppose that the fitness payoffs are as in Table IV (the poison is not fatal):

TABLE IV  
Objective Fitness and Frequencies for Mushroom Eating (case 1)

	Nutritious	Poisonous	
	0.5	0.5	← objective frequencies
$A_1$ : Eat	6	0	
$A_2$ : Don't eat	2	2	

In this situation, neither  $A_1$  nor  $A_2$  is a dominant act. If this case if  $w_{ij} = U_{ij}$  for all  $i, j = 1, 2$ , then the agent will do the right thing so long as  $q > 1/3$ . In this problem, the organism can assign any degree of belief to  $q$  provided it thinks that there is at least a 1 in 3 chance that the mushrooms are nutritious. This leaves room for considerable deviation from the reliabilist ideal, although the situation is more restrictive than the one described before in which there is a dominant act.

If, on the other hand, the organism believes the true (has subjective degrees of belief that match the objective frequencies,

that is,  $p = q$ ), then, in order to satisfy  $(\alpha)$ , it must assign subjective utilities such that  $(U_{11} - U_{21}) > (U_{22} - U_{12})$ . Surprisingly, this allows the organism to assign the highest value to eating poison and the lowest to eating nutritious food.

#### SYSTEMATIC IMPORTANCE

We have seen situations in which an organism can have fairly unreliable beliefs or desires. Now, we will consider some alternative kinds of cases. Suppose that the organism encounters a second problem in which the mushrooms are less nutritious than they were in the previous example, although the frequency of nutritious and poisonous mushrooms remains the same. We can imagine that because the organism migrates between a winter and summer habitat it encounters *both* types of cases from time to time.<sup>13</sup>

TABLE V  
Objective Fitness and Frequencies for Mushroom Eating (case 2)

	Nutritious	Poisonous	
	0.5	0.5	← objective frequencies
$A_1$ : Eat	3	0	
$A_2$ : Don't eat	2	2	

Here  $p = 0.5$ , but  $(w_{22} - w_{12})/(w_{11} - w_{21}) = 2$ . This means that not eating the mushrooms is the fitter act. Consequently, if the organism's subjective utilities are equal to the objective utilities, the organism's subjective probability  $q$  must be less than  $2/3$  if the organism is to do the right thing, because when  $q < 2/3$ ,  $w(A_2) > w(A_1)$ . In this case, the organism can assign any probability to the statement that the nutritious state obtains as long as it is less than  $2/3$ . An organism that assigned a higher value would conclude that  $w(A_1) > w(A_2)$ , and would choose the wrong action.

From the two mushroom-eating situations just described, we have both *upper* and *lower* bounds on the value of  $q$  :  $1/3 < q < 2/3$ . Notice that the degree to which a non-reliabilist can

deviate from perfect reliability (and still perform the actions that maximize fitness) shrinks if the relevant state-proposition is what I call *systematically important*. Systematically important propositions are defined by a special class of decision problems. More precisely, we can say that:

A proposition has greater *systematic importance* to the degree that:

- (i) Multiple decision problems exist that require a belief about whether the proposition is true in order for the organism to make the best decision, *and*
- (ii) Among the decision problems specified in (i), the objective expected importances of the relevant states that characterize the organism's decision problems vary from one problem to the next.
- (iii) Among the decision problems specified in (i) and (ii), which state-proposition has the higher objective expected importance changes as the decision problem changes.

Condition (iii) entails (ii) but not vice versa. The first two conditions are necessary for a minimal amount of systematic importance. If condition (iii) is met, there will be both upper and lower bounds on the degree of belief that can be assigned to the relevant proposition, if the organism is to maximize its fitness. If a proposition is relevant in multiple problems, but the expected importance is the same in every case, then the proposition will not be systematically important in the relevant sense.

In general, if we consider two different decision problems, there will be both a lower and upper bound on the organism's subjective probability assignments to the propositions  $S_1$  and  $S_2$  if a reversal occurs in which proposition has the higher objective expected importance. A reversal in which proposition has the higher (objective) expected importance entails a reversal in which action is best. In mushroom case 1, when  $A_1$  is favored, if the organism desires the good, the organism can have any degree of belief in the state with the higher expected importance above some minimum threshold. In general, whichever act is favored yields a minimum probability threshold on the probability assigned to the corresponding proposition with the higher expected importance. So when a reversal occurs in which proposition has the higher expected

importance, this provides *sandwiching* – an upper *and* a lower bound on the probability assignments the organism must make.

Analogous remarks apply when considering a condition that suffices for there to be upper and lower bounds on the value of  $U_{ij}$ . If we assume that the organism *believes the true*, then as long as the organism assigns the appropriate lower bound to the proposition with the higher objective expected importance, it will meet the optimality condition described above. Once again, if there is a reversal in which proposition has the higher (objective) expected importance, this will result in a corresponding bound from the other direction on the value of the proposition's importance.

We can further clarify the phenomenon of *sandwiching* by considering the question: of two payoff matrices that bound the subjective probability from the same direction (both from above or both from below), what condition suffices for one of them to provide the more restrictive bound? For example, consider the following two problems in Table VI:

TABLE VI

Problem (1)		Problem (2)				
	0.8	0.2				
$A_1$	3	0	$A_3$	5	0	← frequencies
$A_2$	1	4	$A_4$	2	5	

In the first problem, the objective expected importance of  $S_1$  is  $(0.8)(3 - 1) = 1.6$ ; the objective expected importance of  $S_2$  is  $(0.2)(4 - 0) = 0.8$ . In the second problem, the objective expected importance of  $S_1$  is  $(0.8)(5 - 2) = 2.4$ ; the objective expected importance of  $S_2$  is  $(0.2)(5 - 0) = 1$ . In each problem,  $S_1$  has a higher objective expected importance than  $S_2$ , so  $A_1$  is favored (objectively) over  $A_2$  and  $A_3$  is favored over  $A_4$ . Notice that the *ratio* of the objective expected importances of  $S_1$  to  $S_2$  is closer to one in the first problem than in the second ( $1.6/0.8 = 2$  versus  $2.4/1 = 2.4$ ). This means that the first problem provides a more restrictive bound on the value of the subjective probab-

ility that the organism must assign to the propositions. Precisely analogous reasoning applies to the assignment of subjective utilities that deviate from the objective values when the organism believes the true ( $p = q$ ). The closer the ratio of the objective expected importances is to one, the more restrictive the bound will be in any given direction.<sup>14</sup>

Finally, consider two distinct pairs of beliefs, each ‘sandwiched’ from above and below by its own pair of decision problems represented in Table VII. What condition suffices to make one pair have a tighter interval on the value assigned to  $q$ ?

TABLE VII

First decision pair					Second decision pair										
		$S_1$	$S_2$			$S_3$	$S_4$								
		0.3	0.7			0.9	0.1								
$A_1$	<table border="1"><tr><td>6</td><td>0</td></tr></table>	6	0	$A_3$	<table border="1"><tr><td>4</td><td>0</td></tr></table>	4	0	$A_5$	<table border="1"><tr><td>4</td><td>2</td></tr></table>	4	2	$A_7$	<table border="1"><tr><td>2</td><td>2</td></tr></table>	2	2
6	0														
4	0														
4	2														
2	2														
$A_2$	<table border="1"><tr><td>1</td><td>2</td></tr></table>	1	2	$A_4$	<table border="1"><tr><td>2</td><td>2</td></tr></table>	2	2	$A_6$	<table border="1"><tr><td>0</td><td>3</td></tr></table>	0	3	$A_8$	<table border="1"><tr><td>1</td><td>12</td></tr></table>	1	12
1	2														
2	2														
0	3														
1	12														

Once again, the crucial factor is the ratio of the expected importances. Each decision problem has its own ratio of expected importances, so a *pair* of decision problems can be characterized by a *ratio* of two ratios. For instance, the ratio of objective expected importances of  $S_1$  to  $S_2$  in the first problem is  $(0.3)(6 - 1)/(0.7)(2 - 0) = 1.5/1.4 \approx 1.07$ . The ratio of the objective expected importances of  $S_1$  to  $S_2$  in the second problem is equal to  $0.6/1.4 \approx 0.43$ . The ratio of objective expected importances of  $S_3$  to  $S_4$  in the third problem is  $3.6/0.1 = 36$ ; the ratio of objective expected importances of  $S_3$  to  $S_4$  in the fourth problem is  $0.9/1 = 0.9$ . The ratio of the first pair of decision problems (involving  $S_1$  and  $S_2$ ) is  $1.07/0.43$ , which is much closer to one than the ratio of the second pair of decision problems (involving  $S_3$  and  $S_4$ ), which is  $36/0.9$ . As a result, the first pair of decision problems provides a tighter interval (‘more sandwiching’) on the value of the subjective degree of belief  $q$ . Analogous remarks apply to the assignment of subjective utilities.

Notice that a large *systematic* expected importance is very different in kind from simply having a large expected importance.

Consider Pascal's wager;<sup>15</sup> traditionally, the payoffs are as in Table VIII:<sup>16</sup>

TABLE VIII

	( $S_1$ ) God exists $p \approx 0.001$ (very small)	( $S_2$ ) God does not exist $(1 - p) \approx 0.999$ (very large)
Believe in God	$+\infty$	small loss
Don't believe in God	$-\infty$	small gain

In Pascal's Wager, the expected importance of  $S_1$  is *huge*, whereas the expected importance of  $S_2$  is *tiny*. The proposition that *God exists*, however, is not likely to be *systematically* important. This will be true if the other decision problems in which the proposition that *God exists* is used are like Pascal's wager in that the importance of that proposition is much bigger than that of the other one. Because the difference in expected importances is so great, it doesn't matter what degree of belief one assigns to the proposition that God exists – as long as it meets some minimal threshold (in this case, as long as it is non-zero, assuming we can make sense of infinite utilities.)

The previous pair of mushroom examples (Tables IV and V) is one in which the best action changes because the value of one of the acts changes (the new environment is such that the food is not nutritious). This is one way in which a proposition (that the mushrooms are nutritious) can be systematically important. This same state is relevant to making a decision about two actions, but in different situations the utilities of the actions change, even though the probabilities of  $S_1$  and  $S_2$  do not.

Alternatively, the proposition (that the mushrooms are nutritious) might be systematically important because it is used in multiple decision problems where the *choices* change. So, we can imagine that instead of deciding between eating or not eating the mushrooms, the organism must decide between eating the mushrooms itself and giving them to its offspring. This scenario is represented in Table IX:

TABLE IX  
Objective Fitness and Frequencies for Mushrooms (case 3)

	Nutritious	Poisonous	
	0.5	0.5	← objective frequencies
$A_1$ : Eat	6	0	
$A_3$ : Give to offspring	8	-5	

If the offspring is less able to metabolize the poison than the parent, then it is more sensitive to poisonous mushrooms; the decision is correspondingly more critical.

If the payoffs and probabilities are as above, then  $A_1$  is the better action. If the organism desires the good, then the organism's subjective degree of belief must be:  $6q > 8q + -5(1 - q)$ , which reduces to:  $q < 5/7$  for the organism to decide to do  $A_1$ . Notice that since these decision problems may come in any particular order, there is still a sense in which propositions in which the lower bound simply increases from problem to problem are systematically important. When an organism desires the good, natural selection favors believing the true for propositions that are systematically important. Similarly, when an organism believes the true, natural selection favors desiring the good for systematically important propositions.

SIMULTANEOUS DEVIATIONS FROM DESIRING THE GOOD AND BELIEVING THE TRUE

Until now, I have considered cases in which the organism desires the good (or believes the true), and then examined what happens if the organism deviates from believing the true (desiring the good). So far, I have argued that the more systematically important a proposition is, the closer the organism must come to believing the true, if it desires the good, and vice versa.

The question remains: what about simultaneous deviation from both the true *and* the good? What about organisms that have belief formation policies that are skewed from believing the true and desire

formation policies that are skewed from desiring the good? I now will tackle this more general problem.

Recall that the objective criterion for  $A_1$  to be the best action is

$$(\alpha) \quad p/(1-p) > (w_{22} - w_{12})/(w_{11} - w_{21})$$

whereas the subjective decision procedure for  $A_1$  to be the best action is

$$(\beta) \quad q/(1-q) > (U_{22} - U_{12})/(U_{11} - U_{21}).$$

To simplify matters, I will abbreviate the ratios of importances as  $R_o$  (ratio of objective importances) =  $(w_{22} - w_{12})/(w_{11} - w_{21})$ , and  $R_s$  (ratio of subjective importances) =  $(U_{22} - U_{12})/(U_{11} - U_{21})$ .

Now, let  $q/(1-q) = p/(1-p) + X$  and let  $R_s = R_o + Y$ , where  $X, Y$  can be positive, negative, or zero. Thus  $X$  and  $Y$  represent the difference between what the organism believes and what is true ( $X$ ) and what the organism desires and what is good ( $Y$ ). In this case, the optimality criterion  $(\alpha)$  iff  $(\beta)$  becomes:  $p/(1-p) > R_o$  if and only if  $p/(1-p) + X > R_o + Y$ .

Suppose that action  $A_1$  is (objectively) favored over  $A_2$ , which means that:  $p/(1-p) > R_o$ . In this case, the organism will perform the fitter action if and only if  $p/(1-p) - R_o > Y - X$ . If, on the other hand,  $p/(1-p) < R_o$ , then the optimality criterion will be true if and only if  $p/(1-p) - R_o < Y - X$ . In the first case, since  $p/(1-p) > R_o$ ,  $p/(1-p) - R_o$  will always be positive. Consequently, this will provide an upper bound on the value of  $(Y - X)$ . In the second case,  $p/(1-p) - R_o$  will always be negative, and hence will provide a lower bound on the value of  $(Y - X)$ .

Once again, when there's the kind of systematicity in which there is a reversal in which state has the highest objective expected importance, there are both upper and lower bounds on the value of  $(Y - X)$ . In general, as the value of the objective expected importance of a state varies from decision problem to decision problem, there will be more 'sandwiching' on the value of  $(Y - X)$ . Consequently  $(Y - X)$  must get closer to zero the more systematically important the state.

Now, *one* way for an organism to meet the optimality criterion is for it to happen that both  $X$  and  $Y = 0$ . Recall that  $X$  and  $Y$  represent the subjective deviations from the objective probability and utility (fitness) ratios. As previously noted, the organism that believes the true and desires the good ( $X = Y = 0$ ) will meet the

optimality criterion. However, there are *infinitely* many other ways for the organism to ensure that it meets the optimality criterion. Let  $X$  be any number, as long as it is equal to  $Y$ . In that case the value of  $(Y - X) = 0$ . What reason, if any, is there for natural selection to favor believing the true and desiring the good over these other alternatives?

Here is a kind of ‘engineering’ argument. It is reasonable to suppose that in many (perhaps most) cases, the objective probabilities (frequencies) and the objective utilities (fitness values) vary independently in an organism’s environment. The organism needs one mechanism for detecting and tracking the probabilities, and another mechanism for estimating utilities. One way to get an organism to meet the optimality criterion *without* believing the true and desiring the good would be to first determine the true and the good, but then modify these by adding  $X$  to one estimate and  $Y$  to the other, where  $X = Y$ . Although this would meet the above criterion, we can argue against it on grounds of ‘internal’ fitness. Any organism that first estimates accurate values for  $p$  and  $w_{ij}$  and *then* calculates a modification is taking an ‘extra step.’ This extra step is likely to cost the organism in terms of time and (mental) energy, and hence is at a fitness disadvantage when compared to the more direct method.

On the other hand, perhaps the biased estimates can be made without this extra step. Consider an analogy: suppose I have a scale that weighs objects and the readings are always five pounds too heavy. The scale does *not* first get the true weight and then add five pounds to the reading. It has just one mechanism. In a similar fashion, a biased device for estimating  $p/(1 - p)$  or  $w_{ij}$  might not require an extra step.

This scale analogy, however, is imperfect. In the case of the scale, only one value is being estimated. In our situation, both the frequencies and the utilities must be estimated, and they must be ‘coordinated’ in just the right way. Suppose we assume that the organism must have *one* device for estimating all its probabilities and *one* device for estimating all its utilities. In this case, if we also assume that in many situations the objective probabilities (frequencies) and the objective utilities (fitness values) vary independently in an organism’s environment, and that different propositions have

different probabilities and that outcomes in different decision problems have different utilities, there will be no way for the organism to ‘coordinate’ in advance just the right biases in its estimations of the probabilities and utilities. The same set of beliefs may interact with a variety of distinct desires; unless all these different evaluations of different outcomes in different circumstances have exactly the same bias built into them, they won’t cancel out the bias built into the probability judgment. If this engineering argument is correct, then we can conclude that selection favors reliable belief and desire formation mechanisms regarding systematically important propositions.

MULTIPLE ACT ALTERNATIVES AT A TIME

So far we’ve considered sandwiching as something that arises *between* decision problems. However, it also can arise *within* a single problem when there are three or more act alternatives. *This can occur even when the propositions describing the relevant states are not systematically important in the sense described above.*

Consider Table X, in which an organism has three different choices:

TABLE X

	$S_1$	$S_2$	
objective frequencies →	$p$	$1 - p$	Objective expected values
Act alternatives $A_1$	$w_{11}$	$w_{12}$	$pw_{11} + (1 - p)w_{12}$
$A_2$	$w_{21}$	$w_{22}$	$pw_{21} + (1 - p)w_{22}$
$A_3$	$w_{31}$	$w_{32}$	$pw_{31} + (1 - p)w_{32}$

In order for  $A_1$  to be the best action, its objective expected value must be higher than both  $A_2$ ’s and  $A_3$ ’s. That is:

- (1)  $pw_{11} + (1 - p)w_{12} > pw_{21} + (1 - p)w_{22}$  and
- (2)  $pw_{11} + (1 - p)w_{12} > pw_{31} + (1 - p)w_{32}$

In order to evaluate this kind of situation, we need a new notion of importance. With two actions, the expected importance of a proposi-

tion is determined by the difference between the fitness outcomes of the two acts (given the state obtains) multiplied by the probability of the state. With more than two actions, we will use a notion of expected importance of a state *relative to* a given pair of actions:

The objective expected importance of  $S_j$  relative to  $A_i$  and  $A_k$  (where  $i \neq k$ ) =  $Pr(S_j)(w_{ij} - w_{kj})$ .<sup>17</sup>

For example, the objective expected importance of  $S_1$  relative to  $A_1$  and  $A_2$  is  $p(w_{11} - w_{21})$ . By defining a notion of expected importance in this way, we can apply the lessons from our study of systematic importance to the issue of *multiple act alternatives at a time* in a natural way.

First of all, if a given action's outcome is as good or better than every alternative action in both possible states of the world, we once again have a case of dominance; as long as the organism assigns subjective utilities that induce the same ordering as the objective fitness values, it does not matter what degrees of belief it assigns to the states.

On the other hand, if one action (*e.g.*,  $A_1$ ) is better when one state obtains and one or more other actions (*e.g.*,  $A_2$ ,  $A_3$ ) are better when the other state of the world obtains, then there is no dominance. In such cases, one action will have the best (or tied with the best) objective expected value. Suppose  $A_1$  has the highest expected value. Suppose further that  $A_1$  does better than either  $A_2$  or  $A_3$  in state 1 but worse than both  $A_2$  and  $A_3$  in state 2. In this case, both  $A_2$  and  $A_3$  provide a lower bound on the value the organism can assign to  $q$  (the probability of the state in which  $A_1$  does better). In order to determine which of  $A_2$  or  $A_3$  provides the more restrictive bound, we need to compare the ratios of the objective expected importances of  $S_1$  and  $S_2$  relative to  $A_1$  and  $A_2$  and relative to  $A_1$  and  $A_3$ . If  $[Pr(S_1)(w_{11} - w_{21})]/[Pr(S_2)(w_{22} - w_{12})]$  is closer to one than  $[Pr(S_1)(w_{11} - w_{31})]/[Pr(S_3)(w_{32} - w_{12})]$ , then it is  $A_2$  that provides a more restrictive lower bound on the value of  $q$ .

In order for the organism's subjective degree of belief to be bounded in both directions, however, it is necessary that the action with the highest objective expected value do worse than one alternative in one state and worse than a different alternative in another state. For instance, consider an infant vervet monkey that must decide what to do when it hears a warning call from an adult in

its troop. Since vervets are not born with the ability to recognize such calls, there is a time during which infants are not very good at distinguishing the distinct warning calls for leopards, eagles, and snakes.<sup>18</sup> Suppose the infant thinks the call is either an eagle ( $S_1$ ) or a leopard ( $S_2$ ) call, but it is not sure which (for the moment, assume that the mother makes the call only when there is an eagle or leopard nearby). In the past, when the infant heard warning calls, half were for eagles, and half were for leopards. Thus we have objective  $Pr(\text{Eagle Above}|\text{Call}) = Pr(\text{Leopard in the bushes}|\text{Call}) = 0.5$ .

Eagles hunt vervets effectively in the trees; a vervet's best defense against an eagle is to run into the bushes. On the other hand, since leopards hunt from the bushes, a vervet is better off in a tree when a leopard is nearby. So there is a high cost to either kind of mistake. Consequently, it may be best for the infant to simply run to its mother, rather than make a costly mistake. Suppose the fitness payoffs are as in Table XI:

TABLE XI. Objective Fitness Matrix

	$S_1$ (Eagle) $Pr(S_1 \text{Call}) = 0.5$	$S_2$ (Leopard) $Pr(S_2 \text{Call}) = 0.5$	Objective expected value of acts
$A_1$ : Run into bushes	10	-10	0
$A_2$ : Run into tree	-10	10	0
$A_3$ : Run to mother	1	1	1

So running to its mother is a kind of compromise for the infant – it doesn't afford the best protection against either kind of predator, but it avoids disaster.

Now consider a monkey whose decision procedure is such that it does not believe exactly what the evidence indicates about the relative frequencies of leopards and eagles in the past (given a certain kind of warning call). Suppose that its subjective  $Pr(\text{Eagle}|\text{Call}) = 0.6$  and its subjective  $Pr(\text{Leopard}|\text{Call}) = 0.4$ . Then the expected subjective utility of  $A_1$  would be 2 whereas the expected subjective utility of  $A_2$  would be -2 but the value of  $A_3$  remains 1. So this vervet would think that running into the bushes is the best action. Notice that in the long run, this would

be worse for the vervet than running to its mother. Similarly, if the vervet's beliefs are skewed in the other direction (*e.g.*, subjective  $Pr(\text{Eagle}|\text{Call}) = 0.4$  and  $Pr(\text{Leopard}|\text{Call}) = 0.6$ ), then it would choose  $A_2$ , which also would be to its disadvantage. In general, the possible deviation from the reliabilist ideal is  $\pm 0.05$  in order for the vervet's decision procedure to lead it to do the best action.<sup>19</sup> Here we see an example in which there are significant constraints on the extent to which an organism can deviate from the ideal of perfect reliability even if the proposition is not systematically important.<sup>20</sup>

In general, these results are similar to the ones we saw in connection with the issue of systematic importance. Suppose that the best act overall is not the best in either state by itself (as in the infant vervet example above). Consider first the ratio of objective expected importances of  $S_1 : S_2$  relative to the best act (in this case  $A_3$ ) and relative to the action that does best in  $S_1$ . Call this first ratio  $K$ . Compare that to the ratio of the objective expected importances of  $S_1 : S_2$  relative to the best act and relative to the act that does the best in  $S_2$ . Call this second ratio  $L$ . The closer the ratio of  $K : L$  is to one, the more sandwiching there will be on the degree of belief that the organism can assign to  $q$  if it desires the good.

What if the vervet changes its desire forming policies or its decision rule? Suppose it has a desire forming strategy such that it assigns values to the utilities so that they *do not* match the objective fitness values. Assume the objective values for belief and desire are as before, and let subjective  $Pr(\text{Eagle}|\text{Call}) = 0.6$  and  $Pr(\text{Leopard}|\text{Call}) = 0.4$ .

The subjective utility matrix is represented in Table XII:

TABLE XII

	Eagle above $Pr(\text{Eagle} \text{Call}) = 0.6$	Leopard in bushes $Pr(\text{Leopard} \text{Call}) = 0.4$
$A_1$ : Run into bushes	10	-20
$A_2$ : Run into trees	-20	10
$A_3$ : Run to mother	1	1

If the monkey overestimates the danger of a mistake, then a greater range of degrees of belief will lead it to make the correct decision

(A<sub>3</sub>).<sup>21</sup> In general, the results for an organism that believes the true but does not desire the good are precisely analogous to those of the organism that desires the good but does not believe the true. Similarly, the engineering argument sketched in the last section also applies to cases in which there are multiple act alternatives at a time.

### CONCLUSIONS

Previous discussions of the evolution of rationality have been too sketchy and abstract.

Philosophers such as Godfrey-Smith and Sober began to develop detailed models, but these models are oversimplified in important ways. We need models that represent the organism's subjective beliefs (probabilities) and desires (utilities) as well as the objective values they reflect to greater or lesser degrees. Models of phenotypic plasticity, though interesting, are not sufficient to model organisms with minds that manipulate representations. Furthermore, previous models of the evolution of rationality have failed to account for the *systematic* quality of our beliefs and desires. I argued here that systematicity plays a crucial role in understanding the adaptive value of various belief and desire formation policies. In particular, we should expect that when an organism desires the good, natural selection favors believing the true for propositions that are systematically important. Similarly, when an organism believes the true, natural selection favors desiring the good for systematically important propositions.

In arguing against the selective value of true beliefs, Stephen Stich writes that one might try to argue that true belief does "a better job *in general, or in the long run*. Well perhaps it does. But to show this requires an argument, and as far as I know, no one has any inkling of how that argument might go" (1990, p. 124).<sup>22</sup>

I hope that the patient reader now has more than a mere inkling of how that argument might go. Up to now, philosophers have tended to answer the question of whether natural selection favors true beliefs in a simple *yes-or-no* fashion. I have argued here that the answer to this question is *neither yes nor no*. One is better off asking the question, for what kinds of propositions and under what kinds of conditions will natural selection favor true beliefs?

I have presented a new model of the co-evolution of belief and desire formation policies. The model developed here indicates that natural selection will not always favor true beliefs. However, non-reliable belief formation policies will get the organism into trouble in two basic kinds of cases: ones in which the relevant propositions are systematically important, and cases in which there are several possible actions that the organism must consider, each of which has important consequences. Although not all problems are like this, many (perhaps most), are. I conclude that my analysis has a great deal of generality.

At the same time, it is important not to misinterpret the model given here. The model provides an analysis of the conditions under which reliable belief and desire formation rules will be functionally useful. This leaves open the extent to which that functional utility was important in the overall evolution of the trait. As Orzack and Sober (1994) point out, to test adaptive hypotheses one can develop a model that assumes that natural selection is the primary force in the evolution of the trait in question. If such a model fails to make accurate predictions, this may indicate that natural selection is not the only important force relevant to the evolution of the trait. This method of testing adaptationism allays some of the worries that Gould and Lewontin (1978) have about the falsifiability of adaptationist hypotheses.

Philosophers such as Cherniak (1986) have emphasized the importance of various cognitive limitations due to problems of combinatorial explosion and limited memory capacity.<sup>23</sup> It is of course not feasible for an organism to track all features of its environment in such a way that all its beliefs and desires are reliable indicators of the true and the good. However, some propositions have such a high degree of systematic importance – they occur in such a wide variety of differently structured decision problems – that we should expect organisms to be designed by natural selection to have reasonably accurate beliefs about them. Other propositions are insignificant – and it is here that we should expect to find pockets of irrationality.

Also, I have not attempted to incorporate learning into my model; the model is about the sense in which we should expect organisms to have reliable ways of finding out about what is true and good, but I

have not given details as to how the organism might do that. It would be useful to consider a model that considered learning explicitly, though I suspect that my results about systematic importance would hold in such a model.

Finally, I should mention that some philosophers<sup>24</sup> have suggested that in the extreme, global counter-reliabilism and globally desiring the bad seems to be behaviorally equivalent to the reliabilist alternative. Indeed, one might think that for each combination of beliefs and desires, there is a kind of mirror image where one has beliefs that deviate from the true and desires that deviate from the good in such a way that these pairings are behaviorally equivalent. Others have argued that it is conceptually impossible for an organism to have mostly (or all) false beliefs.<sup>25</sup> I have argued that although a certain kind of *act* claim may be true – *viz.*, for any given problem one can find a combination of beliefs and desires which are biased, but will get the organism to do the right thing – there is reason to expect that *rules* that lead to more accurate beliefs and desires will be selected for over rules that lead to less accurate ones. Even within the limited range of considerations we have examined here, we have seen that there is no simple monolithic answer to the question of whether natural selection favors reliable belief and desire formation policies. And we have seen, in addition, what the structural issues are that are relevant to answering our original question in specific contexts.<sup>26</sup>

#### NOTES

<sup>1</sup> Quine (1969), Ruse (1986), Sober (1980), Dennett (1978, 1987), Fodor (1981), Millikan (1984, 1987), Goldman (1986), Lycan (1988), Sorensen (1992), Nozick (1993) and Papineau (1987, 1993) all defend various claims linking evolution with true belief or rationality.

<sup>2</sup> See, for example, Stich (1985, 1990), Godfrey-Smith (1991, 1996), Fodor (1987, 1998), Sober (1994), Stein (1996) and Plantinga (1993). Stich (1985, 1990) and Stein (1996) are primarily concerned to argue that many of the philosophers' arguments in the first camp are unsuccessful; they do not offer (or claim to offer) detailed models that show that belief formation mechanisms that lead to irrational or false beliefs were in fact selected for. See also Feldman (1988) and Fitelson and Sober (1998) for useful discussion.

<sup>3</sup> Besides Stich (1990), Sober (1994) and Godfrey-Smith (1991, 1996) develop

models that rely, in one way or another, on the better-safe-than-sorry argument. I discuss their accounts in the next section.

<sup>4</sup> Similar models of phenotypic plasticity can also be found in Moran (1992) and Stephens (1989, 1991). For an introduction to signal detection theory, see Coombs, Dawes, and Tversky (1970).

<sup>5</sup> Kitcher (1993) uses multiple decision matrices in order to model the evolution of altruism in humans. In his model, one matrix represents the payoffs that the agent assigns based on his or her desires, and another matrix determines what will count as reproductive success. See also Sterelny (1999) for an interesting discussion about the evolution of agency, as opposed to mere plasticity.

<sup>6</sup> One might think that natural selection would favor organisms that have ultimate desires only for reproductive success. Three kinds of considerations suggest otherwise. First, organisms have limited minds and resources. As Sober and Wilson (1998, pp. 309–310) point out, having desires about reproductive success requires a sophisticated mind that can understand and represent the concept of reproduction; it is likely that some ultimate desires, even in organisms like us, evolved prior to the development of the ability to represent such concepts. Second, even if an organism has the relevant concept of reproduction, the desire for reproductive success is evolutionarily good only in so far as it actually leads to an increase in fitness. Just as a hedonist might get the most pleasure by not always trying to pursue pleasure directly, an organism might maximize fitness by not directly pursuing the maximization of fitness. Finally, whether or not desiring the good is the best policy depends on the organism's belief formation policies and decision rules.

<sup>7</sup> The term 'good' obviously has moral connotations as well. I do not mean to suggest that one should, from the moral point of view, have desires for ends that are fitness enhancing. The model developed here is about the extent to which an organism should have goals for what is fitness enhancing, from the 'point of view', so to speak, of natural selection.

<sup>8</sup> Here and throughout I use *state-propositions*, which are propositions that describe the possible states of the world in the decision problems that the organism confronts. If the reader has qualms about propositions, substitute 'statement' for 'proposition'. There is an interesting issue about the exact nature of the kind of entity that probabilities attach to (*e.g.*, propositions or events? See Eells (1982) and Jeffrey (1983) for details). Nothing in my paper hinges on the outcome of this debate; I use proposition-talk purely for expository convenience.

<sup>9</sup> I assume that the utilities of different actions are not exactly identical (*e.g.*, that  $(w_{11} - w_{21}) \neq 0$ ) and that  $p \neq 1$ , so that the optimality criterion is well defined.

<sup>10</sup> Sober (1994) and Godfrey-Smith (1996) use the term 'importance' in a similar manner. However, since on their models, there is a close connection between belief and action, they say that the importance of a proposition is determined by how much difference it makes whether the organism *believes* a given proposition is true, if it is true.

<sup>11</sup> It is arbitrary which state-proposition is described as  $S_1$  and which action is described as  $A_1$ ; however, what is crucial for my definition of importance is

that the order in which one subtracts the fitness outcomes if a given proposition is true must be *reversed* when one considers the importance of the other state-proposition. One reason for this is that in many interesting cases, one action will do better if one proposition is true whereas another action is better if a different proposition is true.

<sup>12</sup> One might also interpret the content of the organism's belief as "the frequency of  $S_1$  is 0.1." That is, the content might *include* a reference to the frequency. Although this approach is closer to a literal interpretation of the claim that  $p = q$ , I interpret the content of the organism's belief so that it does not include reference to probability. This is because such an organism would need to have a mind that is sophisticated enough to have the notion of probability. I assume that this ability occurs later in the evolutionary process. I am indebted to Peter Godfrey-Smith and Wayne Riggs for pressing me to clarify this.

<sup>13</sup> One might think that the organism could develop two different beliefs. One about the frequency of poisonous mushrooms in its summer habitat, and one about the frequency of poisonous mushrooms in its winter habitat. If so, then the unqualified belief about the frequency of poisonous mushrooms would cease to be systematic; since it would be replaced by two different beliefs. It is likely, however, that there are at least *some* decision problems where the organism must rely on the *same* belief. An organism will not always have the cognitive resources to form a new kind of belief for every possible situation. The mushroom example described here is meant to be one of those situations.

<sup>14</sup> There is an important exception to the idea that as the expected importance of  $S_1$  approaches that of  $S_2$ , the bound on  $q$  or  $U$  becomes more restrictive. In the limit, that is, if the expected importance of  $S_1$  is *exactly equal* to that of  $S_2$ , then *there is no bound at all*, since it does not matter which action is performed.

<sup>15</sup> This is not, of course, part of our evolutionary model. Presumably beliefs about God come along much later in the evolutionary scheme of things.

<sup>16</sup> The payoffs for  $S_1$  do not have to involve infinity – they can be 'very big' positive and 'very small' negative numbers, and much of the argument goes through. However, Pascal's wager has several problems, see Mougin and Sober (1994) for details.

<sup>17</sup> In order to compare the expected importances of any given pair of state-propositions, one must reverse the order in which one subtracts the relevant fitness values. So if the expected importance of  $S_1$  relative to  $A_1$  and  $A_2$  is determined by subtracting  $w(A_1) - w(A_2)$  if  $S_1$  is true, then the expected importance of  $S_2$  relative to  $A_1$  and  $A_2$  is determined by subtracting  $w(A_2) - w(A_1)$  if  $S_2$  is true.

<sup>18</sup> See Cheney and Seyfarth (1990).

<sup>19</sup> If we assume that the organism desires the good (its utilities match the objective fitness values), then for  $EU(A_3) > EU(A_1)$ ,  $q + (1 - q) > 10q + -10(1 - q)$  which means  $q < 11/20$ . For  $EU(A_3) > EU(A_2)$ , then  $q + (1 - q) > -10q + 10(1 - q)$  which means  $q > 9/20$ . Since  $p = 0.5$ , this means that if the organism's subjective utilities match the objective fitness values, the organism's subjective degree of belief can only deviate from the true value by  $\pm 1/20$ .

<sup>20</sup> Of course, it is likely that propositions about the probability that various predators are around are also *systematically* important.

<sup>21</sup> Examples in which the organism must decide between three or more actions can easily be multiplied. In foraging, an organism must decide how long to stay in a particular patch before moving on; there is a continuum of possible act alternatives in such a situation.

<sup>22</sup> Stich (1990) elaborates on a couple of distinct worries in his book. One is that no one has successfully shown that true beliefs do better than false beliefs on average, or in the long run. Another is that no one has successfully argued that true beliefs do better than TRUE\* or TRUE\*\* beliefs, where the starred truth functions are ones that map mental states not to their truth conditions but in various different ways that make the beliefs false. I have tried to address the former worry in this paper; the latter worry, however, rests on Stich's (1990) general skepticism about standard notions of reference (see chapter 5). I have not deal with these more general skeptical issues here, although it is worth noting that Stich (1996) has apparently since given up some of these more extreme skeptical views.

<sup>23</sup> The psychological literature on human reasoning suggests that humans use various kinds of heuristics that are accurate in a wide range of circumstances but fail in others. For example, see Kahneman, Slovic, and Tversky (1982), Gigerenzer (1991), Evans and Over (1996) and Stanovich (1999).

<sup>24</sup> See Plantinga (1993).

<sup>25</sup> Daniel Dennett (1987) and Donald Davidson (1974)) have reach argued, for various reasons, that it is conceptually impossible for organisms to have mostly false beliefs. Similarly, defenders of evolutionary theories of content such as Papineau (1987, 1993) sometimes speak as if it is a conceptual truth that reliable belief formation policies will be selected for. Even if the teleosemantics project is vindicated, and it can be shown that evolution determines semantic content in some way, it remains to ask about the evolutionary history of various belief formation policies, as I did here. Given that evolution has determined the content of mental states, why not then have the policy of believing the content which is false, or which otherwise deviates from what the evidence indicates?

<sup>26</sup> I thank Robin Andreasen, André Ariew, Thomas Bontly, Ellery Eells, Berent Enç, Peter Godfrey-Smith, Dan Hausman, Deborah Kallmann, Kelli Kiyomi Kadokawa, David Lorvick, Laura Sizer, Kim Sterelny and especially Elliott Sober for helpful comments on previous drafts. Earlier versions of this paper were presented at the International Society for the History, Philosophy and Social Studies of Biology in the summer of 1999, as well as to audiences at Simon Fraser University, the University of Wisconsin-Madison and the University of Oklahoma. Thanks to all the participants for useful discussion.

## REFERENCES

Cheney, D.L. and Seyfarth, R. (1990): *How Monkeys See the World*, Chicago, University of Chicago Press.

- Cherniak, C. (1986): *Minimal Rationality*, Cambridge, MA, MIT Press.
- Coombs, C.H., Dawes, R.M. and Tversky, A. (1970): *Mathematical Psychology: An Elementary Introduction*, Englewood Cliffs, Prentice-Hall.
- Davidson, D. (1974): "On the Very Idea of a Conceptual Scheme," *Proceedings and Addresses of the American Philosophical Association* 47, pp. 5–20.
- Dennett, D. (1978): *Brainstorms*, Cambridge, MA, MIT Press.
- Dennett, D. (1987): *The Intentional Stance*, Cambridge, MA, MIT Press.
- Eells, E. (1982): *Rational Decision and Causality*, Cambridge, Cambridge University Press.
- Evans, J. and Over, D. (1996): *Rationality and Reasoning*, Psychology Press.
- Feldman, R. (1988): "Rationality, Reliability, and Natural Selection," *Philosophy of Science* 55, pp. 218–227.
- Fitelson, B. and Sober, E. (1998): "Plantinga's Probability Arguments Against Evolutionary Naturalism," *Pacific Philosophical Quarterly* 79, pp. 115–129.
- Fodor, J.A. (1981): "Three Cheers for Propositional Attitudes," in *Representations*, Cambridge, MA, MIT Press.
- Fodor, J.A. (1987): *Psychosemantics*, Cambridge, MA, MIT Press.
- Fodor, J.A. (1999): "Is Science Biologically Possible?" in *In Critical Condition*, Cambridge, MA, MIT Press.
- Gigerenzer, G. (1991): "How to Make Cognitive Illusions Disappear: Beyond 'Heuristics and Biases'," *European Review of Social Psychology* 2, pp. 83–115.
- Goldman, A. (1986): *Epistemology and Cognition*, Cambridge, MA, MIT Press.
- Gould, S. and Lewontin, R. (1978): "The Spandrels of San Marcos and the Panglossian Paradigm," in E. Sober (1994) (ed.), *Conceptual Issues in Evolutionary Biology*, Cambridge MA, MIT Press.
- Godfrey-Smith, P. (1991): "Signal, Detection, Action," *Journal of Philosophy* 88, pp. 709–722.
- Godfrey-Smith, P. (1996): *Complexity and the Function of Mind in Nature*, Cambridge, Cambridge University Press.
- Jeffery, R. (1983): *The Logic of Decision*, second edition, The University of Chicago Press.
- Kahneman, D., P. Slovic, and A. Tversky (1982): *Judgment under Uncertainty: Heuristics and Biases*, Cambridge, Cambridge University Press.
- Kitcher, P. (1993): "The Evolution of Human Altruism," *Journal of Philosophy* 90, pp. 497–516.
- Lycan, W. (1988): *Judgement and Justification*, Cambridge, Cambridge University Press.
- Millikan, R. (1984): "Naturalist Reflections on Knowledge," *Pacific Philosophical Quarterly* 65, pp. 315–334.
- Millikan, R. (1987): *Language, Thought and Other Biological Categories*, Cambridge MA, MIT Press.
- Moran, N. (1992): "The Evolutionary Maintenance of Alternative Phenotypes," *American Naturalist* 139, pp. 971–989.
- Mougin, G. and Sober, E. (1994): "Betting Against Pascal's Wager," *Nous* 28, pp. 382–395.

- Nozick, R. (1993): *The Nature of Rationality*, Princeton, N.J., Princeton University Press.
- Orzack, S. and Sober, E. (1994): "Optimality Models and the Long-run test of Adaptationism," *American Naturalist* 143, pp. 361–380.
- Papineau, D. (1987): *Reality and Representation*, Oxford, Basil Blackwell.
- Papineau, D. (1993): *Philosophical Naturalism*, Blackwell Publishers.
- Plantinga, A. (1993): *Warrant and Proper Function*, Oxford, Oxford University Press.
- Quine, W. (1969): *Ontological Relativity and Other Essays*, New York, Columbia University Press.
- Resnik, M. (1987): *Choices: An Introduction to Decision Theory*, Minneapolis, University of Minnesota Press.
- Ruse, M. (1986): *Taking Darwin Seriously*, Oxford, Blackwell.
- Sober, E. (1980): "The Evolution of Rationality," *Synthese* 46, pp. 95–120.
- Sober, E. (1993): *Philosophy of Biology*, Boulder, CO., Westview Press.
- Sober, E. (1994): "The Adaptive Advantage of Learning and A Priori Prejudice," in *From a Biological Point of View*, Cambridge University Press, pp. 50–69.
- Sober, E. and Wilson, D.S. (1998): *Unto Others: the Evolution and Psychology of Unselfish Behavior*, Cambridge, MA, Harvard University Press.
- Sorensen, R. (1992): *Thought Experiments*, New York, Oxford University Press.
- Stanovich, K. (1999): *Who is Rational? Studies of Individual Differences in Reasoning*, Lawrence Erlbaum Associates, Inc., Publishers.
- Sterelny, K. (1999): "The Evolution of Agency," in Valerie Gray Hardcastle (ed.), *Where Biology meets Psychology: Philosophical Essays*, Cambridge, MA, MIT Press.
- Stein, E. (1996): *Without Good Reason*, Oxford University Press.
- Stephens, D. (1989): "Variance and the Value of Information," *American Naturalist* 134, pp. 128–140.
- Stephens, D. (1991): "Change, Regularity, and Value in the Evolution of Animal Learning", *Behavioral Ecology* 2, pp. 77–89.
- Stich, S. (1985): "Could Man be an irrational Animal?" *Synthese* 64(1).
- Stich, S. (1990): *The Fragmentation of Reason*, Cambridge, MA, MIT Press.
- Stich, S. (1996): *Deconstructing the Mind*, New York, Oxford University Press.

*Department of Philosophy*  
*University of Oklahoma*  
*455 W. Lindsey, Room 605*  
*Norman, OK 73019*  
*USA*  
*E-mail: cstephens@ou.edu*

