Toward a Unified Model of Deception

Donald D. Rose

Irvine Computational Intelligence Project
Department of Information and Computer Science
University of California, Irvine, 92717

Abstract

We will first argue that ignoring possible deception in multi-agent scenarios can lead to planning failures; specifically, we show how standard deduction may be able to solve the Wise Man Problem, but not a variant where some agents are deceptive (i.e., the Wise-Yet-Deceitful Man Problem, or W-Y-D). Second, we will show how to avoid planning failures in scenarios such as W-Y-D, by developing models of both (1) the deceptive tendencies of other agents, and (2) how these other agents *themselves* reason about deception; the concepts of *best-case* and *worst-case* deceptive agents will be introduced as examples. Third, we will show how to represent *deception axioms* within Konolige's deduction model of belief, and the fourth section will more closely analyze how one solves W-Y-D. Finally, we will suggest how the new model developed for this problem can be generalized into a more unified model of deception.

Introduction

This paper deals with increasing the sophistication with which agents reason about *other* agents' beliefs, as well as their own. The particular enhancement of interest here is allowing agents to plan not only for cases where an agent is truthful, but also for cases where he is deceitful (to others, to himself, or to both). Thus, we will relax the assumption that what an agent *says* he believes always equals what he is *actually* believing (i.e., if other agents are fooled by agent *d*, then *d*'s beliefs will not be the same as the others' beliefs *about d*'s beliefs). We begin by illustrating the Wise Man Problem, considered "a good test of the *competence* of any model of belief" [KONOLIGE, 84]. Using W, B and U to represent the possible responses (i.e., black, white, or don't know), here is:

THE WISE MAN PROBLEM. *A king, wishing to know which of his three advisors is the wisest, paints a white dot on each of their foreheads, tells them there is at least one white dot, and asks them to tell the color of their own spots. After a while the first replies U; the second, on hearing this, also replies U. The third then responds W.*

The problem is whether No. 3 can ascertain his color, based solely on seeing the others' dots, and *reasoning about their beliefs.* His reasoning proceeds as follows: "Suppose my spot were black. Then No. 2 would know that his own spot was white (since, if it were black, the first of us would have seen two black spots and thus would have known his own spot's color). Since both answered U, my spot must be white" [KONOLIGE, 84]. Note, however, that No. 3 did not consider the possibility that No. 1 or No. 2 (or both) might give a response that they *did not believe themselves.* Nor does No. 3 address a more subtle question: whether or not *No. 2 plans for deception* as well. Thus, the standard Wise Man Problem led to a correct answer by No. 3. But suppose the following scenario ensued:

THE WISE-YET-DECEITFUL MAN PROBLEM. *The King, after No. 3's success, decides to test No. 3's abilities further. He replaces the other wise men with two evil men and repaints all the dots. The first man, experienced in deceit, gets white; his protege, a naive deceiver, gets black. The third again gets white. The same rules apply: there's at least one white dot, and each must tell the King his color. (The King warns the third of the others' deceptive abilities; the third dismisses this since no one has ever lied before the King.) The first replies W; his protege says B. Then the third man incorrectly responds B. The King banishes him from his court.*

This sad ending could have been avoided if the warning had been properly reasoned with. If No. 3 had codified the other men's deceptive abilities into rules, then added these to his existing rules about how he and the others draw inferences, this planning failure would not have happened. We now look at why No. 3 gave an incorrect answer, and how his reasoning should proceed in order to solve this new problem.

Solving the Wise-Yet-Deceitful Man Problem

Let us use Pi to represent the proposition "*i* has a white dot" (and ¬Pi for black). Now, the way No. 3 reasoned in trying to solve our new problem was the same as for the standard Wise Man Problem. Recalling initial configuration (P1 ∧ ¬P2 ∧ P3), No. 3 reasons: "Suppose I was white. Then No. 1's response would have been U, because he could not have known his color unless No. 2 and I were both black. Since he said W, I must be black." Note that No. 2's response seems to further confirm No. 3's reasoning. Upon seeing No. 3 is black, and hearing No. 1 say W, No. 2 would immediately conclude he was black (again, since seeing two black dots should make No. 1 say W). Since No. 2 *did* say B, No. 3 might feel secure in the conclusion that he's black.

The dilemma here is that agents No. 1 and No. 2 are modelled as completely truthful. However, in real-world environments, agents often interract with other agents that not only might possibly lie, but can often hide such deception. The types of behavior we need to account for here are (1) whether or not an agent is being deceitful, and (2) whether or not that agent himself considers that others may try to deceive *him*. So how does No. 3 solve the Wise-Yet-Deceitful Man Problem? First, we must take into account the King's warning about No. 1 and No. 2. Thus No. 3 will model No. 1 as a *best-case* deceptive agent - i.e., No. 1 may or may not lie, and he believes others may or may not do the same. No. 2 will be modelled as a *worst-case* deceptive agent - i.e., No. 2 *always* lies, yet believes other agents always tell the truth. Finally, everyone knows these traits of both No. 1 and No. 2. Thus, one can see why No. 2's is "worst-case"; in a deceptive world, he's infinitely gullible, yet his infinite deceit is always recognized. No. 1, however, trusts no one (and, since everyone knows it, others think twice before trying to deceive him); in addition, everyone knows his responses are unpredictable (i.e., he may or may not be lying). Thus No. 1 is in the "best" position to capitalise in a deceptive world.

Here is how No. 3 should solve the Wise-Yet-Deceitful Man Problem (W-Y-D). Remembering the initial dot configuration (P1 $\wedge$ $\neg$P2 $\wedge$ P3), a more wary No. 3 reasons: "There are only two possible assumptions: either *No. 1 is truthful and I'm black*, or *No. 1 is lying and I'm white* (see informal proof section for why this is so). Suppose the first assumption holds. Then No. 2, who always believes everyone is truthful, would believe No. 1 when he said W; hence *No. 2 believes he's black* (since seeing two black dots would, in No. 2's view, lead to No. 1's W response). Thus I know that No. 2 would've said anything *except* B at this point, since everyone knows No. 2 always lies. However, No. 2 *did* say B; hence *No. 2 must not believe he's black*. Thus, my first assumption is false, and so the second must be true. Hence, I tell the King I'm W, I expose the fact that No. 1 lied before his majesty – and I keep my job." (Note that telling the King about No. 2's lie is no revelation, since everyone already knows No. 2 always lies.)

Representing Axioms of Deception

The next step is to formally represent axioms that model *all* forms of reasoning an agent may go through in a possibly deceptive environment (i.e., the model should account for any response, and any belief about the truthfulness of the agent responding). In addition to these axioms will be axioms for modelling *specific* agents (e.g. No. 2). All axioms are based on Konolige's deduction model of belief [KONOLIGE, 84]. First, some notation: the *belief operator* $[S_i]$ is used to indicate whether agent i believes a certain proposition. For example, [S3]P3 says that agent No. 3 believes the proposition P3 ("No. 3's dot is white"); [S3]$\neg$P3 says he believes $\neg$P3 ("No. 3's dot is black"). In short, $[S_i]x$ is true if x is in i's belief set, for any proposition or belief x [KONOLIGE, 84].

Now, the final W-Y-D conclusion was that No. 3 believed not only that his dot was white, but also that No. 1 was lying; if L_i is the proposition "agent i lied", then No. 3's final conclusion would be stated as [S3](P3 $\wedge$ L1). In short, the simplest kind of proposition has no belief operator; more complex propositions start with a belief operator (indicating who holds the proposition), and may or may not have belief operators after it, depending on the degree of nesting being represented. (A final note: [S0] means that whatever follows it is a "common belief", and would be held by any agent). Thus, the first four axioms we desire should capture the most basic elements of our W-Y-D scenario: (1) No. 2's dot is black, the others are white; (2) it's a common belief that there's at least one white dot; (3) when an agent has a white dot, all others know it's white (and this rule itself is a common belief); and (4) is the same as (3), but for black dots. Using the formal language of the deduction model of belief, we have:

W1 P1 $\wedge$ $\neg$P2 $\wedge$ P3

W2 [S0](P1 $\vee$ P2 $\vee$ P3)

W3 $(P_i \supset [S_j]P_i) \wedge [S0](P_i \supset [S_j]P_i)$ $\quad i \neq j,\ j \neq 0$

W4 $(\neg P_i \supset [S_j]\neg P_i) \wedge [S0](\neg P_i \supset [S_j]\neg P_i)$ $\quad i \neq j,\ j \neq 0.$

In a world where deception is *not* modelled, the next actions taken would occur after each agent's response, when axioms are asserted stating that (1) the agent believes what he said, and (2) all other agents believed him as well. For example, if No. 1 had responded U ("I don't know") in our problem (where his dot happens to be W - i.e., where P1 is true), the non-deceptive model creates:

W5 $\neg$[S1]P1 $\wedge$ [S0]$\neg$[S1]P1.

This reads: "it is *not* the case that No. 1 believes P1, and it is a common belief that this is so." This is exactly the axiom created when the deduction model is used to solve Konolige's variation of the Wise Man Problem (the Not-So-Wise Man problem, mentioned later; see [KONOLIGE, 84], p. 48). However, this is exactly the type of axiom we must *avoid* if we start planning for the possibility of dishonest responses; *it should not be a universal given that what an agent believes is also believed by everyone else.* Thus, instead of asserting this **W5** axiom (and a similar **W6** after No. 2 responds), we add to the initial axioms six *deception axioms.* Each has the same theme concerning a possibly-deceptive agent i: depending on what i said (and whether or not agent j believes i is lying), j will make some deduction about *what agent i believes about his own dot's color.* For example, if j believes i is truthful when he says W, then j believes that i believes he's white (**W5'** below). In **W10**, however, j believes i is lying when he says he doesn't know his color (i.e., U); thus, j believes i actually *either believes he's white, or believes he's black.* The first three axioms are of interest when j believes i is *not* lying; the last three for when j feels i *is* lying. Remembering that Wi being true means that "i said W", etc., the new axioms are:

W5' $([S_j](\neg L_i \wedge W_i) \supset [S_j][S_i]P_i) \wedge [S0](...)$

W6' $([S_j](\neg L_i \wedge B_i) \supset [S_j][S_i]\neg P_i) \wedge [S0](...)$

W7 $([S_j](\neg L_i \wedge U_i) \supset [S_j](\neg[S_i]P_i \wedge \neg[S_i]\neg P_i)) \wedge [S0](...)$

W8 $([S_j](L_i \wedge W_i) \supset [S_j]\neg[S_i]P_i) \wedge [S0](...)$

W9 $([S_j](L_i \wedge B_i) \supset [S_j]\neg[S_i]\neg P_i) \wedge [S0](...)$

W10 $([S_j](L_i \wedge U_i) \supset [S_j]([S_i]P_i \vee [S_i]\neg P_i)) \wedge [S0](...)$.

([S0](...) abbreviates the fact that the left side of the conjunction is a common belief.) Now, since we assume all six axioms are known to all agents, how can we model an incomplete (less-than-ideal) agent - one who is not cognizant of possible deception - if deception abilities are common beliefs? There are actually two ways in which to model an agent who is naive about deception. First, the agent's rule set may be incomplete because he has left certain deception axioms out of his set of "relevant problem-solving information". Thus, the heart of the axiom, as well as the fact it is common knowledge, would *never be used in the agent's reasoning.* This behavior is called *circumscriptive ignorance* [KONOLIGE, 82]. In short, one way an agent might not plan for possible deception is to exhibit *relevance incompleteness* - by ignoring axioms that are essential to solving a problem, an agent may become ignorant of some of the logical consequences of his beliefs. However, we model naive deceptive agents a second way; we allow all agents to have, and use, the six deception axioms, but define two specific axioms which model No. 2's specific behavior. The first axiom captures No. 2's belief that no agent ever lies; the second states that everyone knows No. 2 always lies:

W11 $([S2]\neg L_i) \wedge [S0]([S2]\neg L_i) \quad i \neq 2$

W12 [S0]L2.

This approach is useful for agents who must reason about No. 2's beliefs: if No. 3 desires, he can always perform voluntary circumscription and *ignore* **W11** (and speculate: "perhaps No. 2 isn't completely gullible"), or can ignore **W12** (i.e., "perhaps No. 2 doesn't always lie"). In the former case, No. 2 would (in No. 3's view) start "recognizing" rules **W8-W10**, because (to No. 2) the possibility of Li (i.e., "agent i lied") now exists. Such circumscriptive ignorance by No. 3 would most likely be done if an answer was not found by other means first; in the W-Y-D scenario, such a step was not necessary.

One final step is needed to model No. 3's reasoning in the W-Y-D scenario: representing the axioms constructed after No. 1, then No. 2, give their responses. After No. 1 looks and sees No. 2 is black and No. 3 is white, he does not know his color (white); however, he lies. Although both these facts are known only to No. 1 (**W13**), what No. 1 *says* is a common belief (**W14**):

W13 $\neg[S1]P1 \wedge L1$

W14 $W1 \wedge [S0]W1$.

Thus, all agents except No. 1 will not have **W13** in their reasoning system. (Remember that our aim was to distinguish between what agents *say* they believe, and what their actual beliefs are; these two axioms do just that.) Now, No. 2, who blindly believes that No. 1 knows he's white, cannot get any information from No. 1's response. Thus, upon seeing two white dots, No. 2 concludes he doesn't know his color (which is black). However, he always lies, so he cannot say U; thus he says B (although a W response would still fit his deceptive behavior pattern):

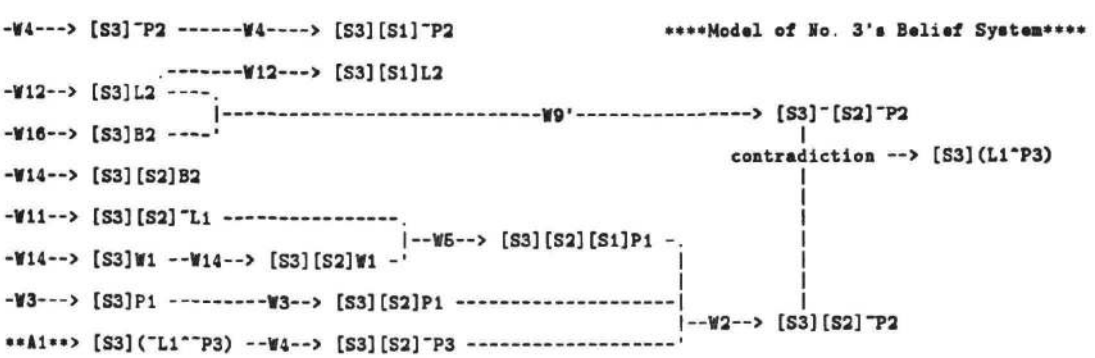
W15 $\neg[S2]\neg P2 \wedge L2$

W16 $B2 \wedge [S0]B2$.

An Informal Proof of the W-Y-D Solution

Let us further illuminate No. 3's reasoning. The diagram shows how No. 3 uses his axioms, his sense (and "given") beliefs, and his beliefs about what *others* believe in order to solve the W-Y-D problem. Each of No. 3's beliefs are generated from one of No. 3's 16 axioms, which completely model what he sees and thinks about other agents (e.g. No. 3 knows No. 2 is lying ($[S3]L2$) from **W12** (which states "it is a common belief that No. 2 always lies".) Note first how No. 3 deduces there are only two possible assumptions. After No. 1's response of W, No. 3 reasons about the four combinations between No. 1 possibly lying, and No. 3's possible colors. ($L1 \wedge P3$): possible (if No. 1 lied, No. 3 would have to be white, assuming no other incompleteness on No. 1's part); ($\neg L1 \wedge \neg P3$): possible (if honest, No. 1 *would* say W upon seeing two black dots); ($L1 \wedge \neg P3$): impossible (saying W upon seeing two Bs isn't lying); and ($\neg L1 \wedge P3$): impossible (if No. 3 is white, there's no way No. 1 can deduce he himself is white; thus, if he said W, No. 1 would be lying). Since only the first two options are possible, we prove the first if the second leads to a contradiction:

"First I assume my dot is black, and that No. 1 told the truth ($[S3](\neg L1 \wedge \neg P3)$). Since No. 1 said he's W ($[S3]W1$), it follows No. 2 heard this ($[S3][S2]W1$). Since No. 2 believes everyone is honest, he believes that *No. 1* actually believes he is white ($[S3][S2][S1]P1$). Now, if my dot is black ($[S3]\neg P3$) then No. 2 saw I am black ($[S3][S2]\neg P3$). Also, No. 2 and I see No. 1 is white ($[S3]P1$ and $[S3][S2]P1$). These last three beliefs about No. 2's beliefs (i.e., $[S3][S2]...$) lead to the conclusion that No. 2 believes he is black ($[S3][S2]\neg P2$). This makes sense: if No. 2 believes that I'm black, that No. 1's white, and that No. 1 actually believes he's white - then No. 2 must now believe he's black (see bottom of "contradiction"). Continuing: since I believe that No. 2 always lies ($[S3]L2$), and I heard his response of B ($[S3]B2$), deception axiom **W9** leads me to conclude that No. 2 *cannot* be holding the belief that he's black. In other words, rules **W8-W10** tell me that when someone lies, I must deduce that he actually believes the *negation* of what he *said* he believed. In this case, I thus believe the negation of $[S2]\neg P2$ ($[S3]\neg[S2]\neg P2$) (see top of contradiction). Since a contradiction has been found, my initial assumption must be false, so I deduce the "opposite" of $(\neg L1 \wedge \neg P3)$" - i.e., $[S3](L1 \wedge P3)$.



The Four Classes of Deceptive Behavior

Up to this point, our discussion has focused on generalizing the deduction model of belief to allow for the case where one or more agents exhibit deceptive behavior (making other agents believe what is not true). The behavior discussed so far we call *intentional deception* (or "misleading" behavior), because one deliberately attempts to mislead others. We propose generalizing this notion of deception into a larger framework, adding three other basic types of deception: *unintentional deception* ("misinterpreted" behavior); *unintentional self-deception* ("naive" behavior); and *intentional self-deception* ("irrational" behavior). Our first example illustrates intentional deception:

I went out with Agnes because I wanted Linda to get jealous and thus bake me a cake. Linda thought I was losing interest, so when I came home there was a big cake on the table. It worked.

Here, a deceiver *deliberately tries* to make Linda think he used a certain line of reasoning (e.g. "I'm fed up with Linda so I'm going to start dating other girls"), when in fact such reasoning was not used at all by the deceiver. Yet a slight variation results in unintentional behavior:

I went out with Agnes, a cousin I had not seen in years. Linda thought I was losing interest, so when I came home there was a big cake on the table. I was totally surprised.

Here, Linda assumed the "deceiver" was using the same line of reasoning laid out above, and so took the same action. The difference is that there is *no intent* to deceive here; Linda misinterpreted her boyfriend's behavior as loss of interest. Thus, the most basic forms of deception are (1) intentional: one tries to fool another and succeeds; and (2) unintentional: one does *not* try to fool another, yet succeeds in doing so anyway. The last two forms involve the notion of *deceiving oneself.* In unintentional self-deception, one doesn't try to fool oneself, yet (unfortunately) succeeds:

I did not know the bear's growling was relevant to my petting him, so I petted him. He bit off my hand.

Thus the classification "naive behavior". Another explanation is that one holds a belief, even though one's belief system would deduce the opposite. The key is that you are not cognizant of this latter fact, and hence *not aware that your behavior was illogical* - i.e., you don't know where your reasoning was naive. However, even if one *is* aware of certain deduced beliefs, one might still continue to believe the opposite:

I knew *the bear's growling meant he was angry, but I petted him anyway. He bit off my hand.*

Why one would try to *deliberately deceive oneself* seems difficult to understand. One explanation: *emotion* may interfere on the logical reasoning process; someone wanting to pet a bear may love bears so much that rational danger signs are acknowledged, but disobeyed. Fear can also interfere with rational behavior; one may *know* that the odds of accident are less in planes than in cars, yet still refuse to take a plane for no valid, logical reason. In short, anyone who has performed an activity even though "logic dictated otherwise" has exhibited intentional self-deception. Agents do not always believe deductions that they make and are aware of, and often accept the opposite without the need for justification. People can go even further and "rationalize" their irrational behavior - acknowledging *reasons* for non-justified beliefs *that do not actually exist* in their belief system. These four proposed classes merit closer analysis than this paper provides; we claim all four can be modelled by slightly modifying the basic deception axioms.

Conclusion

This paper has four main conclusions. First, we found that being able to model possible deceptive behavior is an essential ability for problem-solving agents, and constructed an example where reasoning about deception was the only way to solve the problem. Specifically, in our variant of the Wise Man Problem (the Wise-Yet-Deceitful Man Problem) we showed how reasoning about the abilities of "worst-case" and "best-case" deceptive agents allowed agent No. 3 to deduce not only that his color was white, but that agent No. 1 *must* have been lying. Second, we showed how fundamental deception axioms could be represented in Konolige's deduction model of belief, and how reasoning with these axioms can model different types of deceptive behavior. Third, we suggested a taxonomy of deceptive behavior, where standard intentional deception is shown to be related to three other types: unintentional deception, unintentional self-deception ("naivete"), and intentional self-deception ("irrationality"). The final conclusion is that Konolige's notion of incompleteness, and our modelling of different facets of deceptive behavior, are actually quite closely related. We've seen that when an agent does not plan for possible deception by other agents, he is exhibiting a form of *relevance incompleteness* (i.e., ignoring deception axioms can lead to incorrect deductions about other agents). Yet, reversing our point of view, we see that when an agent exhibits non-deceptive incompleteness, another agent (e.g. No. 3) *viewing* this behavior is often deceived if such incompleteness is not explicitly revealed (and, in real-world situations, it often is not). Thus, being fooled because another agent exhibits incompleteness is an instance of unintentional deception (because another agent did not use lines of reasoning *you assumed he was capable of using*). This paper, on the other hand, has focused on how to model both intentional and unintentional deception with explicit axioms. Thus, we feel that adding such explicit deception axioms to Konolige's paradigm, plus using his model's ability to reason with incomplete agents, will result in a more unified model of belief in general.

References

Konolige, K. (1982) Circumscriptive Ignorance. *Proceedings of the Second National Conference on Artificial Intelligence*, Carnegie-Mellon University, Pittsburgh, PA.

Konolige, K. (1984) Belief and Incompleteness. *Center for the Study of Language and Information Report No. CSLI-84-4*, Stanford University, Stanford CA.