Detection Biases in Bluffing - Theory and Experiments

Holm, Jerker

2004

Document Version:
Other version

Detection Biases in Bluffing

By Håkan J. Holm*
Department of Economics
School of Economics and Management
Lund University, Sweden

This version: February 2007.

Abstract: Beliefs in signals that reveal lies or truths are widespread. These signals may lead to a truth or lie detection bias if the probability that such a signal is perceived by the receiver is contingent on the truth value of the sender's message. Such detection biases are analyzed theoretically in a bluffing game. The detection bias shrinks the equilibrium set to a unique non-pooling equilibrium, in which the better a player is at detecting lies the more often the opponent player will lie. With proper deception techniques such biases can in principle be used to extract hidden information.

Keywords: Bluffing, Game theory, Truth detection, Lie detection, Detection bias.

JEL-codes: C72, D82.

* An earlier longer version of this paper also contained an experiment. After receiving comments, the author decided to develop the experimental part more and to divide the material into two separate papers. I have received valuable comments from Hans Carlsson, Toshiji Kawagoe and Tetsuo Yamamori. The original version of the paper was presented at the Economic Science Association's World Meeting in Amsterdam in June and at the international conference "Experiments in Economic Sciences" in Kyoto in December 2004. It has also been presented in seminars at the Department of Economics at Lund University and at the Department of Economics and Statistics at Göteborg University in Sweden. I am grateful for comments from the participants at these conferences and seminars. Financial support from The Bank of Sweden Tercentenary Foundation and the Swedish Competition Authority is gratefully acknowledged.

1 Introduction

"Your poker face needs work my friend. It took me several seconds , but I can see now that you are lying."
Dan Brown, The Da Vinci Code (2004, p. 553).

A recurring theme in fiction is that a character believes that others can see through his lies or that he can tell if someone is lying just by looking at or listening to him. In Dostoevsky's "Crime and Punishment" Raskolnikov is haunted by thoughts that police superintendent Nikodim Fomitj can tell that he is lying, and Sir Leigh Teabing in "The Da Vinci Code" believes, as the quote above indicates, that he has the ability to call a bluff. Furthermore, it is not uncommon that people in real life claim that they can see through a lie, while others claim that they are quite bad at bluffing.[1] In addition to this, psychological research suggests that although people in general are relatively bad at distinguishing lies from truths, they tend to believe that certain observable non-verbal cues indicate lies (see e.g., Vrij, 2000, p.58). This psychological inclination motivates further investigations into the theoretical implications of lie detection cues in strategic situations. This paper will suggest a simple way to model lie detection and demonstrate that lie detection cues may have important theoretical consequences in situations of pure conflict of interest.

Lying in strategic situations has received little attention in economic theory.[2] The standard assumption is that if a player does not want to, communication does not necessarily disclose his type or his intentions. Thus, lying is possible and costless.

[1] Beliefs about such abilities might have had some historical impact. One example is when Hitler in a meeting before WWII lied to the ex-British Prime Minister Chamberlain about his intentions to invade Czechoslovakia. Chamberlain, obviously with a certain confidence in his abilities to identify liars, wrote to his sister: "in spite of the hardness and ruthlessness I thought I saw in his (i.e., Hitler's) face, I got the impression that here was a man that could be relied upon when he had given his word" (Ekman, 1992, pp15-16, parenthesis is mine.) In the Parliament Chamberlain said that he was convinced that Hitler did not try to deceive him. Czechoslovakia was invaded by Germany a few weeks later.

[2] There is however, a small number of interesting papers that experimentally investigate lying or issues closely related to it (see e.g., Frank et al., 1993, Ockenfels and Selten 2000, Brosig, 2002, Brandts and Charness, 2003, Charness and Dufwenberg, 2006, Gneezy, 2005). However, the approaches in these papers differ from the present one in that they involve reasoning about consequences, intentions, social preferences, or preferences for lying. Furthermore, the game analysed in these papers also differs from the pure conflict of interest game studied here.

Without this assumption the literature on asymmetric information ought to be fundamentally modified.

As recognized by Crawford (2003) and Hendricks and McAfee (2003) many strategic situations involve a stage where one party has the opportunity to misrepresent information. One such example of pure conflict of interest is where to attack in a war. Since disclosure of real intentions can be exploited by the counterpart, game theory (see Crawford and Sobel, 1982) predicts that zero-cost messages sent in such games are not informative (i.e., cheap talk). However, Crawford (2003) argues that there are many examples in reality where such messages are actually sent and appear to have an effect. If players are heterogeneous with respect to their reasoning capacity, Crawford (2003) has demonstrated that misrepresentation in such games may matter. However, there is also a more direct and psychological explanation. Players may actually be able, or believe that they are able, to recognize signals that are directly observed and related to the act of lying or truth-telling. Now, if such signals are present, there is no obvious reason to assume that "lie signals" are as accurate as "truth signals". If there is a difference between an individual's (imagined or real) ability to detect a lie and his ability to detect a truth, there is a detection bias. Mathematically, such a bias implies that the likelihood of detecting a lie is conditional on if the message is a lie or a truth, and furthermore, that the conditional likelihood of detecting a lie differs from that of detecting a truth. The aim of this paper is to analyze the implication of such a bias.

Psychological research indicates that detection biases are possible. For instance, a review based on some 40 studies (see Vrij, 2000, p.69) noted a 67 percent average accuracy rate for detecting truths. The corresponding accuracy rate for detecting lies was only 44 percent. One possible explanation is a truth detection bias.[3] Furthermore, lie and truth signals obtained by new techniques may also involve detection biases. One

promising finding in neuroscience (see Langleben et al., 2002) is that certain simple lies (which involve inhibitory processes) are associated with higher activity in certain areas of the brain. The increased activity is detectable by functional magnetic resonance imaging of the brain.[4]

In this paper it is not important if people in general are better at recognizing truths than lies or vice versa, the important thing being that beliefs about the recognizing capacity might differ and be conditional on whether the truth is told. One possibility is detection biases that are relationship specific. For instance, a man may know (or believe) that his wife sometimes with certainty can tell when he is lying and both are fully aware of this. The reason might simply be that the man is a hopelessly poor liar and that the wife has become a good lie detector after learning some observable cues associated with her husband's lies. When lying, the man is not in full control of when he emits these cues.

The issue of detection biases is certainly (to the author's knowledge) new in economics and game theory, and it would appear that analyzing detection biases with game theoretical tools is something new in psychology. To investigate the effects of detection bias, a simple signaling game is introduced and analyzed in section 2. Implications of the results are discussed in section 3 and the paper ends with some concluding remarks.

[3] It should be mentioned that this difference can be explained in different ways. One important part of the difference is that people have a tendency to judge other's statements as truthful.

[4] For a popular treatment on evidence that lying and truth-telling generate physiologically different effects that are observable with e.g., modern brain-scanning techniques (see The Economist, 2004, July 10th, "Lie Detection, Making Windows in men's souls, p. 71-2). If lying creates physiologically different reactions from truth-telling, it is not impossible that some of these physiological differences also generate differences in behavior that are i) observable ii) more easily recognizable for truths (or lies) than for lies (truths). Recent voice stress analyzing software is partly based on this idea. Despite the fact that its reliability has been debated, this software has already been used by British insurance companies to screen telephone claims in the hope of detecting fraud (see New York Times, 2004, July 1st, "It's the Way You Say It, Truth Be Told", Technology section).

2 Theory

This section will analyze detection biases theoretically. The game, denoted as the Bluffing game, can be presented as follows. In period 1 Nature selects the state variable $s \in \{R, B\}$ with probability $p_N = 1/2$. The S-player then observes a perfect signal of *s*. In period 2, S makes a statement $m \in \{B, R\}$ to the R-player about *s*. R is then to make a guess, $g \in \{T, F\}$, as to whether the statement is true ($g = T$) or false ($g = F$). A statement is said to be true if $m = s$ and false otherwise. R wins 1, if her guess correct, that is if $g = T$ and $m = s$, or if $g = F$ and $m \neq s$. In that case S gets zero. If R's guess is incorrect S wins 1 and R gets zero.[5]

There is an infinite number of mixed (perfect Bayesian) equilibria in this game. These equilibria can be characterized by $p_B \in [0,1]$ and $p_T = 1/2$, where p_B and p_T denote the probability of player S stating *B* (i.e., $m = B$) and the probability of player R guessing that the statement is true (i.e., $g = T$), respectively.[6] Furthermore, R believes that the probability of being in either node in the non-singleton information sets is $1/2$.[7] There are also pure pooling equilibria, where S makes a statement (*R* or *B*) with probability one and player R chooses for each statement either *T* or *F* with probability one.

2.1. Detection Bias in a Game Theoretical Setting

We will phrase the detection bias as a truth detection bias and then show that lie detection biases can be modeled in a symmetrical way. Incorporating a truth detection bias in this

[5] This game can be interpreted as a signalling game (see e.g., Gibbons, 1992), where Nature draws S's (i.e., the sender's) type after which he sends a message to R (the receiver) who forms beliefs about S's type and chooses an action contingent on these beliefs. The game in its extensive form is given in Figure A1 in the appendix.

[6] Note, in the full extension of the game S's strategy is obviously contingent on the state, but in any mixed equilibrium the probability of S sending a certain message will be the same for both states. Similarly, the probability of R making a certain guess is contingent on the statement, but in any mixed equilibrium this probability is the same for both statements. To simplify the notation these probabilities are written as if they are unconditional.

paper involves adding a stage in the game where R with a certain probability $\pi \in (0,1)$ observes a perfect signal if the statement is true.[8] This stage takes place after S has made his statement and only if S tells the truth (i.e., $m = s$). In the proof of Proposition 1 we show that as soon as a truth detection bias is introduced the pooling equilibria vanish and the game can be reduced to the one in Figure 1. Since the decision to lie or tell the truth is crucial for S in the remaining equilibrium, the analysis can do without referring to the state variable in this symmetrical game. In the reduced form of the game we simply say that S chooses between telling the truth (*T*) or lying (*L*), where *T* means $m = s$ and *F* means and $m \neq s$.

Before the equilibrium is derived some variables will be introduced. Let *a* denote the probability of S choosing *T*, and let μ denote R's beliefs that she is at node R_2 in Figure 1 when in this information set. Finally, let *b* denote the probability of R choosing *T* in the information set that is not singleton (i.e., consisting of nodes R_2 and R_3) and let *c* be the corresponding probability for the singleton node (R_1).

Proposition 1: *Introducing a truth bias as described above leads to a unique perfect Bayesian equilibrium in the Bluffing game. The equilibrium is characterized by* $a^* = 1/(2-\pi)$, $b^* = (1-\pi)/(2-\pi)$, $c^* = 1$ *and* $\mu^* = 1/2$.

Proof: Let us start by ruling out all pooling equilibria. Suppose, in the non-reduced game, that e.g., B is played with a certain strictly positive probability that is state independent (i.e., pooling), then R's updated belief concerning the probability of being in either node of the non-singleton information set will differ. Since R has not received a truth signal, it is more probable that S has lied about the state. This makes it optimal for R to set $b = 0$. Clearly, this cannot be consistent with equilibrium, since S will expect this and react to it.

[7] Se Figure A1.

Let us now derive the non-pooling mixed equilibrium by analyzing R's decision in the last stage, where there are two different information sets corresponding to node R_1 and the two nodes R_2 and R_3 in Figure 1.[9] At node R_1, R has received a perfect signal, which means that it is optimal to set $c^* = 1$. At nodes R_2 and R_3, R's conditional expected payoff from choosing *T* and *F* will be $a(1-\pi)/(1-a\pi)$ and $(1-a)/(1-a\pi)$, respectively. These payoffs must be equal in a mixed equilibrium, which require that $a^* = 1/(2-\pi)$. Similarly, S's payoff from the pure strategies of *T* and *L* are $(1-\pi)(1-b)$ and *b*, respectively and must also balance. Hence, $b^* = (1-\pi)/(2-\pi)$. Furthermore, consistent beliefs then require that $\mu = (1-\pi)/\left(a^{*-1} - \pi\right) = 1/2$.

Uniqueness should be clear from the following observations. First, $\mu \neq 1/2$ cannot be consistent with equilibrium since it would then be optimal for R to select a pure strategy in this information set. A pure separating strategy obviously cannot form an equilibrium in this "matching pennies" like game. Furthermore, if $\mu = 1/2$, then no strategy combination other than the one chosen will balance the expected payoff of the pure strategies. Together with the non-existence of any pooling equilibrium this establishes the uniqueness result. Q.E.D.

To see that the reasoning behind Proposition 1 also applies to a lie detection bias, assume instead that π is the probability of R observing the perfect signal if the statement is false. This stage takes place after S has made his statement if and only if S lies. Furthermore, let *a* denote the probability of S choosing *F*, let μ denote R's belief that she is at the node where S has lied but no perfect signal has been emitted from Nature

[8] This signal could be a twinkle, a tick, a blush, a particular brainwave recognizable by EEG or a specific signal that the R player has learnt to associate with truth-telling. Furthermore, S knows this.
[9] This equilibrium is weakly separating in the sense that the probability of S sending a certain message (B or R) in the non-reduced game is type dependent.

and let *b* denote the probability of R choosing *F* in the non-singleton information set. Finally, let *c* be the corresponding probability for the singleton node.

Note, irrespectively of whether the bias is in terms of lie detection or truth detection, the introduction of it has a theoretical quality in that the equilibrium set is refined. Furthermore, the actual existence of true detection signals is not crucial. What is crucial is that the players believe in them. This has important implications for the possibility of extracting the likely truth, as will be explained in the next section.

It should be stressed that the flavor of Proposition 1 is retained even if both truth and lie signals exist simultaneously in the game. This is proven in Proposition 2 in the Appendix. The only additional assumption required is that the probability of receiving a truth signal is not the same as the probability of receiving a lie signal.[10]

3. Implications

Let us now use Proposition 1 to make some observations concerning the implied equilibrium behavior when a detection bias is at hand. The message contingent probability for the perfect signal, π, can be interpreted as a measure of how good R is at detecting truths (or lies). The first part of Proposition 1 stating that $a^* = 1/(2-\pi)$ then implies that the better R is at recognizing truths, the more often S will tell the truth. This observation may seem counter-intuitive since one might expect S to be more wary of telling the truth the better the opponent is at tracing it. Perhaps it appears even more counter-intuitive when expressed in terms of lie detection; the better R is at detecting lies the more often will S lie. The intuition for the result is that the better R is at detecting lies the more informative it will also be for her not to receive a signal. Hence, the stronger the lie detection bias the more likely it is that S is telling the truth if no lie signal is observed. R will exploit this by guessing more often that S is telling the truth in this situation, which

will make the truth strategy less profitable to S. Thus, one might say that S is squeezed by either lying at the risk of being revealed by the perfect lie detection signal or being outguessed when telling the truth.

One interesting implication of Proposition 1 is that if R could make S believe in the detection bias game, she would command a powerful tool to improve her information about the true state even if no perfect signals actually exist.[11] To give a concrete example, suppose, before a large invasion by an S-army, a high-ranked officer in that army is captured by the R-army. Both armies know that S will soon attack, but only the S-officers know exactly where. There are only two feasible places to attack, Redding and Blackburn. R wins if and only if its officers succeed in deploying its army in the city attacked; the S-army wins otherwise. Now, assume that the captured S-officer only cares about the future success of his army and is brought in to be questioned by the R-officers about the place S plans to attack. Thus, the strategic situation is similar to the one in the bluffing game. However, before he is confronted with the crucial question he is successfully deceived into believing in a lie detection device possessed by the R-army and, also that the R-officers actually believe in it. He is informed that the probability that the lie detector will emit a perfect signal if he tells a lie is very high. When he is confronted with the crucial question Proposition 1 predicts that the officer will lie with high probability. Hence, if the officer's answer to the crucial question is Redding, the R-officers will know that Blackburn is the probable target for the attack.[12]

---

[10] In fact, the class of games analyzed in Proposition 2 in the appendix contains the class analyzed in Proposition 1. Hence, the former can be regarded as a generalization of the latter.

[11] Some professionally used lie detection methods rely on deceiving subjects into believing in the lie detection device. After the deception, physiological reactions (blood pressure, palm sweating, respiration etc.) are measured when certain questions are posed (see e.g., Vrij, 2000, p.179). What is argued here is that in certain strategic situations, as long as the deception works, the data from such measurements may be unnecessary from a game theoretic perspective!

[12] It should be mentioned that it is important that the players are sufficiently rational, but not too smart. For instance, if the captured officer realizes the deception scheme and can feint that he believes in it, he can use a meta-feinting strategy by telling the truth since he knows it will be taken as a lie. Obviously, there is no end to such a meta reasoning.

4. Concluding Remarks

The aim of the paper is to contribute to the study of bluffing. It is noted that beliefs in signals revealing lies or truths are widespread, and that neuroscience has recently suggested new methods for detecting such signals. These signals lead to a truth or lie detection bias if the probability that such a signal is perceived by the receiver is contingent on the truth value of the sender's message. The theoretical implication of lie and/or truth detection bias in bluffing appears to be an entirely new question in the literature and is given a first theoretical account here. The analysis is accomplished by developing a simple bluffing signaling game with conflicts of interest. In such a game it is shown that if the detection bias is modeled as a perfect signal from Nature, sent with a certain probability conditional on the truth value of the sender's message, the bias can be incorporated into a game theoretical equilibrium analysis. The result of the analysis is that the equilibrium set shrinks to a unique non-pooling equilibrium, in which the better the receiver is at detecting lies the more often the sender will lie. This somewhat counter-intuitive result could, in principle, be used to improve predictions about hidden information if the uninformed receiver is able to make the informed sender believe in the detection bias.

Literature

Bolton G. E., and Ockenfels, A., 2000, "A Theory of Equity, Reciprocity and Competition," American Economic Review, vol. 90(1), pp. 166-93.

Brandts, J., and Charness, G., 2003, "Truth or Consequences: An Experiment", Management Science, 49(1), 116-130.

Brosig, J., 2002, "Identifying Cooperative Behavior: some experimental results in a prisoner's dilemma game," Journal of Economic Behavior and Organization,¨47(2), pp. 275-290.

Charness G., and M. Dufwenberg, 2006, "Promises & Partnership," Econometrica 74, 1579-1601.

Crawford, V., and Sobel, J., 1982, "Strategic Information Transmission", Econometrica, 50(6), 1431-51.

Crawford, V., 1998, "A Survey of Experiments on Communication via Cheap Talk," Journal of Economic Theory, 78(2), 286-98.

Crawford, V., 2003, "Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions," American Economic Review, 93:1, 133-149.

Ekman, P. and O'Sullivan, M., 1991, "Who Can Catch a Liar?," American Psychologist, 46, 913-920.

Ekman, P., 1992, "Telling lies: clues to deceit in the market place, politics and marriage." New York: W.W. Norton.

Ekman, P., O'Sullivan, M., and Frank, M., 1999, "A Few Can Catch a Liar", Psychological Science, 10, 263-266.

Frank, R., 1987,"If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience?" American Economic Review, 77:4, 593-604.

Frank, R., Gilovich, T., and Regan, D., 1993, "The Evolution of One-Shot Cooperation: An Experiment," Ethology & Sociobiology., 14, 247-256.

Gibbons, Robert (1992): A Primer in Game Theory, Harvester-Wheatsheaf, Hempstead, Herfordshire.

Gneezy, U., 2005, "The Role of Consequences," American Economic Review, 95(1), 384-394.

Hendricks, K., and McAfee, P., "Feints", 2003, University of Texas, unpublished manuscript.

Langleben D., Schroeder, L., Maldjian, J., Gur, R., McDonald, S.,Radland, J., O'Brien, C., and Childress, A., 2002, "Brain Activity during Simulated Deception: An Event-Related Functional Magnetic Resonance Study", Neuroimage 15, 727-732.

Ockenfels, A., and Selten, R., 2000, "An Experiment on the Hypothesis of Involuntary Truth-Signalling in Bargaining,", Games and Economic Behavior, 33, 90-116.

Vrij, A., 2000,"Detecting Lies and Deceit, The Psychology of Lying and the Implications for the Professional Practice," John Wiley& Sons, Chichester.

Appendix

Detection bias with both truth and lie signals

In this appendix it will be demonstrated that with a slight modification Proposition 1 also holds in the case where there are possibilities for both truth and lie signals in the game. Let $\sigma \in [0,1]$ be the probability of R receiving a perfect lie signal when S is lying. Furthermore, let $d \in [0,1]$ be the probability of R choosing $g = T$ if she has received the lie signal (i.e., she is at R_4 in Figure A3). The other variables are defined in section 2. The following proposition is a generalization of Proposition 1.

Proposition 2: If $\pi \neq \sigma$ then there exists a unique perfect Bayesian equilibrium characterized by $a^* = \frac{1-\sigma}{2-\pi-\sigma}$, $b^* = \frac{1-\pi}{2-\pi-\sigma}$, $c^* = 1$, $d^* = 0$ *and* $\mu^* = 1/2$.

Proof: In the subgames where R has received a perfect signal it should be obvious that $c^* = 1$ and $d^* = 0$. Now, let us start with the case where $\pi > \sigma$ (i.e., there is a relative truth detection bias). For the same reasons as stated in the proof of Proposition 1, all pooling equilibria disappear. A pooling equilibrium (in the non-reduced game) implies that $a^* = 1/2$. The updated probability of being in node R_3 would then be larger than the corresponding probability of being in R_2. This cannot be consistent with equilibrium, otherwise it would be optimal for R to set $b = 0$. Obviously, S has to react to that, meaning that $a^* \neq 1/2$. Thus, pooling equilibria can be disregarded.

Let us now concentrate on the non-pooling equilibria. As noted in the proof of Proposition 1, in a mixed non-pooling equilibrium R's expected payoff from $g = T$

and $g = F$ must be the same in the non-singleton information set. The conditional payoffs of these strategies are $\frac{a(1-\pi)}{1-a(\pi-\sigma)-\sigma}$ and $\frac{a(1-\pi)}{1-a(\pi-\sigma)-\sigma}$. To balance, it is necessary that $a = \frac{1-\sigma}{2-\pi-\sigma}$. Similarly, R will choose a strategy so that S's payoffs from the pure strategies balance. Hence, $(1-\pi)(1-b) = (1-\sigma)b \Rightarrow b = \frac{1-\pi}{2-\pi-\sigma}$. Furthermore, as noted in the proof of Proposition 1 any separating equilibrium in pure strategies can be ruled out in this "matching pennies" like game.

Finally, the reasoning above would be entirely symmetric if it instead was assumed that $\pi < \sigma$. Q.E.D.

To see that Proposition 2 is a generalization of Proposition 1, let $\sigma = 0$.

Figures

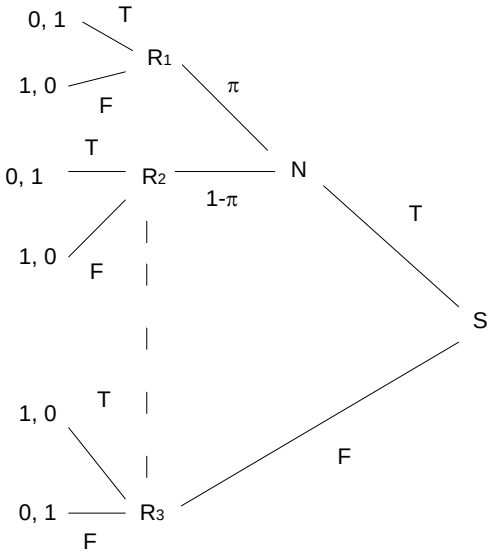

Figure 1: Reduced bluffing game with detection bias.

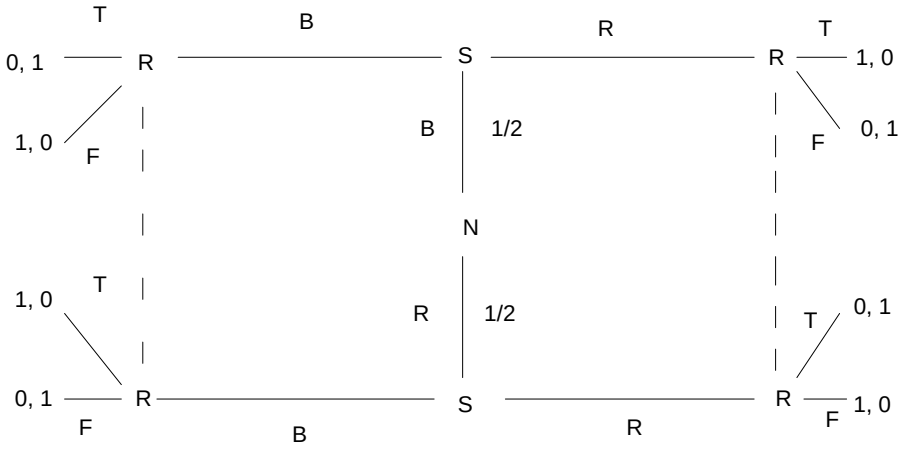

Figure A.1: Bluffing game without detection bias.

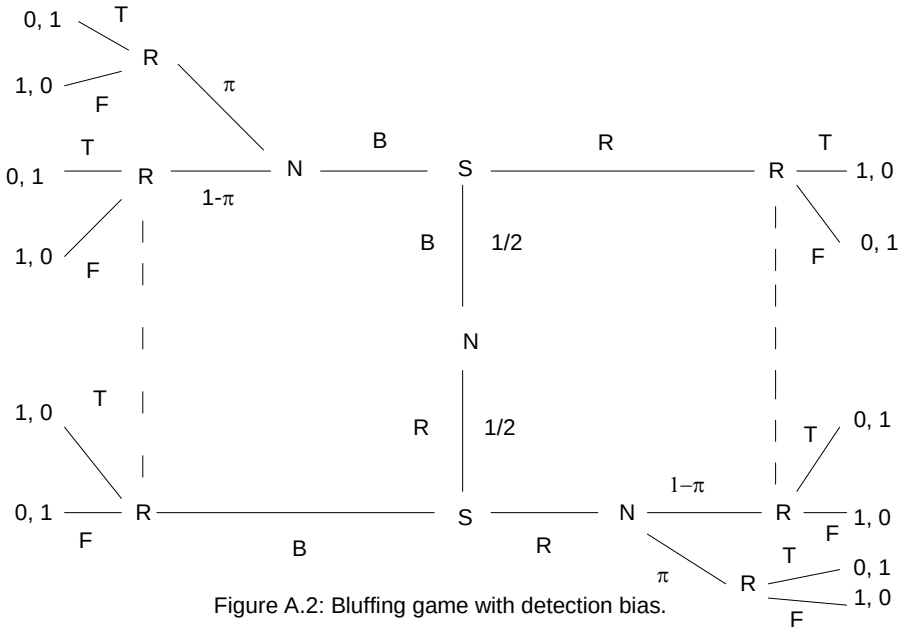

Figure A.2: Bluffing game with detection bias.

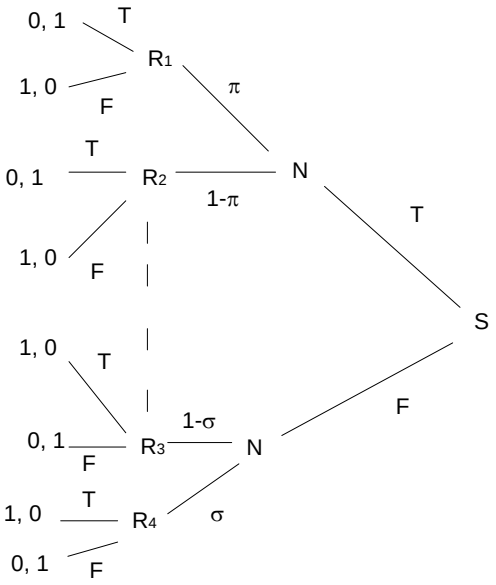

Figure A3: Reduced bluffing game with both truth and lie signals.