An analysis of psychological experiments on non-monotonic reasoning

Edward Hoenkamp

University of Nijmegen, The Netherlands*

ABSTRACT

Since hardly any of people's everyday decisions are made with certainty, it is often necessary to retract earlier conclusions on the basis of new input. This aspect of common-sense reasoning in humans is often cited as a raison d'être for non-monotonic theories. Going beyond this intuitive notion, this paper is based on well-documented psychological experiments. In these experiments it turns out that inferences are often remarkably unresponsive to new input even if the original basis for the inferences is discredited. The focus in the present paper, therefore, is on modeling this pervasive, yet counter-intuitive retraction behavior.

0. INTRODUCTION

In everyday life we frequently draw conclusions which on the basis of further information may have to be retracted. Artificial Intelligence has been interested since its early days in the reasoning underlying belief for which there is no proof, and it has advanced several theories to deal with its non-monotonic character. The paradigmatic example is about Tweety, who is a bird and therefore can fly, but who is subsequently found to be an ostrich. In real life, conclusions are not always so quickly withdrawn. Since common sense reasoning in people is the prevalent motivation for non-monotonic theories (e.g. [Winograd, 1980; McCarthy, 1980; McCarthy, 1986]**), this paper goes beyond artificial examples and looks at well-documented experiments in which non-monotonic reasoning takes place. The present paper is divided in two parts. The first part reviews experiments on belief perseverence, the second part shows how the notions found there can be formalized to give a more coherent and transparent framework for the experimental paradigm.

*Author's address: Psychological Laboratory, Montessorilaan 3, Nijmegen, The Netherlands. I am grateful to Anthony Jameson, Eric Meyer and Peter Shell for discussions about the work reported.

**Five out of seven uses mentioned by McCarthy [1986] are psychologically motivated.

1. A CASE FOR NON-MONOTONIC REASONING IN HUMANS: 'DEBRIEFING' AFTER DECEPTION EXPERIMENTS.

A speaker at a conference may be heartened afterwards by someone from the audience who congratulates him for his interesting and clear exposition. If later he finds that this person mistook him for a potential referee, his self-esteem may decline again somewhat but it will probably not sink all the way back to its original level. In contrast to the AI examples of belief revision, people are often reluctant to adjust their opinion after the original evidence is descredited. This phenomenon has received special attention in connnection with psychological experiments in which subjects are deceived about the true nature of the setting, and are later 'debriefed' about the manipulation [Baumrind, 1964; Holmes, 1976; Ullman & Jackson, 1982]. The subject must be convinced that the information was fraudulent, i.e. he must be *dehoaxed.* Sometimes, the subject's feelings about himself (e.g. due to having behaved unethically [Milgram, 1963]) must be altered. This aspect concerns the *subject's behavior*, which cannot be refuted. Dehoaxing on the other hand, concerns the *experimenter's deception.* For this aspect, conclusions accepted earlier during the experiment can be disproved. Therefore, in the context of non-monotonic reasoning the dehoaxing aspect of debriefing is the more appropriate one to study.

1.1 EXPERIMENTS ON DEHOAXING PER SE

The aspect of dehoaxing would be difficult to isolate from the experiments, since these differ greatly in the nature and degree of deception. Fortunately enough, many experiments have been conducted on dehoaxing per se, employing different designs, different domains, and varying degrees of external validity. A

Source	Domain	Deception	Debriefing information	Perseverence after discrediting
Valins 1966	S watches slides taken from Playboy Magazine	hartbeat feedback changing rate with some slides	feedback was prerecorded sound tape	S prefers 'reinforced' slides
Walster et al. 1967	S fills out a 'Social Aptitude Achievement Test'	high (vs. low) scores are reported	no such test exists	S rates herself as similar to person with high (low) score
Holmes 1973	Instructions on tape informs S he will receive electric shocks during subsequent period	(no shocks are administered)	experimenter interups and tells electrode is fake	arousal remains
Ross et al. 1975	S discriminates authentic from unauthentic suicide notes	report of success (vs. failure)	ratings were prepared in advance	S rates herself according to original feedback
Andersen et al. 1980	S examines relationship between risk-taking and success as firefighter	data suggestive of positive (vs. neg.) relationship	data on ability are manufactured	S perseveres in estimates for new cases
Caretta et al. 1982	1972 Nixon voters were selected (vs. McGovern voters)	(not applicable)	Watergate hearings	Nixon voters retain positive feelings

Table 1. A representative selection of experiments on debriefing, with an approximate account of the setup. 'Vs.' indicates, where applicable, the success vs. failure manipulation. 'S' refers to the subject.

sample of representative studies is summarized in Table 1. As the paradigmatic example I will use an experiment by Ross, Lepper, and Hubbard [1975], which was very carefully designed and has been replicated many times. In it, the subject was presented with cards containing pairs of suicide notes. She was told that one note in each pair was genuine, the other bogus, and she was asked to indicate the genuine one. In addition she was informed about the average score in a pretest. The subject received false feedback indicating success or failure after each card. After completion of the task she was informed that the feedback had been determined prior to the experiment, and that it was not related to her actual performance (this was called *outcome debriefing*). Nevertheless, the greater the apparent initial success, the higher she estimated her scores for past and future performances. In short, subjects showed a substantial perseverence of the initial, erroneous impressions. Only after the process underlying the perseverence was explicitly discussed was the initial perception abandoned (the *process debriefing*).

Ross et al.'s explanation for the phenomenon has essentially two parts. The first part stems from the literature on attribution theory: An individual who witnesses a surprising (or extreme) outcome generates (searches for) confirmatory evidence capable of explaining the observed outcome. Second, if the original evidence for the outcome is removed, these antecedents may survive to give independent evidence for the outcome. For example, a subject may attribute her success on the discrimintion task to the fact that she was once personally acquainted with a suicide victim.

It may be argued that Ross et al.'s experiment does not in itself prove the presence of self-generated confirmation-biased evidence. Independent support for its presence has been found in various ways, however. For example, enhancing the possibility of producing such evidence increases perseverence [Anderson, Lepper, Ross, 1980]. On the other hand, when an interference task (e.g. counting backwards from 200 by 3) prevented the subject from engaging in explanations, no perseverence effect could be established [Fleming & Arrowood, 1979; Barefoot & Straub, 1971].

2. A MODEL FOR THE PROCESS OF DEBRIEFING

2.1 DEBRIEFING MODELED USING TMS

To introduce the model for Ross et.al.'s experiment I will use Doyle's [1979] TMS, a technique that allows for non-monotonic reasoning***. In this technique every assertion entering the data base is represented by a node. A record is kept of the dependency of nodes on inferential steps, i.e. the *justifications* of a node. This way the inference steps can be retraced to maintain consistency in a system. An assertion that is believed is called IN. An assertion that depends on the fact that another assertion is not believed (i.e. is OUT), is called an *assumption*. In the debriefing example someone else's assertions are believed as long as it is not believed that the other person is lying. If the latter belief comes IN, the assumption will go OUT (is not believed anymore). The debriefing experiment is depicted in Table 2. with the different stadia in terms of TMS. The subject starts out with no particular beliefs about the task. When the experimenter says the subject has performed well (**c**), she infers that this is the case (**b**). From this she generalizes to the belief that she is generally good at recognizing real suicide notes (**a**). This can be probed, e.g., by asking a subject how she would score in the future, or how she thinks she compares to other subjects. At the same time she generates confirmatory evidence (**e**), which comes IN. This evidence itself is an additional justification for belief **a**.The debriefing takes place by informing the subject about the deception (**d**). Since **b** depends on **d** being OUT, **b** goes OUT when **d** comes IN. But when asked, the subject will still believe **a**, on the basis of the independent support **e**. The process debriefing consists of an elaborate discussion of the perseverence phenomenon itself. The subject becomes aware of the self-generated confirmatory evidence she used, and leaves this out of the argument, i.e. **e** goes OUT, and as a consequence (**a**) goes out as well.

Belief	Dependencies in	Dependencies out	I	II	III	IV
a. I am good at this kind of task	b, e			●	●	
b. I performed well on this task	c	d		●		
c. E said I performed well				●	●	●
d. E provided bogus information					●	●
e. [self-generated confirm. evidence]	c			●	●	

Table 2. The debriefing experiment by Ross et al. The colums labeled 'dependencies' show how beliefs depend on other beliefs. A black dot indicates the belief is IN for the situation before the experiment (**I**), after feedback (**II**), after outcome debriefing (**III**), and after process debriefing (**IV**).

***The example is too simple to justify a comparison with other techniques. TMS was chosen because it is the most general technique proposed to date.

There is more to say about factors that are conducive to belief perseverance (see e.g. [Schul & Burnstein, 1985]). Keeping things simple however, consider a variation of the experiment by Ross et.al. One could start the system with **d** IN. In other words, the subject is told in advance that feedback will not be genuine. What will happen? We will come back to this after I have taken a closer look at the states of belief involved in the experiment.

2.2 STATES OF BELIEF AS ADMISSIBLE EXTENSIONS

The model developed so far describes the intended behavior (i.e. in the Ross et al.'s experiment) by showing how the subject gets from one state of belief to another. To ensure that the system represents the intended model, however, it must also rule out behavior not found in (or falsified by) the experiments. A way to find this out is by examining what belief states the system is capable of generating. To this end I will cast Ross et al.'s experiment in the form of a non-monotonic theory. Remember that at the heart of the model lies the notion of 'assumption.' Specifically, the subject believes the experimenter in the absence of reasons not to. The reason for believing C in the presence of A and in the absence of B will be denoted as:

$A \parallel B \Vdash C$. If B is empty (Ø), the inference from A to C is like an ordinary implication****. If B has the form 'Defeated(R)', the reason R is called *defeasable*. Let us describe the dependencies from Table 2 as a set of reasons **R** (indexed by consequent):

$$\mathbf{R} = \{ r_{a1}, r_{a2}, r_b, r_c, r_d, r_e \}, \text{ with}$$

$$r_{a1} = b \parallel \emptyset \Vdash a, \quad r_{a2} = e \parallel \emptyset \Vdash a, \quad r_b = c \parallel d \Vdash b,$$

$$r_c = \emptyset \parallel \emptyset \Vdash c, \quad r_d = \emptyset \parallel \emptyset \Vdash d, \quad r_e = c \parallel \emptyset \Vdash e.$$

From a set of reasons a set of *extensions* can be derived. They are the analogue of the 'deductive closure' in ordinary logic, and represent the internally consistent beliefs. Computing the closure (e.g. [Etherington, 1987]) $\mathbf{R}^*$ of **R** gives two extensions:

$$\mathbf{R}_1^* = \mathbf{R} \cup \{a, b, c, e\}$$

$$\mathbf{R}_2^* = \mathbf{R} \cup \{a, c, d, e\}$$

which are precisely the statements believed before and after debriefing. Now, where does the process debriefing come in? Statement **a** perseveres via r_c, r_e and r_{a2}. At least one of these reasons is apparently attacked by E (the experimenter). Reason r_c cannot be refuted since **c** is a fact. So, by discussing the perseverence process itself, the experimenter either defeats r_e, or

****But note: a reason is used to justify a belief. It is not an inference rule. So while **e** in Table 2 is justified by **c**, this does not mean it is logically implied by it; in general it isn't.

r_{a2}. Let us first assume the former. This can be formalized by rewriting r_e (and **R** changing accordingly):

$$r_{e1} = c \parallel \text{Defeated}(r_{e1}) \parallel\!\!- e$$

$$r_{e2} = \emptyset \parallel \emptyset \parallel\!\!- \text{Defeated}(r_{e1})$$

Now, in addition to $\mathbf{R_1}^*$ and $\mathbf{R_2}^*$, two new extensions result:

$$\mathbf{R_3}^* = \mathbf{R} \cup \{a, b, c, \text{Defeated}(r_{e1})\}$$

$$\mathbf{R_4}^* = \mathbf{R} \cup \{d, c, \text{Defeated}(r_{e1})\}$$

where $\mathbf{R_4}^*$ gives the belief state after process debriefing. $\mathbf{R_3}^*$ shows the efficacy of the process debriefing, i.e. as measured by the subject's prediction of her future performance on a similar experiment. Another interpretation is that a subject may be forwarned not to generate confirmatory evidence, i.e. to have r_{e2} ready in advance. A natural setting where this could occur is the courtroom. Indeed, in such a situation subjects are much easier to debrief [Hatvany & Strack, 1980]. Independent support to propose r_{e1} and r_{e2} stem from the experiments with an interference task. The interference effectively blocks the generation of confirmatory evidence, or formally, defeats r_{e1}. In this case $\mathbf{R_3}^*$ and $\mathbf{R_4}^*$ represent the states of belief before and after outcome debriefing in the interference task. Now r_{e1} and r_{e2} have been sufficiently justified, it remains to discuss the role of r_{a2}. It could be that this reason is defeated during process debriefing, although this cannot be ascertained on the basis of the experimental evidence currently available. In any case, it can be fromalized in a manner analoguous to our treatment of r_e above.

2.3 The model is neither too weak nor too strong

Since the model proposed above is based on an existing formalism for non-monotonic reasoning, I want to relate it to a criticism that has been advanced concerning such formalisms. Recently, Hanks and McDermott [1986] questioned whether these formalisms produce the expected results. They provide axioms for a simple problem (the 'Yale shooting scenario') and show that a well-established technique (c.q. predicate-circumscription) produces not only the intended extension, but in addition one that is counter-intuitive. Now, whereas Hanks and McDermott could have chosen between attacking either the axioms or the inference technique, they chose the latter. For this reason, in the section above I generated all the extensions of the proposed axioms for the Ross et al. experiment, and checked if they indeed belonged to the states of belief I wanted to model. They did. So the model is guaranteed neither too weak, nor too strong in generating states of belief. Yet, this is not enough to ensure the same holds for the intended behavior, i.e. for the *sequence* of states. To see this, suppose in the Ross et al. experiment the subject is briefed *in advance* that the feedback will not be genuine. That is, we start in Table 2 with **d** IN. Following through the experiment we will see that the same behavior ensues as before. In other words, the subject believes she performs well on the experiment even knowing beforehand that the feedback is bogus. This surely runs counter to our intuition. A similar reasoning as in the 'shooting scenario' therefore leads us to believe that our model, as defined by **R**, is too weak (it predicts unintended behavior). Yet, let us stay in the vein of this paper, and see if the prediction can be tested. In fact this has been done already by Wegner, Coulton and Wenzlaff [1985] who briefed the subjects in advance with the same words that were used by Ross et al. during debriefing. They found the same perseverence phenomenon, on the basis of which they rejected the theory of Ross et al., and formulated a principle of *transparency of denial*. This principle basically says that when people encounter denied information, that information is available despite the denial. However, in their experiment Wegner et al. tell the subject in advance that the information she *will* obtain is false, i.e. the information is not available at that time. In my opinion it is not necessary to introduce a new principle. Using our terminology, Wegner et al. seem to think they defeat reason r_{e1}, whereas in fact they produce **d**, so that **b** cannot be derived but **c** can. Whichever may be the case, the experiment confirms the counter-intuitive behavior predicted by our model.

3. Conclusions

People often cling to their initial beliefs more strongly than appears warranted. Based on a wide variety of experimental settings, this paper undertook a formalization of this phenomenon. It may contribute to psychology and to AI in the following way:

For psychology:

1. It formalizes and specifies experiments in the debriefing paradigm, and thus
2. It offers a better framework for analyzing these experiments and comparing explanations

For AI:

3. It demonstrates how truth maintenance techniques and non-monotonic theories can be used in modeling experiments on belief revision
4. It calls attention to the role of self-generated confirmatory evidence as an important factor in human non-monotonic reasoning
5. It demonstrates that intuitive appeal of predictions is not a reliable criterion for evaluating a descriptive model.

References

Anderson, C., Lepper, M., Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology*, 39, 1037-1049.

Barefoot, J. & Straub, R. (1971). Opportunity for information search and the effect of false heart-rate feedback. *Journal of Personality and Social Psychology*, 17, 154-157.

Baumrind, D. (1964). Some thoughts on ethics of research: After reading Milgram's "Behavioral study of obedience". *American Psychologist, 19*, 421-423.

Doyle, J. (1979). A truth maintenance system. *Artificial Intelligence*, 12, 231-272.

Etherington, D. (1987). Formalizing nonmonotonic reasoning systems. *Artificial Intelligence, 31*, 41-85.

Hanks, S. & McDermott, D. (1986). Default reasoning, nonmonotonic logics, and the frame problem. *Proceedings of AAAI-86*. Los Altos: Morgan Kaufmann.

Hatvany, N. & Strack, F. (1980). The impact of a discredited key witness. *Journal of Applied Social Psychology, 10*, 490-509.

Holmes, D. (1976). Debriefing after psychological experiments. I. Effectiveness of postdeception dehoaxing. II. Effectiveness of postexperimental desensitizing. *Amercian Psychologist, 31*, 858-875.

Fleming, J. & Arrowood, A. (1979). Information processing and the perseverance of discredited self-perception. *Personality and Social Psychology Bulletin, 5*, 201-205.

McCarthy, J. (1980). Addendum: Circumscription and other non-monotonic formalisms. *Artificial Intelligence*, 13, 171-172.

McCarthy, J. (1986). Applications of circumscription to formalizing common-sense knowledge. *Artificial Intelligence*, 28, 89-116.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Anormal and Social Psychology, 67*, 371-378.

Ross, L., Lepper, M., Hubbard, M. (1975). Perseverence in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology*, 32, 880-892.

Schul, Y., Burnstein, E. (1985). When discounting fails: Conditions under which individuals use dicredited information in making a judgment. *Journal of Personality and Social Psychology*, 49, 894-903.

Ullman, D. & Jackson, T. (1982). Researchers's ethical conscience: Debriefing from 1960-1980. *Amercian Psychologist, 37*, 972-973.

Wegner, D., Coulton, G. & Wenzlaff, R. (1985). The transparancy of denial: Briefing in the debriefing paradigm. *Journal of Personality and Social Psychology*, 49, 338-346.

Winograd, T. (1980). Extended inference modes in reasoning by computer systems. *Artificial Intelligence*, 13, 5-26.

Valins, S. (1966). Cognitive effects of false heart-rate feedback. *Journal of Personality and Social Psychology*, 4, 400-408.