A Model of Assertion Evaluation

Dr Antoni Diller
School of Computer Science
University of Birmingham
Birmingham
B15 2TT
England

A.R.Diller@cs.bham.ac.uk

November 2002

Abstract

In this paper a model of belief-acquisition is presented which encorporates an element that deals with the ways in which people obtain information by accepting other people's assertions. I focus on this component in so much detail that, without too much extra work, a computer program could be written to implement it.

1 Introduction

A substantial number of scientists and engineers in Japan, the United States and various European Countries are putting a great deal of effort into the task of designing and building an android. The ultimate goal of all this research is to manufacture a machine which looks and behaves like a human being and which can interact meaningfully with human beings. Although we are still a long way from achieving this goal, some progress has already been made. Roboticists have managed to build androids which are able to walk on two legs, to climb stairs, to grasp objects without breaking them, to imitate what they see human beings doing, to recognise various physical objects and so on. For example, Rodney Brooks's robot Cog can identify various sorts of physical object [8, pp. 58–65]. It can also distinguish between living and non-living things and it can imitate what it sees people doing. At present, Cog is just a torso with mechanical arms and a movable head with video camers for eyes, but the intention is to eventually add mechanical legs.

One of the most advanced Japanese androids is the Honda P3 which was designed and built at Honda's Wako Research and Development Laboratory by a team led by Masato Hirose [8, pp. 43–45]. This looks like a man wearing a spacesuit, but it is made entirely out of electronic and mechanical components. The P3 is able to walk, to climb and descend stairs and to open doors.

It is hard to build robots that can do these things and I have no desire to belittle the considerable scientific achievements that have already been made. Unfortunately, without significantly more research in the area of belief-acquisition, I feel that, despite their technological sophistication, none of these high-profile projects will succeed in producing an android that can interact meaningfully with a human being and engage in intelligent conversation with that person. The reason for this is that the people involved in these projects have certain mistaken assumptions about knowledge. Most of the scientists involved in building androids assume that perception is fundamental to learning about the world. John Pollock speaks for them when he writes [9, p. 52], 'The starting point for belief formation is perception. Perception is a causal process that produces beliefs about an agent's surroundings.' I do not deny that we obtain some of our beliefs by means of perception, but the overwhelming majority of our beliefs are acquired by accepting what other people say and what they have written. I have argued for this elsewhere [5], but here I will present another argument.

A common rationale given for undertaking research into android construction is that the machines that are built will be able to act as carers for people who are unable to look after themselves. Imagine an infirm and housebound elderly person being looked after by a mechanical carer. The android would be required to carry out many tasks. For example, it would have to clean and tidy the infirm person's house, it would need to pick up objects that the elderly person could not pick up themselves and it would have to make meals for its elderly charge. These activities require the android to be able to perceive its environment and act appropriately. As well as being able to do these things, however, the android would also have to have the ability to evaluate the assertions that it heard and act on them if appropriate. For example, imagine that one day the doorbell rings and the android answers the door. It sees a man who says, 'Hello. I am from the local water company. There is going to be a disruption to the water supply tomorrow. There will be no water available between midnight and noon.' In this situation the android has to make several decisions. For example, it has to decide whether or not to accept the information about the non-availability of water the following day. This is because, if it is correct, the android would have to make preparations to ensure that the infirm person it is looking after had enough water during that time for washing, drinking and cooking purposes. Whether or not the android accepts this information is likely to depend on whether or not it accepts the man's assertion that he is an employee of the local water company. If there have been several incidents recently of burglars posing as employees of the water company, then the android might ask to see the man's identification, especially if he also says that he would like to enter the property in order to check the water meter. (If there have been a number of burglaries recently, then the android can only know this as the result of accepting what it heard on the television, say, or read in a newspaper. This further illustrates the importance of the ability to learn from written and spoken sources.) If the man proves to be a reliable source of information, the android will accept the information the man gives and act appropriately. This example shows that, not only would the android need to have the ability to learn about its environment through perception, it would also need to have the ability to assess what it hears people

say and decide whether or not to accept the information heard.

The ability to learn from other people's assertions would also be useful socially. Being a carer involves not only helping with a person's physical needs, but also developing some sort of relationship with the person being cared for. Being looked after by a machine with which you could not communicate would not be a pleasant way to live. Engaging in meaningful conversation with another person involves, in addition to the ability to understand language, the ability to acquire knowledge from what another person says. An android that could not learn in this way would be a poor companion. For example, elderly people like to talk about the weather, their relatives, their friends and other people, how things were in the past and what a dangerous world we live in today! The cared-for person is likely to be upset if the android carer could not remember the names of her children and grandchildren, what their occupations are or what schools they attend and other basic facts about them. In fact, even knowing the relationship that exists between one of her relatives and the elderly person is something that can only be known by believing some assertion or other. It is not a piece of information that can be gained by perception or observation.

I hope that I have convinced the reader that the ability to learn from testimony and tradition is one that is essential for an android to have if it is to function adequately in human society. It is not an optional extra that adds nothing fundamental to the android's ability to live and work amongst human beings.

In this paper I present a model of how human beings evaluate the information that they encounter in their daily lives. I do not want to suggest that this model is the last word on the subject. It is, however, the best that I have been able to come up with so far. I present it as a conjectural solution to the problem of how people acquire knowledge. As a hypothetical solution I invite the reader of this paper to criticise it as best he can so that I or another person can refine it into something better.

The purpose of the model presented here is to help in the design and eventual construction of an android that can live in human society and which can interact meaningfully with human beings. It should be noted that if an android is to be built to act as the carer, say, of an elderly person, then it will need many cognitive abilities. For example, it would have to be able to acquire knowledge through its various senses. It would need to have the ability to revise its belief-system if it discovered that two or more of its beliefs formed an inconsistent set. It would have to be able to make plans and organise its time. It would have to be able to prioritise its various goals and so on. By not discussing these abilities at length in this paper, I do not want to suggest that they are unimportant. On the contrary, all of these abilities are essential in designing and building a fully-functioning android and, furthermore, I am well aware that their implementation is no trivial matter. The ability to learn by believing other people's assertions, however, is just as important as these other abilities, if not more so, and yet it has not attracted as much attention as they have. One of my aims is to remedy this situation and to convince other people that it is vitally important to study this ability if we are ever to achieve the ultimate goal of AI.

It should also be noted that in this paper I am only concerned with how agents acquire factual information. People acquire much more that just factual knowledge as

they are enculturated into society. They also acquire, for example, ethical principles and aesthetic sensibilities and maybe also a religious faith. An android, to live in human society, would also need ethical principles and other non-factual 'information'. How these are acquired and modified will have to be understood one day before they can be programmed into a computer, but that is not my concern here. It is difficult enough to model our acquisition of knowledge by believing other people's assertions.

2 Modelling Human Assertion Assessment

Elsewhere I have argued that our acceptance of other people's assertions is governed by the defeasible rule to believe them [5]. This can be seen as a modern way of expressing Thomas Reid's principle of credulity, which states that we have 'a disposition to confide in the veracity of others, and to believe what they tell us' [11, p. 194], and Henry Price's maxim or methodological rule [10, p. 124], 'Believe what you are told by others unless or until you have reasons for doubting it.' Reid's principle and Price's maxim help us to understand how human beings acquire knowledge from testimony and tradition, but they are not detailed enough in order for us to be able to model this ability on a computer. The same could be said about the defeasible rule, 'Believe what you hear or read.' What I do in this paper is to unpack this defeasible rule so that, without too much extra work, it could serve as the basis of a computer program which can evaluate the assertions that are given to it and decide sensibly which of them to accept and which of them to reject.

My model of belief-acquisition is shown in Fig. 1. A person gains information in two main ways, namely by making judgements about his perceptual environment and also by accepting some of the assertions he reads or hears other people make. The way in which we come to have beliefs about our surroundings through the use of our senses is complicated, but it is not my concern in this paper. So, I will just make use of the fact that we do have the ability to make judgements about our immediate neighbourhood and I will not attempt to analyse it. It should be noted, however, that this ability is not an infallible one. People do make mistakes about what they perceive. Thus, it is possible to acquire false beliefs through observation. Though, of course, it is impossible to acquire a false belief in this way knowing it to be false [15]. We may come to know that some of our beliefs acquired through perception are false because they conflict with other beliefs acquired through perception that we regard as being more reliable. We may come to know that some of our beliefs acquired through perception are false because they conflict with other of our beliefs acquired by believing other people. It is, however, also possible that we may go through our entire life not knowing that some of our beliefs acquired through perception are false.

Most of the knowledge a person has has been acquired by accepting other people's assertions, but nobody believes everything that he reads or hears. As already mentioned, my proposal is that a person's acceptance or rejection of the assertions that he encounters is governed by the defeasible rule to believe them [5]. In the model these are the only two options available. A person can either accept an assertion that he hears or reads and add it to his belief-system or he can reject it. Assertions that we

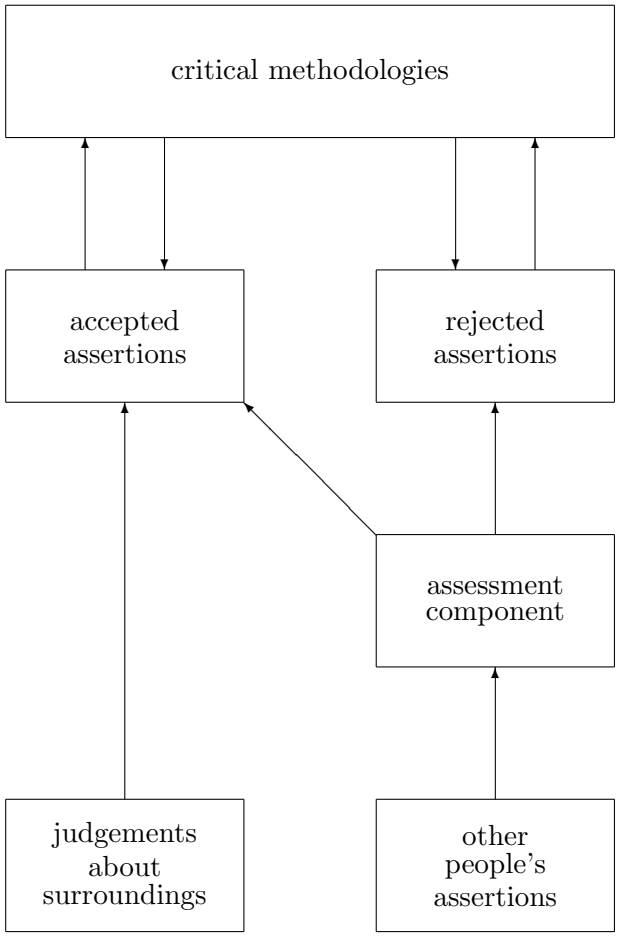

Figure 1: A two-stage model of belief-acquisition.

accept are, in reality, not all treated in the same way. In practice, we do not accept everything that we assess to be worth believing with the same degree of conviction. We do believe some things more strongly than others. Similarly, the assertions that we reject are, in reality, not all treated in the same way. For example, although a person may not accept an assertion, he may decide to entertain it in order to work out its consequences or he may flag it up as something that he will investigate thoroughly later. My model is, therefore, a simplification of or a first approximation to the actual way in which people deal with assertions. Such a simplification is justifiable, however, in order to gain a better understanding of how people evaluate the assertions that they encounter in vast numbers on a daily basis.

I will now briefly characterise the assessment component of the model, but, as this is the main topic of this paper, it is treated more fully below. The way in which I unpack the defeasible rule to believe other people's assertions is to represent it as an ordered set of rules, all of which except the last are conditional ones. The last rule

in the ordered set is not conditional and it is the non-defeasible rule to believe the assertion in question. There are many reasons why someone may decide not to accept an encountered assertion and each such reason becomes the antecedent of one of the conditional rules in the assessment component of the model. For example, a play is a work of fiction and so we do not normally believe the various assertions that the actors performing it make. This can be captured by adding the following conditional rule to the assessment component, 'If the assertion X is uttered by an actor during the performance of a stage play, then reject X.' (This is a simplification of the actual way in which we treat assertions by actors, since sometimes they are made to say things that are true in the real world.)

So far what I have been describing is the first stage of belief-acquisition. The judgements that we make about our perceptual environment and the evaluations that we carry out concerning the assertions we encounter have to be done in real time. A person driving a car, for example, has to continually make, and act upon, many judgements about what he sees other road users doing and he cannot afford the time to ascertain the complete accuracy of every judgement he makes. Similarly, a person reading a book cannot stop after each assertion that he reads in order to undertake a detailed investigation of its truth or falsity. The decision has to be made virtually instantaneously whether to accept the assertion in question or reject it. As, in both cases, the decisions involved have to be made in real time the assessment that takes place cannot be very sophisticated. As a result of this people will come to have some false beliefs and they will also reject some assertions which, as a matter of fact, are true. In my model, therefore, I propose a second stage of belief-acquisition in which a small number of a person's beliefs are subjected to a thorough and possibly time-consuming investigation of their truth or falsity. In the second stage a person can either re-evaluate something that he actually believes or he can re-consider an assertion that he previously rejected. When someone re-evaluates one of his current beliefs, he may either decide that he was correct in having that belief or he may conclude that on this occasion he made a mistake and reject the belief in question. Similarly, when someone re-considers an assertion that he rejected on an earlier occasion, he may change his mind about its truth and add it to his belief-system or he may decide that he was correct in rejecting it and not alter its classification. When a person does change his mind in either of these ways, he will probably need to make further revisions to his belief-system in order to remove any possible inconsistencies arising out of the change of mind.

Unfortunately, there is not a single methodology that can be used to check the correctness of every kind of factual assertion. Different methods are used for different kinds of assertion. For example, the way in which we check a historical assertion, such as 'The Battle of Hastings took place in 1066', is different from the way in which we would go about checking the truth of a physical assertion, such as 'The speed of light in a vacuum is 299,792,458 metres per second'. Furthermore, not everybody has the ability to check every kind of belief that he has. A historian would know how to ascertain the date of the Battle of Hastings and a physicist would know how the value of the speed of light is determined, but the historian is unlikely to have the knowledge

to work out what the speed of light is and the physicist is unlikely to be able to conduct historical research into issues relating to the dating of past events.

People re-evaluate some of their actual beliefs for a variety of reasons. They may, for example, become aware of an inconsistency between several of their beliefs and desire to remove this. Another reason could be that they come into contact with someone with radically different views from their own and find that when challenged to explain why they believe what they do they may not be able to answer adequately. My aim in this paper is to elaborate a model of belief-acquisition, and especially to describe the assessment component of this model, and so I am unable to look in any great detail at the reasons why people re-evaluate their beliefs or the means that they use in order to do this.

It should not be thought that an agent's assessment component undergoes no changes during that agent's life. There are several reasons why an agent may decide to alter his assessment component and there are three ways in which such a change can be made. The agent may decide to add an entirely new rule to his assessment component or to delete an existing rule or to modify an existing rule. Like the re-assessment of beliefs that occurs in the second stage of the model being developed, changes to a person's assessment component occur during reflective periods of that person's life when he is not actively engaged in gathering new information. Because changes to an agent's assessment component impact so directly on that agent's evaluation of other people's assertions, I discuss such possible changes more fully below.

3 The Assessment Component

An example of part of the assessment component of the first stage of the two-stage model that I am proposing can be found in Fig. 2. The rules comprising this component are ordered and they have to be applied from the top down. The ordered set of rules comprising the assessment component unpacks the notion that a person's acceptance or rejection of an assertion that he encounters is governed by the defeasible rule to believe that assertion. Thus, rule (∞) in Fig. 2 must be understood as being a non-defeasible rule and it must always come last in the ordered set of rules comprising the assessment component.

There are huge differences in the ways in which different people assess the quality of testimony and tradition. In presenting an example of part of an assessment component in Fig. 2 I do not want to suggest that these are the rules that everybody uses in evaluating other people's assertions. I am not trying to model as accurately as possible *the* way in which we assess other's assertions, because there is no single correct way in which this is done. What I am doing is to give some indication of what an assessment component looks like. The *form*, if you like, that an assessment component takes. The specific *content* varies from individual to individual. That is to say, different people would have different ordered sets of conditional rules, but everyone would have some such set.

In other papers I have considered in general terms the various factors that may cause us to overrule the defeasible rule to believe other people's assertions. We receive

(1) If the assertion X is uttered by an actor during the performance of a stage play, then reject X.

(2) If the assertion X is uttered in the course of a role-play exercise during counselling training, then reject X.

(3) If the assertion X is uttered by a person known to be unreliable, then reject X.

(4) If the assertion X is about a topic that the assertor does not know a lot, then reject X.

(5) If the assertion X is made by someone whose ideology is radically different from mine, then reject X.

(6) If the person speaking to us has as his main purpose in talking to us something other that the communication of the truth, then reject what he says.

(7) If the assertor of the assertion X makes a lot of meaningless hand movements when asserting X, then reject X.

(8) If the assertion X is obviously incoherent or inconsistent, then reject X.

(9) If the assertion X is about a topic that people often lie about, then reject X.

(10) If the content of the assertion X is out of the ordinary, then reject X.

(11) If the assertion X is straightforwardly inconsistent with my existing knowledge, then reject X.

(12) If the acceptance of the assertion X is going to have a profound effect on my current plans, life-style or belief-system, then reject X.

(∞) Believe the assertion X.

Figure 2: An example of part of the assessment component.

information from several sources and I have investigated our acquisition of knowledge from other people [3], from the reading of books [4] and from the study of journal articles [2]. To illustrate how the assessment component works I will make use of examples of how we acquire information by listening to other people in the flesh. Clearly, some of these rules are also relevant when we see someone talking on the television or in a film, but other factors apply there as well, which do not apply when we are listening to the person in the flesh, and it therefore seems sensible to treat the various cases separately.

I have grouped the factors that may cause us to override the defeasible rule to believe other people's assertions when we are listening to another person into five categories. These relate to the social context in which the assertion is made, to any information that we may have about the assertor, to the manner in which the assertion is made, to the content of the message and the final category relates to facts about the recipient of the message.

In the example shown in Fig. 2, rules (1) and (2) are examples of rules relating to the social context in which an assertion is made. There are a large number of social situations in which assertions can be made. For example, assertions can be made during the performance of a stage play. They can also be made by a member of a religious sect as he tries to convince us of the truth of his faith and attempts to get us to join his organisation. They can be made by a politician as he talks to us on our doorstep during an election campaign trying to get us to believe his message and vote for his party. They can be made by either the client or the psychotherapist during a counselling session. They can be made by a friend or a stranger in the course of an informal conversation in a pub. (Additional information about the role that the social context in which an assertion is made plays in helping us to evaluate the correctness of that assertion can be found elsewhere [3, pp. 227–228], where additional examples are also given.) Rules (1) and (2) cover only two of these cases. A more realistic example of an assessment component would need to have rules to cover assertions made in other contexts. However, these two rules give an idea of what the other rules would look like and nothing of theoretical interest would be gained by giving lots of examples of this kind of rule here. If we were involved in the practical task of writing a computer program to model the human ability to learn from others, then, clearly, we would need to include rules to cover all the social situations in which the agent is likely to encounter assertions.

Rule (1) comes into play when the agent is present at the performance of a stage play in a theatre. On the whole, when we are watching a play we do not expect to acquire a lot of new factual information. If we want to learn about a particular topic, we would go to a lecture about that topic or read a relevant book. Our first thought would not be that we had to go and see a relevant stage play. It is not easy to say what the purpose of the theatre is, but it is clear that is not to convey factual information.

To illustrate the operation of rule (1), consider the case of someone going to see a performance of George Bernard Shaw's play *Man and Superman*. In the fourth act of this play Malone says to Violet, after she has mentioned Hector [13, p. 184], 'His grandmother was a barefooted Irish girl that nursed me by a turf fire.' Upon

hearing this a person does not add to his real-world belief-system the fact that Hector's grandmother nursed Malone. This is not a fact, because there are no such people. Similarly, we would not want an android that attended this play to leave the theatre believing that Helen was a real person who had a grandmother that nursed Malone. If this android had rule (1) as part of its assessment component, then indeed it would not have these factual beliefs.

In some plays, however, there are characters who do make assertions that are true in the real world. The agent I am considering, who is only equipped with rule (1) to cover this situation, could not learn any new factual information at the theatre by listening to the characters. A more sophisticated agent could have a rule more complicated than (1) in order to allow him to acquire new factual beliefs in this way.

Another social situation in which people do not tell the truth occurs in role-play exercises that form part of the training of some counsellors and psychotherapists. In some forms of counselling training students are explicitly told not to use material from their own lives in role-play exercises. When playing the part of a client, they are told to invent a scenario and a problem (or they may be given a fictional scenario and told to elaborate on it). In these circumstances the student playing the part of the therapist would not believe the assertions uttered by the student playing the part or the client (or vice versa) and this case is covered by rule (2).

Rules (3), (4), (5) and (6) are examples of rules that apply to what the agent knows about the person making the assertion he is assessing. This information can be of several different kinds. For example, some of it could be about the character of the speaker, some of could be about his reputation, some of it could be about his allegiances and some of it could be about any possible agenda that the speaker may have in talking to us.

We do take a person's character into account, if we know anything about this, in assessing the correctness of the information that he is imparting to us. To take an extreme case, we are very wary of believing what a person, who is known to us as a habitual liar, tells us. Fortunately, not many people are habitual liars, but we do have a tendency to rank people according to how reliable or trustworthy they are in conveying information. When we have classified someone as being unreliable, then we are reluctant to accept what they tell us. This reluctance is captured in rule (3). This rule is perfectly acceptable as an example of a rule belonging to a person's assessment component, but most people would have a more sophisticated version of this rule which would take into account the fact that sometimes we do actually believe what an unreliable person tells us. This is because unreliable people are not unreliable about everything they say. They are not like the knaves in Smullyan's logical puzzles who always lie [14, p. 3]. Unreliable people, as well as making incorrect statements, also make some correct ones. For example, a person lacking credibility may tell us that the Renaissance was a cultural and artistic movement that originated in the 14th century AD in the city states of northern Italy. On this occasion, the unreliable person has got his facts right. An android equipped with rule (3), however, would not be severely disadvantaged in its task of acquiring knowledge from other people. It would occasionally reject true information, as in this situation, but it would not

acquire any false beliefs because it used rule (3) to assess other people's assertions. This is because rejecting an assertion is not the same as accepting its negation.

We take a person's reputation into account when evaluating the assertions that he makes. We also think of some people as being experts on particular subjects and if we regard a person as an expert then we are likely to accept what they say about their area of expertise. Conversely, if a person talks about a topic of which he knows very little about, we are unlikely to accept what he says. This tendency is captured by rule (4).

Sometimes we are in possession of information about a person's allegiances and what organisations he is a member of. One way in which this information may influence our assessment of a person's assertions is that, on the whole, we tend to be wary of assertions made by people belonging to a very different ideology. Before the fall of communism, for example, most of us living in the western democracies were very wary of anything a leader of a Communist country said. Rule (5) captures in an extreme form the reluctance we have to believe assertions made by people who we know subscribe to a very different ideology from our own.

Sometimes we know that a person speaking to us has a definite agenda in mind. Such a person does not talk to us just to pass the time of day with us. He wants something specific from us. A double-glazing salesman, for example, may begin by talking about the weather after we open the door to him, but we know that he is really trying to get us to buy something and that is the main purpose he has in mind as he talks to us. Similarly, when a Jehovah's Witness calls on us and asks what we think of the state of the world, we know that they are not interested in a free exchange of opinions. What they want is to convert us to their way of thinking and seeing the world. Rule (6) is one way of capturing the fact that people who talk to us with a specific agenda in mind are likely to sacrifice truth to the project of achieving their goal.

It has been known for a long time that a person's body language and the way in which he conveys information may indicate to us that that person is not telling us the whole truth. David Hume, for example, was aware of this. In section X of his book *An Enquiry Concerning Human Understanding* (1748) he says, amongst other things, that we are reluctant to accept the testimony of a person if he either hesitates when giving it or presents it with 'too violent asseverations'. Rule (7) is an example of a rule that applies to the manner in which an assertion is made. People do sometimes lie and in evaluating an assertion it is necessary to take this possibility into account. We, therefore, tend to look for signs that another person is lying. Many people, however, are quite bad at spotting a lie. Rule (7) is meant to capture one aspect of a person's behaviour that might indicate that he is lying. In reality, people are likely to look for various signs that a person is lying when he is talking to them. I am aware that research has been carried out in order to discover what behavioural characterists usually accompany lying. However, people looked for signs of lying before such research was undertaken and many people are unaware of this research and still look for signs of lying. Rule (7) is meant to be an example of a rule that the agent has before he has studied this research in any detail. The agent may decide to modify this rule if he

agrees with the findings of the research into lying.

Our evaluation of the correctness of an assertion sometimes depends on the content of that assertion. Rules (8), (9) and (10) are examples of rules that apply to the content of an assertion. The most straightforward example of how the content of an assertion effects our assessment of its likelihood to be true arises when the statement in question is obviously inconsistent or incoherent. Rule (8) applies to this situation. It does not happen very often that someone utters a self-contradictory assertion, such as 'It is raining and it is not raining', but we need a rule preventing us from accepting such a statement on those rare occasions when we happen to encounter it.

There are a number of statements that are uttered insincerely so often that someone hearing them may need convincing that they are being asserted genuinely. Many of these occur in relationships between men and women. For example, Hollander lists 101 lies that men often tell women [7]. These include the following statements: 'Honestly honey, it's just for the guys—none of the wives go to the conference', 'I'll call you' and 'I've got to work late at the office tonight'. Rule (9) applies to this sort of assertion.

On the whole, we are reluctant to believe things that people tell us that are out of the ordinary. The 17th century philosopher John Locke relates the story of a Dutch diplomat who informed the King of Siam that in the Netherlands lakes sometimes froze in winter and became so hard that men could walk across them. The King replied, 'Hitherto I have believed the strange things you have told me, because I look upon you as a sober, fair man: but now *I* am sure you lie.' (See his *Essay Concerning Human Understanding* (1690), book IV, chapter XV, section 5.) The Dutch diplomat was reporting what he himself had seen and experienced. To him his statements were as certain as they could possibly be and yet the King of Siam did not accept them because they conflicted so forcefully with his own experience of the weather in his own country.

A more recent example of people's disinclination to believe astonishing reports involves the pioneering deep-sea explorer William Beebe. In 1934 he made the deepest dive that had been made up to that time. His primitive bathysphere dived to a depth of half a mile. Beebe carefully described in his diary the strange creatures that he observed, but the life-forms that he wrote about were thought so outrageous by the scientific community that his observations were discounted. He gave many public lectures about what he had seen. Although many members of the general public were fascinated by his accounts of deep-sea creatures, the scientific community of his day was dismissive of his claims. Only in recent years, when more people have seen the same creatures that he saw, has his reputation been restored. (An account of Beebe's dive can be found in his book *Half Mile Down* [1].)

Both the King of Siam and the members of the scientific community in Beebe's day refused to accept the assertions that they heard because they accepted a rule like rule (10). I am not suggesting that people should adopt rule (10) as a means of dealing with out-of-the-ordinary information. All I am saying is that, as a matter of fact, some people do accept this rule (and it has led them to reject true information).

An agent's assessment of the reliability of an assertion that he hears is sometimes influenced by factors relating to him. For example, whether or not the agent accepts

an assertion could depend on its importance to him. It could also depend on his pre-existing knowledge or on the consequences involved in accepting the assertion in question. Rule (11) is an example of a rule that shows how an agent evaluates an assertion because of its relation to his pre-existing knowledge and rule (12) is an example of a rule that illustrates how an agent evaluates an assertion because of the possible consequences its acceptance may have on his plans, life-style or belief-system.

Adult human beings, with very few exceptions, have a vast amount of information. They have a large number of beliefs. There will be occasions when a person hears something that conflicts with his pre-exiting knowledge. This can happen, for example, about a topic that the person has a special interest in. A person may have an interest in ancient Egyptian history and spends a lot of his spare time reading about this ancient civilization. He goes to Egypt for his holidays and enjoys visiting all the historic sites. In the course of his studies he has acquired the belief that the 18th dynasty pharaoh Amenhotep IV, also known as Akhenaten, reigned from 1352 BC until 1336 BC. His belief in the truth of this has been confirmed because roughly the same dates are given in all the major reference works about the ancient Egyptians. Then, one day a friend takes him to hear a lecture by David Rohl about his ideas on the New Chronology of ancient Egypt. Rohl talks about the 18th dynasty and puts forward his view that Akhenaten actually reigned from 1022 BC until 1006 BC. (This information can also be found in Rohl's book *A Test of Time* [12, pp. 21 and 241].) Our amateur Egyptologist now faces a problem, since the content of Rohl's assertion clashes with virtually everything that he knows about the 18th dynasty. If his assessment component contains rule (11), then he will simply reject Rohl's assertion. There are, no doubt, some people who do accept rule (11). This is a very conservative rule which makes it difficult for them to alter their beliefs. There are other people, however, who are attracted by ideas that challenge their assumptions. It is unlikely, however, that anyone would have a rule like, 'If the assertion X contradicts my existing knowledge, then believe it'. Some people with an open mind would decide that inconvenient information should be looked at a bit more closely using techniques belonging to the second-stage of belief-acquisition.

Mature adults tend to live busy lives. They have goals and make plans. They have set ways of doing things in order to save time. If they come across information which would force them to make a big change in how they live their lives or would force them to drastically alter their current plans, then they will be faced with a difficult decision to make. Do they accept the information and make the requisite changes or do they reject it and carry on as they have learnt to live? An agent equipped with rule (12) would reject such life-changing information. There are people who accept rule (12). Studies of disasters, such as the fire at the Kings Cross Underground station in London on the 18th of November 1987, show this. Some people lost their lives in this fire, because they refused to act on the information supplied to them by employees of the London Underground. To act on such information would have forced those people to alter their current plans and they were not prepared to do that. (More information about the Kings Cross fire can be found in the official report on the tragedy [6].)

In this section I have given examples of rules that may occur in an agent's assess-

ment component. The particular rules that I have chosen are only meant to illustrate the sorts of rule that people actually use. I am confident that there are great differences between people in the actual rules that they employ and so it is pointless to try to devise a single set of rules that everybody uses. I think that the information I have provided could serve as the foundation of a project to construct a comprehensive assessment component that would enable an android to evaluate the assertions that it hears people make. Such an assessment component would be of great practical use, but little of theoretical interest would be gained by actually building such a component and my main interest in this paper is in the theory of assertion assessment.

4 Truth Maximisation and Error Minimisation

People want to have true beliefs and they want to avoid having false beliefs. My two-stage model of belief-acquisition takes into account the fact that people do, unfortunately, manage to acquire false beliefs and reject true information. Thoroughly checking out a belief that you have in order to decide if it is true and investigating a claim you have rejected in order to see if it is in fact true are both of them very time-consuming processes. It would be better neither to acquire false beliefs nor reject true assertions in the first place. Although it is impossible to devise a system that never lets in false beliefs and never rejects true information, it is sensible to refine your assessment component and revise your belief-system so that your evaluation of other people's assertions is more accurate. The two main principles that an agent uses in order to improve the way in which he evaluates information are the following:

(A) The agent wants to maximise the number of beliefs that he acquires as a result of hearing assertions that are in fact true. In other words, the agent wants to minimise the number of true assertions that he rejects.

(B) The agent wants to minimise the number of beliefs that he acquires as a result of hearing assertions that are in fact false. In other words, the agent wants to maximise the number of false assertions that he rejects.

I will first show how these two principles result in an agent's assessment component being improved and then I will consider how they may influence an agent to change his belief-system.

4.1 Altering the Assessment Component

It should not be thought that the rules comprising an agent's assessment component are created intact at that agent's birth and remain unchanged during his entire life. It is possible for the rules to be modified or removed entirely and it is possible for new rules to be added. This raises an interesting question: Why should an individual decide to make changes to his assessment component? There are several reasons why an agent may decide to make changes to his assessment component. To begin with I will consider how the principles (A) and (B) influence such changes. In order to use these principles to alter his assessment component the agent must, to some extent, keep track

of how successful the rules that he employs are at preventing him from acquiring false beliefs and at stopping him from rejecting true statements. An individual will consider whether or not to revise one of his assessment-component rules when he becomes aware that it has caused him either to acquire one or more false beliefs or to reject one or more true pieces of information.

When an agent discovers that he is acquiring a number of significant false beliefs, this means that there are one or more factors applying to the assertions that he hears that should cause him to reject those assertions, but he is currently ignoring them. This situation is remedied by the agent either adding a new conditional rule to his assessment component or by adding an additional clause to the antecedent of an existing rule. Consider, for example, rule (7). This tells an agent to reject information from a person making a large number of meaningless hand movements. The rationale behind this rule is that a person behaving in this way is likely to be lying. The agent may find, however, that he is failing to detect when some people are lying because their lies are not accompanied by strange gesticulations. The agent may reflect on some of the occasions when he acquired false beliefs as a result of believing lies and realise that on these occasions the speaker looked him straight in the eye and did not avert his gaze for the duration of the lie. The agent can incorporate this realisation by either altering rule (7) or by adding a new rule to his assessment component. If he decides to change rule (7), he will replace it with rule (7′):

(7′) If the assertor of the assertion X makes a lot of meaningless hand movements when asserting X or if the assertor looks me straight in the eye and does not avert his gaze while asserting X, then reject X.

If the agent decides to add a new rule to his assessment component, this would take the following form:

(7a) If the assertor looks me straight in the eye and does not avert his gaze while asserting X, then reject X.

I think that adding a completely new rule, like (7a), is preferable to modifying an existing rule, because it is the simpler of the two options.

There are two ways in which the agent can alter his assessment component in order to remedy the situation in which he finds that he is rejecting a significant number of true pieces of information. He can either alter one of his existing rules or he can delete a rule altogether.

An agent may reject an assertion because it satisfies one of the conditions contained in the antecedent of one of his rules. This may sometimes lead him to reject true information. For example, this may happen with rule (5), because someone belonging to a radically different ideology does not distort the truth as we see it in the light of his ideology all the time. When he is talking about something unrelated to his ideology, then we should judge his assertions as we would judge anybody else's. This can be captured in the following rule:

(5′) If the assertion X is made by someone whose ideology is radically different from mine, then reject X, unless the content of X is unlikely to be contaminated by that different ideology.

It is more difficult to think of sensible examples of someone deciding to delete one of his assessment-component rules altogether, rather than just modifying it. However, it is possible to imagine a fairly realistic scenario in which someone decides to give up rule (10). That rule relates to information that is out of the ordinary. Many people do accept such a rule. Above I discussed the King of Siam's rejection of the assertion that lakes in the Netherlands sometimes freeze solid and the scientific community's rejection of Willima Beebe's accounts of exotic sea-creatures. Both the King of Siam and the scientific community is Beebe's day were making use of rule (10) or a very similar rule to it. Someone who became aware of the fact that rule (10) causes people to reject true information may decide to get rid of rule (10) altogether. He feels that he does not want to be in the position of either the King of Siam or the scientific community is Beebe's day. I am sure that there are people who would reject rule (10) when they read of cases when out-of-the-ordinary information proves to be correct. A more thoughtful person, however, would probably try to remedy the defects of rule (10) by modifying it in some way. This is because, without rule (10) in our assessment component, we will quickly acquire very strange beliefs, especially if we happen to hear someone talk about having been abducted by aliens!

Although principles (A) and (B) are very important in helping us to improve our assessment components, they are not the only considerations which may force us to alter one or more of our assessment-component rules. A person, for example, may tolerate the possibility that he may acquire some false beliefs or reject some true pieces of information if the alternative is to lose his life or suffer some other sort of tragedy. Consider, for example, rule (12). This tells the agent to reject an assertion if it is going to greatly inconvenience him. In many situations this is a reasonable policy to adopt. This is especially true for people leading busy lives. If a person has to achieve many things in a day, he does not want to be distracted. He is, therefore, likely to reject any piece of information that hinders him from achieving his goals that day. However, if there is a chance that the rejection of a piece of information may lead to the agent's death, then it does not make sense to reject it. An agent, hearing about the death of people in the Kings Cross fire, may reflect that had he been present, equipped with rule (12), he too would have lost his life. Thus, he decides to modify this rule to something along the following lines:

(12′) If the acceptance of the assertion X is going to have a profound effect on my current plans, life-style or belief-system, then reject X, unless there is a reasonable chance that the rejection of X will cost me my life.

In the fire in Kings Cross Underground station people lost their lives because they disregarded the information that there was a fire in the station's ticket-hall. This case is quite complicated, however, because it could be argued that not only rule (12) came into play, but also rule (10). It is very unusual for there to be a fire in an Underground railway station and so this is a piece of information that is out of the ordinary. An agent may, therefore, also decide to alter rule (10) into something like the following:

(10′) If the content of the assertion X is out of the ordinary, then reject X, unless there is a reasonable chance that the rejection of X will cost me my life.

4.2 Belief-system Revision

I have discussed how principles (A) and (B) may cause an agent to make changes to his assessment component. I want now to turn to the issue of how these same principles may encourage a person to revise his belief-system. First, I will look at the operation of rule (A) in this process and then rule (B).

It should be noted that some, at least, of the rules comprising the assessment component make use of information in the agent's belief-system. In order to employ rule (3), for example, the agent needs to know whether or not a speaker lacks credibility. Consider the situation in which the agent believes Jones to be unreliable and Jones has asserted that he has recently won £1,000 by betting on a horse race. Rule (3) tells the agent to reject assertions made by unreliable people. The agent, therefore, does not add the belief that Jones won £1,000 by betting on a horse race to his belief-system. However, if other reliable sources confirm this information and other things that Jones says also prove to be true, then the agent may decide to revise his opinion of Jones's reliability. This would involve revising his belief-system by deleting the belief that Jones is unreliable and adding the belief that he is generally reliable.

I will now consider how principle (B) may encourage an agent to modify his belief-system. This principle comes into play when a person notices that he is acquiring a significant number of false beliefs. Consider the following scenario: While working at home, the doorbell rings and the person at the door tells us that he is conducting a survey of the neighbourhood. He asks us various questions and also imparts quite a significant amount of information in the process. He tells us about the price of gas and electricity as supplied by different companies and asks us how much we are paying. So long as you think that this person is a genuine market researcher you are likely to believe what they tell you, but as soon as you realise that they are in fact a salesman you will become very wary of accepting what they say. This is because, when you add the belief to your belief-system that this person is a salesman intent on getting you to change your electricity supplier, rule (6) comes into play and you cease to believe what the salesman says about the relative merits of different electricity companies.

5 Conclusion

In this paper I have put forward a model of how people acquire information by believing other people's assertions. There are several parts to this model and I have concentrated on what I have called the assessment component. Although people acquire information from a variety of sources, including other people, books, the media, the internet and so on, I have focused in this paper on how people acquire beliefs by accepting what they hear other people say in the flesh. As well as presenting the assessment component of my two-stage model of belief-acquisition, I have also indicated how the rules comprising the assessment component may be modified and some of the reasons why an agent should decide to modify them.

Although the model presented here contains enough information for a computer system to be built which implements it, a great deal of work has still to be done before

a realistic version of an assessment component can be produced. This is because the rules that I have chosen as examples are simplifications of the actual rules that people use in order to acquire information supplied by other people. Research, therefore, needs to be done in order to devise better rules.

I think that the goal of building an android with human-like abilities is one that is definitely worth pursuing. I believe that much of the research that is currently being done in order to achieve this goal is very valuable. Although much work is being done on emulating many human abilities, one fundamental ability has largely been overlooked and that is the ability to acquire knowledge by accepting other people's assertions. My research focuses on this ability and I hope the results of my work presented in this paper will encourage others to look more closely at this fascinating topic.

References

[1] William Beebe. *Half Mile Down*. Harcourt Brace, New York, 1934.

[2] Antoni Diller. Evaluating information found in journal articles. In Ángel Nepomuceno, José F. Quesada, and Francisco J. Salguero, editors, *Logic, Language and Information: Proceedings of the First Workshop on Logic and Language: Instituto de Lógica, Lenguaje e Información, Universidad de Sevilla, Sevilla, 29, 30 de noviembre y 1 de diciembre de 2000*, pages 71–78. Kronos, Sevilla, 2000.

[3] Antoni Diller. Everyday belief-acquisition. In Gabriela P. Henning, editor, *Argentine Symposium on Artificial Intelligence (ASAI2000) Proceedings: Tandil, September 5–7, 2000*, pages 221–232. Sociedad Argentina de Informática e Investigación Operativa (SADIO), Buenos Aires, 2000. This paper is also available on the Internet at the following URL: `http://www.cs.bham.ac.uk/~ard/papers/asai2000.html`.

[4] Antoni Diller. Acquiring information from books. In Max Bramer, Alun Preece, and Frans Coenen, editors, *Research and Development in Intelligent Systems XVII: Proceedings of ES2000, the Twentieth SGES International Conference on Knowledge Based Systems and Applied Artificial Intelligence, Cambridge, December 2000*, pages 337–348. Springer, London, 2001. This paper is also available on the Internet at the following URL: `http://www.cs.bham.ac.uk/~ard/papers/es2000.html`.

[5] Antoni Diller. Designing androids. *Philosophy Now*, (42):28–31, July/August 2003.

[6] Desmond Fennell. *Report of the Official Inquiry into the Kings Cross Fire*. HMSO, London, 1988.

[7] Dory Hollander. *101 Lies Men Tell Women and Why Women Believe Them*. HarperPerennial, New York, 1997.

[8] Peter Menzel and Faith D'Aluisio. *Robo Sapiens: Evolution of a New Species.* MIT Press, Cambridge (MA) and London (England), 2000.

[9] John L. Pollock. *Cognitive Carpentry: A Blueprint for How to Build a Person.* MIT Press, Cambridge (MA) and London (England), 1995.

[10] Henry H. Price. *Belief: The Gifford Lectures, 1960.* George Allen and Unwin, London, 1969.

[11] Thomas Reid. *An Inquiry into the Human Mind on the Principles of Common Sense: A Critical Edition.* Edinburgh University Press, Edinburgh, 1997. This book was first published in 1764.

[12] David M. Rohl. *A Test of Time,* volume one*, The Bible—From Myth to History.* Century Ltd., London, 1995.

[13] George Bernard Shaw. *Man and Superman: A Comedy and a Philosophy.* Penguin, London, 1946. This play was first published in 1903.

[14] Raymond Smullyan. *Satan, Cantor, and Infinity and Other Mind-boggling Puzzles.* Oxford University Press, Oxford, 1993.

[15] Bernard Williams. Deciding to believe. In *Problems of the Self: Philosophical Papers 1956–1972*, chapter 9, pages 136–151. Cambridge University Press, London, 1973.