Technical Report

UCAM-CL-TR-193
ISSN 1476-2986

Number 193

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Belief revision and a theory of communication

Julia Rose Galliers

May 1990

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

https://www.cl.cam.ac.uk/

Technical reports published by the University of Cambridge
Computer Laboratory are freely available via the Internet:

https://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

Belief Revision and a Theory of Communication*

Julia Rose Galliers
University of Cambridge Computer Laboratory
Cambridge CB2 3QG, ENGLAND
jrg@cl.cam.ac.uk

Abstract

This report concerns choices about changing belief. It describes research to establish and model a principled theoretical basis by which rational agents autonomously choose whether, as well as how to revise their beliefs. Aspects of the various problems in belief revision are discussed, and solved in the context of an AI tool for reason maintenance extended to cover situations of new evidence as not assumed 'truth'. Primarily this results from the inclusion of a non-numeric theory of strength of belief, which relates strength to persistence in the context of challenge. Such autonomous belief revision is presented as the basis of a theory of communication, as a special case of reasoning about change in an uncertain world with incomplete information, comprising others similarly constrained.

*Research supported by a SERC postdoctoral IT fellowship.

Contents

1 Introduction

The primary purpose of this report is to describe a model of autonomous belief revision. The model discriminates between possible alternative belief sets in the context of change. Its theoretical basis concerns relative persistence or comparative strengths of the alternative cognitive states. It incorporates an existing AI mechanism for belief revision, an assumption-based truth maintenance system or ATMS (de Kleer, 1986), with endorsements (Cohen, 1985) attached to foundational assumptions, and associated discriminatory reasoning machinery.

The motivation for the work is as a component of a model of communication between agents, in which agents can choose whether and how to revise their beliefs. This is an important aspect of design for multi-agent contexts as open environments (Hewitt, 1986), in which no one element can be in possession of complete information of all parts of the system at all times. Communicated information cannot therefore be assumed to be reliable and fully informed.

The model of autonomous belief revision as presented here represents the completion of a first phase in the development of the computational model of communication. The theory underlying these models is of communication as both determining and determined by belief revision; this is explicated in section 2. Belief revision is presented as a fundamental aspect of rational interaction with the world or environment. Communication as a type of rational interaction which involves the world comprised of other agents, is a development of this.

Section 3 follows with an outline of various problems inherent in current approaches and models of belief revision. These all relate to choices between alternative revisions. Aspects of the strength of belief issue, such as representation, origin, context and modelling are all raised.

Section 4 describes the proposed model of autonomous belief revision, and its resolutions of the issues raised in section 3. An ATMS generates alternative environments for reflection about potential revisions. The comparison entails reasoning with a non-numeric theory of strength of belief in which all beliefs are certain but variably corrigible and hence relatively more or less persistent in the context of their overall coherence with other beliefs. The emphasis is on the nature of the combinations of assumptions which underlie reasons for a belief, as opposed to the supporting reasons themselves. A detailed example is given to illustrate the theory, and to demonstrate its relevance to modelling communication.

2 Belief revision as the basis of a theory of communication

Many researchers in AI are concerned with the design of automated systems which can plan and execute actions. These actions should be appropriate to the goals of the system, and its context or environment. In this sense they are rational

behaviours. They are determined according to the constraints imposed by the system's cognitive architecture (Rosenschein, 1988) comprising three related and dependent components of perception, belief and desire, and action. And in the sense of being determined thus, and by past and present experience of the world as opposed to inflexible, imposed assumptions of the designer (Russell, 1989), a system can be autonomous. Being an autonomous, rational agent then is about having a basis upon which to reason about relations and behaviour appropriate to self and the world. And that world includes other agents, who similarly reason in order to act autonomously and rationally.

Primary in this reasoning are representations; beliefs or cognitive states generated through perception and inference, and related to desires and action according to the rules of rationality encoded into the system. But these cognitive states are inevitably constantly changing. The world is dynamic. Expansion and contraction of a belief set occurs as new data is perceived or inferred, and old data is lost over time or in the light of new evidence. Often expansion and contraction occur together. This is belief revision (Gärdenfors, 1988, 1989); changing one's cognitive state.

New data can be perceived directly from the world. It can also be communicated via another agent. An utterance is a perceived event that conveys an intention; the communicating agent's intention that the attending agent recognise an intention for a particular *change* in the attendee's cognitive states. This last phrase may be more traditionally presented as the intention to *induce* a particular mental state. The point being made is that agents always have some mental state. Any change in the environment, including the recognition of a communicative intention via an utterance, *changes* that mental state, and can be dealt with as an incidence of revision. In fact knowing this as a principle of rational behaviour makes revision of another's cognitive state the motivating force for communicative behaviour in the form of utterance planning. The principles of belief revision can be viewed as basic elements for interactive, cooperative rationality, removing the need for separate explicit statements or assumptions about cooperation and sincerity. The accepted basis upon which one belief set is preferable to another, for example in cases of contradiction yet logical equivalence (see section 3.1), can be equally applied in a context of contradiction arising as a consequence of an utterance. There is no need for separate axioms describing helpful agents as those that always adopt other's recognised goals, for example to believe P, unless they conflict with one in existence, such as already believing not P (Cohen and Levesque, 1987, Perrault, 1987). There is no need to dictate either adoption or persistence, or to treat contradictions in any way as a special case. A basic system of preference is laid down and understood, general enough to encompass change of beliefs as expansion, contraction or revision wherever in the world the new evidence comes from. What is being considered is: Which is the more coherent state given my current cognitive state and this change in the environment which has just occurred?

The principle of rationality or property of agents which is embodied within this, is that agents are autonomous over their mental states. Changes of mental state are guided by general principles of belief change, relevant to communicative and non-communicative contexts. Autonomous agents may or may not comply with the recognised intended effects of an utterance on their cognitive states. There are no specialised rules dictating what is a cooperative response. Rational communicative action must therefore be planned not only as purposive, but as *strategic*(Galliers, 1988,1989).

Strategic interaction acknowledges all participants as sharing control over the effects of a communication. Strategic action is that which maximises one's own outcome. Maximising one's own outcome in a situation of shared control, is a matter of it being maximal for the other party(s) also. Achieving a desired change in another's belief states is therefore a matter of creating a context such that the general rules of rational belief revision would dictate that change anyway. The aim therefore in utterance planning, is to determine one's own actions according to one's own goals and the context. This context includes the other agent and her presumed existing mental states, and a prediction of the changed context which will result in her preferring the intended belief state according to the principles of rational, autonomous belief revision. Cooperative behaviour can emerge autonomously, without being imposed from explicitly stated descriptions.

The need for a lack of imposed 'helpfulness' and associated assumptions about agents as reliable and informed and hence 'knowing what they are talking about', is because multi-agent environments are 'open environments' (Hewitt, 1986). No agent can know everything about its environment. No agent can *know* another's belief states. Such a state of affairs would not even be desirable as there would be unnecessary bottlenecks of information processing (Hewitt, 1986, Gasser, 1989, Galliers, 1990). Hence the use above of words such as 'presumed' and 'predicted' in phrases referring to others' mental states. This lack of complete information, together with the dynamic nature of both the physical and multi-agent world, is the background within which belief revision is viewed as fundamental to rational interaction. It is also the background to collaborative dialogue in which multiple utterances comprise information offered and requested, or presented and then confirmed as accepted or not. Dialogue is a series of negotiated or mutually accepted belief revisions.

To summarise the main points:

- An agent always has a cognitive state. Perceiving and inferring new beliefs changes this state by expansion, contraction and revision. There are principles which determine how this takes place; principles of rational belief change.

- Communication is a special case of belief change which involves other agents. Agents plan to revise, and perceive (recognise) other's plans as such.

- Agents are autonomous. They determine their own cognitive states accord-

ing to the principles of rational belief change. There are no special axioms dictating helpfulness and promoting cooperation when in a multi-agent context.

- Multi-agent communication must therefore be strategically driven. Plans to effect changes to another's cognitive state can only be successful if they take into account the principles whereby rational and autonomous agents change their beliefs. Maximising one's own outcome with respect to another agent is dependent upon them maximising theirs.

- Cooperative behaviour emerges from general principles of belief revision and agent autonomy.

The relationship between principles of autonomous belief revision and utterance planning, and the notion of dialogue as a series of negotiated revisions are expanded further in section 4.

3 Belief revision and strength of belief

3.1 Belief revision in AI

Belief revision in AI is associated with nonmonotonic reasoning; reasoning with inferences potentially withdrawable at some later stage. Doyle specifies two aspects of nonmonotonicity. Firstly, temporal nonmonotonicity in which attitudes are lost and gained over time, and secondly logical nonmonotonicity, in which unsound inferences are made but as a product of sound reasoning, incomplete information and a 'will to believe' (Doyle, 1988). An example of the latter is default reasoning.

Reason maintenance systems (RMS's) are AI's mechanisms for belief revision. They maintain consistent sets of beliefs in the light of new evidence. DeKleer's ATMS(1986) maintains various consistent sets of beliefs appropriate to different assumptions or contexts, whereas the RMS's of Doyle (1979) and McAllester (1980) maintain just one.

But new evidence may be accomodated into a belief set in alternative ways, and all of these maintain consistency. This is known as the 'multiple extensions' problem. For example:

$$(a) P \vee Q \quad (b) R \supset Q \quad (c) P \vee R$$

$$\text{new evidence:} \neg P \wedge \neg Q$$

Incorporating the new evidence results in two logically equivalent extensions. These are (b) and the new evidence, or (c) and the new evidence, because (a) is inconsistent with the new evidence, and *either* (b) *or* (c) are consistent with it, but not both (Rescher, 1964).

Alternatively again, the new evidence can be rejected if it is not assumed as 'truth' in which case the third possible extension is (a), (b) and (c). This latter

alternative is a possibility for example in communication, as long as there are no assumptions regarding the communicator's omniscience and/or sincerity.

The only way of determining a preferred option from these kinds of possibilities is to incorporate some factor other than consistency. This factor should be the basis for *ordering or prioritising the various alternative combinations of belief.* The following section deals with various aspects, problems and solutions to this issue as one of strength of belief as a determiner of preference in belief revision.

3.2 Preference in belief revision. Solving the multiple extensions problem

This section comprises various parts. Each is a question with alternative answers, all directed at the problem of ordering or assigning priorities in the belief revision context. The basis for this ordering is varying strengths or attachments to belief. Advantages and disadvantages to each different approach are considered. Section 4 then follows with a description of the proposed model of autonomous belief revision, in which a particular stand has been taken on each of the issues raised here.

3.2.1 Issue 1. Strength of belief.

In general, AI approaches to non-monotonic reasoning do not consider beliefs to vary in strength. All beliefs are equal for the purposes of inference and decision. Strength of belief is an accepted notion within inductive logic, however. It can involve acceptance theories comprising sets of confirmation functions and acceptance rules. Alternatively, Jeffrey's theory of partial belief (Jeffrey, 1983) assigns degrees to beliefs as subjective probabilities computed using Bayes' theorem from a set of evidence hypotheses. Some AI approaches similarly assign numbers as probabilities to every belief. For example, certainty factors in expert systems, and Dempster/Shafer theory (Shafer, 1976). In these cases, individual beliefs are differentiated in a manner which provides a ranking or order. The values assigned to new beliefs inferred from old or as evidence is gained or lost, reflects the combinations of values from their multiple sources.

Some AI approaches maintain beliefs as equal but differentiate the rules which generate those beliefs. It is a kind of preemptive approach whereby beliefs that would be inferred on the basis of less preferred rules are not inferred in the first place. Examples are systems employing prioritised competing default rules, [1] such as in HAEL (Hierarchic AutoEpistemic Logic) (Konolige, 1988). In this, the belief set is divided into a hierarchy of evidence spaces. Sentences in lower spaces are considered stronger evidence in being more specific, than those higher

[1] (Poole, 1985) and (Reichgelt, 1989) have a different approach to defaults in which default rules are expressed as propositions. New propositions added on the basis of these are marked as such. In Reichgelt, 1989, instantiations of default propositions are rejected in favour of instantiations of non-default propositions in circumstances of competition.

up. Inferences drawn from rules situated lower in the hierarchy override potential inferences higher up. An individual bat for example, can be inferred to fly even though the following two default rules contradict each other:

1. Normally mammals do not fly
2. Bats are mammals which do fly

The latter default rule relies on the more specific information, and is thus placed lower in the hierarchy. The bat as a mammal that cannot fly is not less preferred; it is never inferred in the first place. Whether priorities should be structured into the belief set in this way, is an aspect of the next issue for discussion.

3.2.2 Issue 2. Reasoning about strength of belief.

In the example of HAEL above (Konolige, 1988), priorities are structured into the belief system. The priorities are reasoned with, but not something to be reasoned *about*. In contrast, Cohen (1985) deals explicitly with the importance of reasoning about uncertainty. He develops a representation suitable for expressing reasons for believing and disbelieving within an expert system. Gärdenfors (1988, 1989) describes a formal model of belief revision, in which a number of rationality postulates operate as constraints from which the sentences within a belief set can be ordered. The ordering relates to the logical properties of the belief set and describes 'epistemic entrenchment'. This determines the sentences given up when a belief set is revised or contracted. In both of these latter cases, what is being provided is a basis for assessment, as opposed to a fixed measurement or structure. That basis is qualitative.

The primary limitation of fixed structural ordering is its inaccessibility and inflexibility. The latter issue is dealt with by Doyle, who refers to Konolige's specification of the hierarchy in HAEL as 'dictatorial' and violating the modularity principle, critical to successful construction of complex structures such as commonsense knowledge bases (Doyle, 1989). Modularity offers general rules of combination applied as the need arises, as opposed to employing a 'sovereign authority' whose task of resolving all potential conflicts is in any case infeasible with a large set of criteria. In addition, new criteria would necessitate a complete restructuring of the preference order.

The issue of inaccessibility is dealt with by Carver (1988) and Cohen (1985). If it is impossible to reason why a particular fixed ordering has been set, it is impossible to revise satisfactorily and flexibly in the light of new evidence (Carver, 1988). This is especially the case with numeric representations.

3.2.3 Issue 3. Representing strength of belief

Numeric representations of strength of belief are used with Bayes' theorem to provide a means of computing the probability of a conclusion given the numeric probability or degree of belief attached to each evidence hypothesis. There are

various problems with this 'conditionalization' approach (Jeffrey, 1983). Firstly, for every proposition who's probability is to be updated in the light of new evidence, there must be already assigned probabilities to various conjunctions of the proposition and one or more of the possible evidence propositions and/or their denials. This leads to a combinatorial explosion. The number of conjunctions is an exponential function of the number of possibly relevant evidence propositions (Harman, 1986).

In addition, once the number has been set, it's rationale in terms of the multitude of factors from which it is comprised, is submerged. There is no means of distinguishing between ignorance and uncertainty, for example (Carver, 1988). A low number could imply a lack of evidence or alternatively plenty of dubious evidence. Dempster/Shafer is a numeric approach which does not suffer from this latter problem in representing both a belief's support and its plausibility (Shafer, 1976). However, Cohen and Carver prefer non-numeric representations attached both to data and to rules, to represent all the various aspects appropriate to reasoning about uncertainty. Cohen refers to these as *endorsements.*

The advantage of numbers is ease of manipulation and combination. But for determination of preferred belief states for 'real' problems, the calculation must be based on more than probabilities of truth. As pointed out by both Doyle and Harman, however probable and well supported or plausible a tautology is, it has little utility (Doyle, 1988, Harman, 1986). In contrast, epistemic entrenchments are an indication of *explanatory power and informational value* (Gärdenfors 1988, 1989). Associated with such an emphasis on the utility of belief as opposed to its certainty, is a very particular viewpoint on the nature of strength of belief, which is described below.

3.2.4 Issue 4. The nature of strength of belief

The probability approach described above considers beliefs as variably certain. Only fully accepted or certain beliefs have a probability of 1. An alternative viewpoint is to consider all beliefs as accepted sentences, fully believed with a probability of 1, but not all of these may be equally *corrigible* in the sense of being more or less 'vulnerable to removal' (Levi, 1984).

> 'It is tempting to correlate these grades of corrigibility with grades of certainty or probability. According to the view I advocate, this would be a mistake. All items in the initial corpus L which is to be contracted are, from X's initial point of view, certainly and infallibly true. They all bear probability 1' (Levi, 1984).

What distinguishes them then is their persistence; their relative ease of disbelief. For Gärdenfors(1988) this is related to their usefulness in inquiry and deliberation. He offers an example from modern chemical theory. Knowledge about combining weights is more important than colour or taste; it has more explanatory power. If chemists change their opinion over the combining weight of two substances,

this would have more radical effects on chemical theory than if they changed their opinions over tastes. Beliefs about weights are therefore less corrigible or more entrenched than knowledge about tastes, although knowledge about both is certain. As Harman would explain it, it is harder to revise.

> 'I am inclined to suppose that these varying strengths [of explicit belief] are implicit in a system of beliefs one accepts in a yes/no fashion. My guess is that they are to be explained as a kind of epiphenomenon resulting from the operation of rules of revision. For example, it may be that P is believed more strongly than Q if it would be harder to stop believing P than to stop believing Q, perhaps because it would require more of a revision of one's view to stop believing P than to stop believing Q'(Harman, 1986).

This view expressed by Harman, Levi and Gärdenfors, is that accepted beliefs are certain but variably corrigible, as opposed to all beliefs being variably certain. This is an important component of the model of autonomous belief revision described in the next section. In this model, beliefs are held or not held in a yes/no fashion, but strength as a pragmatic and purely comparative notion is entertained at the point when such a held belief is challenged. It is a facet of revision. Preference of cognitive state in the light of a particular change is assessed according to relative persistence or *comparative hardness of revision.* What the origins of this varied corrigibility or persistence are, is the next issue for discussion.

3.2.5 Issue 5. The origins of strength of belief

What makes one belief more entrenched, or harder to revise, or more persistent than another? In the discussion of issue 4 above, it is suggested that this aspect of belief does not relate to varying certainty or probability of truth, but perhaps to utility in terms of explanatory power and informational value. So what is the basis of this explanatory power or informational value?

The specificity/generality distinction referred to in section 3.2.1 as the basis of HAEL (Konolige, 1988) is one candidate. A specific belief is preferable over a generality (Poole, 1985). It has more explanatory power and informational value. This notion is also incorporated into inferential distance algorithms for inheritance systems (Etherington, 1987, Touretzky, 1986).

A wider approach in this vein is to look generally at the source of beliefs or the evidences from which they were concluded. As well as being specific or general, perceived beliefs can be the result of first hand experience via sensory apparatus, or they may be the result of second hand communications via other agents or documentation. Cohen attaches various endorsements to data, one type of which is based on source information (Cohen, 1985). A representation of such endorsements and related set of heuristics regarding combinations of endorsements is outlined in the following sections's description of the model of autonomous belief revision. The intuition is that there are general rules with respect to sources of assumptions

underpinning beliefs, such as whether information came from a reliable source or was the subject of gossip for example, which are an important factor in determining relative persistence i.e. relative explanatory power and informational value.

An assumption of all the above is that it is feasible to deploy general domain-independent principles related to the properties of individual beliefs or the belief set, in order to assess priorities for revision. Another example is the ordering of epistemic entrenchments referred to in earlier sections, which is according to purely logical properties of the belief set (Gärdenfors, 1988, 1989). However, Gärdenfors, (1988) refers additionally to pragmatic factors such as conversational context in debates, and cites Levi in suggesting that informational value is context dependent, where context includes more than the agent's beliefs.

Some recent work by Konolige refutes the use of generalities in favour of 'knowledge-intensive heuristics tailored to a domain':

> '...any general domain-independent principles will be very weak, ... information from the semantics of the domain will be the most important way of deciding among competing arguments' (Konolige, 1989).

Issues 6 and 7 deal with the context of belief revision within which any ordering, whether on the basis of general domain-independent principles or not, may operate.

3.2.6 Issue 6. The context for strength. Alternative theories of belief revision

There are currently two competing theories of rational belief change. These form the alternative contexts within which any ordering or system of priorities for revision would have to be accomodated. They are *foundation theory* and *coherence theory*. Foundation theory considers new beliefs are only to be added on the basis of other justified beliefs, and beliefs no longer justified are abandoned. An example of this approach in practice is the truth (reason) maintenance system of Doyle (1979). Foundation theory takes its name from the emphasis on justification for belief, which obviously is not infinite. Where it ends up is in beliefs which are justified by themselves, and which then justify or are *foundational* to others. These are self-evident beliefs, for example an observation, as in the epistemology of positivists.

Coherence theory on the other hand, represents a conservatism whereby justification is only a requisite condition of believing if there is a special reason to doubt a belief:

- *The Principle of Conservatism*:
  current fully accepted beliefs are justified in the absence of any challenge to them (Harman, 1986).

If there is such a challenge, for example a new belief making one's belief set inconsistent, the guiding principles are those of *minimal change* and *maximal coherence*:

- *The Principle of Minimal Changes*:
  In revising one's view one should make minimal changes in both adding new beliefs and eliminating old ones (Harman, 1986).

The notion of changes of state being restricted to keep as much as possible of the previous state, is generally accepted as a good thing, both in philosophy and AI. The competing notion is coherence. This prevents such conservatism resulting in tenacity of belief regardless of evidence to the contrary:

> '...changes are allowed only to the extent that they yield sufficient increases in coherence' (Harman,1986).

Coherent beliefs are mutually supporting. P can be justified because it coheres with Q and Q be justified because it coheres with P. But the *nature* of this mutual support is of interest. According to Harman, coherence includes not only a consistency relation, but relations of implication and explanation too. Coherence is connections, and the connections are of *intelligibility*, in particular intelligible deductive and non-deductive explanation of why or how it is that something is the case. For example, if one believes P, Q and R, but also R *because P and Q*. Part of one's view makes it intelligible why some other part should be true. The 'because' can be deductive in P and Q implying R, or it could be statistical as in P and Q generally implies R 'if other things are equal', or it could be based in commonsense psychology (Harman, 1986). Believing R is explained by the beliefs P and Q. The connection offers intelligibility and makes the set more coherent than if P, Q and R were consistent but unrelated.

3.2.7 Modelling the context

Associated with a choice of context or theory for revision, is the issue of how these are to be modelled. There are various formal models of belief revision (Nebel 1989, Gärdenfors, 1988, 1989, Rao and Foo 1989, Martins and Shapiro, 1988). Those which model coherence theory model minimal change amongst sets of consistent beliefs with no justification relations. Maximal coherence is the retention of the maximum possible *logically consistent* beliefs during belief change. These approaches therefore leave out much of Harman's intuitions on the nature and role of coherence. They cannot express that some beliefs are reasons for or explanations of others. However, Gärdenfors' (1988, 1989) epistemic entrenchments are an attempt to include some of the justificational information available in foundation theory into a formal coherence model.

There are also various computational models of belief revision, such as TMS (Doyle, 1979), ATMS (de Kleer, 1986), CMS or Clause Maintenance System (Reiter and de Kleer, 1987) and MBR or Multiple Belief Reasoner (Martins and Shapiro, 1988). The prevalent theory in these is foundation theory. Both Harman (1986) and Gärdenfors (1989) cite debriefing studies however which demonstrate experimentally that people do not keep track of the justifications for their beliefs.

It may therefore not be known when sole reasons for a belief have been discredited, and as a consequence unjustified beliefs are retained. Disregarding psychological plausibility, it is also the case that the benefits from keeping track of justifications are outweighed by the computational costs. This view is borne out by RMS's being very inefficient (Rao and Foo, 1989b). Justifications are important however. The conclusions from the debriefing studies were that in people, beliefs will eventually be abandoned, but only on the basis of *positive* beliefs about lack of good *reasons* for them. Harman correspondingly expands the principle of conservatism as follows:

- *The Principle of Positive Undermining*:
  only stop believing a current belief if there are positive reasons to do so, and this does not include an absence of justification for that belief (Harman, 1986).

Positive reasons are believing one's reasons for believing the belief to be nogood. This is stated as:

> 'It is incoherent to believe both p and also that all one's reasons for believing p relied crucially on false assumptions' (Harman,1986).

Harman and Doyle criticise the use of logic in models of belief revision, claiming it to be insufficient, or even of no special relevance to theories of *reasoned* belief revision.

> 'As Harman (1986) puts it, inference is not implication: reasoning and inference are activities, *reasoned* changes in view, while proofs in a logic are not activities but atemporal structures of a formal system, distinct from the activity of constructing proofs. Thus logic is not, and cannot be, the standard for reasoning. (Doyle, 1988).

But RMS's can also be criticised. RMS's lack semantic theory; they cannot interpret new sentences added to node names because they do not understand what the nodes stand for. And the logic of propositions is lost in this representation of beliefs, with logical inferences having to be reintroduced as special systems of justifications (Gärdenfors, 1988, 1989). There is also an emphasis on the programmer, for example in the assignment of assumptions, which promotes a fairly ad hoc basis for belief revision.

The following section describes a model for autonomous belief revision which incorporates coherence principles into a primarily foundational model. The preferred cognitive state is determined by reasoning about relative persistence of the alternatives. The more persistent is the hardest to revise in terms of offering maximal coherence, given beliefs all held with probability 1. What this means is that it is the state with greatest justificatory backup for its component beliefs and also which offers the greatest explanatory power. This is not merely in terms of numbers of such relations but also takes account of the combinations of sources of the

underlying assumptions upon which the reasons for believing all the component beliefs are founded.

4 A model of autonomous belief revision

The model of autonomous belief revision described here, determines preferred cognitive states at times of change. Of particular interest are instances of change caused by communicative acts, and where the content of an utterance contradicts an existing belief. In such cases, the principles upon which preferred cognitive states are determined are employed to reason about *whether* to adopt a recognised intended belief via an utterance in preference to retaining an existing one, as well as *how* to do this in terms of which alternative cognitive state incorporating the new belief is preferred from the logically equivalent possibilities. As discussed in section 3.2.4, beliefs are represented in an all-or-nothing manner, but compared at times of challenge on the basis of relative persistence or comparative 'hardness' of revision.

- *The Principle of Persistence*:
  A belief's persistence is determined relative to the particular challenge. The persistent belief is the one which is harder to revise, in that context.

This means that if an existing belief is contradicted via an utterance, the belief is 'stronger evidence' than the utterance if it is considered harder to revise that belief than it would be to revise the belief entailed by the utterance, if the latter were currently believed.

In order to make this assessment, the alternative contexts are set up to compare revisions. This is discussed in detail in section 4.2. First, the ATMS as the chosen tool for this comparison is explained below.

4.1 Using an ATMS, with endorsements

Reason (truth) maintenance systems work in conjunction with problem solvers. The RMS is the bookkeeper maintaining conclusions drawn by the problem solver, as a consistent set of beliefs. De Kleer's ATMS (de Kleer,1986), is an ideal tool for comparisons of alternative possible revisions because it maintains various such consistent sets according to *context*, or different sets of underlying assumptions.

Beliefs are stored as ATMS nodes. Each comprises a description, label and justifications. The ATMS primarily maintains the labels, or alternative environments (sets of assumptions) from which each description may be derived. They are maintained as consistent, sound, complete and minimal with respect to that belief's justifications or support inferences encountered so far. The theoretical basis of the ATMS is therefore foundational [2]. Assumptions are represented as nodes,

[2]described in section 3.2.6

distinguished by being beliefs justified simply by their own existence. These are the foundational beliefs from which others are derived.

I have adapted the ATMS slightly by including an extra element in the assumption nodes. Assumptions as foundational, self-evident beliefs are variously endorsed [3] according to their source. The possibilities are:

1. *communicated*, either *first-hand* (sensory information) or *second-hand* (via another agent or text). These assumptions are also very roughly graded as 'pos' if they are communicated with conviction or from a very reliable source, or 'neg' if they are communicated from a spurious source or without conviction. The possibilities are represented in the example in section 4.3 as: [1c-pos], [1c-neg], [2c-pos] or [2c-neg].

2. *given*, either as *specific* information widely believed and without any particular source, for example 'Thatcher is currently prime minister', or as *default* generalities similarly widely believed. For example, 'birds fly'. Alternatively, given assumptions may be *values* denoting a notion of goodness which may be linked with desires. Values can also be 'pos' or 'neg' as a rough grading scheme between those more persistent in being considered a 'very good thing' and those just considered 'a good thing'. These are obviously subjective to the individual being modelled, although generally accepted (culturally held) values such as it being good to have money or to be conscientious or trustworthy can be incorporated as defaults. The given possibilities are represented in the example in section 4.3. as: [spec], [def], [value-pos] or [value-neg].

3. *hypothetical*, with no evidence at all other than as a possible grounding for a belief under consideration [hypoth].

The intuition behind these is that source information is relevant to credibility and hence persistence of ideas. I am more loathe to give up a notion read about in a respected scientific journal than one read about in the Sun newspaper, for example. However, most notions are multiply endorsed, and dealing with combinations of endorsements is not a matter of applying Bayesian principles of combinations as with numeric probabilities. A set of a few general heuristics have been devised for this purpose which are outlined later in the next section.

Assumptions can be additionally justified in some contexts, by beliefs other than themselves. The assumption is then alternatively potentially explained. For example, in the example described in detail in section 4.3 and represented diagramatically in Figure 1, a car owner is told she has no bill to pay after leaving her car at a garage for the day. For reasons explained in section 4.3, this is endorsed as a [1c-neg] assumption in this particular case. Experiencing the lack of bill is justification for believing there is nothing to pay; it founds or justifies this latter belief. But in addition the lack of bill can be *explained* by believing that the mechanic found no fault with the car, or alternatively that he found a fault

[3] see section 3.2.5

but was being generous! The environments supporting such inferences make the experienced lack of bill more coherent in being more intelligible; they are potential explanations of endorsed experience. Environments consistent with the postulated notion of the mechanic as a crook and out for what he can get, for example, are less coherent. In other words, I have adopted Harman's notion of coherence as intelligible explanation of why or how it is that something is the case. [4]

The ATMS makes available all *maximal* contexts or *extensions* (de Kleer, 1986). Extensions are maximal in the sense that including any further assumptions would cause a contradiction. Extensions are computed from the labels of each node by *interpretation construction.*

- Extensions are maximal consistent subsets of beliefs.
- Interpretations are the maximal consistent subsets of assumptions from which the extensions are derived. An interpretation is an extension's characterising environment (Kelleher, 1988).

When a contradiction is reported to the ATMS by the problem solver, all environments corresponding to labels on the contradictory data are recorded as *nogood.* This affects interpretation construction and the eventual content of the extensions such that no interpretation and hence extension contains a conjunction of assumptions from which a contradiction is derivable.

4.2 Belief Revision and Coherence

The model of autonomous belief revision is foundational in using an ATMS which represents justifications along with beliefs, justification ultimately being founded in self-justified beliefs or assumptions. As described above, these assumptions are represented as variously endorsed. The beliefs they found are related to other beliefs and assumptions by justification *and* as potential explanations, built up from deductive and non-deductive rules of reasoning. How these relations and their variously endorsed grounding assumptions determine *coherence*, which in turn determines a belief's *persistence* in that context, is described in section 4.2.1 below.

The advantage of a foundational mechanism is the accessibility of relational information between individual beliefs; information about justificatory and explanatory relations which, as explained in section 3.2.6, is relevant to notions of coherence as intelligibility as envisaged by Harman. However, there are practical concerns related to the effort involved in maintaining these relations, and also theoretical concerns in the existence of belief being *wholly* dependent on them. In contrast, coherence models offer sets of mutually supporting beliefs, some of which may be related by justification or by explanation, but these relations are unrecognised. The support or coherence is purely by virtue of their consistency, which is

[4] discussed at the end of section 3.2.6

not the desired notion of coherence as intelligibility. Beliefs can be retained even without justification, but relevant relational information is just unavailable.

The precise blend of the two theories in the model of autonomous belief revision is described below.

4.2.1 The mechanism of Persistence

As described at the end of section 4.1, the ATMS computes *extensions* or maximal consistent subset of beliefs, from the various alternative *interpretations* as feasible combinations of consistent sets of assumptions. Extensions are important in this model of autonomous belief revision as *coherent* sets of beliefs. They are coherent maximal consistent sets of belief, in the sense that they are mutually supporting; they comprise beliefs, some of which are related by justification, some by potential explanation and some merely by being consistent. The interpretations of these extensions are the assumptions which ground the justificatory, explanatory or merely consistent relations. They ground the coherent set.

In revision, coherence should be maximised. The 'harder' belief to revise is the belief in the maximally coherent extension. The persistence comparison therefore relates to extensions - alternative coherent belief *states* where particular competing beliefs are true - and not the beliefs themselves. And maximal coherence relates to the assumptions comprising the interpretations. The reason for this is based in Harman's principle of positive undermining. To stop believing a belief, the reasons for that belief must be believed grounded upon false assumptions. The more assumptions there are to be believed false, the more changes there are to be made, which makes the belief harder to revise. But changes are justified by increased coherence. In this model, it is the nature or combined endorsements of the assumptions comprising each interpretation (and thus grounding both explanatory and justificatory relations) that gives a belief its coherence in that context. If believing an assumption false generates a context which for example, additionally justifies another assumption otherwise held but only as self-justified, this context has additional coherence. For example, I may have been told that Annie has a sister, but I believe I have only ever seen Annie going into her house, next door to mine. I believe she has a sister because I have been told by a reliable source, but this is merely a self-justified assumption; I have no other justification for it. On discovering via another friend that Annie is a twin, I realise my belief that I have only ever seen Annie is probably false. Believing this assumption now to be false and that there is someone else I have seen who looks just like Annie, means that the sister evidence is additionally justified. The context in which Annie has a twin who is a sister is now the more coherent one; it is now the harder to disbelieve.

The preferred cognitive state is the most persistent extension, determined by reasoning about comparative coherence for the required changes involved in dropping a belief i.e. in believing all the assumptions grounding its reasons to be false. The maximally coherent state has the greater number of better endorsed assumptions

than the other competing states. These would be harder to believe false. The preferred state will comprise the 'stronger' of the beliefs in question.

Coherence is determined according to a limited set of heuristics governing the combinations of endorsements in an interpretation. For example,

1. Beliefs founded upon first-hand evidence are harder to disbelieve than those founded on any other combination of assumptions. (This does not take the possibility of faulty sensors into account).

2. The more positive communicated assumptions or specific assumptions, that ground a belief, the harder it is to disbelieve, regardless of the number of 'neg' or default or value assumptions.

3. Combinations of 'neg' endorsed assumptions and defaults can be relatively ranked, and values can enhance these. Believing it would be good to believe something does additionally endorse its belief. However, values are only compared when in conjunction with other endorsements. For example, however much it may be believed that it is good to win the pools, this can only endorse and make more persistent the belief state in which I believe I have won the pools if I have some other even vague, evidence for this. The ranking orders [1c-neg] as relatively more persistent than either [2c-neg] or [def] which are equivalent. These all supercede [value-pos] which offers slightly more persistence support than [value-neg].

Thagard uses some similar heuristics for determining explanatory coherence pertinent to the acceptance and rejection of hypotheses (Thagard, 1989). For example,

> 'From past experience, we know that our observations are very likely to be true, so we should believe them unless there is substantial reason not to. Similarly, at a very different level, we have some confidence in the reliability of descriptions of experimental results in carefully refereed scientific journals' (Thagard, 1989).

The numbers issue is contained within the heuristics for coherence. The cognitive state or extension founded by the most assumptions endorsed as first hand evidence, is the one preferred. If there is more than one of these or none of them, then the cognitive state founded by the most assumptions endorsed as positive communicated assumptions or specific assumptions is preferred. Minimal change is ensured in the role of the assumptions endorsed as 'hypothesis' in the heuristics. An assumption believed false in order to make a comparison with an alternative belief set, but there being no other endorsement, is endorsed as 'hypothesis'. 'Hypothesis' assumptions are passively negative in coherence assessments because the more there are implies the more change; that more assumptions are having to be believed false and with no good backup.

Implicit in the emphasis on assumptions underpinning reasons for belief as opposed to the reasons themselves is a measure where if one believes P and Q, and P is one's only reason for believing Q, then giving up P and Q counts as only one change, not two. This adheres to the principle of undermining, but does not protect the revision of crucial parts of one's reasons for many other beliefs as does 'The Simple Measure' where each explicit belief given up or added is counted (Harman, 1986).

All this is best demonstrated via an example.

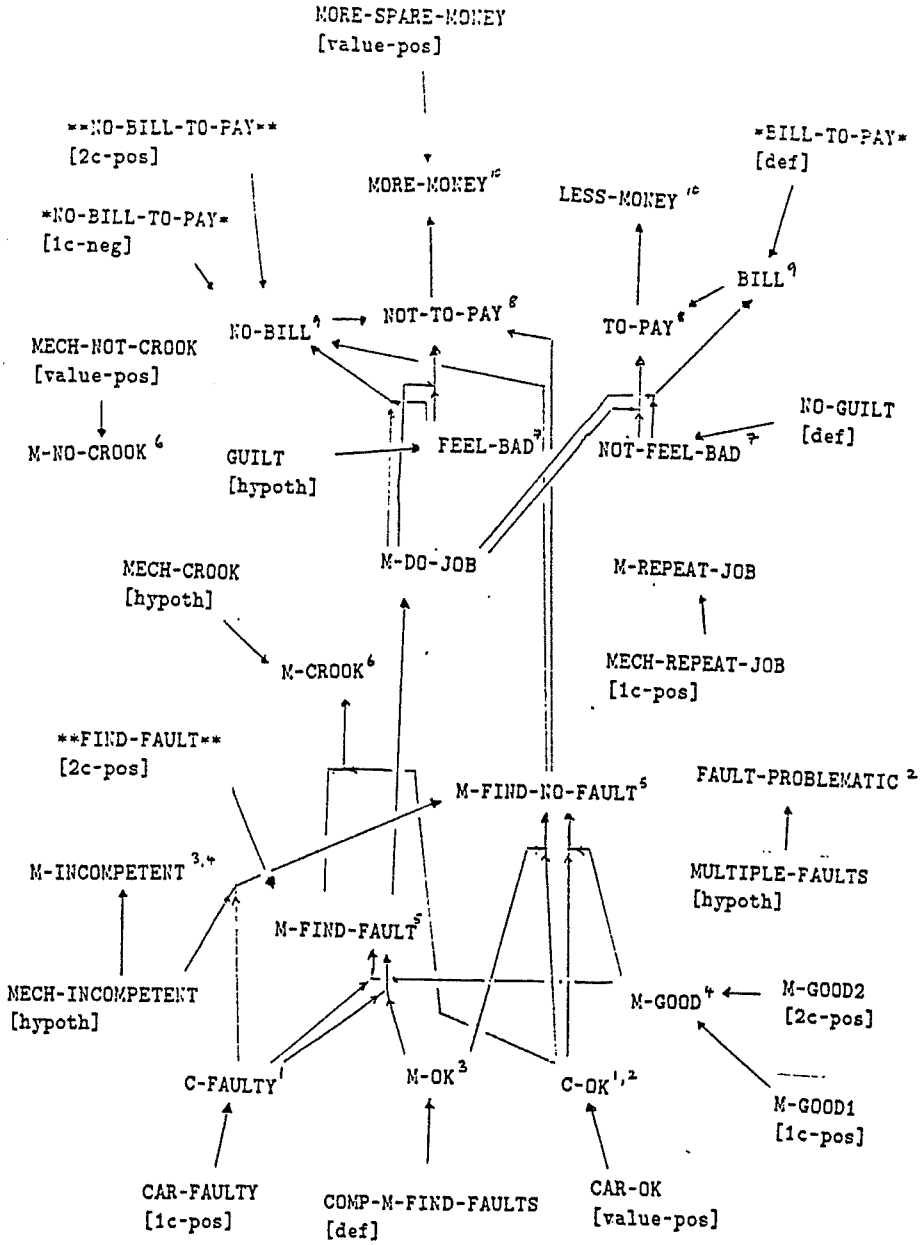
4.3 An example. The garage.

The ATMS is set up to model a scenario I experienced recently. It is set up with an agent's beliefs, assumptions and justificatory and explanatory relations at an initial stage of an interactive context involving a car owner, a mechanic and the mechanic's assistant. This is shown in Figure 1.

It is also informed of the environments unsatisfactory as support for inference due to inconsistency. These 'nogoods' are indicated in the figure with numbers. The endorsements associated with each assumption are indicated. They show that the car owner experiences the car as faulty [1c-pos]. She has beliefs about mechanics in general (competent mechanics find faults [def]) and this mechanic in particular as being a good mechanic. This is endorsed by her past experience [1c-pos], and also because she has been informed of this from a reliable source [2c-pos]. She also has experienced that this mechanic has attempted to rectify the same fault on several previous occasions (mech-repeat-job [1c-pos]). She has beliefs about paying for jobs and not paying for jobs and why. Not paying means more money in the pocket which is good [value-pos], but with no reason to believe the mechanic should do the job for nothing [def], the only explanation for this would be if he found nothing wrong with the car, which might mean that the assessments of him as a competent mechanic were incorrect. So, a potential explanation is that he is incompetent[hypoth] which is also the assumption that he is competent believed false. The setup includes *all* the believed assumptions believed false, and in the absence of any other endorsement, these are endorsed as [hypoth]. These are included because revision of an existing belief requires believing its reasons to be based upon false assumptions.

The ATMS of this scenario generates 18 possible extensions. These extensions are alternative, consistent, and *coherent* belief sets. But some are more coherent, more intelligible, than others. What makes them more coherent is the assumptions grounding *all* the supportive relations of justification, possible explanation or just consistency for this entire cognitive state. They are the set of assumptions which are *hardest to believe false.* They comprise that preferred extension's interpretation.

The preferred extension with respect to the issue of payment includes the belief that there is money to pay. This is on the basis of the car as faulty, the mechanic as good, but not a crook, (he may have demanded payment but found nothing



numbers indicate nogoods
the basic diagram represents all initial assumptions and beliefs
* indicates the assumptions added after the first utterance
** indicates the assumptions added after the second utterance

Figure 1: The garage example

wrong with the car), and so the mechanic having done the job which implies he requires payment. The preferred extension looks like this:

> (INITIALISED-PREMISE-NODE CAR-FAULTY COMP-M-FIND-FAULTS MULTIPLE-FAULTS M-GOOD2 M-GOOD1 MECH-NOT-CROOK MECH-REPEAT-JOB NO-GUILT C-FAULTY M-OK M-GOOD M-FIND-FAULT FAULT-PROBLEMATIC M-NO-CROOK M-DO-JOB M-REPEAT-JOB NOT-FEEL-BAD TO-PAY LESS-MONEY)

The determination of this as the preferred of 18 possible extensions involves a comparison of the combinations of endorsements associated with the interpretation of each extension, according to the heuristics described in section 4.2. An example comparison follows with just one of the other extensions, as a demonstration. The extension chosen for this demonstration comparison, is the preferred (according to the heuristics) of the subset of extensions which include the belief 'not-to-pay'.

The interpretation of the preferred extension including the belief 'to-pay' comprises the following assumptions:

> (NO-GUILT MECH-REPEAT-JOB MECH-NOT-CROOK M-GOOD1 M-GOOD2 MULTIPLE-FAULTS COMP-M-FIND-FAULTS CAR-FAULTY)

with endorsements as follows:

> ((DEF) (1C-POS) (VALUE-POS) (1C-POS) (2C-POS) (HYPOTH) (DEF) (1C-POS))

The interpretation for the preferred extension *if only considering those extensions including the belief 'not-to-pay'* comprises the following assumptions:

> (GUILT MORE-SPARE-MONEY MECH-REPEAT-JOB MECH-NOT-CROOK M-GOOD1 M-GOOD2 MULTIPLE-FAULTS COMP-M-FIND-FAULTS CAR-FAULTY)

and endorsements:

> ((HYPOTH) (VALUE-POS) (1C-POS) (VALUE-POS) (1C-POS) (2C-POS) (HYPOTH) (DEF) (1C-POS))

It can be seen that the above two interpretations are mostly the same. The difference is that believing 'to-pay' in this most coherent extension for the belief 'to-pay', involves a default or generality about mechanics not generally suffering from guilt, which is harder to 'believe false according to the heuristics, than a [hypoth] and a [value-pos] about more money being a good thing. [Hypoth]s offer no resistance to being believed false; they offer no persistence support. In fact, it is just the opposite. As explained in the previous section, their presence implies unendorsed change, which is not encouraged. And values *assist* belief retention, but are not evidence as such. Wanting to believe I don't have to pay does not make this belief harder to give up than the generality that garage mechanics tend to want paying for their work.

The belief that there is money to pay is challenged by the first utterance. The above demonstrates that until this first utterance, the belief that there is money to pay is inferred in preference to its negation in the given context, and in the absence of any particular challenge. The preferred extension in which it resides is the existing belief state. The utterance challenges this. It is made by the mechanic's assistant and conveys the proposition that there is nothing to be paid. It is endorsed as [1c-neg] because it is experienced in the sense that the car is driven away without payment (as opposed to merely being told over the phone for example [2c-neg]) and the assistant is not considered very knowledgeable or able to deal with the financial arrangements. The assumptions added to the initial set-up on the basis of this first utterance, are indicated in Figure 1 with one asterisk. (Those added after the second utterance are indicated with two asterisks.)

The effect on the ATMS is that the number of possible extensions expands to 29, and there are *two* preferred extensions with respect to the payment issue. These are determined according to the heuristics and exactly as described above. They are equivalently and preferentially coherent belief sets in the light of the new evidence. One is as before with the car faulty, the mechanic good and a bill to pay because of default generalities regarding bills to pay and the lack of any reason to assume the mechanic would do a job for nothing. The other also has the car as faulty and the mechanic as good but no bill to pay because of the utterance of the assistant enhanced by the value of having more money. The mechanic feeling guilty about his previous failed attempts is the preferred potential explanation for the utterance.

> (NO-BILL-TO-PAY GUILT MORE-SPARE-MONEY MECH-REPEAT-JOB MECH-NOT-CROOK M-GOOD1 M-GOOD2 MULTIPLE-FAULTS COMP-M-FIND-FAULTS CAR-FAULTY)

and the endorsements:

> ((1C-NEG) (HYPOTH) (VALUE-POS) (1C-POS) (VALUE-POS) (1C-POS) (2C-POS) (HYPOTH) (DEF) (1C-POS))

and:

> (BILL-TO-PAY NO-GUILT MECH-REPEAT-JOB MECH-NOT-CROOK M-GOOD1 M-GOOD2 MULTIPLE-FAULTS COMP-M-FIND-FAULTS CAR-FAULTY)

and the endorsements:

> ((DEF) (DEF) (1C-POS) (VALUE-POS) (1C-POS) (2C-POS) (HYPOTH) (DEF) (1C-POS))

Again these interpretations are much the same, but the two [def]s grounding the extension in which 'to-pay' is believed, offer no more or less coherence according to the heuristics, than the [1c-neg] utterance, [value-pos] and [hypoth] of the preferred extension where 'not-to-pay' is believed.

In other words, the belief that there is no bill to pay is now as coherent as that there is not. Either can cohere with other beliefs in a belief set grounded in equivalent combinations of endorsed assumptions, which is therefore equally as persistent or 'hard' to revise. It is as preferred to believe one way as the other resulting therefore in no overall belief with respect to the issue of payment. And in reality, this is exactly what occurred. At this stage, I had no idea whether I was to pay or not. There was a lack of information for resolving the conflict. This can motivate the planning of an appropriate utterance. Alternatively, the situation can remain and there be a lack of belief, given the existence of two equally possible preferred belief contexts. In reality, I made a plan to find out by adding more assumptions; changing the context of my belief states such that one extension would be preferred and hence one belief. The plan was to speak to the mechanic.

The second utterance comes from the planned conversation and is endorsed as a [2c-pos]. It reiterates that there is no bill to pay, but from the mechanic, over the phone. He also confirms that he found a fault [2c-pos]. The effect of these two endorsed assumptions on the ATMS is that the number of possible extensions expands to 36, but there is now one preferred extension with respect to the issue of paying. This shows that there is nothing to pay, on the basis of the car as faulty, the mechanic competent, but feeling guilty about his previous failures.

> (NO-BILL-TO-PAY2 FIND-FAULT NO-BILL-TO-PAY GUILT MORE-SPARE-MONEY MECH-REPEAT-JOB MECH-NOT-CROOK M-GOOD1 M-GOOD2 MULTIPLE-FAULTS COMP-M-FIND-FAULTS CAR-FAULTY)

and the endorsements:

> ((2C-POS) (2C-POS) (1C-NEG) (HYPOTH) (VALUE-POS) (1C-POS) (VALUE-POS) (1C-POS) (2C-POS) (HYPOTH) (DEF) (1C-POS))

It should be noted that the examples are with respect to the beliefs of 'to-pay' and 'not-to-pay'. The comparison of extensions is oriented around those in which 'to-pay' is coherent, and those in which 'not-to-pay' is coherent. There may be different preferred extensions relevant to believing a different belief.

It should also be noted that this example deals with a conflict between an existing belief and that recognised via an utterance. It is equally as applicable however, in cases where the taking on of a new belief is not in direct conflict with one in existence. The comparison is between the belief state which exists without the new belief, and that which would exist after taking the new belief on. The question is which would be the most persistent? It may be that the new belief adds extra coherence to what is currently believed, or it may not. In either case, it is preferred that it be adopted. What would prevent its adoption, would be if its adoption offered only a belief set *less* coherent than before. The belief set would be more persistent without it.

For example, take a context between a librarian and a student. The librarian believes this student to be a medical student. The student asks for a book on house plants to assist with her project work. The librarian has no contradictory

belief in which the student does not want a book on house plants for her project, but this belief does not cohere with his view of the student as a medical student and what medical students do projects on. Incorporating this belief would involve abandoning some well endorsed beliefs. It would be easier not to take this new one on. However, in reality it is most likely that faced with this contradiction, the librarian would additionally generate a plan to establish some justification for his existing beliefs or taking on the new one. For example, asking why a medical student would be doing a project on house plants. The answer may justify the new belief and make it cohere with the existing, seemingly contradictory beliefs. For example, the medical student may be doing her project on allergic responses to house plants. Alternatively, the librarian may end up revising his views about this student's background; she may have switched from medicine to botany.

4.3.1 Issues

One criticism of the model of autonomous belief revision as outlined here is its practicability. The performance of the ATMS reflects a problem's complexity and the worst case performance is NP-complete (Bowen, 1989). With increasing numbers of assumptions, it takes more and more time for the ATMS machinery to generate all the extensions prior to the reasoning about them. This could be considered as an implementation issue, and Bowen (1989) has developed a version of an ATMS called CMS within a REASON system for belief revision, in which approximations are made to make the computation of extensions more tractable.

It should also however be considered as a theoretical issue. For example, some of the beliefs and assumptions in the garage example were irrelevant to the eventual comparisons, such as whether the car was suffering from multiple faults. In addition, many competing extensions were comprised of assumption sets where only one or two were different. These were then the only relevant ones for comparison. In other words, some notion of *relevance* could well be employed to constrain processing. Harman distinguishes beliefs which are 'of interest'. These relate to the immediate environment or facilitate further theoretical or practical reasoning (Harman, 1986). How to go about practically distinguishing beliefs 'of interest' in a reasoning system such as this however, is a research issue as yet unsolved. The simplifying assumptions in the example above and generally in RMS's are firstly that all beliefs, including those implicit in one's beliefs, are explicit. Secondly, all beliefs are equally 'of interest', or relevant.

Another important point is that there is no assurance with this model, that the preferred extension is the 'right' set of beliefs. There is no notion of 'rightness'. The determination of relative persistence is according to how hard it is to believe false the assumptions grounding a coherent belief set. This 'hardness' is based on the combination of individual assumptions as examples of a few general endorsement types, and some 'rough and ready' heuristics about their combinations. The endorsement types include communications. As discussed in section 2, for an agent in an open environment where so much information is inferred from

generalities and past experience and very little is certain, communications can never be considered as incontrovertible. There is no intention in this research to model communication or reasoning about communication, as certain. The study of human or non-human interaction cannot be an exact science.

4.4 Autonomous belief revision and utterance planning

It has been suggested in this report, that knowledge of the principles of autonomous belief revision drives a communicator's selection of appropriate intended belief states. If the participants in the dialogue understand the primary importance of assumptions grounding explanatory and justificatory reasons for beliefs, then the job of assisting the other to revise their beliefs is to find out or predict upon what assumptions their existing beliefs are based. Believing such assumptions false leads to dropping a belief, and the theory describes the basis upon which combinations of endorsed assumptions are dropped in favour of others. In addition, assumptions can be suggested which would imply or explain (cohere better with) other data, and which would then lead to intended belief changes. This is a strategic approach to dialogue. It is strategic because it emphasises maximising one's outcome in knowledge of the other party who shares control of that outcome. The dialogue is a negotiation about the change of belief states. It is aimed at changing those belief sets to mutual satisfaction.

To demonstrate this I shall develop the description of the interaction between a librarian and medical student at the end of section 4.3. Let's say the student has told the librarian she is doing her project on allergens and house plants. The librarian knows that the house plant books in the library refer only to growing conditions; they would not be useful to the student. He plans to alter the student's cognitive states such that she drop the desire for a house plant book and adopt one for a more general book, but on allergens. Knowing the basis for dropping a belief, he offers some explanation and/or justification which he believes will render the suggestion and rejection of existing view, more coherent for the student without any imposed acceptance of the librarian as superior or herself as just being 'helpful'. Their belief states and decision are mutually agreed.

The above concerns utterance planning related to intended changes in another's cognitive states. Of course, dialogue may also be planned with respect to intended changes to one's own cognitive states. An example of this arose in section 4.3, in the description of the effect of the mechanic's assistants utterance on the car owner's mental states. The result was two equally preferred cognitive states, one in which the car owner had to pay and the other in which she did not have to pay. This situation motivated a plan to generate an utterance, the response to which would alter further the context of her own mental states such that the conflict perhaps be resolved. In the librarian example described at the end of section 4.3, the librarian was also motivated by desired changes to his own belief states in his questioning why a medical student would be doing a project on house plants. In this case there was no actual conflict; there was one preferred cognitive

state. The librarian chose not to adopt the recognised intended belief that the student wanted a book on house plants for her project. However, given the source of the seemingly incoherent utterance, a plan was also generated to gain more well endorsed assumptions. In a context such as this, it is most likely that many of the assumptions grounding various beliefs will be default assumptions. These assumptions are therefore not particularly well endorsed and much of the dialogue may be planned as a means of better endorsing or grounding reasons for believing, prior to then planning to alter another's belief states. For example, the generally assumed nature of first year student projects, may not cohere well with a request for a very specialised journal report. It may be that the student does not realise that she would be better off with something more general, but it may be that she has a very particular use for this journal report, and is aware of the strangeness of the request. Appropriate action to alter her belief states requires well endorsed confirmation first by the librarian, of his own.

Further research into the application of this theory to strategic utterance planning for cooperative task-oriented dialogue is to begin at the end of this year (ESRC/MRC/SERC Cognitive Science/HCI Initiative Project ID: 90/CS42), under the joint direction of myself and Dr. Karen Spark Jones. The context is the interaction between a library user and a librarian. It is hoped to generate successively more complex versions of an 'automatic librarian' to establish this theory of cooperative dialogue as better than those currently proffered, and lay a foundation for an eventual real automated library interface. The study context was chosen as one in which neither participant in the dialogues is dominant in terms of knowledge. The librarian knows more about the library system and the user knows more about themself and their overall objective, but only together can they achieve the knowledge required for appropriate document retrieval. In addition, a detailed model developed by Belkin (1983), Brooks (1986) and Daniels (1987) is an available starting point for the automated librarian design, developed from a collection of documented interviews between real librarians and library users.

5 Conclusions

This report concerns a theory of communication. One in which belief revision is considered a fundamental property of rationality, communication being a special case of this. Communicating agents recognise each others intentions to *change* their cognitive state. Such observed communicative actions alter a cognitive state which already exists, as do observations of the natural world.

Agents which are autonomous in their actions and reactions to the world, share control over the changes induced by each others communicative actions. Communicative actions are determined according to a strategic rationality which takes account of this autonomy. This is an important aspect of interaction in open, multi-agent environments where no one agent can be in possession of the 'truth' and prescribed behaviours imposing cooperation as benevolence, may therefore be

inappropriate. Cooperative behaviour falls out of the strategic approach in the attainment of *mutually* satisfactory belief states.

The model for autonomous belief revision described here is a first stage towards an implementation of a dialogue system based upon the above theory of communication. It determines preferred cognitive states at times of change. And as described above, knowledge of the principles whereby cognitive states are preferred over others directs appropriate utterance planning; it directs the selection of appropriate intended belief change:
The preferred cognitive state is the most persistent. It is the 'hardest' state to revise, in terms of offering maximal coherence for minimal change, with respect to the particular challenge. Coherence is determined according to the number and nature (in terms of source) of assumptions grounding the justificatory and explanatory relations of a cognitive state. These would have to be believed false in order to revise one's view.

And finally, although this research is oriented towards the design of a computational model of dialogue, the theory is also more generally applicable. The issue of strength of belief and preferred cognitive states applies equally in situations of competing inference, as exemplified by the Nixon diamond problem, for example. In this, alternative belief states or extensions can be generated in which Nixon is inferred to be a pacifist because he is a quaker, or Nixon is inferred to be a hawk on the basis of being a republican. Being a hawk and a pacifist are contradictory. Which is the preferred extension?

As in the dialogue contexts, source information endorsing ground level assumptions can combine together here in the alternative coherent extensions, to provide a basis for choice. Additional information about Nixon may be potentially explained by beliefs about him as a hawk or a pacifist, such as that he is known [spec] to have been involved in the Vietnam war. These additionally offer coherence or intelligibility to the relevant extension. With this example, the one where he is a hawk is the more coherent. This would be the preferred extension in the absence of any further and better endorsed information of him as a pacifist.

6 Acknowledgements

I would like to thank Gerry Kelleher especially for his implementation of de Kleer's ATMS. Also Han Reichgelt, Innes Ferguson and Victor Poznanski for helpful advice and support.

References

[1] Belkin N.J., Seeger, T. and Wersig G. Distributed expert problem treatment as a model for information systems analysis and design. Journal of Information Science 5. 1983.

[2] Bowen J. The Design, Implementation and Evaluation of a Truth Maintenance System. PhD thesis. Psychology Dept., University of Sheffield. 1989.

[3] Brooks, H.M. An intelligent interface for document retrieval systems. PhD thesis, City University, 1986.

[4] Carver N. Evidence-Based Plan Recognition. COINS Tech. Report 88-13, Computer and Information Science Dept., University of Massachusetts at Amherst, 1988.

[5] Cohen P.R. Heuristic Reasoning about Uncertainty: an Artificial Intelligence Approach, Pitman, Boston, 1985.

[6] Cohen, P. and Levesque H. Rational Interaction as the basis for Communication. Technical report No. 89, Centre for the Study of Language and Information, Stanford University, California, U.S.A., 1987.

[7] Daniels P.J. Developing the user modelling function of an intelligent interface for document retrieval systems. PhD thesis, City University, 1987.

[8] De Kleer J. An Assumption-based TMS. Artificial Intelligence Vol 28 No. 2 pp127-162,1986

[9] Doyle J. A Truth Maintenance System. Artificial Intelligence Vol. 12, pp232-272, 1979.

[10] Doyle J. Reasoned Assumptions and Pareto Optimality. Proceedings of IJCAI, 1985.

[11] Doyle J. AI and Rational Self-Government. Tech Report CMU-CS-88-124, Carnegie-Mellon, Computer Science Dept., 1988.

[12] Doyle J. On Universal Theories of Defaults. Carnegie-Mellon Computer Science Tech. Report. No. CMU-CS-88-111. March, 1988.

[13] Doyle J. and Wellman M. P. Impediments to Universal Preference-Based Default Theories. Proceedings of First International Conference on Knowledge Representation and Reasoning. Toronto. 1989.

[14] Etherington D.W. Formalizing Nonmonotonic Reasoning Systems, Artificial Theories and Inferential Distance. Proc. AAAI, 1987.

[15] Galliers J.R. The Positive Role of Conflict in Cooperative Systems. In eds: Demazeau Y. and Muller J-P. '89 Decentralized Artificial Intelligence. Proceedings of the 1st European Workshop on Modelling Autonomous Agents in a Multi-Agent World. Elsevier, Amsterdam, 1990.

[16] Galliers, J.R. A Theoretical Framework for Computer Models of Cooperative Dialogue, Acknowledging Multi-Agent Conflict. PhD thesis. Cambridge University Computer Lab. Tech Report No. 172, and HCRL, Open University Tech Report No. 51, 1989.

[17] Galliers, J.R. A Strategic Framework for Multi-Agent Cooperative Dialogue. Proceedings of the Eighth European Conference on Artificial Intelligence, Munich, pp 415-420, August, 1988.

[18] Gärdenfors P. Knowledge in Flux: Modeling the Dynamics of Epistemic States. MIT Press, 1988.

[19] Gärdenfors P. The Dynamics of Belief Systems: Foundations vs Coherence Theories. To appear in Revue Internationale de Philosophie, 1989.

[20] Gasser L. et al. Representing and Using Organizational Knowledge in Distributed AI Systems. In eds: Gasser and Huhns. DAI Volume 2. Pitman, London, 1989.

[21] Harman G. Change in View - Principles in Reasoning. Bradford Book, MIT Press, Camb., Mass. 1986

[22] Hewitt, K. Offices are Open Systems. ACM Transactions on Office Information Systems, 4(3) pp 271-287, 1986.

[23] Jeffrey R.C. The Logic of Decision. Univ. of Chicago Press, Chicago, 1983.

[24] Kelleher G. An Overview of Assumption Based Reason Maintenance. Technical Report CBLU-ULE/043, Computer Based Learning Unit, Univ. of Leeds, 1988.

[25] Konolige K. Hierarchic Autoepistemic Theories for Nonmonotonic Reasoning. Tech. Note 446, SRI International, California, 1988.

[26] Konolige K. Defeasible argumentation in reasoning about events. Proceedings of the International Symposium on Machine Intelligence and Systems. Italy 1989

[27] Levi I. Truth, Fallibility and the Growth of Knowledge. in Decisions and Revisions, Cambridge University Press, 1984.

[28] Martins J.P. and Shapiro S.C. A Model for Belief Revision. Artificial Intelligence, Vol. 35 No. 1, pp 25-79, 1988.

[29] McAllister D.A. An Outlook on Truth Maintenance. AI Lab, AIM no. 551, MIT, Camb. Mass. 1982.

[30] Nebel. B. A Knowledge Level Analysis of Belief Revision. in Proceedings of 1st Conference on Principles of Knowledge Representation and Reasoning. Toronto, Canada. 1989.

[31] Perrault, C.R. An application of Default logic to Speech Act Theory. Report No. CSLI 87-90, CLSI, SRI International, California, U.S.A., 1987.

[32] Poole D.L. On the Comparison of Theories; Preferring the most Specific Explanation. Proc. IJCAI, Los Angeles, pp144-147, 1985.

[33] Rao A.S. and Foo N.Y. Minimal Change and Maximal Coherence: A Basis for Belief Revision and Reasoning about Actions. in Proceedings IJCAI '89, Detroit, U.S.A. 1989a.

[34] Rao A.S. and Foo N.Y. Formal Theories of Belief Revision. Proceedings of 1st Conference on Principles of Knowledge Representation and Reasoning. Toronto, Canada. 1989b.

[35] Reichgelt H. The Place of Defaults in a Reasoning System. in Kelleher G. and Smith B.(eds) Reason Maintenance Systems and Their Applications. Ellis Horwood, 1988.

[36] Reiter R. A Logic for Default Reasoning. Artificial Intelligence Vol. 13, pp81-132, 1980.

[37] Reiter R. and de Kleer J. Formal foundations of assumption-based truth maintenance systems: preliminary report. Proceedings of 6th AAAI, pp183-187, Seattle, Washington. 1987.

[38] Rescher N. Hypothetical Reasoning. North-Holland Publ. Co., Amsterdam, 1964

[39] Rosenschein, S. A Cognitive Architecture for Rational Agents. SRI Report, 1988

[40] Russell S.J. Execution Architectures and Compilation. IJCAI '89, Detroit, U.S.A. 1989.

[41] Shafer G. A Mathematical Theory of Evidence. Princeton Univ. Press, Princeton, NJ, 1976.

[42] Swain M. ed. Induction, Acceptance and Rational Belief", Reidel, Dordrecht, 1970.

[43] Thagard P. Explanatory Coherence. Behavioural and Brain Sciences Vol. 12 No. 3. 1989.

[44] Thost M. Generating facts from Opinions with Information Source Models. in Proceedings of IJCAI '89, Detroit, U.S.A. 1989.

[45] Touretzky D.S. The Mathematics of Inheritance Systems. Pitmans Research Notes in Artificial Intelligence, Pitman Publ. Ltd., London, 1986.