

The Deception of the Greeks: Generalizing the Information Structure of Extensive Form Games*

Oliver Board[†]

`oliver.board@economics.ox.ac.uk`

November 2002

Abstract

The standard model of an extensive form game rules out an important phenomenon in situations of strategic interaction: deception. Using examples from the world of ancient Greece and from modern-day Wall Street, we show how the model can be generalized to incorporate this phenomenon. Deception takes place when the action observed by a player is different from the action actually taken. The standard model does allow imperfect information (modeled by non-singleton information sets), but not deception: the actual action taken is never ruled out. Our extension of extensive form games relaxes the assumption that the information sets partition the set of nodes, so that the set of nodes considered possible after a certain action is taken might not include the actual node. We discuss the implications of this relaxation, and show that in certain games deception is inconsistent with common knowledge of rationality even along the backward induction path.

“You are to hear now how the Greeks tricked us. From this one proof of their perfidy you may understand them all” (Aeneas).

1 Deception in games

Although the Trojan war pre-dates the formal study of games by almost three thousand years, the Greek generals clearly possessed a sound understanding of the basic principles of game theory.

*I am grateful to two anonymous referees, and to Michael Bacharach, Alexandru Baltag, Steven Brams, Laurence Emmett, Erik Eyster, Alexander Gümbel, Meg Meyer, Llewelyn Morgan, Rami Spiegler and Daniel Zizzo, as well as participants at the Fifth Conference on Logic and the Foundations of Game and Decision Theory, the Fifth Spanish Game Theory Meeting, Nuffield College and the Computing Laboratory, Oxford for helpful advice.

[†]Department of Economics, Univeristy of Oxford, Manor Road, Oxford, OX1 3UQ, UK.

Odysseus, in particular, was a master: his dealings with the Sirens, for example, provide an excellent illustration of the value of commitment; returning home at last, he disguises himself as a beggar to collect information about his wife's suitors, only revealing his true identity when the time is right, thus trading short-run losses for long-run gains. It would be possible to write an entire game theory text book using the Greek myths as a basis. The current aim is more modest: to construct a game-theoretic model of a single incident at the end of the Trojan war, when the Greeks tricked the Trojans by abandoning their camp and sailing behind the island of Tenedos. We claim that the standard definition of an extensive form game is too restrictive to capture an important feature of this story, namely *deception*.

Deception takes place when one player tricks another into believing that she has done something other than what she actually did. In this case, the Greeks remain in the vicinity of Troy but out of sight behind Tenedos so that the Trojans believe they have sailed home. This phenomenon is ruled out by the way information is modeled in extensive form games. The standard structure of an extensive form game does allow actions to be uninformative (whenever information sets are non-singleton), in that it is not revealed which of several actions has been taken. But they cannot be deceptive: the actual action taken is never ruled out. Relaxing the assumption that the information sets partition the set of nodes allows deception to take place. In particular, the set of nodes considered possible after a certain action is taken might not include the actual node. In section 2 we show how a game with a non-partitional information structure can be used to represent the story above, and consider a more recent example of deception.

Recent work on decision problems of imperfect recall such as the absent-minded driver problem (introduced by Piccione and Rubinstein [15]) has suggested that the interpretation of information sets in extensive form games is not straightforward. If non-partitional information structures are allowed, things become even less clear. Section 3 reviews the standard definition of an extensive form game, and shows how the information structure can be generalized. In section 4 we comment on various issues of interpretation in the generalized model. The notion of equilibrium in these games is also discussed. Section 5 reviews related literature, and some conclusions are offered in section 6.

2 Two examples

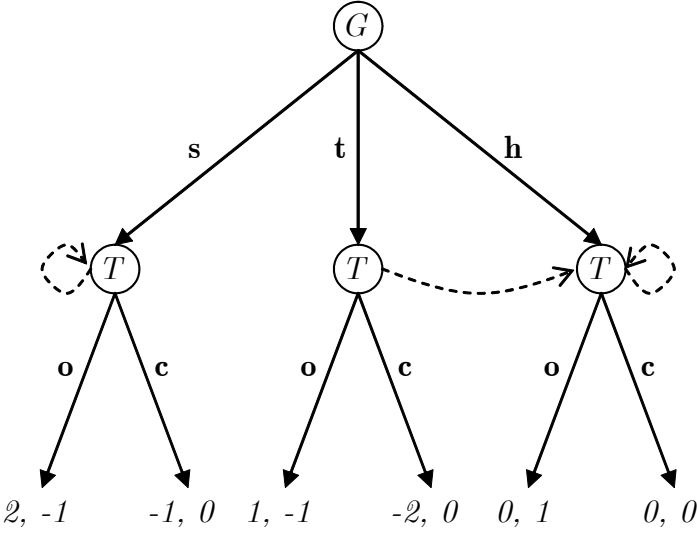
2.1 The Trojan war

“Within sight of Troy is the island of Tenedos. In the days of Priam’s Empire it had wealth and power and was well known and famous, but there is nothing there now, except the curve of the bay affording its treacherous anchorage. The Greeks put to sea as far as Tenedos, and hid from sight on its lonely beaches. We thought they had sailed for Mycenae before the wind and gone home. So all the land of Troy relaxed after its years of unhappiness. We flung the gates open and we enjoyed going to look at the unoccupied, deserted space along the shore where the Greek camp had been”. (Aeneas, quoted in *The Aeneid* Book II [19].)

In Book II of *The Aeneid*, Vergil tells the story of how the Greeks gained entry to the city of Troy by means of a trick. After ten years waging an unsuccessful war, the Greeks considered their options: to go home and give up the war or to stay and attempt to sack Troy. The latter seemed hopeless until one of their number, Prylis, suggested the following plan: they should sail their ships out of sight behind the island of Tenedos and leave a gigantic wooden horse in front of the city. Believing the Greeks had really gone home, the Trojans accepted the horse as a gift and broke down their walls to wheel it into Troy. The Greeks then leapt out of the horse and successfully sacked the city. Deception was essential for the success of this plan. The Trojans were highly suspicious of the wooden horse and would not have accepted the it into their city unless they really believed the Greeks had gone home.

To model the deception of the Greeks, we allow them *three* action choices at the beginning of the game: to go home (**h**); to sail behind the island of Tenedos (**t**); and to stay put (**s**). In each case, the Trojans can choose to open up their gates and accept the wooden horse (**o**), or to keep them closed and reject it (**c**). We represent the information structure of the game in the usual way, by information sets: if player i is on move at node x , then $\mathcal{I}(x)$ lists the set of nodes she considers possible given the information at that time (i.e. $\mathcal{I}(x)$ is the smallest set of nodes in which she is sure to find the actual node). The Trojans’ information sets at their three decision nodes are: $\mathcal{I}(s) = \{s\}$; $\mathcal{I}(h) = \{h\}$; and $\mathcal{I}(t) = \{h\}$ (using the obvious notation). In other words, if the Greeks stay, the Trojans can see that they have stayed; if the Greeks sail away, the Trojans can see that they have sailed away; but if the Greeks sail to Tenedos, it seems to the Trojans as if they

have sailed away. It is clear these information sets do not partition the set of decision nodes of the Trojans, and hence this game does not fit the standard definition of an extensive form game. Using dotted arrows to represent the information structure, the game is shown in the diagram below. Payoffs to the Greeks are given first. They prefer to stay if and only if the Trojans open the gates, and suffer minor inconvenience from sailing behind the island. The Trojans prefer to open the gates if and only if the Greeks go home.



The Trojan War

The game can be solved by backward induction. At node *s*, the Trojans know they are at node *s* and will keep the gates closed; at nodes *t* and *h*, the Trojans believe they are at node *h* and will open the gates. If the Greeks know the Trojans are rational, they will be aware of this, and will choose to sail behind the island. And indeed, this is what happened.

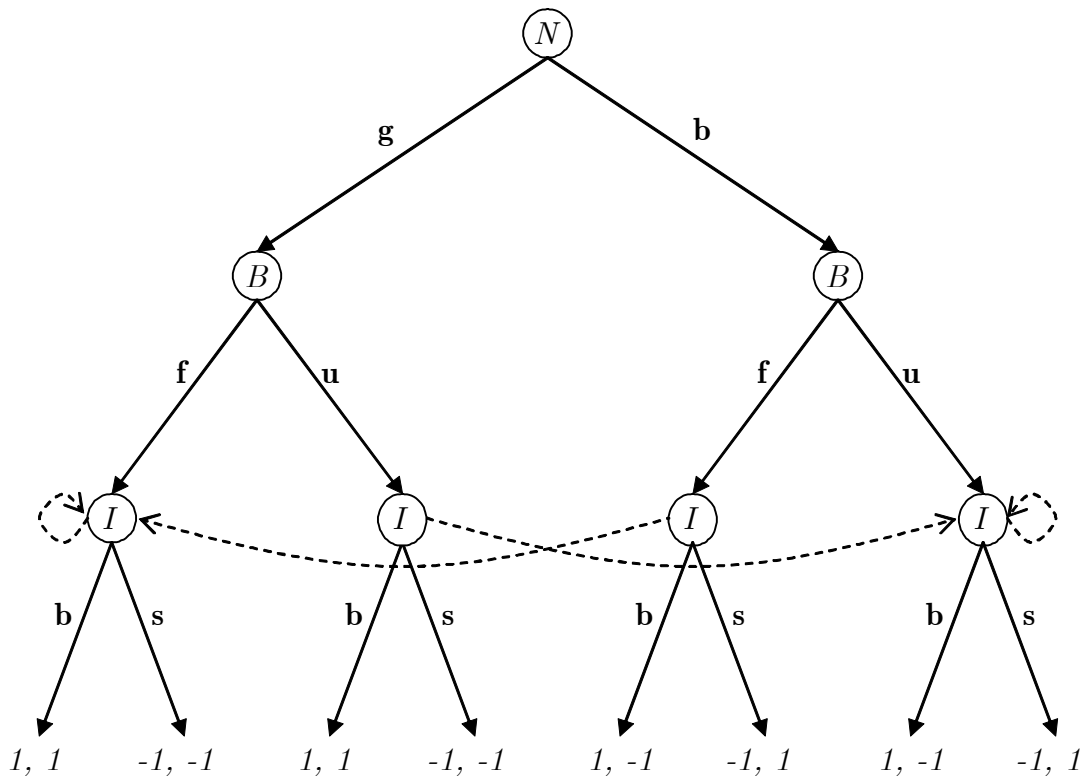
2.2 Investment advice

On 22nd May 2002, *The Times* newspaper ran the following story

Merrill Lynch agreed today to pay fines of \$100 million (£68.7 million) to help settle charges that it was telling clients to buy stocks it secretly believed were “junk”... The deal followed an investigation by New York Attorney General Eliot Spitzer alleging Merrill Lynch gave overly optimistic opinions of companies to win them over as investment banking clients.

The diagram below gives a stylized representation of this situation. Nature (*N*) moves first and determines whether the stock is good (**g**) or bad (**b**); the bank (*B*) observes the quality of the

stock, and makes a report to the potential investor (I), which can be favorable (f) or unfavorable (u); finally the investor, having heard the report but in ignorance of the true quality of the stock, decides whether to buy (b) or sell (s). The payoff structure is very simple: the bank cares only about creating business for itself, so prefers for the investor to buy; the investor wants to buy if and only if the stock is actually good. As before, information is represented by dotted arrows: the investor always believes the bank's report, so that she thinks the stock is good whenever the bank gives a favorable report, and that the stock is bad whenever the bank gives an unfavorable report. Letting gf , gu , bf , and bu denote the investor's decision nodes, her information sets are $\mathcal{I}(gf) = \mathcal{I}(bf) = \{gf\}$; and $\mathcal{I}(gu) = \mathcal{I}(bu) = \{bu\}$.



Investment Advice

Again we use backward induction to solve the game. At nodes gf and bf , the investor believes she is at node gf , and plays b ; at nodes gu and bu , the investor believes she is at node bu and plays s . The bank will therefore play f whatever the move by nature at the start of the game. Favorable reports will be issued even for stocks which are known to be bad.

2.3 Deceived or simply mistaken?

Is it really the case that standard extensive form games with partitional information sets cannot represent either of the situations described above? It could be argued that the Trojans and the clients of Merrill Lynch were not deceived; rather, they were simply mistaken about which of two possible nodes had been reached. We consider each case in turn.

Adopting this line of argument, the information sets in *The Trojan War* would be $\mathcal{I}(s) = \{s\}$ and $\mathcal{I}(h) = \mathcal{I}(t) = \{h, t\}$. These sets partition the Trojans' decision nodes. It is easily verified that there is a unique perfect Bayesian equilibrium in which the Greeks mix between going home and sailing behind the island, and the Trojans mix between opening the gates and keeping them closed. But the historical outcome of the game cannot be explained as a particular realization of these mixed strategies. For in this mixed strategy equilibrium, Bayesian updating requires that the Trojans assign equal probability to nodes h and t , since the Greeks are mixing 50–50. Yet we are told that they “thought they had sailed for Mycenae before the wind and gone home”.

An alternative explanation is that we are observing out-of-equilibrium behavior: the Trojans were simply mistaken about the Greeks' strategy choice and failed to play a best response. In fact, the observed outcome is rationalizable (i.e. consistent with common knowledge of rationality) in the game with standard information sets. The beliefs that rationalize this outcome are as follows:

G The Greeks believe that the Trojans will open their gates, and that the Trojans believe the Greeks have gone home;

T The Trojans believe that the Greeks have gone home, and that the Greeks believe that the Trojans will keep their gates closed.

But this story can give no explanation of why the Trojans came to have these beliefs: many other beliefs are also rationalizable. Indeed they were warned by the priest Laocoön that the Greeks had not gone home: “Do you really believe that your enemies have sailed away?... I still fear Greeks, even when they offer gifts”. They ignored his advice because they were deceived: “we gave Sinon [one of the Greeks] our trust, tricked by his blasphemy and cunning”. A related point is that rationalizability as a solution concept has little predictive power in the standard game; in the game of deception, on the other hand, there is a unique rationalizable outcome.

It is even harder to tell a coherent story about what is going on in the investment advice game using standard information sets. The information sets of the investor would be: $\mathcal{I}(gf) = \mathcal{I}(bf) =$

$\{gf, bf\}$; and $\mathcal{I}(bf) = \mathcal{I}(bu) = \{bf, bu\}$. There is a mixed-strategy perfect Bayesian equilibrium in which false advice is given and followed with positive probability, but as before this does not reflect what actually happened (i.e. that the investor believed *all* of the bank's reports). The observed outcome is rationalizable, but only by very odd beliefs on the part of the investor:

B The bank believes that the investor will invest if and only if the report is favorable, and that the investor believes the bank will give a favorable report if and only if the stock is good;

I The investor believes that the bank will give a favorable report if and only if the stock is good, and that the bank believes the investor will ignore all reports.

The reason is that the bank cares only about persuading the investor to buy rather than sell; thus a conditional reporting strategy (one in which the report issued depends on whether the stock is good or bad) can be rational only if the content of the report does not affect the investor's action. Furthermore, once again rationalizability allows a whole range of alternative outcomes in the standard game, in contrast to the unique prediction in the game of deception.

A more general defence of the standard representation of extensive form games is discussed in section 4.3.

3 A generalization of extensive form games

The games discussed in section 2 differ from standard extensive form games only in their information structure. The assumption that information sets partition the decision nodes is relaxed. The following definition is adapted from Osborne & Rubinstein [14]. Note that here we consider only finite games; the extension to the infinite case is straightforward.

Definition 1 *An extensive form game with generalized information structure is a tuple*

$$\langle N, H, P, f_c, \mathcal{I}, (u_i)_{i \in N} \rangle,$$

where

- *N is a finite set of players*
- *H is a finite set of sequences such that (a) $\emptyset \in H$; and (b) if $(a^k)_{k=1, \dots, K} \in H$ and $L < H$, then $(a^k)_{k=1, \dots, L} \in H$*

Each member of H is a history; each component of a history is an action taken by a player. A history $(a^k)_{k=1,\dots,K} \in H$ is terminal if there is no a^{K+1} such that $(a^k)_{k=1,\dots,K+1} \in H$. The set of actions available after the nonterminal history h is denoted $A(h) = \{a : (h, a) \in H\}$ and the set of terminal histories is denoted Z .

- P is a function which assigns to each nonterminal history a member of $N \cup \{c\}$. P is the player function, and $P(h)$ is the player who takes an action after history h ; if $P(h) = c$ then chance determines which action is taken after history h .
- f_c is a function which associates with every history h for which $P(h) = c$ a probability measure $f_c(\cdot | h)$ on $A(h)$. $f_c(a | h)$ is the probability that a occurs after history h . Each probability measure is independent of every other such measure.
- \mathcal{I} is a function which assigns to each nonterminal history a nonempty set of nonterminal histories such that if $h' \in \mathcal{I}(h)$, then (a) $P(h) = P(h')$; and (b) $A(h) = A(h')$. \mathcal{I} is the information function, and $\mathcal{I}(h)$ is the set of histories that player $P(h)$ considers possible if the true history is h . Condition (a) says that a player always knows when she is on move, and condition (b) says that she also knows what actions are available to her.
- u_i is a function from Z to \mathbb{R} , the utility function of player i .

An information function is more general than an information partition, in the sense that every information partition can be represented by an information function, but not every information function can be represented by an information partition unless we impose additional constraints on the form of \mathcal{I} . More precisely, if we assume that (c) $h \in \mathcal{I}(h)$, and (d) if $h' \in \mathcal{I}(h)$, then $\mathcal{I}(h) = \mathcal{I}(h')$, then we can find a partition which represents \mathcal{I} . Condition (c) rules out the possibility of deception (see Definition 2 below); the interpretation of condition (d) is discussed in Section 4.2.

We can use the information function to provide a taxonomy of a player's information whenever she is on move.

Definition 2 *The player $P(h)$ on move after history h is:*

- (a) perfectly informed if $\{h\} = \mathcal{I}(h)$;
- (b) imperfectly informed if $h \in \mathcal{I}(h)$ and $|\mathcal{I}(h)| > 1$;
- (c) deceived if $h \notin \mathcal{I}(h)$.

4 Comments

In this section, we discuss various issues of interpretation concerning games of deception, and address some potential criticisms. First, we question whether the standard assumption that the structure of the game is common knowledge is coherent in these games. Next, we argue that a deceived player need not be irrational, and discuss various forms of bounded rationality. We then reevaluate the claim that standard extensive form games are adequate for modeling deception. Finally, after asking how it is that deception might arise, we consider what might be an appropriate solution concept for games of deception.

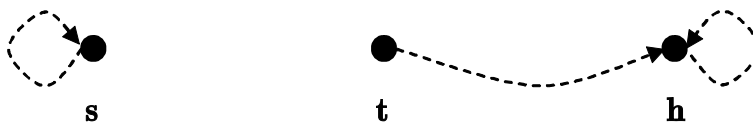
4.1 Can there be common knowledge of the game?

Game theorists standardly assume that the structure of the game, i.e. everything specified in Definition 1 above, is common knowledge among the players. This assumption is crucial as we are to make sense of standard solution concepts. Indeed the very notion of rationality as expected utility maximization presupposes that the rational player knows which options are available to her and what her utility function is. But surely it is not possible for a player to know that she has been deceived? And surely if a player knows that she *will* be deceived, this will undermine the deception? In fact, the first statement is true but the second need not be, and neither rules out common knowledge of the game. A systematic analysis of these issues requires a formal model of the player’s knowledge and beliefs.

Let us consider again the first example, *The Trojan War*. There is nothing incoherent in the assumption that both the players have common knowledge of the structure of the game (tree, information and utilities) before any moves are made, although it might seem strange that the Trojans know the Greeks have the option of sailing behind the island and at the same time know that they will be deceived if the Greeks do so. Reinterpreting \mathbf{t} as “trick” rather than “Tenedos” removes this awkwardness: if the Trojans do not know the exact form of the trick, it is more reasonable to suppose that they will be taken in by it. But what about the Trojans knowledge and beliefs when they come to move? At node t , is it possible for them to be deceived, while retaining knowledge of the structure of the game? We shall construct an *epistemic model* for the Trojans which shows that it is.

An epistemic model for a player tells us what that player knows and believes at a certain point in the game. It consists of a set of *states*, W ; a *history function*, $\mathcal{H} : W \rightarrow H$, which tells us which

history has been reached at each state; and an *accessibility relation* R , which tells us which states the player considers possible (i.e. if wRw' , then if w is the true state, the player considers state w' possible)¹. A player believes something if it is true at every state she considers possible, and we shall assume (rather crudely) that she knows something if she believes it and it is true. Let B_i and K_i stand for “player i believes that ...” and “player i knows that...” respectively. Consider an epistemic model with three states, $W = \{1, 2, 3\}$, with $\mathcal{H}(1) = \mathbf{s}$, $\mathcal{H}(2) = \mathbf{t}$, $\mathcal{H}(3) = \mathbf{h}$; and $1R1$, $2R3$, $3R3$. A diagrammatic representation is given below (note that here the arrows represent the accessibility relation, not the information function of the game as before).



An epistemic model for the Trojans

To see how the epistemic model works, suppose that the true state is 2, i.e. that Greeks have sailed behind the island of Tenedos. Then the only state the Trojans consider possible is state 3, in which the Greeks have gone home. Thus at state 2, the following sentences are true: \mathbf{t} , $B_T\mathbf{h}$. The information structure of the game can be summarized by the following three sentences: $\mathbf{s} \rightarrow B_T\mathbf{s}$, $\mathbf{t} \rightarrow B_T\mathbf{h}$, and $\mathbf{h} \rightarrow B_T\mathbf{h}$. It is easy to check that these sentences are true at *every* state. In particular, they are true at state 2, the true state, and at state 3, the only state the Trojans consider possible. Thus at state 2, the Trojans know that all three sentences are true: $K_T((\mathbf{s} \rightarrow B_T\mathbf{s}) \& (\mathbf{t} \rightarrow B_T\mathbf{h}) \& (\mathbf{h} \rightarrow B_T\mathbf{h}))$. They have been deceived, and yet retain knowledge of the structure of the game. Indeed, they know that they know it, and know that they know that they know it, and so on. Intuitively, although the Trojans know that *if* the Greeks play \mathbf{t} they will be deceived, this does not prevent them from being deceived when it actually happens.

A more detailed investigation of what the players can know and believe about each other and about the structure of the game would require a more sophisticated epistemic model, representing the beliefs of all the players at every stage of the game. Such models can be found in Board [3]. But the current aim is merely to convince the reader that deception is not inconsistent with common knowledge of the game. It is hoped that the toy model above is sufficient for this purpose.

¹Readers familiar with modal logic will recognize that an epistemic model is essentially a Kripke structure, with the history function playing the role of the interpretation; those not are referred to Fagin *et al.* [10] for a very detailed explanation. Stalnaker [18] shows how epistemic models can be used to analyze rational play in games, and to provide a systematic evaluation of game-theoretic solution concepts.

4.2 Deception, lack of introspection, and unawareness

We have shown that it is not incoherent to assume that the Trojans knew the structure of the game and yet were still deceived. But is a deceived player necessarily an irrational one? In this section we show that the answer to this question is no, and distinguish between three forms of bounded rationality.

The claim that the Trojans must be irrational could be based on the following argument: if the Trojans know the structure of the game, then they know that whether the Greeks play **t** or whether they play **h**, they will believe that the Greeks have played **h**. So if they find themselves believing that the Greeks have played **h**, they should remain open to the possibility that the Greeks actually played **t**. But this argument merely denies that the Trojans were deceived, and contradicts the structure of the game. There is nothing irrational in realizing that something could have happened for two reasons, but ruling out the first.

A more subtle argument could be based on the additional premise that the Trojans knew that the Greeks were rational, and that the Greeks knew that the Trojans were rational. If this is so, they should be able to carry out the backward induction argument we used to solve the game in section 2.1, and conclude that the Greeks would play **t**. This is perfectly true, but all it tells us is that the additional premise is inconsistent with node *t* being reached. At this node common knowledge of rationality breaks down². The idea that common knowledge of rationality may not survive along every path through an extensive form game is not a new one, and is discussed in detail by Reny [16] and many others. In games of deception it is possible that common knowledge of rationality cannot survive along *any* path. This is true of *The Trojan War* but not of *Investment Advice*.

There is a sense, however, in which a player who is deceived must be only boundedly rational. If we think of the set of states in an epistemic model as representing every possible contingency in a particular situation, and the accessibility relation as describing what signal a player receives in each state, then a fully rational player should be able to invert that signal to figure out what state it might have come from. This inversion process will generate a new accessibility relation which

²If this feature of *The Trojan War* is thought to be unpalatable, a modification of the game allows the Trojans to retain their knowledge that the Greeks are rational and know that they are rational, even when deceived. Simply add a move by nature to the beginning of the game, according to which it is determined whether the Greeks have a choice between **s**, **t**, and **h** or just a choice between **s** and **h**. The information structure is such that whenever **t** is played or **h** is played in either game, the Trojans believe that **h** is played *in the smaller subgame*. Common knowledge of rationality can survive at this node. Intuitively, the Trojans are unsure whether the Greeks have a trick they can play or not; when the Greeks actually play the trick, they assume that it was not available.

partitions the set of states. If beliefs are defined in the same way as before, then everything this fully rational player believes must be true. Of course people in the real world do have false beliefs: it may be that there are simply too many contingencies for us to be able to consider every one of them. This idea is developed further in the next section when we consider how deception might arise.

Information functions which satisfy conditions (c) and (d) can be represented by standard information partitions. We have seen that relaxing (c) allows us to model deception. Relaxing (d) gives another form of bounded rationality, *lack of introspection*. A player who lacks introspection does not know all of her own beliefs, i.e. there must be something she believes but does not know she believes; or something she does not believe but does not know she does not believe. The link between (d) and introspection follows from a well-known theorem in modal logic (Theorem 3.1.5 in Fagin *et al.* [10]). For the present purposes, it is sufficient to point out that conditions (c) and (d) are logically distinct: a player may be deceived but introspective (as is the case in both of the examples above), or lack introspection even if not deceived.

A final form of bounded rationality is *unawareness*. In many cases, a very plausible explanation for deception is that the deceived was not aware of the possibility of a particular move being made³. To model unawareness we would need a richer framework than that discussed above. The assumption of common knowledge of the game must be relaxed and a distinction made between the actual game that is being played and the game as it appears to each player. A more detailed discussion of the difficulties in modelling unawareness can be found in Dekel *et al.* [9], who show that there is no way of representing a plausible notion of unawareness using standard epistemic models and propose an alternative approach.

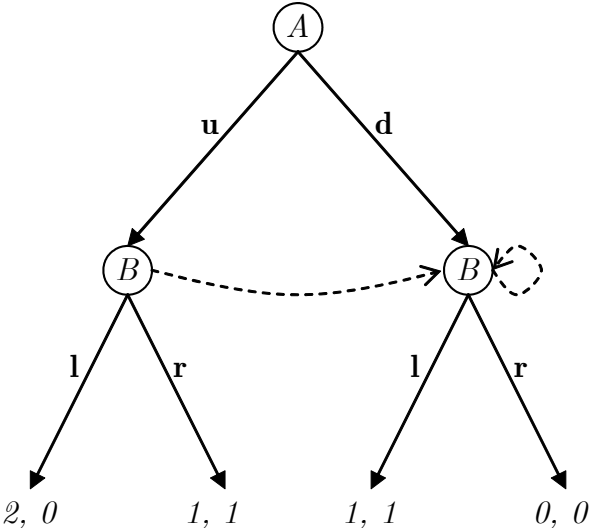
4.3 Exogenous and endogenous mistakes

In section 2.3 we argued that games of deception offer a more reasonable representation of scenarios such as the Trojan war and the Merrill Lynch scandal than can standard extensive form games. We showed that in each case the observed outcome was not an equilibrium of the appropriate standard game, while it was the (unique) backward induction outcome of the game of deception. But in both cases the observed outcome *was* rationalizable in the standard game. We now return to this issue.

³This does not explain why the Trojans were deceived by the Greeks. The Trojans were certainly aware of the existence of the island of Tenedos, and of the fact that the Greeks might trick them (Laocoön made this quite clear to them).

Against the charge that the rationalizing beliefs were rather *ad hoc*, designed specifically to sustain the desired outcome, it could be countered that the same is true of the information structures in the games of deception. Both approaches violate standard equilibrium analysis and make *ad hoc* assumptions about the beliefs of the players. The difference is that in the standard approach, these assumptions are endogenous, part of the solution concept applied to the game; and in the new approach, they are exogenous, imposed by the information structure. Viewed in these terms the advantages of the new approach are not clear.

To defend ourselves against this critique we make two points. The first is that the solution to a game of deception is not always a rationalizable outcome in the corresponding standard game. Consider for example the following game:



Simple Deceit

This game has a unique backward induction solution: (\mathbf{u}, \mathbf{l}) . But consider the standard version of this game, where player B 's information is given by $\mathcal{I}(u) = \mathcal{I}(d) = \{u, d\}$. This game is dominance solvable and thus has a unique rationalizable outcome: (\mathbf{u}, \mathbf{r}) . The rationalizable outcome in the standard game and the backward induction outcome in the game of deception do not coincide. Of course there are beliefs that are compatible with the standard game which can explain the outcome (\mathbf{u}, \mathbf{l}) : precisely the same beliefs which are imposed by the information structure of the game of deception will serve this purpose. But rationalizability is the weakest solution concept that is commonly used in economic applications of game theory. If the observed outcome is not rationalizable, perhaps we have the wrong model.

The second point is a methodological one. In standard game theory, the structure of a game

specifies four elements⁴:

- (i) the players (who is involved?);
- (ii) the rules (who moves when? what do they know when they move? what can they do?);
- (iii) the outcomes (for each possible set of actions, what is the outcome of the game?);
- (iv) the payoffs (what are the players' preferences over the possible outcomes?).

These are the exogenous parameters. The actual actions chosen by the players, and their beliefs about the choices of others not already determined by the rules, are endogenous variables to be explained or predicted by the application of some solution concept. The issue is whether the fact that a player is deceived is thought of more properly as part of the rules of the game or as part of the solution concept. Here we argue for the former. Even in standard games the rules describe each player's beliefs about past moves. It is a very natural extension to allow these beliefs to be false and accept that the fact of deception is part of the description of the situation rather than a facet of the way the game is played.

4.4 How and why does deception occur?

We have so far begged the questions of how and why a player might be deceived. Games of deception encode into the information structure itself the fact that deception occurs. We have argued that these games provide a more satisfactory model of certain situations than games with standard information structures. But the fact that deception occurs is a modelling assumption, and not something derived endogenously. It is therefore important to understand when it might be a reasonable assumption. This is perhaps as much a question for the psychologist as the economist, but we offer two suggestions.

The first is that in the absence of any contradictory evidence, people tend to take their observations at face value. Thus when the Trojans see the Greeks abandon their camp and sail away out of sight, it is natural for them to believe that they have gone home. Similarly, when a bank tells someone that a stock is good, it is natural for them to believe that the stock really is good. Brams [5] adopts this line of argument when he writes "Since Deceived is in possession of no information — in particular, information that would conflict with Deceiver's announcement — there is no reason

⁴This list is borrowed from Mas-Colell *et al.* [13].

for him not to believe Deceiver". (This work is discussed in section 5.) Of course in the cases discussed above it could be argued that there *are* good reasons to disbelieve ones eyes and ears. Deception will fail if these reasons outweigh the natural tendency to believe.

The second suggestion is based on complexity considerations and the boundedness of human minds. There are many possible explanations for the epistemic inputs a particular individual receives, and typically only a small subset can be considered. If this subset excludes certain actions that might be made by others, then the individual will form mistaken beliefs if these actions are taken. In the context of *The Trojan War*, it is plausible that although the Trojans knew the Greeks might play a trick on them, they did not know the exact form the trick might take: they were not able to consider every possible trick, and in particular, they did not think about the Greeks sailing behind the island. This idea can also explain why people are not usually taken in by the same trick twice, and can shed light on the nature of recent corporate deceptions such as the WorldCom and Enron scandals, in which complex networks of companies were set up to hide costs and losses. If this line of argument is taken, our games of deception could be thought of as simplifications of a more complex games in which some players are unaware of some moves. This simplification allows us to model the essential features of a given situation while retaining the standard game-theoretic assumption of common knowledge of the game.

4.5 Solution concepts for games of deception

The two examples we considered in section 2 were simple enough for backward induction to yield unique solutions. But in more complex games this will not be the case, and it is important to discuss what solution concepts might be appropriate in the general case. An obvious proposal is to use sequential equilibrium, with strategies defined as functions from information sets to actions in the normal way. But there is a problem with the interpretation of the consistency requirement of sequential equilibrium in games of deception. In *The Trojan War*, for example, the Trojans have a dominant strategy of (**c** if $\{s\}$; **o** if $\{h\}$), and the Greeks' best response to this strategy is to play **t**. Thus these strategies must be played in any equilibrium. Yet the information structure of the game dictates that the Trojans must believe that they are at node h after the Greeks have played **t**. Although this belief is consistent in the formal sense (since given *any* strictly mixed strategy for the Greeks, Bayesian updating will assign a probability of one to node h , the only node in the Trojans' information set), it is not in the spirit of equilibrium analysis, which assumes that everyone knows what everyone else is doing. Of course there can be no equilibrium in this sense, since the very

nature of deception is that the deceived player does not know what the deceiver is doing! We do not however advocate rejecting the use of sequential equilibrium in games of deception; we merely wish to point out that careful analysis is required to clarify its implications. The framework developed by Board [3] provides an ideal tool for this purpose.

5 Related literature

In the economics literature, discussions of deception can be divided into three main strands. In the first we find *cheap talk* games, introduced by Crawford and Sobel [8]. In these games players communicate by means of costless messages, and can lie in order to gain strategic advantage. But deception can never succeed in equilibrium, where the sender's strategy is assumed to be known, and any potentially harmful message will be ignored.

The second strand is represented by Sobel [17] and Benabou and Laroque [2], among others. Again messages are costless to send, but here deception can succeed because of the existence of *honest* types (who never lie) alongside the standard *opportunistic* types (who can lie). If the proportion of honest types is high enough, it may be worthwhile to believe messages even though there is a chance they could be deceptive.

Crawford [7] is an example of the final strand. Here the messages are no longer cheap talk (and so could be interpreted as actions of any form rather than just statements) and no-one is inherently honest, but successful deception can take place because there are *mortal* as well as *sophisticated* players. In equilibrium, sophisticated players are assumed to know each other's strategies as usual, but mortal players can have arbitrary beliefs about what everyone else is doing. The upshot is that even sophisticated players can deceive each other, if each thinks it sufficiently likely that she is facing an mortal opponent.

While there is much to be learned from all of these stories, we think it is an advantage of our approach that deception can be modeled in a parsimonious framework, without recourse to artificial devices such as hypothetical types of player.

More closely related to the current project is work of the political scientist Brams ([5], [6] and elsewhere). Brams models deception in normal form games, and assumes that it takes the form of a misrepresentation of preferences by one agent (Deceiver) to another (Deceived). More precisely, he assumes that "Deceived has no *a priori* information about the preferences of Deceiver. . . [and] Deceived believes Deceiver's announcement of his preferences (true or misrepresented), and

Deceived knows that he does” (Brams [5]). He goes on to show that 33 of the 78 2×2 games⁵ are *deception-vulnerable*, in the sense that Deceiver can obtain a better outcome if he misrepresents his preferences than if he reports them truthfully. A distinction is made between *tacit* deception, when Deceiver’s strategy choice does not reveal his deceit (i.e. his choice is consistent with his stated preferences), and *revealed* deception, when it does. Brams provides a detailed analysis of the Cuban Missile Crisis within this framework.

The generalization of extensive form games described in section 3 provides a way to formalize Brams’ assumption that statements about preferences are always believed. Uncertainty about Deceiver’s preferences can be represented by a move by Nature at the start of the game, in which the true preferences are chosen. Deceived does not observe this move, but does observe a statement made Deceiver, and the information structure is such that only subgames consistent with Deceiver’s statement are considered possible by Deceived. Whenever that statement is false, this will not include the actual subgame, hence the information function will not satisfy condition (c). Note that in Brams’ framework deceptive actions are always cheap talk: the original statement by Deceiver does not affect actual payoffs. But it does affect Deceiver’s beliefs about these payoffs, and so deception is possible even in equilibrium notwithstanding the results of Crawford and Sobel [8].

6 Conclusions

There can be little doubt that deception is an important feature of strategic interaction. The proliferation of corporate scandals at some of America’s highest-profile firms in recent months (including Enron, Global Crossing, Tyco, Qwest and WorldCom) has prompted George W. Bush to accuse executives of “breaching trust and abusing power”, and he has pledged to “end the days of cooking the books, shading the truth and breaking our laws”. But deceit is by no means a new phenomenon in the financial world. Benabou and Laroque [2] tell a story of the banker Nathan Rothschild. Rothschild had a network of carrier pigeons which gave him superior information from France, and in 1815 during the battle of Waterloo he walked around the city of London looking dejected, spreading the news that the battle was going badly, and arranging for his agents to make a public display of selling British government securities. At the same time Rothschild was secretly

⁵There are 78 combinations of pairs of strict preferences orderings over the four outcomes of a 2×2 game, modulo permutations of player and strategy labels.

buying much larger quantities of these securities at the depressed price, waiting for the time when news of the victory would finally reach the masses.

In the military world, the Greeks set a trend that has lasted until the present day. Herodotus [11] describes how Zopyrus mutilated himself, cutting off his nose and ears, in order to convince the Babylonians that he was a deserter from the Persian army; the lies he told facilitated the Persian capture of Babylon, and Zopyrus was made Governor as a reward. More recent examples include the Allied invasion of Normandy on D-Day, June 6 1944, after a feint at Calais had convinced the Germans they would land there (see Kemp [12]); and the misrepresentation of American preferences by John F. Kennedy during the Cuban Missile Crisis in 1962 (analyzed by Brams [5]). Political life is a rich source of further examples: George Bush Senior's 1988 campaign promise, "Read my lips: no new taxes" is one of the more obvious.

The aim of this paper has been to argue that the standard information structure of extensive form games is not able to capture the notion of deception, and to show how replacing information partitions with the more general information functions can provide a solution to this problem. We conclude with a brief discussion of several alternative motivations for considering for this generalization.

Bonanno [4] gives the example (which he attributes to van Bentham) of an individual sitting in a bar who correctly believes that if he has a drink it will be unsafe to drive. After drinking, however, he becomes more confident and believes it is safe to drive. His mistaken belief is the result of alcoholic confusion and not active deception by an opponent, but the implications for the information function are the same: condition (c) must be relaxed and therefore no partitional representation is possible. Absent-mindedness can provide a reason to relax condition (d). Aumann *et al.* [1] consider a more complex version of the absent-minded driver problem in which the driver has three junctions to contend with rather than the usual two. In that paper it is assumed that the driver cannot tell at all which of the three junctions he is at, so all three are contained in a single information set. But it not implausible to suppose that he might remember passing at least one junction when he is at the third; and when he is at the first, he might be sure that he has not passed as many as two: he knows where he is on the road to within plus or minus one junction. This generates three distinct and overlapping information sets, which can be represented by an information function which satisfies condition (c) but not condition (d). As discussed in section 4.2, this player displays a lack of introspection, i.e. he does not know all of his own beliefs. If he did, assuming he also knows the structure of the game, he could invert the information function

and figure out exactly where he was. These additional examples confirm the importance of the information function approach proposed in this paper.

References

- [1] AUMANN, R. J., S. HART, & M. PERRY (1997), “The Absent-Minded Driver”, *Games and Economic Behavior* **20**, 102–116.
- [2] BENABOU, R. & G. LAROQUE (1992), “Using Privileged Information to Manipulate Markets: Insiders, Gurus, and Credibility”, *Quarterly Journal of Economics* **107**, 921–958.
- [3] BOARD, O. J. (1998), “Belief Revision and Rationalizability”, *Theoretical Aspects of Rationality and Knowledge*, Proceedings of the Seventh Conference, ed. by I. Gilboa.
- [4] BONANNO, G. (2002), “Memory of Past Beliefs and Actions”, Working Paper, Department of Economics, University of California, Davis.
- [5] BRAMS, S. J. (1977), “Deception in 2×2 Games”, *Journal of Peace Science* **2**, 171–203.
- [6] BRAMS, S. J. (1994), *Theory of Moves*. Cambridge University Press, Cambridge.
- [7] CRAWFORD, V. P. (2001), “Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions”. Discussion Paper 2001-16, University of California, San Diego.
- [8] CRAWFORD, V. P. & J. SOBEL (1982), “Strategic Information Transmission”, *Econometrica* **50**, 1431–1451.
- [9] DEKEL, E., B. LIPMAN, AND A. RUSTICHINI (1998), “Standard State-Space Models preclude Unawareness”, *Econometrica* **66**, 159–173.
- [10] FAGIN, R., J. Y. HALPERN, Y. MOSES, AND M. Y. VARDI (1995), *Reasoning About Knowledge*. The MIT Press, Cambridge, MA.
- [11] HERODOTUS (1954), *The Histories*, translated by A. de Sélincourt. Penguin Books, Harmondsworth, Middlesex, England.
- [12] KEMP, A. (1994), *D-Day and the Invasion of Normandy*. Harry N. Abrams, New York.

- [13] MAS-COLELL, A., M. D. WHINSTON AND J. R. GREEN (1995), *Microeconomic Theory*. Oxford University Press, Oxford, England.
- [14] OSBORNE, M. J. & A. RUBINSTEIN (1994), *A Course in Game Theory*. The MIT Press, Cambridge, MA.
- [15] PICCIONE, M. AND A. RUBINSTEIN (1997), “On the Interpretation of Decision Problems with Imperfect Recall”, *Games and Economic Behavior* **20**, 3–24.
- [16] RENY, P. J. (1992), “Rationality in Extensive-Form Games”, *Journal of Economics Perspectives* **6** 103–118.
- [17] SOBEL, J. (1985), “A Theory of Credibility”, *Review of Economic Studies* **52**, 557–573.
- [18] STALNAKER, R. (1994), “On the Evaluation of Solution Concepts”, *Theory and Decision* **37**, 49–73.
- [19] VERGIL (1956), *The Aeneid*, translated by W. F. J. Knight. Penguin Books, London, England.