Reset reproduction of CMU Computer Science report CMU-CS-85-121. Published in *IJCAI-85*, pp. 87-90 (1985). Reprinted July 1994. Current address: MIT Laboratory for Computer Science, Cambridge, Massachusetts.

Reasoned Assumptions and Pareto Optimality

Jon Doyle

Computer Science Department
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213
U.S.A.

Abstract: Default and non-monotonic inference rules are not really epistemological statements, but are instead desires or preferences of the agent about the makeup of its own mental state (epistemic or otherwise). The fundamental relation in non-monotonic logic is not so much self-knowledge as self-choice or self-determination, and the fundamental justification of the interpretations and structures involved come from decision theory and economics rather than from logic and epistemology.

December 17, 1984
Revised January 29, 1985

For presentation at IJCAI-85.

This paper is based on a talk presented at the CSLI Workshop on Planning and Practical Reasoning, Stanford University, June 1984. I thank Joseph Schatz for his advice and comments, and Monnett Hanvey and one of the IJCAI referees for discovering an error in an earlier draft. This research was supported by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 3597, monitored by the Air Force Avionics Laboratory under Contract F33615-81-K-1539. The views and conclusions contained in this document are those of the author, and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Government of the United States of America.

Introduction

Non-monotonic logic and other formulations of non-monotonic reasoning schemes have usually been presented as studies of a logical or epistemological topic, that is, concerned with belief, self-knowledge, and self-consistency. While some topics in the literature are properly logical ones (for example circumscription—see [Doyle 1984]), recent work on reasoned assumptions ([Doyle 1982]) indicates that non-monotonic logic and reason maintenance are more psychological topics than logical or epistemological ones, and are closer to economics and decision theory than to logic. This paper reviews the work on reasoned assumptions to point out these connections. Related considerations appear to motivate recent work by Borgida and Imielinski [1984], who relate non-monotonic logic to committee decisions but do not draw on the decision-making literature.

To summarize the present view, default rules and non-monotonic inferences are not really epistemological statements (such as generalizations and inductive hypotheses from which one draws conclusions in a statistical or inductive logic), but are instead desires or preferences of the agent about the makeup of its own mental state (epistemic or otherwise). The fundamental relation is not so much self-knowledge as self-choice or self-determination, and the fundamental justification of the interpretations and structures involved come from decision theory and economics rather than from logic and epistemology. Self-consistency is prominent in non-monotonic logic simply by accident, because it is one convenient encoding of presence and absence of beliefs within the logical language, not because the issues involve it. Other encodings do not involve self-consistency at all (see [Doyle 1982]).

What are reasoned assumptions for?

One basic problem faced in thinking about what to do is that there are too many things to think about, and not enough (or too much) known about them. One approach to this problem taken in artificial intelligence is to reduce the required effort and circumstantial uncertainty by using "habitual" rules of thumb. Standard assumptions are used to remove uncertainties (rules of anti-agnosticism), and standard decisions are used to resolve conflicts (rules of anti-confusion). These rules of thumb are either formulated ahead of time and applied when needed to get the assumptions and decisions, or are formulated and applied when needed, or both. These rules of thumb have many applications in artificial intelligence systems, especially as sources of defaults and expectations in plans, frames, and inheritance hierarchies, but we will not go into any applications here.

What are reasoned assumptions?

We consider here only a very simple sort of rule of thumb, the *simple reason*. A simple reason $A \setminus\setminus B \mid\!\!\sim C$ (read "A without B gives C") means, in the case that A, B, and C are sets of potential beliefs, that the items in C should be believed if each of those in A are believed and none of those in B are believed. Conclusions drawn from simple reasons are called *reasoned assumptions*, reasoned because they are derived from reasons, and assumptions because they may depend on the absence of defeating or qualifying information. Simple reasons have corresponding meanings in case other mental elements like desires, intentions, concepts, fears, etc. are involved, but we treat only the case of belief to simplify the discussion.

For example, one simple reason might be

$$\{BEL(\text{Fred is a bird})\} \setminus\setminus \{BEL(\text{Fred cannot fly})\} \mid\!\!\sim \{BEL(\text{Fred can fly})\}.$$

The intent of this reason is to make the assumption that Fred can fly as long as there are reasons for believing that Fred is a bird and no known reasons for believing that Fred can't fly.

The precise interpretation of a simple reason has two principal parts. The first partial interpretation is as one of the agent's self-specifications, as a possibly non-monotonic "closure condition" on

the agent's set of beliefs. We say that a mental state is *admissible* (with respect to simple reasons) iff it satisfies each of its component simple reasons. That is, a state S is admissible just in case for each simple reason $r \in S$, if $A(r) \subseteq S$ and $B(r) \cap S = \emptyset$, then $C(r) \subseteq S$.

For example, let $S_0 = \{\emptyset \setminus\setminus \{1\} \mid\!\!\vdash \{2\}, \quad \emptyset \setminus\setminus \{2\} \mid\!\!\vdash \{1\}\}$, where 1 and 2 stand for possible beliefs. Then S_0 is not admissible, but its three supersets $S_1 = S_0 \cup \{1\}$, $S_2 = S_0 \cup \{2\}$, and $S_{12} = S_0 \cup \{1, 2\}$ are admissible. S_1, S_2, and S_{12} are called *expansions* of S_0. (Expansions are just admissible supersets. [Doyle 1982] uses the term "extension" instead of expansion.)

The second partial interpretation of a simple reason is as one of the agent's restrictions on derivability or arguability of conclusions, that is, as non-monotonic "inference records." We say that a state E is an *admissible expansion* of a set S iff each element of the state is "grounded" in S. There are several different interesting notions of groundedness. The two principal ones are *local* groundedness, in which each element has an immediate argument from S and E, and *strict* groundedness, in which each element has a noncircular argument from S alone. Formally, E is locally grounded in S just in case for each $x \in E$, either $x \in S$ or $x \in C(r)$ for some $r \in E$ with $A(r) \subseteq E \subseteq [B(r)]^c$. E is strictly grounded in S just in case for each $x \in E$ there is a finite sequence $\langle g_0, \ldots, g_n \rangle$ of elements of E such that $x = g_n$ and for each $i \leq n$, either $g_i \in S$ or there is some $j < i$ such that (1) $g_i \in C(g_j)$, (2) for each $y \in A(g_j)$, $y = g_k$ for some $k < j$, and (3) $z \notin E$ for each $z \in B(g_j)$.

Continuing the example above, S_1 and S_2 are each locally and strictly grounded in S_0, while S_{12} is not locally grounded or strictly grounded in S_0, but is locally grounded and strictly grounded in itself. Let $S_0' = \{\{1\} \setminus\setminus \emptyset \mid\!\!\vdash \{2\}, \quad \{2\} \setminus\setminus \emptyset \mid\!\!\vdash \{1\}\}$, $S_1' = S_0' \cup \{1\}$, $S_2' = S_0' \cup \{2\}$, and $S_{12}' = S_0' \cup \{1, 2\}$. Then only S_{12}' is an expansion of S_0', and it is locally but not strictly grounded in S_0'. Moreover, S_{12}' is strictly grounded in each of S_1' and S_2'.

The theory of simple reasons, elements of which are developed in [Doyle 1982], reveals several important facts about the structure of admissible expansions. Recall that one can intuitively interpret each simple reason as a sort of closure condition. One theorem is that these "local" closure conditions can be combined into one "global" one. That is, the strictly grounded expansions of a set are exactly the fixed points of a natural "closure" operator. These fixed points are not always limits of the operator, and non-limits are sometimes non-computable (as in non-monotonic logic). But conceptually, this result permits one to think of strictly grounded expansions as equilibria. From this point of view, they are reminiscent of the more abstract notion of "reflective equilibrium" described by Rawls [1971]—the conscious agreement of principles and judgments—and of the decision-theoretic notion of "ratifiable" decision proposed by Jeffrey [1983]—decisions rationally chosen for the person one expects to be once one has chosen. In the following, we will see that these similarities to nominally political and economic theories are no accident.

Group decision making

The fundamental problem in the theory of group decision-making is that of combining individual attitudes (beliefs, preferences, etc.) to get the corresponding attitudes of the group. Pollsters, elections, markets, and committees are some of the most respected means commonly used; bureaucrats, politicians, dictators, and prophets are other means. The underlying theory developed by economists abstracts from the hurly-burly of actual decision-making in groups of people by studying "decision rules," mathematical functions that describe the input-output behavior of decision processes, the inputs being the attitudes of the individuals, the outputs being the resulting group attitudes. Due to the abstractions commonly used, economists typically discuss only the case of combining individual preferences to get group preferences, and we will also concentrate on that case here. (See [Arrow 1963] and [Debreu 1959] for classic treatments.)

Economists abstract from humans and other biological species and base their theories on a creature called *homo economicus*, or economic man. Economic men (and presumably economic women and children too) are marvels of consistency and calculation, but their powers of calculation are not relevant here. For our purposes, the most important property of the economist's abstraction is it requirement of

consistent preferences. We write $x < y$ to mean that the agent prefers x to y. If the agent is an economic man, then its preferences are consistent, that is, it does not prefer x to y and y to x. Economic man's preferences are also transitive ($x < y$ and $y < z$ entail $x < z$). If neither $x < y$ nor $y < x$ we say that the agent is indifferent about the choice of x or y (or in the common corruption, "x and y are indifferent").

In the theory of group decision-making, each individual is internally consistent in the above sense, but different individuals may contradict each other. That is, as long as Alice and Bob are individually consistent, we may have in some case that $x <_{\text{Alice}} y$ and $y <_{\text{Bob}} x$. Indeed, if we never did, group choice would not be a very interesting topic. But people do disagree, and group choice is difficult for the theory requires that group preferences be consistent as well, that is, that the results of choice result in a composite economic man, a "group individual." The conflicts must somehow be resolved.

In mathematical form, the theory of group decision-making considers functions (decision rules) from sets of consistent sets of preferences (those of the group members) to consistent sets of preferences (those of the group as a whole). In addition to restricting the input and output preference sets to be consistent, the theory also restricts the sorts of functions allowed. This is done because there are many decision rules the economist deems uninteresting or undesirable. For example, one such rule is the rule of "apathy," namely the rule yielding the constant result of the empty set of preferences. Having no preferences at all is perfectly consistent, and this rule says that all groups are indifferent about all choices. This sort of rule is not very useful in practice, and to avoid it (among others) the economist restricts consideration of decision rules to those satisfying the *Pareto condition* (named after Vilfredo Pareto, an Italian economist of the last century). The Pareto condition is that if all members prefer x to y, then the group prefers x to y, that is, the group agrees with unanimous opinion. The rule of apathy fails to satisfy this condition, since it admits no group preferences, even when the individuals are unanimous.

Pareto optimality is an even stronger condition than the Pareto condition. Pareto optimality means the group preferences agree with as many individual preferences as possible, or more precisely, that any change to the group preferences that better satisfies one member must satisfy another member less. Satisfying the Pareto condition implies Pareto optimality when agents are never indifferent. To see this, suppose the group preferences include $x < y$, and that Alice has the opposite preference, $y <_{\text{Alice}} x$. Since the group preferences satisfy the Pareto condition, we know that some other member must disagree with Alice. Otherwise, the members would be unanimous about $y < x$, and the group preferences would have to agree. Let us assume $x <_{\text{Bob}} y$. Then it is easy to see that if we changed the group preference to $y < x$ to make Alice happier, we would violate Bob's preference and make him unhappier. That's Pareto optimality.

Reasoned assumptions and Pareto optimality

We are now in a position to indicate the link between these topics. Consider a group decision concerning invitations to a party. We assume that each person p has a very simple set of preferences, which we abbreviate as $A(p) \setminus\setminus B(p) \parallel\!\!- C(p)$, meaning that p prefers sets S of invitations that satisfy the condition $A(p) \subseteq S \subseteq [B(p)]^c \supset C(p) \subseteq S$ to those sets not satisfying this condition, and is indifferent to choices within these two classes. We might express p's preferences as "if we invite the A's and don't invite the B's, we should invite the C's." What are reasonable choices for invitations if we take the preferences of the guests as seriously as the preferences of the hosts?

This decision about party invitations should look familiar. It is not hard to prove that with these individual preferences, the Pareto optimal sets are exactly the admissible states defined earlier. ([Doyle 1982] employs the term "satisfaction optimal" for this result.) Put another way, interpreting simple reasons as preferences about mental states leads naturally to the idea of admissible states, and to the interpretation of mental states as decisions by the agent about what to be. The moral is that reasoned assumptions are economic entities. (This does not exclude them from being logical or other sorts of entities as well.)

It is worth noting that these interpretations take some liberties with the standard formulations of group decision-making. In our applications, the group membership may vary with the choice, and preferences about preferences (reasons about reasons) are basic. These differences lead to an interesting extension of the usual economic theory (in preparation).

One may carry the economic interpretation of reasons further. Consider agents like the party-goers before, except that among the sets of invitees satisfying the individual conditions, these folks prefer "validating" satisfying sets to "invalidating" satisfying sets, that is, p prefers sets S such that $A(p) \subseteq S \subseteq [B(p)]^c$ to those failing this condition. With these more refined preferences, one can prove that E is Pareto optimal among the strictly grounded expansions of S whenever E is a strictly grounded expansion of S. (See [Doyle 1982] on validity optimality.) I do not know if the converse is true, but the moral here extends the previous one. Not only are the admissible mental states natural economic entities, but the derivability relation of strictly grounded expansion is an economic notion as well.

Reflections on economic theory

The preceding has indicated the economic nature of some important notions in artificial intelligence. In this section, we draw some conclusions about the role of economic theory in artificial intelligence, and about the possible influence of artificial intelligence on economic theory.

First, the identification of reasoned assumptions as the results of group choices constitutes a substantial regression of the usual theory of individual choice. Current economics is based on the economic man idealization of humans. Since humans fall short of this ideal, some people are tempted to dismiss economic theory as useless if not wrong. But the previous discussion suggests that whatever their value as a model of humans, the standard idealizations of economics may be quite useful at a lower level than the whole human. The elementary building blocks (e.g. reasons, procedures, etc.) may all be consistent, if simple, individuals, with more complex mental faculties and the mind as a whole stemming from the activity of these elements, that is, as group choice (see [Doyle 1983b]). Minsky has urged something like this view for years with his "society of mind." It is instructive to find a very similar proposal (even including some neurodevelopmental epistemology) in the works of the economist Robert Mundell [1968, Ch. 18]. Once one grasps this point of view, it becomes clear that much of artificial intelligence has labored under an enormously restrictive methodology, that of seeking purely bureaucratic organizations. With few exceptions (notably Minsky's "heterarchy" and society theories, perceptron-class theories, and more recently the work of Thomas Malone), artificial intelligence has ignored market-like structures, which appear to have great information-transmission advantages over bureaucracies in many circumstances. Bureaucratic organization may be best for large numbers of specific computational problems, but it seems unreasonable to ignore other organizations for tasks as complex as thinking until that is shown to be reasonable. This ignorance is especially dangerous if one attempts to adapt artificial intelligence ideas to the design of human organizations.

If the economic interpretation of reasoned assumptions provides a role for standard economic theory in artificial intelligence, other artificial intelligence notions might be used to broaden economic theory. For example, non-monotonic rules of reasoning are pervasive in artificial intelligence, but violate most of the economist's idealizing axioms. Specifically, an agent based on the theory of simple reasons violates axioms about the consistency and transitivity of preferences and about the independence of irrelevant alternatives. A simple reasons agent may easily accommodate conflicting preferences (defaults), and come to different (possibly inconsistent) conclusions depending on the complete set of reasons involved. This isn't some undesirable side-effect of the artificial intelligence ideas. These properties are what non-monotonicity is all about. Eventually artificial intelligence may provide new idealized psychologies on which to base economics, psychologies as precisely formulated as that of *homo economicus*, but much closer to the informational, computational, and logical limitations of humans. I am studying such possibilities in my work on rational psychology, and a comprehensive treatment is in preparation.

Conclusion

We have seen how the important artificial intelligence concept of reasoned assumption has a natural economic interpretation, in which reasons and defaults express preferences of the agent about the composition of its own state of mind, and in which the admissible states of mind are the results of decisions based on these preferences.

As a final note, we remark that because artificial intelligence agents are non-atomic and these decisions are group decisions, the results may be ambiguous in the sense that several incomparable outcomes are possible. (In fact, except in dictatorial organizations, such ambiguities may be hard to avoid.) [Doyle 1982] shows how one may develop a theory of degree of belief or subjective probability by measuring these ambiguities. This "epiphenomenal" role for subjective probabilities appears to fit the practical needs of artificial intelligence much better than the "foundational" role given to them in standard applications of decision theory, as the qualitative information used in reasons is often easier to obtain and to modify than the quantitative information derived from it. See [Doyle 1983a] and [Doyle 1983c] for more on this idea.

References

Arrow, K. J., 1963. *Social Choice and Individual Values*, second edition, New Haven: Yale University Press.

Borgida, A., and Imielinski, T., 1984. Decision-making in committees—a framework for dealing with inconsistency and non-monotonicity (extended abstract), *AAAI Workshop on Non-Monotonic Reasoning*, 21-32.

Debreu, G., 1959. *Theory of Value: an axiomatic analysis of economic equilibrium*, New Haven: Yale University Press.

Doyle, J., 1982. Some theories of reasoned assumptions: an essay in rational psychology, Pittsburgh: Department of Computer Science, Carnegie-Mellon University. See especially sections 3-8 and 25-27.

Doyle, J., 1983a. Methodological simplicity in expert system construction: the case of judgments and reasoned assumptions, *AI Magazine* **3**, #2, 39-43.

Doyle, J., 1983b. A society of mind: multiple perspectives, reasoned assumptions, and virtual copies, *Eighth International Joint Conference on Artificial Intelligence*.

Doyle, J., 1983c. What should AI want from the supercomputers?, *AI Magazine*, V. 4, No. 4, 33-35, 31.

Doyle, J., 1984. Circumscription and implicit definability, *AAAI Workshop on Non-Monotonic Reasoning*, 57-69.

Jeffrey, R. C., 1983. *The Logic of Decision*, second edition, Chicago: University of Chicago Press.

Mundell, R. A., 1968. *Man* and *Economics*, New York: McGraw-Hill. See especially chapter 18.

Rawls, J., 1971. *A Theory of Justice*, Cambridge: Harvard University Press.