ELSEVIER

Evolution and Human Behavior 22 (2001) 11–29

Evolution and Human Behavior

The psychology of social chess and the evolution of attribution mechanisms: explaining the fundamental attribution error

Paul W. Andrews

Department of Biology, Castetter Hall, University of New Mexico, Albuquerque, NM 87131-1091, USA

Received 21 October 1999; accepted 2 August 2000

Abstract

Theory of mind is the field devoted to understanding how organisms discern the mental states of others. Because mental states are not directly observable, they can only be inferred from observable features of the actor (such as behavior) and the situational context that the actor is in. Social psychologists, who study theory of mind processes under the rubric of attribution research, have shown that people often make a logical error of inference: The ''fundamental attribution error'' (FAE) is the tendency to assume that an actor's behavior and mental state correspond to a degree that is logically unwarranted by the situation. The social environment in which theory of mind capacities evolved may have influenced attributional processing in ways that could explain the error. In particular, the error could be caused by a psyche that is designed (1) to consider only those noncorresponding mental states (such as deception) that could have fitness consequences to the mind reader; (2) to bias inferences in a way that reduces the costs of erroneous inferences; or (3) to bias inferences in a way that yields reputational benefits. The existing literature is reviewed in light of these hypotheses.

Keywords: Altruism; Bystander intervention effect; Correspondence bias; Deception; Fundamental attribution error; Handicap; Honesty; Mind reading; Signal; Theory of mind

1. Introduction

In a classic article, Humphrey (1976) suggested that many of the unique aspects of human intelligence were the products of a co-evolutionary arms race for predicting and manipulating

E-mail address: pandrews@unm.edu (P.W. Andrews).

1090-5138/01/$ – see front matter
PII: S1090-5138(00)00059-3

the behavior of other humans. He likened the social dynamic to a game of chess because outcomes depend on the tactics that both players choose and how they respond to each other. In the game of social chess, a successful player must be able to anticipate the possible moves of his opponent and mentally play out alternative scenarios to determine possible outcomes, while simultaneously taking into account the fact that his opponent is doing the same thing. Since behavioral choices are the output of these mental processes, the advantages accruing to those who can successfully predict behavior strongly favor those who can discern the mental states of their opponents (Baron-Cohen, 1995; Whiten & Byrne, 1988).

Research into the cognitive skills for mind reading is referred to as "theory of mind" (Premack & Woodruff, 1978).[1] Evolutionary psychologists understand relatively little about the cognitive processes that people use to make inferences about the mental states of others (but see Baron-Cohen, 1995; Cosmides & Tooby, 1992; Haselton & Buss, 2000). Of course, if mind reading is one of the most difficult cognitive tasks facing people, then figuring out how they do the task is likely to be difficult as well. Yet, given that deception is an important part of the dynamic of social chess, it is surprising that the body of theory developed by behavioral ecologists on honest and deceptive communication has not been incorporated into theory of mind research.

In contrast, social psychologists have studied the processes by which people attribute mental states to others for over half a century under the domain of attribution research (Gilbert & Malone, 1995). Unlike theory of mind, attribution theory has been extensively developed (for a review, see Schneider, Hastorf, & Ellsworth, 1979). A key concept in attribution theory is that of *correspondence* (Jones & Davis, 1965). An actor's behavior and mental state correspond to each other when they can be described by the same, or similar, words (Schneider et al., 1979). An observer makes the *corresponding inference* if he or she infers that the actor's behavior and mental state correspond to each other. For example, if I give a speech on the virtues of Fidel Castro, you will have made the corresponding inference if you infer that I support Castro.

Since mental states are not directly observable, people can only make attributions about an actor's mental state based on observable features of the actor (such as behavior) and the situational context that the actor is in. All attribution theories posit that people use both behavioral and situational information to make inferences about an actor's mental state (Gilbert & Malone, 1995; Schneider et al., 1979). One theory predicts that when the situational context suggests no other reason for the actor's behavior, people will tend to make the corresponding inference. However, if the situational context suggests several possible mental states that could account for the actor's behavior, people should be less likely to make the corresponding inference (Jones & Davis, 1965). Suppose that you were told that I gave the pro-Castro speech as part of a class debate, and that the professor had assigned me the pro-Castro stance. The situational context indicates that I could have one of two possible mental states: (1) I really support Castro (the corresponding mental state), or (2) I do not support Castro but I was forced to take the pro-Castro stance by my professor (a

[1] The term "theory of mind" derives from the assumption that in order to read the mind of others, an individual must have a theory about how the minds of others work (Premack & Woodruff, 1978).

noncorresponding mental state). The prediction is that you will be less likely to make the corresponding inference when given the no-choice situational information than if you are not given that information.

The results from over three decades of attribution research have forced a surprising conclusion. People often bias their attributions towards correspondence more than seems logically warranted (for reviews, see Gilbert & Malone, 1995; Jones, 1979; Nisbett & Ross, 1980; Ross, 1977). When people infer that the actor's behavior and mental state correspond to a degree that is logically unwarranted by the situation, they are said to have made the *fundamental attribution error* (FAE; Choi, Nisbett, & Norenzayan, 1999; Gilbert & Malone, 1995; Jones, 1979). Because it appears as if people generalize from the actor's behavior and ignore the situational context in which behavior occurs, the FAE is often described as a tendency to underattribute the cause of behavior to situations and overattribute it to dispositional traits (Ross, 1977). *Dispositional traits* are relatively stable, internal states (e.g., ability or lack thereof, personality, etc.). Mental states (such as attitudes, beliefs, and intentions) can be the product of underlying dispositions, though they may also be much more ephemeral (Jones & Davis, 1965). As we will see, there are potentially good evolutionary reasons for people to make the FAE even if they do not tend to overattribute the causes of behavior to dispositional traits.

The existence of phenotypic imperfections has historically been one of the most powerful evidence of evolution (Cronin, 1991; Williams, 1992). In this tradition, the existence of errors of logical reasoning has provided some of the most fertile ground for the application of evolutionary principles for understanding the design of the human psyche (e.g., Cosmides, 1989; Cosmides & Tooby, 1992, 1996; Gigerenzer, 1998; Gigerenzer & Hoffrage, 1995; Gigerenzer & Hug, 1992; Haselton & Buss, 2000). Detailed analyses of the precise nature of logical errors often yield great insights into the specific selective forces that have shaped human cognition.

In this paper, I argue that the game of social chess influenced the evolution of attribution mechanisms in several ways that could explain the FAE. I first review theoretical and empirical work within social psychology that led to the discovery of the FAE. Next, I discuss how the social chess dynamic can explain why attributional processes may be necessary for predicting the behaviors of others. In the remaining sections of the paper, I discuss the ways in which the social chess dynamic could have influenced the evolution of attribution mechanisms and how they may shed light on the FAE.

2. Prior theoretical and empirical research on attribution processes within social psychology

In the attribution literature, the key players are actors and observers. *Actors* are people who produce behavior. *Observers* are people who perceive behavior and attribute mental states to the actor. Every behavior is the result of the actor's internal state interacting with external stimuli. Because mental states are not directly observable, observers can only make inferences about the mental state of an actor based on his or her behavior and the situational context in which it occurs. Attribution theorists attempt to understand how observers make attributions

about a specific internal cause of behavior when they are given information about the situation in which the behavior occurs.

2.1. The principle of noncommon effects

Jones and Davis (1965) described a principle by which observers could make inferences about the mental state of an actor when information about the actor and the situation is incomplete. They assumed that the actor usually intends the consequences of his actions. The *principle of noncommon effects* allows an observer to make use of the fact that an actor, placed in a given situation, may have multiple behaviors he or she could choose from, each of which may have different consequences. According to the principle, an internal state is inferred from the actor's chosen behavior, the alternative behaviors that were not chosen, and the effects that were unique to the chosen behavior. For example, if a woman has a choice of three dresses that are identical except for color, and she decides to buy the green one, the principle of noncommon effects would suggest that the most likely reason for her decision is that she prefers green to the other two colors. A necessary corollary to the principle of noncommon effects is the *discounting principle* (Kelley, 1972). When multiple mental states are equally possible, the discounting principle states that an observer cannot attribute any single mental state to the actor. For instance, if the green dress also had different buttons than the other two, then either button-type or color could have been the predominant factor influencing the woman's decision, and the existing information will not allow us to determine which was likely to be more important.

To test these predictions, Jones and Harris (1967) gave subjects either a pro-Castro essay or an anti-Castro essay purportedly written by a student. In one set of variants (the *choice variants*), subjects were merely told that the student wrote the essay as part of a class. In the other variants (the *no-choice variants*), subjects were told that the professor assigned the student either the pro-Castro or anti-Castro stance. In the choice variants, the principle of noncommon effects suggests that the essay's stance corresponds to the writer's mental state (e.g., the writer supports Castro when he wrote the pro-Castro essay). In the no-choice variants, the discounting principle predicted that the type of essay would not affect people's attributions. In other words, if the student really had no choice in writing the essay, then people should be just as likely to attribute a pro-Castro attitude regardless of whether the student wrote a pro-Castro or an anti-Castro essay. Instead, the evidence suggested logically inadequate discounting. People were more likely to attribute a pro-Castro attitude when the student was assigned the pro-Castro essay than when the student was assigned the anti-Castro essay.

Subsequent experiments have bolstered the claim of erroneous attribution in the no-choice conditions. Compliance with the assignment does not appear to be very diagnostic of the writer's actual mental state. When asked to actually write an essay on a designated position on any of several topics (e.g., Castro, abortion, federal provision of free medical care), no subject has refused the assignment (Snyder & Jones, 1974, Experiments 1, 2, 4, and 5). Moreover, the strength and quality of essays written by subjects is unrelated to their true attitudes (Snyder & Jones, 1974, Experiments 4 and 5). Other experiments indicate that the strength and quality of essays written by subjects have little influence on the tendency of

observers to make corresponding attributions (Snyder & Jones, 1974, Experiments 4 and 5). Variation in the degree to which the essay writer's lack of choice is made salient also has little influence on observer attributions (Gilbert & Jones, 1986; Snyder & Jones, 1974, Experiments 1 and 5). Finally, it is very telling that attributions tend to conform more to predictions when people have an incentive to make logically correct inferences (i.e., when money is at stake; Vonk, 1999). Since attribution experiments almost never give people a real incentive to make logically correct attributions, this suggests that deviation from the predicted pattern is indeed erroneous.

Nevertheless, several exceptions to the FAE have been found. People from more interdependent societies are less likely to make the error than people from more independent societies (Choi et al., 1999), and depressives are less likely to make the FAE than normals (Yost & Weary, 1996). As noted above, having a personal stake in accurately assessing the actor's mental state also appears to facilitate accurate attribution (Vonk, 1999). Finally, people tend to make more accurate attributions when the actor has an apparent incentive to deceive others about his true mental state (Fein, 1996; Fein, Hilton, & Miller, 1990; Hilton, Fein, & Miller, 1993; Vonk, 1998).

3. The function of mental states

Attributions are mental representations of the mental states of other people. To understand why people make attributions about each other requires an understanding of why they experience mental states in the first place. While mental states are generally thought to be involved in causing behavior, it is not clear why they are useful for this purpose. Would it not be possible for selection to design people so that, like biological automatons, they adaptively responded to stimuli without experiencing mental states (Dennett, 1991)? The study of consciousness has a long history in philosophy, psychology, and neuroscience (Dennett, 1991; Hobson, 1999; Humphrey, 1992), and a review is beyond the scope of this paper. Rather, I discuss how the insights of Humphrey (1976) generate a working hypothesis that seems to underlie most theory of mind research and which forms the basis for the rest of the paper.

All organisms were faced with problems in their evolutionary past in which they had to choose a response. For some problems, the type of response that maximized fitness depended heavily on subtle variations in context; for other problems, the best response was much less dependent upon subtle variations. However, there is no way to design a nervous system that simply calculates the fitness consequences of different actions to determine the option that maximizes reproduction (Symons, 1992; Tooby & Cosmides, 1990).

Still, the organism must have a nervous system that allows it to make good choices. Selection could imbue the nervous system with a suite of hard-wired response rules, each tailored to a particular context. If the organism is merely a collection of hard-wired rules, it will respond to environmental stimuli without necessarily experiencing mental states. However, hard-wired solutions are less tenable for problems in which the optimal response varies dramatically with subtle changes in context. To deal with a nearly infinite number of subtly varying contexts that could be encountered, the biological automaton must come

equipped with a nearly infinite number of hard-wired rules, each of which is invoked in a slightly different context.

Selection could also imbue the nervous system with a suite of goals that approximated fitness in ancestral environments (e.g., acquisition of sugar or sexual experience) and a means for tracking those goals. For example, each time a particular situation is encountered, the nervous system could respond to it differently until it found the best response. Such post hoc learning may require the organism to experience some mental states (such as pain and pleasure) that allow it to evaluate whether a choice produced a good outcome. In any event, post hoc learning will not be very efficient for situations in which the adaptive response varies with subtle changes in context because it only finds good responses to contexts that have already been encountered.

Alternatively, selection could design a nervous system that allows the organism, before making a choice, to internally simulate the likely outcomes of behavioral options and predict which ones best satisfy internal goals (Alexander, 1989). Presumably, internal simulation would be less cumbersome than storing a large number of hard-wired rules and more efficient than post hoc learning. It would require the ability to internally represent the self and the external environment, including other actors if their behaviors must be simulated. Since actors have goals, internal simulation would also require representation of the motivational systems of the actors in the simulation. The organism must then make behavioral decisions on the basis of the outcomes of the internal simulation, which will require some internal standard of utility for identifying and comparing desirable and undesirable outcomes (e.g., aesthetic experience). In short, mental states (such as perceptions, beliefs, emotions, intentions, etc.) may allow an organism to identify a behavioral option (from a large suite of options) that approximates an optimal solution to a problem posed by the environment.

Social chess is very amenable to internal simulation. The best behavioral option is often dependent on subtle changes in the social environment, especially the option chosen by one's opponent (Humphrey, 1976; Maynard Smith, 1982). Moreover, as outcomes depend on the actions taken by both parties, social chess favors those who can predict and influence their opponents' behavior (Humphrey, 1976). But if the behavioral decisions made by one's opponent are also based on internal simulation, prediction may require internal simulation from the opponent's perspective. People may make attributions about the mental states of others because, in ancestral environments, they had to predict the behavior of those who also made behavioral decisions by internal simulation.

4. Attribution as a mind-reading process

A *mind reader* is an individual who tries to predict the behavior of an actor by making an inference about the actor's mental state. Assume that a mind reader and an actor are poised for potential social interaction, and that the actor produces behavior *B*. It will be useful to categorize the behavior according to whether or not it was produced to communicate information about the actor's mental state to the mind reader. If the behavior was produced to communicate such information then it is a *signal* (Krebs & Davies, 1993). The signal is *honest* if the actor's behavior and mental state are corresponding, and it is

deceptive if they are noncorresponding. (Of course, the actor need not be consciously aware of the behavior's purpose for it to serve a communicative function.) If the behavior was not produced for communicative purposes, and the behavior and mental state are corresponding, then it is a *cue*; if the behavior is not communicative, and it does not correspond to mental state, then it is a *miscue*.

Because behavior and mental state correspond if they can be described by the same (or similar) words (Schneider et al., 1979), there are only a limited number of ways behavior can correspond to the actor's mental state (and often only one way). Conversely, behavior and mental state can be dissimilar in a potentially infinite number of ways, and so there are many more ways in which they can be noncorresponding. Logically accurate attribution requires the mind reader to consider, and rule out, all the possible corresponding and noncorresponding mental states before attributing a particular mental state to an actor. This will almost always be an impossible task because the mind reader will rarely (if ever) have enough information about an actor to completely rule out all the alternatives. Moreover, the more alternatives that must be considered, the more cognitive effort it will take to make an attribution. Clearly, the mind reader will only be able to consider a limited number of possible mental states, and will often have to settle for a probabilistic assessment of the actor's mental state. Consequently, the ancestral mind reader faced two fundamental problems: (1) What mental states should be considered? (2) How much weight should be given to each mental state? The solutions to both problems depended not only on the mind reader's assessment of what mental states were most likely, but also on the fitness consequences that were at stake for the mind reader.

4.1. What mental states should the mind reader consider?

Avoiding the FAE is a cognitively demanding task because it requires the mind reader to consider the possible noncorresponding mental states that the actor could have in addition to the corresponding one. Indeed, studies indicate that those who avoid the FAE spend more cognitive effort than those who make it (Vonk, 1999; Yost & Weary, 1996). A mind reader may often find it too unprofitable to put out the cognitive effort needed to consider noncorresponding mental states. Since behaviors are the product of mental decision-making processes, an actor's mental state and behavior will often correspond to each other. A heuristic of inferring correspondence will then be quick and reasonably effective even if it does lead to attributional errors. Thus, one potential explanation for the FAE is that, in the typical attribution experiment, people have little incentive to put out the cognitive effort needed to consider noncorresponding mental states and avoid making the FAE. If so, then people should be less likely to make the FAE when they have an incentive to consider noncorresponding mental states.

4.1.1. Deception

Deception is a noncorresponding mental state of which mind readers should be particularly wary. Actors produce deceptive signals to gain a benefit (or avoid a loss) often at the expense of the mind reader (Fein, 1996). In ancestral games of social chess, there were probably few other reasons for an actor to produce noncorresponding behavior. Moreover, the ancestral

costs to mind readers for failing to consider deception were probably high because (1) they would be unable to identify undesirable social partners and avoid interacting with them; and (2) it put them at risk of exploitation when interacting with others (Cosmides & Tooby, 1992).

Many attribution experiments have shown that people are less likely to make the FAE when the situation suggests that the actor has a possible incentive to deceive (Fein, 1996; Fein et al., 1990; Hilton et al., 1993; Vonk, 1998). These experiments also indicate that detecting a deceptive mental state is typically easier for people to do than detecting other noncorresponding mental states (Hilton et al., 1993). This suggests that ancestral human beings infrequently needed to detect other noncorresponding mental states (either because these other mental states were unlikely or because the costs of erroneous attribution were low).

4.1.2. Other noncorresponding mental states

Nevertheless, mind readers will consider other non-corresponding mental states when their outcomes depend on it. A recent study has shown that subjects made more accurate attributions about the writer of an assigned essay when they were dependent on the writer's behavior, and they could predict the writer's behavior from his essay (Vonk, 1999). Moreover, people who are highly dependent upon others may even be cognitively primed to make more accurate attributions because they have a greater incentive to predict and influence behavior. For instance, depression is strongly related to social dependency (Coyne & Whiffen, 1995), and depressed people are less likely than normals to make the FAE in the assigned essay paradigm even when their outcomes do not directly depend on it (McCaul, 1983; Yost & Weary, 1996). Similarly, people from more interdependent societies are less likely to make the FAE than people from less interdependent societies (Choi et al., 1999).

4.2. How much weight should a mind reader give to a mental state?

The next major problem facing a mind reader is to determine the weight that should be given to the corresponding and noncorresponding mental states under consideration. The mind reader's inferences should be influenced by the fact that the actor had behavioral choices. For instance, suppose the actor had the choice of producing either behavior *B* or alternative activity *A*. The actor presumably has adaptations for evaluating and choosing the option that yields the highest fitness payoff (at least in ancestral environments). If the actor chooses *B* over *A*, the principle of noncommon effects would suggest that the actor perceived *B* to be more advantageous than *A*. However, if the actor's choice is to be used to predict the actor's future behavior, the mind reader may still need to consider the possible reasons why the actor chose *B* over *A* (i.e., to honestly communicate, to deceive, or for noncommunicative reasons [making the behavior a cue or a miscue]). The situational context in which the behavior occurs may contain information that can help the mind reader diagnose the actor's mental state. However, for the information to be diagnostic, it must influence the likelihood of one mental state relative to another.

4.2.1. Distinguishing cues from signals

Since actors usually have little reason to produce miscues, the actor's behavior is likely to be a cue if it is not a signal. As a result, any situational information suggesting that the actor's

behavior is not a signal should increase the mind reader's assessment that the actor's behavior and mental state are corresponding.

Signals are only produced to influence observers. An actor who did not detect any observers could still produce a signal just in case he or she was being observed. Even so, mind readers should be more likely to believe that a behavior is a cue (and that it corresponds to mental state) if they believe that the actor is not aware of being observed. I know of no studies that have tested this prediction. However, it assumes that observers influence the behavior of actors — an assumption that has received a good deal of empirical support (McCann & Higgins, 1992; Schwartz & Gottlieb, 1976, 1980). For example, in one series of studies, subjects were more likely to help a victim in need when they had reason to believe that others were aware of their presence or that others could monitor their behavior (Schwartz & Gottlieb, 1976, 1980).

Many behaviors are obviously signals because they serve little or no purpose outside of communication (e.g., language). Other behaviors are not so clear. For instance, acts of altruism could serve either signaling or nonsignaling functions (Alexander, 1987; Frank, 1988). Relative to more ambiguous behaviors, signals should generally decrease the mind reader's confidence that the actor's behavior and mental state are corresponding. Signals cannot be cues or miscues; yet, since a miscue is already an unlikely possibility, the net effect of a signal is to rule out a major reason for believing that mental state and behavior correspond to each other (the cue).

4.2.2. A motive to deceive

Deception will become more plausible relative to honesty when the actor has an apparent incentive to engage in deceptive behavior. Moreover, the behavior of an actor with a motive to deceive is less likely to be a cue. For both of these reasons, the actor with a motive to deceive will decrease the mind reader's perception of correspondence. Conversely, since behavior and mental state will often correspond to each other in the absence of deception, the mind reader's perception of correspondence should be augmented when the actor has no apparent motive to deceive. Many studies have found that people tend to avoid the FAE (i.e., they are less likely to infer correspondence) when the actor has an apparent incentive to engage in deceptive behavior (Fein, 1996; Fein et al., 1990; Hilton et al., 1993; Vonk, 1998).

An actor with an apparent motive to deceive is often perceived with suspicion (Fein et al., 1990; Hilton et al., 1993). *Suspicion* is a state of suspended judgment (or agnosticism) about the actor's mental state. In other words, a motive to deceive does decrease perceptions of correspondence, but suspicious people are generally no more likely to infer a noncorresponding mental state than they are to infer a corresponding mental state. A likely reason for this agnosticism is that successful deception often requires convincing the mind reader that the actor's incentive to behave honestly is greater than the incentive to deceive. Thus, the successful deceiver must usually have an apparent incentive to communicate honestly as well, and it will often be difficult for the mind reader to discern the actor's true mental state.

For instance, in one study (reviewed in Hilton et al., 1993), a man of moderate means courts a wealthy woman. He tells her that he is in love with her, he gives her flowers and candy, and he asks her to marry him. The man has a motive to deceive the woman about his affection (i.e., to gain access to her money by convincing her to marry him), but he also has

an incentive to tell the truth (i.e., if the man really does love the woman, he has an incentive to tell her so). As a result, the man's actual mental state could either be corresponding (i.e., he may really love her) or it could be noncorresponding (i.e., he may not love her at all). Consistent with the analysis, subjects were neutral in their attributions about whether or not the man loved the woman (i.e., they avoided the FAE). However, the other attributional issue in the study was whether the man really wanted to marry the woman. The man had no apparent motive to deceive the woman about wishing to marry her, and subjects did tend to believe that he wanted to marry her.

4.2.3. Handicapping signals

Behavioral ecologists have developed a large body of empirical and theoretical work on honest and deceptive communication (see Andersson, 1994; Bradbury & Vehrencamp, 1998; Johnstone, 1995, 1997; Wiley, 1994 for reviews). In situations where actors have a potential motive to deceive mind readers (Maynard Smith, 1991, 1994), this research suggests that mind readers will evolve to rely on *handicapping signals* — signals that impose costs on the actor (Enquist, 1985; Godfray, 1991; Grafen, 1990; Zahavi, 1975, 1977). Costly signals reveal information to observers because the marginal gains are greater for honest signalers than for deceivers (Getty, 1998). A signal's cost provides mind readers with a degree of assurance that the signal is not deceptive, and, by inference, that it is honest. Costly signals should then tend to augment perceptions of correspondence.

Consider a simple model by which mind readers can use information about a signal's cost to help diagnose the actor's mental state. Assume that a mind reader and actor are poised for potential social interaction. The actor produces a costly signal B (designed to entice the mind reader into social interaction) and foregoes the opportunity to pursue some alternative activity A (see Fig. 1). In the absence of mistakes, the actor will have chosen the option that yields the greatest likely net benefit (at least in ancestral environments). Assume that the mind reader knows something about the incentives influencing the actor's decision. V_H is the incentive for the actor to produce an honest signal; V_D is the incentive for the actor to produce a deceptive signal; V_A is the incentive for the actor to pursue the alternative activity; and c is the cost of the signal. Also assume that these values are all in the same currency. If the signal is honest, the actor's apparent payoff is $P_H = V_H - c$. If the signal is deceptive, the actor's apparent payoff is $P_D = V_D - c$. Finally, the apparent payoff that the actor could have received for pursuing the alternative activity is $P_A = V_A$.

There are two conditions that could increase the mind reader's confidence that the actor's signal was honest. Both require showing that there is some strategy with a higher payoff than the deceptive strategy. This other strategy must be a true alternative to deception (i.e., one that could have been done in its place). I will return to this point shortly.

4.2.4. Direct inference

If the honest and deceptive signaling strategies are true alternatives to each other, the mind reader's confidence that the signal is honest will increase if the incentive to communicate honestly is greater than the incentive to deceive (i.e., $P_H > P_D$). This yields:

$$V_H > V_D. \tag{1}$$

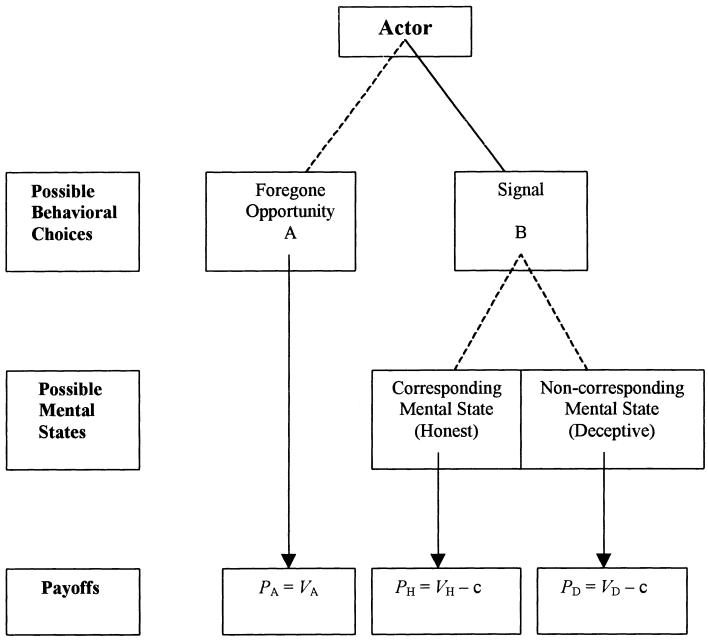

Fig. 1. How a mind reader can infer that a signal of mental state is honest. See the text for details.

Condition (1) indicates that the degree of cost of producing a signal should not influence observer attributions. The reason for this prediction is that the costs of producing the signal are cancelled out because they are on both sides of the inequality. More generally, if a mind reader has enough information to determine that honesty is more profitable than deception, a costly signal provides no new information about the actor's true intentions.

Often, honest and deceptive communication are not true alternatives to each other. If not, direct inference may lead to an erroneous conclusion about the actor's veracity. Suppose that a woman named Susan has asked me out for a date, and that I could be interested in pursuing either a long-term (investing) relationship with her or a short-term (exploitive) relationship. If you use direct inference to assess my mental state, then you will tend to believe that I am interested in a long-term relationship with her. Because I can derive higher fitness (more children) from Susan by having a long-term relationship than by having a short fling, it will appear that honest communication is more profitable than deception. However, the long-term and short-term strategies are not on the same time scale. A long-term relationship may last a lifetime whereas a short fling may last as little as a few minutes. They are not true alternatives to each other because I could not have a long-term relationship in the same time frame that I could have a short fling. Indeed, as an extremely handsome man, it is possible that I could have short flings with many different women in the same time that I could have one long-term relationship with Susan. The total fitness value of pursuing multiple short flings may exceed what I could get from the long-term relationship with Susan. If so, then direct inference will lead to an erroneous conclusion about my mental state.

4.2.5. Indirect inference — deduction by a process of elimination

With indirect inference, an observer can infer that an actor is pursuing honest communication provided that the motive to deceive is less profitable than the apparent incentive to pursue some other strategy (such as the alternative activity) that is a true alternative to deception (see Fig. 1). In that event, the observer can deduce that the actor is not pursuing the deceptive strategy (because he would have done better by pursuing the alternative activity). The actor also cannot be pursuing the alternative activity (else he would not have produced the signal). By eliminating the alternatives, the mind reader can deduce that the signal is honest. The condition for inferring honesty by indirect inference is when $P_A > P_D$, which yields the following:

$$c + V_A > V_D. \tag{2}$$

The cost of producing the signal and the value of the foregone opportunity are on the same side of the inequality to represent the fact that the total cost of a signal must include the opportunities that are foregone as a result of making the signal. Thus, one implication of indirect inference is that foregone opportunities will augment the perception of correspondence. In one study, subjects read about a man who wrote a speech but could have done so to ingratiate himself to a superior (Fein et al., 1990, Study 3). Because the speechwriter had an incentive to deceive the superior about his true beliefs, subjects were initially agnostic about whether the man really believed in the speech. However, subsequent information revealed that the speechwriter had foregone additional opportunities to ingratiate himself to the superior. Subjects then tended to infer that the writer's belief in the speech was real (i.e., foregoing the additional opportunities to ingratiate tended to augment the corresponding inference).

Another implication of indirect inference (and of direct inference) is the existence of threshold effects in the mind readers' perceptions of the honesty of a signal. This prediction follows directly from Condition (2). If the total apparent cost of the signal is greater than the motive to deceive, then the mind reader can infer that the actor is not pursuing the motive to deceive (at least the perceived motive), and the signal should tend to increase the perception of honesty. If the cost is less than the motive for deception, then the signal is not diagnostic, and mind readers should treat the signal with suspicion. The threshold point is where the cost equals the motive to deceive.

There are at least two reasons why threshold effects may be difficult to detect. First, mind readers will often not know the incentives affecting the actor's behavioral decisions with certainty; rather, they will have probabilistic assessments. Second, these incentives will often be in different currencies, making it difficult to compare them. For these reasons, threshold effects will often be obscured. Similarly, unless the mind reader knows all the actor's incentives with certainty, the fact that deception is apparently less profitable than some foregone opportunity will only be diagnostic of honesty (i.e., it increases confidence in honesty, but does not make it a certainty). These limitations also hold true for direct inference.

With these qualifications in mind, imagine again the scenario in which Susan has asked me out for a date. Suppose that Melinda has also asked me out for a date on the same night, and that Melinda is *less* beautiful than Susan. Faced with this choice, I decide to go out with

Susan. If beauty is an approximate predictor of the fitness consequences of my choice, Condition (2) suggests that you can infer nothing about my intentions. The only cost I appear to have incurred is foregoing the opportunity to date Melinda (i.e., $c = 0$). The value of the foregone opportunity (a possible fling with a less attractive Melinda) appears to be less than the value of deception (a possible fling with a more attractive Susan), and I should be treated with suspicion (i.e., $V_A < V_D$). Now suppose that Melinda is *more* beautiful than Susan, but I still decide to go out with Susan. By giving up a date with a more beautiful woman to go out with Susan, Condition (2) suggests that I am more likely to be interested in a long-term relationship with her. The value of the foregone opportunity appears to be greater than the value of deception, and perceptions of my long-term interest should increase (i.e., $V_A > V_D$). The threshold is predicted to exist at the point where Susan and Melinda are of equal beauty.

Consider one more implication of indirect inference. As the number of people who observe a costly signal increases, there are more individuals that an actor could deceive. Thus, the potential motive to deceive increases with observer number. A costly signal may then be perceived to be less reliable as an indicator of mental state when performed before a large number of observers than it will be when performed before a smaller number of observers. For example, many have suggested that acts of altruism may function as displays of cooperative intent that observers may use in making decisions about social partner choice (e.g., Alexander, 1987; Frank, 1988). A key question regarding an altruistic display is whether it is part of a deceptive con-artist strategy or reflects genuine cooperative intent.

Suppose that an actor behaves altruistically towards a victim in need of help in the presence of a single mind reader, M_1. Through his altruistic act, the actor incurs cost c, and he performs the act to entice M_1 into social interaction. If he is attempting to deceive M_1 about his intentional state, he does so to gain the benefit b_1 from pursuing the short-term exploitive strategy. According to Condition (2), mind readers should only treat the signal as honest if they perceive that the cost of the display is greater than the actor's motive to deceive (i.e., if $c > b_1$).

Suppose now that the actor performs the altruistic act in the presence of two mind readers, M_1 and M_2. The actor still incurs cost c, but now if he attempts to deceive M_1 and M_2, he could gain a total benefit of $b_1 + b_2$. A mind reader can only rule out deception as a motive if $c > b_1 + b_2$. Since the chance that $c > b_1 + b_2$ is less than the chance that $c > b_1$, mind readers should be increasingly unwilling to rule out a deceptive motive for an altruistic display as the total number of observers increases. A consequence is that the probability that a mind reader will trust and respond to a given altruistic display should be inversely related to the total number of observers present. This is an untested prediction. However, this effect may exert an influence on signaler behavior. Individuals may be less likely to perform an altruistic display when there are more observers present because they will realize that observers will be less likely to trust and respond to the display as a signal of cooperative intent.

This could explain the well-known *bystander effect* in which a person in need is less likely to receive help as the number of bystanders increases (Latané, Nida, & Wilson, 1981). The standard social psychology explanation is that responsibility for helping becomes more diffused as bystander number increases, but this explanation fails to address why individuals should help at all. Moreover, many signaling hypotheses for altruistic behavior appear to be inconsistent with the bystander effect because they seem to suggest that helping rates should

increase with observer numbers (e.g., Nowak & Sigmund, 1998). To date, no evolutionary theory of altruism has accounted for the bystander effect. The attribution hypothesis presented here suggests that the greatest incentive to help may exist when observer number is low because observers are then more likely to trust that the altruistic display honestly reflects cooperative intent. The hypothesis requires that the loss of credibility is a steeper function of audience size than the likely gain from having a greater number of observers that one could potentially influence.

4.3. Attribution under uncertainty

When mind readers lack sufficient information to discern with certainty whether a signal is honest or deceptive, there are two incorrect inferences they could make: (a) infer honesty when in fact the actor is deceptive; and (b) infer deception when in fact the actor is honest. If the costs of making the two errors differed in ancestral environments, but there was no asymmetry in the benefits of making correct inferences, the mind reader may have evolved cognitive adaptations for biasing the inference towards the less costly error (Haselton & Buss, 2000).

For instance, an ancestral woman who erroneously inferred that a man was interested in a long-term relationship could have been stuck with raising a child without the help of a committed mate. However, if she erroneously inferred that he was pursuing a short fling, he may have had to work harder (i.e., invest more) to convince her of his commitment. The latter was often the less costly error, and Haselton and Buss (2000) have shown that women often bias their inferences about men's commitment in this direction.

Adaptations for managing the costs of attribution errors could lead to the FAE if the costs of erroneously making the corresponding inference are less than the costs of erroneously making the noncorresponding inference. For instance, prosecutors sometimes try to introduce the prior record of a criminal defendant into a case as evidence that the defendant has a propensity to commit crimes. In the United States, such evidence is typically viewed as relevant to the prosecutor's case, and would normally be admissible under Federal Rule of Evidence (FRE) 402 (Green & Nesson, 1983).[2] However, it is also said to be *unduly prejudicial* to the defendant. Judges and jurors are thought to give more weight to such evidence than they should give if they were just interested in finding out the truth of the defendant's guilt or innocence. Presumably, judges and jurors who hear evidence of a prior record become less willing to take into account situational information indicating that the defendant may be innocent of the crime he is charged with (Wissler & Saks, 1985). Because of these concerns, FRE 404(b) prohibits the introduction of propensity evidence by the prosecution (Green & Nesson, 1983).[3]

---

[2] FRE 402 states "All relevant evidence is admissible, except as otherwise provided by the Constitution of the United States, by Act of Congress, by these rules, or by other rules prescribed by the Supreme Court pursuant to statutory authority. Evidence which is not relevant is not admissible" (Green & Nesson, 1983, p. 3).

[3] FRE 404(b) states "Evidence of other crimes, wrongs, or acts is not admissible to prove the character of a person in order to show that he acted in conformity therewith" (Green & Nesson, 1983, p. 100).

While introduction of a prior record is prejudicial to the criminal defendant (Greene & Dodge, 1995; Wissler & Saks, 1985), conclusive evidence of *undue* (i.e., logically unwarranted) prejudice is lacking. Given that undue prejudice could be demonstrated, the question would still remain as to why the human psyche would give undue weight to such evidence. In the courtroom, jurors are asked to assess guilt in ways that were often not adaptive in ancestral environments (i.e., logically and without regard to their own perceived interests). In the hunter–gatherer groups that we evolved in, an individual with a history of uncooperative behavior was likely to pose a stable threat to the interests of other group members. If such a person was under suspicion for committing another offense, other group members might have had an interest in biasing their assessments towards guilt because the costs of an erroneous inference of guilt were low and the costs of an erroneous inference of innocence were potentially high.

5. Attribution as a means for manipulating others

There may be situations in which people do not make attributions for mind reading, but instead use their attributions to manipulate the attributions of others. Consider how competition for favorable reputations may have affected attributional processes. In ancestral environments, great advantages undoubtedly accrued to those who had reputations for skills, personalities, and other internal traits that were socially desirable to others (Alexander, 1987; Frank, 1988; Gurven, Allen-Arave, Hill, & Hurtado, 2000). To acquire reputation, people must make attributions about the internal traits underlying the actor's behavior, and then communicate them to others. Since communication lends itself to distortion, people should have been under selection to facultatively distort their attributions in ways that enhance their own reputations or derogate the reputations of others.

Evidence for such distortion comes from the differences between the attributions of actors and observers for desirable and undesirable personality traits and for success and failure (Schneider et al., 1979). For instance, actors tend to attribute their successes to ability and their failures to bad luck or lack of effort; observers are relatively more likely to attribute others' successes to good luck or effort and their failures to lack of ability. The difference between the attributions of actors and observers is known as the *self-serving bias*, and it appears to be motivated in large part by strategic self-enhancement (Sedikides, 1993; Sedikides, Campbell, Reeder, & Elliot, 1998).

Strategic self-enhancement could lead to the FAE provided the corresponding inference would either enhance the reputation of the self or derogate the reputations of others. In many of the studies involving the assigned essay paradigm, subjects are often assigned controversial essay stances (e.g., antiabortion, prosegregation, etc.). In these studies, making the FAE is equivalent to making an attribution that could, to a greater or lesser degree, harm the reputation of the writer (at least in certain circles). The tendency to make the error (i.e., to assume that the essay writer was antiabortion or prosegregation even in the no-choice condition) may reflect, in part, design for biasing attributions in ways that derogate the reputations of others.

Because reputations are relative attributes, one acquires a favorable reputation only at the expense of the reputations of others (Miller, 1990). Since this is likely to engender bad feelings, reputational derogation is least advantageous against those whose help one is dependent upon. If so, those who are more dependent on others for help should be less likely to make the FAE. Thus, the greater attributional accuracy of depressives and people from more interdependent societies may not only be because their enhanced dependency gives them a greater incentive to accurately predict behavior. Their enhanced dependency may also leave them cognitively primed to avoid derogating the reputations of others.

6. Conclusion

Evolutionarily informed research on attributional processing presents many potential opportunities to evolutionary and social psychologists who have worked on theory of mind largely independently of each other. Throughout the paper, I have tried to point out some of the interesting attributional problems (discovered in large part by social psychologists) that await evolutionary explanations and how evolutionary theory might give insight into them. Traditional attribution research could profit from treating the brain as an evolved organ that has been designed by selection, in part, to solve specific attributional problems. Similarly, the current evolutionarily oriented research into theory of mind could profit from incorporating the large body of empirical and theoretical literature on honest and deceptive communication developed by behavioral ecologists.

Acknowledgments

I wish to gratefully acknowledge the help of those who, over the years, have influenced the manuscript through comments or conversations: Steve Gangestad, Mike Gurven, Martie Haselton, Kim Hill, Karen Kessler, Moshe Kiflawi, Astrid Kodric-Brown, Randy Thornhill, and Paul Watson. Special thanks are also due to Martin Daly, Margo Wilson, and an anonymous reviewer whose comments greatly improved the manuscript.

References

Alexander, R. D. (1987). *The biology of moral systems*. Hawthorne, NY: Aldine de Gruyter.

Alexander, R. D. (1989). Evolution of the human psyche. In: P. Mellars, & C. Stringer (Eds.), *The human revolution* (pp. 455–513). Edinburgh: University of Edinburgh Press.

Andersson, M. (1994). *Sexual selection*. Princeton, NJ: Princeton Univ. Press.

Baron-Cohen, S. (1995). *Mindblindness*. Cambridge, MA: MIT Press.

Bradbury, J. W., & Vehrencamp, S. L. (1998). *Principles of animal communication*. Sunderland, MA: Sinauer.

Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal attribution across cultures: variation and universality. *Psychological Bulletin*, *125*, 47–63.

Cosmides, L. (1989). The logic of social exchange: has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *31*, 187–276.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In: J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: evolutionary psychology and the generation of culture* (pp. 163–228). Oxford: Oxford Univ. Press.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*, 1–73.

Coyne, J. C., & Whiffen, V. E. (1995). Issues in personality as diathesis for depression: the case of sociotropy-dependency and autonomy-self-criticism. *Psychological Bulletin*, *118*, 358–378.

Cronin, H. (1991). *The ant and the peacock*. Cambridge: Cambridge Univ. Press.

Dennett, D. T. (1991). *Consciousness explained*. Boston, MA: Little, Brown and Company.

Enquist, M. (1985). Communication during aggressive interactions with particular reference to variation in choice of behaviour. *Animal Behaviour*, *33*, 1152–1161.

Fein, S. (1996). Effects of suspicion on attributional thinking and the correspondence bias. *Journal of Personality and Social Psychology*, *70*, 1164–1184.

Fein, S., Hilton, J. L., & Miller, D. T. (1990). Suspicion of ulterior motivation and the correspondence bias. *Journal of Personality and Social Psychology*, *58*, 753–764.

Frank, R. (1988). *Passions within reason: the strategic role of the emotions*. New York: Norton.

Getty, T. (1998). Handicap signaling: when fecundity and viability do not add up. *Animal Behaviour*, *56*, 127–130.

Gigerenzer, G. (1998). Ecological intelligence: an adaptation for frequencies. In: D. D. Cummins, & C. Allen (Eds.), *The evolution of mind* (pp. 9–29). New York: Oxford Univ. Press.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review*, *102*, 684–704.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: social contracts, cheating, and perspective change. *Cognition*, *43*, 127–171.

Gilbert, D. T., & Jones, E. E. (1986). Perceiver-induced constraint: interpretations of self-generated reality. *Journal of Personality and Social Psychology*, *50*, 269–280.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, *117*, 21–38.

Godfray, H. C. J. (1991). Signalling of need between parents and offspring. *Nature*, *352*, 328–330.

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, *144*, 517–546.

Green, E. D., & Nesson, C. R. (1983). *Problems, cases, and materials on evidence*. Boston, MA: Little, Brown and Company.

Greene, E., & Dodge, M. (1995). The influence of prior record evidence on juror decision-making. *Law and Human Behavior*, *19*, 67–78.

Gurven, M., Allen-Arave, W., Hill, M., & Hurtado, M. (2000). It's a wonderful life: signaling generosity among the Ache of Paraguay. *Evolution and Human Behavior*, *21*, 263–282.

Haselton, M., & Buss, D. M. (2000). Error management theory: a new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology*, *78*, 81–91.

Hilton, J. L., Fein, S., & Miller, D. T. (1993). Suspicion and dispositional inference. *Personality and Social Psychology Bulletin*, *19*, 501–512.

Hobson, J. A. (1999). *Consciousness*. New York: Scientific American Library.

Humphrey, N. K. (1976). The social function of intellect. In: P. P. G. Bateson, & R. A. Hinde (Eds.), *Growing points in ethology* (pp. 303–317). Cambridge: Cambridge Univ. Press.

Humphrey, N. K. (1992). *A history of the mind*. New York: Simon & Schuster.

Johnstone, R. A. (1995). Sexual selection, honest advertisement and the handicap principle: reviewing the evidence. *Biological Reviews of the Cambridge Philosophical Society*, *70*, 1–65.

Johnstone, R. A. (1997). The evolution of animal signals. In: J. R. Krebs, & N. B. Davies (Eds.), *Behavioural ecology: an evolutionary approach* (4th ed., pp. 155–178). Oxford: Blackwell.

Jones, E. E. (1979). The rocky road from acts to dispositions. *American Psychologist*, *34*, 107–117.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: the attribution process in person perception. In: L. Berkowitz (Ed.), *Advances in Experimental Psychology, vol. 2* (pp. 219–266). New York: Academic Press.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*, 1–24.

Kelley, H. H. (1972). Attribution in social interaction. In: E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: perceiving the causes of behavior* (pp. 1–26). Morristown, NJ: General Learning Press.

Krebs, J. R., & Davies, N. B. (1993). *An introduction to behavioural ecology* (3rd ed.). Oxford: Blackwell.

Latané, B., Nida, S. A., & Wilson, D. W. (1981). The effects of group size on helping behavior. In: J. P. Rushton, & R. M. Sorrentino (Eds.), *Altruism and helping behavior: social, personality, and developmental perspectives* (pp. 287–313). Hillsdale, NJ: Lawrence Erlbaum.

Maynard Smith, J. (1982). *Evolution and the theory of games*. Cambridge: Cambridge Univ. Press.

Maynard Smith, J. (1991). Honest signalling: the Philip Sydney game. *Animal Behaviour*, *42*, 1034–1035.

Maynard Smith, J. (1994). Must reliable signals always be costly? *Animal Behaviour*, *47*, 1115–1120.

McCann, C. D., & Higgins, E. T. (1992). Personal and contextual factors in communication: a review of the 'communication game'. In: G. R. Semin, & K. Fiedler (Eds.), *Language, interaction and social cognition* (pp. 144–172). London: Sage.

McCaul, K. D. (1983). Observer attributions of depressed students. *Personality and Social Psychology Bulletin*, *9*, 74–82.

Miller, W. I. (1990). *Bloodtaking and peacemaking: feud, law, and society in Saga Iceland*. Chicago, IL: University of Chicago Press.

Nisbett, R. E., & Ross, L. (1980). *Human inference: strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, *393*, 573–577.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a "theory of mind"? *Behavior and Brain Sciences*, *4*, 515–526.

Ross, L. (1977). The intuitive psychologist and his shortcomings: distortions in the attribution process. In: L. Berkowitz (Ed.), *Advances in Experimental Psychology*, *vol. 10* (pp. 173–220). New York: Academic Press.

Schneider, D. J., Hastorf, A. H., & Ellsworth, P. C. (1979). *Person perception* (2nd ed.). Reading, MA: Addison-Wesley.

Schwartz, S. H., & Gottlieb, A. (1976). Bystander reactions to a violent theft: crime in Jerusalem. *Journal of Personality and Social Psychology*, *34*, 1188–1199.

Schwartz, S. H., & Gottlieb, A. (1980). Bystander anonymity and reactions to emergencies. *Journal of Personality and Social Psychology*, *39*, 418–430.

Sedikides, C. (1993). Assessment, enhancement, and verification determinants of the self-evaluation process. *Journal of Personality and Social Psychology*, *65*, 317–338.

Sedikides, C., Campbell, W. K., Reeder, G. D., & Elliot, A. J. (1998). The self-serving bias in relational context. *Journal of Personality and Social Psychology*, *74*, 378–386.

Snyder, M., & Jones, E. E. (1974). Attitude attribution when behavior is constrained. *Journal of Experimental Social Psychology*, *10*, 585–600.

Symons, D. (1992). On the use and mis-use of Darwinism. In: J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: evolutionary psychology and the generation of culture* (pp. 137–159). Oxford: Oxford Univ. Press.

Tooby, J., & Cosmides, L. (1990). The past explains the present: emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology*, *11*, 375–424.

Vonk, R. (1998). The slime effect: suspicion and dislike of likeable behavior toward superiors. *Journal of Personality and Social Psychology*, *74*, 849–864.

Vonk, R. (1999). Effects of outcome dependency on correspondence bias. *Personality and Social Psychology Bulletin*, *25*, 382–389.

Whiten, A., & Byrne, R. W. (1988). Taking Machiavellian intelligence apart: editorial. In: R. W. Byrne, & A. Whiten (Eds.), *Machiavellian intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans* (pp. 50–65). Oxford: Oxford Univ. Press.

Wiley, R. H. (1994). Errors, exaggeration, and deception in animal communication. In: L. A. Real (Ed.), *Behavioral mechanisms in evolutionary ecology* (pp. 157–189). Chicago: Chicago Univ. Press.

Williams, G. C. (1992). *Natural selection: domains, levels, and challenges*. Oxford: Oxford Univ. Press.

Wissler, R. L., & Saks, M. J. (1985). On the inefficacy of limiting instructions: when jurors use prior conviction evidence to decide on guilt. *Law and Human Behavior*, *9*, 37–48.

Yost, J. H., & Weary, G. (1996). Depression and the correspondent inference bias: evidence for more effortful cognitive processing. *Personality and Social Psychology Bulletin*, *22*, 192–200.

Zahavi, A. (1975). Mate selection — a selection for a handicap. *Journal of Theoretical Biology*, *53*, 205–214.

Zahavi, A. (1977). The cost of honesty (further remarks on the handicap principle). *Journal of Theoretical Biology*, *67*, 603–605.