Reset reproduction of CMU Computer Science report CMU-CS-88-111. Reprinted July 1994.
 Current address: MIT Laboratory for Computer Science, Cambridge, Massachusetts.

On Universal Theories of Defaults

Jon Doyle

Department of Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

March 17, 1988

Abstract: Though unifications of some of the numerous theories of default reasoning have been found, we add to doubts about the existence of universal theories by viewing default reasoning from the standpoint of decision theory as a case of rational self-government of inference. Default rules express not only methods for deriving new conclusions from old, but also preferences among sets of possible conclusions. Conflicting default rules, which form the central difficulty in the theories, represent inconsistent preferences about conclusions. These conflicting rules cannot be avoided, as they arise naturally in practice, especially in databases representing the knowledge of several experts. We compare these theories of rational inference with theories of group decision making, and develop doubts about universal theories of the former by considering well-known negative results about the latter.

This research was supported by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 4976, Amendment 19, monitored by the Air Force Avionics Laboratory under Contract F33615-87-C-1499. 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.

1 Introduction

Many of the inferences important in ordinary and specialized reasoning are what artificial intelligence calls *default* inferences, or cases of reasoning by default. For example, two inferences reasonable to most natives of the U.S. are

1. Accountants are mild mannered.
   Bill is an accountant.
   Therefore, Bill is mild mannered.

2. Hell's Angels are rough mannered.
   Spike is a Hell's Angel.
   Therefore, Spike is rough mannered.

In contrast to deductive inferences, these inferences might be incorrect in some circumstances, as might be many inferences from generalizations phrased in terms of plural nouns and verbs. For example, inference (1) might be wrong if Bill is drunk, or if Bill is a top troubleshooter for an aggressive Wall Street corporate-takeover firm, and inference (2) might be wrong if Spike is tranquilized, or more interestingly, once we learn that "Spike" is Bill's nickname. This latter case will be of special interest later.

Artificial intelligence employs many sorts of rules for making many sorts of default assumptions. There are *prototypical* assumptions, such as inferences (1) and (2) above, for drawing conclusions in whole classes of cases; *particular* assumptions, for drawing individual conclusions in specific circumstances (for example, "Ordinarily, Bill is mild mannered; therefore, Bill is mild mannered"); and *systemic* assumptions, such as Occam's Razor and blanket assumptions that what one doesn't know to be true must be false.

Many formal theories have been developed to make precise these ways of making assumptions. Prototypical assumptions were formulated only in computational terms initially as frame defaults [Minsky 1975] and as inheritance networks (e.g., [Fahlman 1979]), and were later formalized in theories of inheritance (e.g., [Touretzky 1986]). Similarly, the first theories of particular assumptions were formulated computationally as nonmonotonic justifications or reasons in reason maintenance systems (RMS's) [Doyle 1979] and later formalized as nonmonotonic logic [McDermott and Doyle 1980], default logic [Reiter 1980], autoepistemic logic [Moore 1983], and simple reasons [Doyle 1983a],

with most of these formalizations generalizing naturally to incorporate some sorts of prototypical assumptions. Systemic assumptions, on the other hand, have been captured in operations like circumscription [McCarthy 1980] and closed world reasoning [Reiter 1979], which began as formal theories (though the closed world assumption describes an older computational practice).

These theories have very different characters. Perhaps the most pronounced difference concerns determinism of inference. Some of these theories, namely circumscription, closed world reasoning, and so-called "skeptical" inheritance [Horty et al. 1987], resemble ordinary logic in that they describe how a set of axioms or rules yields a single set of conclusions closed under inference. But other theories, including nonmonotonic logic, autoepistemic logic, default logic, simple reasons, and "credulous" inheritance [Touretzky et al. 1987], describe how a single set of axioms and rules may yield several different, often incompatible sets of conclusions closed under inference. While some theories have been proposed as unifications or partial unifications of some of these ways of making assumptions (see, for example, [Konolige 1987], [Etherington 1987a], [Shoham 1987]), doubts about the existence of complete unifications have been expressed, notably by Touretzky, Horty, and Thomason [1987], who argue that the gross differences between the theories stem substantial from differences in the underlying intuitions about how to make assumptions. As they put it, the differing theories reflect a "clash of intuitions." Indeed, this clash of intuitions was recognized earlier, as a fact if not a conclusion, by designers of some knowledge representation systems who explicitly surrendered the aim of providing a general inheritance mechanism for the more modest goal of allowing the user to program his own inheritance mechanisms.

Our purpose in this paper is to ask if this clash of intuitions is real, and more generally, if a unified theory of defaults exists. Note that a unified theory may exist even if different intuitions are involved, as long as these different intuitions can be shown to apply to disjoint cases of inferences. In that event, the unified theory is merely a "big switch" or sum of the theories of each of the cases. To investigate this question, we translate these questions about default inference into the context of rational decision making, and ask if the clash is real there. If there is no clash in the more traditional context, the unified theory may be transferred back to the case of default inference. To effect the transfer, we view default reasoning as rational inference, basing our treatment on the original presentation of this view in [Doyle 1983a], and on the

refined presentation in [Doyle 1987]. In particular, we view default reasoning as rational inference in two ways: taking the purpose or selection of default rules as rationally guided, and taking the meaning or interpretation of default rules as rational choice. We discuss these in turn.

2 Rational evaluation of defaults

Thinking often begins with making guesses grounded in one's experience. Guessing, or making assumptions, is often held in disrepute as illogical. In fact, though illogical, it is often quite the rational thing to do. Taking action requires information about the available actions, about their expected consequences, and about the utility of these consequences to the agent. Ordinarily, obtaining such information requires effort, it being costly to acquire the raw data and costly to analyze the data for the information desired. But the first limitation faced in limited reasoning is that one cannot either know or consider everything, and so must ignore most possibilities, relying on reasonable assumptions until they prove wrong. Thus to minimize or avoid information-gathering and inference-making costs, artificial intelligence makes heavy use of heuristics—rules of thumb, defaults, approximately correct generalizations—to guess at the required information, to guess the expected conditions and expected conclusions. These guesses are cheap, thus saving or deferring the acquisition and analysis costs. But because they are guesses, they may be wrong, and so these savings must be weighed against the expected costs of making errors. Many of the cases of default reasoning appearing in artificial intelligence represent judgments that, in each particular case, it is easier to make an informed guess and often be right than to remain agnostic and work to gather the information; that errors will be easily correctable and ultimately inconsequential; and that the true information needed to correct or verify these guesses may well become available later anyway in the ordinary course of things. In other cases, defaults are avoided, either because there is no information available to inform the guess, or because even temporary errors of judgment are considered dangerous.

Thus rationality may be applied as a standard motivating the adoption of individual defaults in a very familiar way, by saying that an assumption or rule of assumption should be adopted if the expected utility of holding it exceeds the expected utility of not holding it. Utility, as it is understood in decision

theory, is a more general notion than simple costs and benefits, which are merely two of the considerations that may enter (possibly in a highly nonlinear way) into an agent's comparisions of defaults, but a complete detailing of all the forms of utility in reasoning is beyond the scope of this paper. Applied to individual assumptions, rational evaluation is a familiar idea, famous under the names of Pascal's wager in the case of religious belief, and James' "will to believe" for the general case of religious and mundane beliefs.

2.1 Pascal's wager

Pascal [1662] framed his problem of belief in God as the following: he can either believe or doubt the existence of God, and God may either exist or not exist. If God exists and Pascal believes, he gains eternal salvation, but if he doubts he suffers eternal damnation. If God does not exist, belief may lead Pascal to forgo a few possible pleasures during his life that doubt would permit him to enjoy. We may summarize these evaluations in a decision matrix

Pascal's decision	God exists	doesn't
Believe	$+\infty$	$-f$
Doubt	$-\infty$	$+f$

where f represents the finite pleasures enjoyed or forgone during Pascal's life. Of course, these same quantities modify the first column as well, but finite modifications to infinities are negligible. As long as God's existence is not judged impossible, the expected utility of belief is $+\infty$, dominating the expected utility of doubt, $-\infty$.

2.2 James' will to believe

James [1897] pointed out that many sorts of rational belief are ubiquitous in mundane reasoning, and explained these as cases of the "will to believe." For example, a mundane parallel to Pascal's belief is that in the morning my habit is to get in my car with my notebooks and start the car, in order to drive into work. Now the car might either be working or broken. It must be working for me to be able to use it to drive to work, but I do not check to see that it is before trying to start it. I simply assume it is working when I plan and pack the car. We can frame my decision in a matrix

My decision	car works	doesn't
believe	$+b-c$	$-c-C$
doubt	$+b-C$	$-C$.

Here we write the benefit from the car starting as b, the cost in effort of packing and starting the car as c, and the cost of checking out the engine, electrical system, transmission, etc. as C, where we assume $C \gg c$. With these utilities, the expected value of believing dominates that of doubting whenever $pC > c$, where p is the probability that the car works. As long as I expect the car to work and $C \gg c$, my assumption is reasonable.

James did not contend that the reasoning by which assumptions are made should be rational calculation. Rather, his point was that skepticism is not always rational, that in many cases it is better to adopt a stance on some issue and risk error than to take no stance at all. The position taken need not be precedential, for along with recognition of the possibility of error, we may also recognize that other or later circumstances raising similar questions may be decided differently. In Tukey's [1960] phrase, we often decide to act for the time being as if something were the case, rather than simply deciding something is the case. But precedential or not, the approach of adopting stances carries with it a commitment to correcting mistakes when they come to light. As James puts it, we might resolve to *Believe Truth!* and to *Shun Error!*, but the preceding suggests the latter resolve is best realized as conscientious correction rather than intellectual cowardice.

2.3 Rational default rules

In artificial intelligence, rules for making default assumptions have been at issue more than individual assumptions, though the two cases may be assimilated since individual default rules may be evaluated as individual assumptions. This view of adoption of default rules has been urged by [Doyle 1983a] and more recently by [Shoham 1987]. By taking utility to be a simple function of application costs and speedup benefits, [Smith 1985], [Langlotz, et al. 1986], and [Minton 1988] have applied rational evaluation to concrete cases of selection of inference rules. Such economic calculations may be made either at the time the information is needed or, as in the default rules prominent in inheritance systems and reason maintenance, in advance. When made in

advance, the rules may be applied to produce the assumptions either when needed during retrieval, as in most inheritance systems, or in advance, as in reason maintenance.

The main problem faced in practical application of this idea is that often our knowledge of the utility of assumptions in general, and of the costs and benefits of assumptions in particular, is incomplete and itself consists mainly of guesses. We can of course look for selections rational with respect to these guesses, but trial and error is the rule with these cases, so we call them heuristics rather than calculated assumptions. In most current artificial intelligence systems, these judgments or calculations are made by the system's designer—the human informants decide what the good guesses are, and these are encoded into the rules that the machine obeys. Alternatively, these considerations may be made by the agent itself as well through reflection and reasoning.

There have been attempts in the artificial intelligence literature to view heuristics or rules for making assumptions purely in probabilistic terms, with a rule of assumption justified as long as its probability exceeds some threshold value, or better yet, as the limiting case of small uncertainties (see [Pearl 1987]). This theory of assumptions is inadequate, first because it ignores preferences about holding beliefs, and second because it is based on the probability of truth of the belief, not on the probabilities of the consequences of belief. For example, tautologies have maximum probability of truth but are generally worthless as beliefs. No agent should waste its effort assuming most tautologies, since tautologically, an assumption is worth the expense of making it only if it is worth it—that is, if the expected utility of making it is high enough. Since the probabilistic theory of assumptions ignores the utility or disutility of assumptions, it is a theory of likely irrelevancies, of tasteless theorizing. Moreover, in some cases it is rational to make assumptions expected to be false. For example, when asking directions of a soldier on an army base, if one cannot read insignias of rank it is advisable to assume the soldier has high rank (such as colonel). Since there many more soldiers of low rank than of high rank, one expects this assumption to be false. But one consequence of error is to have the soldier give his true rank, and the consequence of the expected error is to flatter the soldier, making him more eager to help. More generally, one may judge lying rational just as one may judge honesty rational. Certainly lying to oneself would not be as common as it is if it did not offer some sort of large reward.

A similar but less popular mistake is to base assumptions purely on utilities, assuming something as long as its utility exceeds some threshold, regardless of the probability of its being true. This theory has exactly the same irrational character as the probabilistic theory of assumptions, and has a standard name as well. It is called wishful thinking.

3 Rational interpretation of default rules

The problem of interpreting sets of default rules conceptually is much more difficult than the problem of selection, even though strictly speaking the evaluation needed for rational selection involves interpreting the default rules.

It is natural to think of default rules as having both intentional and preferential content. That is, default rules not only state conditions that sets of conclusions must satisfy (such conditions being intentional content like the content of ordinary axioms, which state beliefs the agent must hold); defaults also indicate which sets of conclusions are preferred to others, that is, which should be held before others if possible. We examine these two interpretations in turn.

3.1 Default rules as constitutive intentions

In their first interpretation, default rules represent constitutive intentions: intentions of the agent about its own constitution or makeup. Each default rule is thus a specification on legal mental states, and legal states must satisfy each of the self-specifications they contain. ([Doyle 1983b] called this sort of semantics for representations "admissible state semantics." Also, when default rules are taken as external specifications rather than the agent's self-specifications, they form what database theorists call "integrity constraints" [Reiter 1988].)

Formally, let R be a set of rules. (We will ignore the question of syntax in this paper, as it is irrelevant to the matters at issue.) For each rule $r \in R$ we write $\mathcal{I}(r)$ to mean the set of sets of conclusions legal according to r. In other words, a set of conclusions S is legal according to r just in case $S \in \mathcal{I}(r)$. For example, one very simple sort of default rule is the propositional default or *simple reason* $A \setminus\setminus B \mid\!\!\sim C$ (read "A without B gives C"), where A, B, and C are sets of propositions or attitudes. (See [Doyle 1983a].) If we write $\mathcal{D}$ to mean the set of all possible propositions or attitudes, then we may view

each state of mind of the agent as a set $S \subseteq \mathcal{D}$. Writing the complement $\mathcal{D} - B$ of B relative to $\mathcal{D}$ as B^c, we may formalize the intentional content of $r = A \setminus\setminus B \mid\!\!\vdash C$ as

$$\mathcal{I}(r) = \{S \subseteq \mathcal{D} \mid A \subseteq S \subseteq B^c \supset C \subseteq S\},$$

that is, S is legal according to r if it contains C whenever it contains A and contains no element of B. For instance, if $r = \{a\} \setminus\setminus \{b\} \mid\!\!\vdash \{c\}$ and $r \in S$, then S is legal according to r if either $a \notin S$, if $b \in S$, or if $c \in S$. That is,

$$\mathcal{I}(r) = \{S \subseteq \mathcal{D} \mid a \notin S \vee b \in S \vee c \in S\}.$$

To consider entire sets of conclusions at once, we say that S is legal if it is legal according to each default rule it contains, that is, if $S \in \mathcal{I}(r)$ for each default rule $r \in S$, which we may write as $S \in \mathcal{I}(R)$. More generally, we can say that S is legal iff $S \in \mathcal{I}(x)$ for each $x \in S$ if we assume that $\mathcal{I}(x) = \mathbf{P}\mathcal{D}$ (the vacuous specification) for every $x \in \mathcal{D}$ that is not a default rule. It is important to note, however, that the intentional contents of rules cannot conflict irretrievably since $\mathcal{D} \in \mathcal{I}(r)$ for every rule r. Thus there is always at least one legal set of conclusions in $\mathcal{I}(R)$.

3.2 Default rules as constitutive preferences

It is easy to see that the intentional content alone is not sufficient to capture the purpose of defaults, since we do not mean defaults to merely state a disjunction. For example, according to the rule r above, each of the sets $\{a, b\}$, $\{a, c\}$, and $\{a, b, c\}$ is a legal set of conclusions. But we mean more than this when we employ the rule, namely that the set of conclusions $\{a, c\}$ be held if a is held and nothing forces b to be held. Another way of putting this is that default rules are also constitutive preferences, expressing the agent's preferences about its own legal states, in this case that states containing the set of conclusions $\{a, c\}$ be preferred to sets containing $\{a, b\}$. Formally, we view the preferential content of each default rule $r = A \setminus\setminus B \mid\!\!\vdash C$ as a strict quasi-order (transitive binary relation) $<_r$ over possible states, where $<_r$ says which states are preferred to others according to r. Specifically, we interpret r as the preference relation $<_r$ such that

$$S <_r S' \equiv \ A \subseteq S' \subseteq B^c \ \wedge \ A \subseteq S \not\subseteq B^c.$$

If we say that r is *valid* in S just in case $A \subseteq S \subseteq B^c$ and that r is *defeated* in S just in case $A \subseteq S \not\subseteq B^c$, then default rules prefer states in which they are valid to states in which they are defeated. Thus for $r = \{a\} \setminus\setminus \{b\} \mid\!\!\sim \{c\}$, r is valid only in $\{a, c\}$ and we have, for example, $\{b\} <_r \{a, c\}$ and $\{a, b, c\} <_r \{a, c\}$. Note that r does not express any preference between, for example, $\{b\}$ and $\{a, b, c\}$. We say that these sets are indifferent in $<_r$, and write $S \leq_r S'$ to mean that either $S <_r S'$ or S and S' are indifferent in $<_r$. Put another way, $S \leq_r S'$ iff $S' \not<_r S$.

Thus if we start with an initial set of reasons R, we use the preferences expressed by R to select from among the conclusions permitted according to the intentions expressed by R. We combine the individual preference orders $\leq_r$ corresponding to each rule $r \in R$ into an overall relation $\leq_R$ by defining $S' \leq_R S$ to hold just in case $S' \leq_r S$ for every $r \in R$, and then select a set of conclusions S such that (1) S is legal and satisfies the intentions expressed by every default rule in R, that is, $S \in \mathcal{I}(S)$ and $S \in \mathcal{I}(R)$, and (2) S is preferred to all other such states by the preferences expressed by reasons in R and S, that is, if $S' \in \mathcal{I}(S')$ and $S' \in \mathcal{I}(R)$, then $S' \leq_R S$ and $S' \leq_S S$. If we think of $\leq_R$ as ranking possible sets of conclusions according to their utility, then this selection maximizes the utility of the selected definitions. In Jeffrey's [1983] terminology, it produces "ratified" rational choices of conclusions.

3.3 Default inference as rational inference

Recognizing the rational nature of many assumptions makes some of the deviant logics formulated in artificial intelligence seem somewhat superfluous, since there are precise theories of both rational decisions and logical deduction extant, and the theory of assumptions is easier to understand when presented as an application of the standard theories. Unfortunately, the dominance of the logical point of view in theoretical artificial intelligence made recognition of the rational basis of assumptions slow in coming. Initially, each of circumscription, the closed world assumption, and nonmonotonic logic (but not reason maintenance) were conceived of in logical terms, and the subsequent treatments of default and autoepistemic logics maintained this conception. The rational elements of nonmonotonic logic were first stated explicitly in [Doyle 1983a] some years after the initial development of the logic. Similarly, when the initial conception of circumscription proved inadequate for applications, "priorities" were introduced into the theory to obtain prioritized circumscrip-

tion (see [McCarthy 1986], [Lifschitz 1986,1987]). Reformulated in rational terms, these priorities are special sorts of preferences about conclusions, making prioritized circumscription a sort of inference to the best explanation, that is, inference rational with respect to these priorities, rather than inference to the logically minimal conclusion. This interpretation comes out clearly in [Shoham 1987]. (See also [Etherington 1987b] and [Russell and Grosof 1987].) In Shoham's treatment, preferences are semantical rather than mental, comparing models of the agent's beliefs rather than the sets of beliefs themselves via a global partial order $\sqsubseteq$ over models (whose sense is the opposite of our order $\leq$), so that the rational conclusions from axioms are just the conclusions true in all $\sqsubseteq$-minimal models of the axioms.

4 Conflicting default rules

With both the intentional and preferential content of reasons set out, the difficulties of interpreting sets of defaults becomes clearer. The reason is that default rules can conflict; they can represent inconsistent preferences. For example, let $R = \{r_1, r_2\}$ where

$$r_1 = \emptyset \setminus\setminus \{a\} \mid\!\!\sim \{b\}$$

and

$$r_2 = \emptyset \setminus\setminus \{b\} \mid\!\!\sim \{a\}.$$

If $a =$ "Bill is rough" and $b =$ "Bill is mild," these two reasons abbreviate our expectations about Bill/Spike the accountant/Hell's Angel. Given these two reasons, the legal states containing r_1 and r_2 are just

$$\{r_1, r_2, a\}, \{r_1, r_2, b\}, \{r_1, r_2, a, b\}$$

since

$$\mathcal{I}(r_1) = \mathcal{I}(r_2) = \{S \subseteq \mathcal{D} \mid a \in S \vee b \in S\}.$$

The preferences among states are then

$$\{r_1, r_2, a\}, \{r_1, r_2, a, b\} <_{r_1} \{r_1, r_2, b\}$$

and

$$\{r_1, r_2, b\}, \{r_1, r_2, a, b\} <_{r_1} \{r_1, r_2, a\}.$$

These orders are incompatible in the sense that they cannot be consistently combined into a full order on states: that is, no strict quasi order extends and agrees with both. For example, the combined relation $\leq_R$ does not extend $<_{r_1}$ and $<_{r_2}$ since neither $\{r_1, r_2, a\} \leq_R \{r_1, r_2, b\}$ nor $\{r_1, r_2, b\} \leq_R \{r_1, r_2, a\}$ holds. Thus no choice of legal state satisfies all the preferences it contains.

Substantial disagreement has existed over whether inconsistent default rules should be accommodated by nontrivial treatment in the theory, and each of the main approaches towards default reasoning in the literature has its own way of addressing the problem of conflicting default rules. Theories like circumscription and the closed world assumption treat representations as fundamentally logical, and so do not tolerate any sort of inconsistency. They forbid the use of inconsistent default rules by forcing the states to be logically inconsistent if the state contains inconsistent preferences. The nondeterministic logics, on the other hand, tolerate inconsistent preferences, though they may still require the agent's beliefs and intentions to be consistent. Skeptical inheritance lies between these extremes, tolerating inconsistencies but drawing no conclusions from them.

Though the purely logical theories have many attractive aspects, they are inadequate for formalizing default reasoning because it is difficult to forgo conflicting defaults. There are several reasons for this. In the first place, attempting to detect conflicts ahead of time defeats the purpose of default rules to some extent, since the point of making guesses is to avoid exhaustive prior analyses.

In the second place, most practical artificial intelligence systems are designed to incorporate all the available knowledge about the relevant subjects. In expert systems, this usually means combining the expertise of several experts, so that differences between these experts must be worked out, either in advance, or while performing. In the simplest case, one might consider encoding each expert's knowledge as a separate set of rules in the system, or as justifications for a subset of the rules which name the expert proferring them. In this case, as Thomason [1986] points out, conflicts between experts become conflicts within the expert system. Of course, the system designer can instead try to reconcile these conflicts at design time, but this may not always be feasible if some conflicts are too subtle to detect, or if the experts themselves knowingly hold mutually irreconcilable opinions. Thus if the system must perform in isolation from the original experts, one must expect it will

sometimes have to deal with conflicts as they arise. For instance, most adults have had the experience of having to administer medications to themselves or to their children while on vacation, only to find that several medications have been prescribed by different doctors or for different symptoms, with each medication contraindicating the others.

In the third place, conflicting defaults arise naturally in commonsense knowledge, since common sense reflects common situations, and does not address what happens in uncommon situations. Accountants who are also Hell's Angels constitute one such uncommon occasion, with the result that our expectations conflict. Another example is the "Nixon diamond" (so called because of the shape of its diagram when written as an inheritance network; perhaps also because it is so hard): Republicans are typically not pacifists, Quakers are typically pacifists, and Nixon is a Republican Quaker. The question is, is Nixon a pacifist or not? More generally, some sets of conflicting defaults can be viewed as cases of what is known as the lottery paradox in statistics. Even though it is rational to assume that each ticket in a large lottery is a loser, taken all together such assumptions for each ticket are inconsistent with the assumption that some ticket will win. (See [Kyburg 1970] for more on this.)

The point is that there are natural examples in reasoning in which holding conflicting defaults seems not merely unavoidable, but perfectly reasonable. Even if questions remain about how they should be interpreted, as long as default rules are seen as independent elements of an agent's knowledge there need not be any way of preferring some to others in cases where they conflict. This is most clearly seen in pure lottery-like cases, in which the only important difference between the conflicting defaults cannot be discerned. Of course, some default rules might express preferences about other preferences, an approach allowed by nonmonotonic and autoepistemic logics. But the problem may return if these higher-order preferences conflict. We can expect this circumstance to arise since nothing requires the agent to have preferences about everything, and more fundamentally, since finite agents cannot have preferences about everything. Put another way, it is not just our knowledge of facts that is incomplete, but also our knowledge of absolute and relative values. Thus in cases like Hell's Angels accountants and Quaker Republicans, the missing preferences seem to be generally unavailable. (See [Levi 1986] for more on this.)

4.1 Reasoning with conflicting defaults

Even if conflicts among defaults are to be tolerated, there are still many ways to proceed. One approach is to ignore conflicts. This path is taken in skeptical inheritance, which draws conclusions only when all rules agree about those conclusions, and which avoids drawing conclusions about questions on which defaults disagree. Other ways of tolerating conflicts are not as conservative as skeptical inheritance, and use some defaults but not others. For example, one might draw conclusions based on a maximal consistent subset of the defaults. Such sets of defaults of course contain the set of mutually consistent defaults upon which skeptical inheritance is based, but may also contain more. This sort of maximal inference corresponds to credulous inheritance and to nonmonotonic, default, and autoepistemic logic. If the disagreements between different maximal consistent sets is worrying, one might instead hold those conclusions that appear no matter which maximal consistent set is chosen. This approach resembles that of skeptical inheritance, but can yield more conclusions. Alternatively, one might simplify inference to the extreme and satisfy only one default at a time, picking one to satisfy and ignoring the others. This is one way of viewing the familiar sorts of sequential production systems, whose "conflict resolution" procedures pick one applicable (valid) rule.

It is not hard to see that each of these methods for reasoning with conflicting defaults can be good or bad depending on the circumstances. For example, a doctor might be sure a patient has one of several diseases but not know which one, and know that the treatments for each of these diseases are contraindicated for the others. In this setting, the specifics of the diseases and treatments call for different ways of reasoning. The deleterious side-effects might be severe or mild, and the patient's prognosis without treatment might be good or bad. If the prognosis is good and the side-effects are severe, it seems sensible to be skeptical rather than risk added injury, but if the side-effects are mild, guessing a treatment is not unreasonable. On the other hand, if the prognosis is bad and the side-effects are severe, both skepticism and the path of applying all treatments are bad, so the best the doctor can do is to risk a guess and hope it is right.

4.2 Defaults and group decisions

Artificial intelligence is fortunate in that the problem of conflicting preferences has already been studied in great detail as the subjects of group decision theory and multi-attribute decision theory. That is, the problem of interpreting sets of default rules to reach a set of conclusions is formally identical to the problem of group choice or decision making, in which the preferences of a group of rational individuals are combined to yield a consistent set of preferences for the group, and to multi-attribute decision making, in which comparisons of each alternative along different dimensions or attributes must be combined into overall comparisons. (See [Arrow and Reynaud 1986].) Symbolically, in theories of defaults we have a set of rules R, preferences $<_r$ and global order $<_R$; in group decision theory, each member m of the group G is taken to have a consistent set of preferences, which may be represented as a strict quasi-order $<_m$ over states of affairs, and the aim is to find an order $<_G$ representing the preferences of each $m \in G$; and in multi-attribute decision theory, we have a set of attributes I, preferences $<_i$ over alternatives corresponding to each attribute $i \in I$, and seek combined preferences $<_I$.

Each of the ways of tolerating conflicts among default rules studied in theories of defaults correspond naturally to principles for making decisions studied in decision theory. The method of skeptical inheritance corresponds to the "unanimity" or *Pareto* principle, namely that $X <_R Y$ whenever $X <_r Y$ for every $r \in R$. The method of credulous inheritance and nonmonotonic logic corresponds to the "maximality" or *Pareto optimality* principle. Informally, this principle stipulates that no unsatisfied default may be satisfied without unsatisfying another. Formally, $<_R$ is Pareto optimal if whenever $X <_R Y$ and $Y <_r X$ for some $r \in R$, there is also some $r' \in R$ with $X <_{r'} Y$. Nonmonotonic logic, default logic, and autoepistemic logic all require conclusion sets to be *grounded* or *stable* with respect to default rules, and [Doyle 1983a, 1985] proves that groundedness implies Pareto optimality, leaving the converse open. (Actually, that proof involved preference orders $<_r$ slightly different from the ones above, orders which preferred validating states to invalidating states. I conjecture that one can prove both the result and its converse with the refined order defined here.) Finally, the method of sequential production systems corresponds to the "pick one" or *dictatorship* principle, in which one default determines the whole order, or formally, $<_R = <_r$ for some $r \in R$. See [Borgida and Imielinski 1984] for further examples and discussion of default

reasoning from the point of view of group decisions.

The combination problems faced in default reasoning and group and multi-attribute decision-making may be formally identical, in that all attempt to combine individually rational but mutually conflicting elements into a collective rational decision, but they need not be substantially identical (though the correspondence is strong when we consider explicitly social theories of mental organization such as Minsky's [1986] society of mind). The preferences $<_r$ expressed by propositional default rules are much simpler than the preferences $<_m$ held by a rational individual human, and that may eventually prove decisive in their analysis. On the other hand, one can imagine more complex sorts of rules involving quantification, conjunction, disjunction, and degrees of belief, and if these are used, the problems of artificial and natural decisions may be more comparable in complexity. But even if the substance of the decisions addressed by each of these theories is of high complexity, the acceptability of decision principles may vary among the domains. That is, principles reasonable for groups of humans need not be reasonable for mental decisions, and vice versa. For example, in the mental case, it sometimes seems reasonable to restore consistency by deleting some beliefs, and many AI systems casually use this method. But when groups of people make decisions, this method corresponds to killing, exiling, or brainwashing dissenters, methods ordinarily considered at least undesirable. Of course, an agent who has internalized the arguments and opinions of several people need not be squeamish about deleting someone's proxies, but the justification of the choice of whose opinions to follow would seem to be the same in both cases, even if the methods must differ. As another example, most inheritance theories use the hierarchical order among prototypes to generate certain sort of preferences about preferences, namely to have defaults in more specific prototypes override defaults in more general prototypes when the two conflict. Again, this method can be controversial when applied to groups of humans, for it corresponds roughly to inequality of political powers among members of the group, of which dictatorship is the most extreme example.

4.3 Arrow's conditions and Arrow's theorem

This formal isomorphism of interpreting default rules with making group decisions casts doubts about the existence of a universally acceptable theory of default reasoning because of negative results about group decisions. The most

famous of these is Arrow's theorem. Though good methods exist for numerous special cases of decision making, Arrow [1963] has shown that under certain mild conditions on acceptable decision rules, there is no acceptable decision rule that works in all cases. More precisely, Arrow showed that if any individual orders $<_m$ are possible, then no decision method satisfies the Pareto principle, nondictatorship, and a more technical condition called independence of irrelevant alternatives, which roughly says that global comparisons of any two alternatives are constant no matter what other alternatives are considered at the same time. We will not go into the precise definitions and proof. The point is that if rules for adopting assumptions can be sufficiently complex, and if we judge resolutions of conflicts between rules as though they were conflicts among human proponents, any universal theory of defaults will have to find some way around Arrow's theorem.

Lacking a definitive comparision of the acceptability of decision principles in the social and mental realms (toward which [Wellman 1986] is a start), we cannot offer here a definitive demonstration of the impossibility of a universal theory of defaults. But the evidence seems to point to impossibility. The multi-attribute decision problem can occur within a single agent's reasoning, and is isomorphic to the group decision problem. Moreover, when we go beyond very simple structures for agents and consider more structured agents, as in many expert systems which combine the knowledge of several human experts, Arrow's result really begins to have force, for how can one decide among several peers? If a universal theory is impossible, the upshot is that we must make do with many special theories, and determine (empirically or theoretically) the circumstances in which each works well or works poorly.

5 Conclusion

We doubt the existence of universally acceptable theories of defaults, since the question seems closely related to the existence of universally acceptable methods of group decision making. To summarize the argument, default rules represent preferences about what conclusions to hold. These preferences come from a variety of expert and common sources, and conflicts among these arise naturally. Having to resolve conflicts among these preferences prior to their use defeats the point of guessing, namely to save the effort of acquiring and analyzing information by simply adopting the expected results. In this setting,

default theories represent methods of resolving conflicts among preferences, and unless special circumstances can be shown to hold, standard results will imply that no universal default theories exist.

Two avenues of further study seem valuable: asking whether the special circumstance of intra-mental reasoning offers an escape from the general impossibility conditions, and transferring the many techniques studied from group decision making to see which work best in different reasoning tasks. For example, there are numerical rules like voting commonly used in human decisions that are little explored in the context of default reasoning. Voting rules (such as those explored by [Doyle 1983a]) yield degrees of belief or strength of attitudes rather than sets of overt conclusions, and these may prove to be natural ways of relating logical and connectionist inference schemes. But many interesting possibilities remain unexplored. Are there, for example, natural applications in the "society of mind" for the iterative voting schemes of Tiedman and Tullock [1976]? (See also [Mueller 1979].)

Acknowledgments

Though intended to stand on its own, this paper is an exposition of some of the material contained in two much longer works, [Doyle 1983a, 1987]. I thank Jaime Carbonell, Ronald Loui, Joseph Schatz, Richmond Thomason, and Michael Wellman for valuable comments and ideas.

References

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

Arrow, K. J., and Raynaud, H., 1986. *Social Choice and Multicriterion Decision-Making*, Cambridge: MIT 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.

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

Doyle, J., 1983a. Some theories of reasoned assumptions: an essay in rational psychology, Pittsburgh: Carnegie-Mellon University, Department of Computer Science, report 83-125.

Doyle, J., 1983b. Admissible state semantics for representational systems, *IEEE Computer*, V. 16, No. 10, 119-123.

Doyle, J., 1985. Reasoned assumptions and Pareto optimality, *Ninth International Joint Conference on Artificial Intelligence* 87-90.

Doyle, J., 1987. Artificial intelligence and rational self-government, Pittsburgh: Carnegie Mellon University, Computer Science Department.

Etherington, D. W., 1987a. Relating default logic and circumscription, *Proc. Tenth Int. Joint Conf. on Artificial Intelligence*, 489-494.

Etherington, D. W., 1987b. A semantics for default logic, *Proc. Tenth Int. Joint Conf. on Artificial Intelligence*, 495-498.

Fahlman, S. E., 1979. *NETL: A System for Representing and Using Real World Knowledge*, Cambridge: MIT Press.

Horty, J. F., Thomason, R. H., and Touretzky, D. S., 1987. A skeptical theory of inheritance in nonmonotonic semantic networks, *Proc. Seventh Nat. Conf. on Artificial Intelligence*, 358-363.

James, W., 1897. *The Will to Believe and other essays in popular philosophy*, New York: Longmans, Green and Co.

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

Konolige, K., 1987. On the relation between default theories and autoepistemic logic, *Proc. Tenth Int. Joint Conf. on Artificial Intelligence*, 394-401.

Kyburg, H. E., Jr., 1970. *Probability and Inductive Logic*, New York: Macmillan.

Langlotz, C. P., Shortliffe, E. H., and Fagan, L. M., 1986. Using decision theory to justify heuristics, *Proc. Fifth National Conference on Artificial Intelligence*, 215-219.

Levi, I., 1986. *Hard Choices: Decision making under unresolved conflict*, Cambridge: Cambridge University Press.

Lifschitz, V., 1986. Pointwise circumscription: preliminary report, *Proc. Fifth Nat. Conf. on Artificial Intelligence*, 406-410.

Lifschitz, V., 1987. Circumscriptive theories: a logic-based framework for knowledge representation, *Proc. Sixth Nat. Conf. on Artificial Intelligence*, 364-368.

McCarthy, J., 1980. Circumscription—a form of non-monotonic reasoning, *Artificial Intelligence* **13**, 27-39. Reprinted in *Readings in Artificial Intelligence* (B. L. Webber and N. J. Nilsson, eds.), Los Altos: Morgan-Kaufmann (1981), 466-472.

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

McDermott, D., and Doyle, J., 1980. Non-monotonic logic—I, *Artificial Intelligence* **13**, 41-72.

Minsky, M., 1975. A framework for representing knowledge, *The Psychology of Computer Vision* (P. Winston, ed.), New York: McGraw-Hill. Appendix in MIT AI Laboratory Memo 306.

Minsky, M., 1986. *The Society of Mind*, New York: Simon and Schuster.

Moore, R. C., 1983. Semantical considerations on nonmonotonic logic, *Eighth International Joint Conference on Artificial Intelligence*, 272-279.

Mueller, D. C., 1979. *Public Choice*, Cambridge: Cambridge University Press.

Pascal, B., 1662. *Pensées sur la religion et sur quelques autres sujets* (tr. M Turnell), London: Harvill, 1962.

Pearl, J., 1987. Probabilistic semantics for inheritance hierarchies with exceptions, Los Angeles: UCLA Cognitive Systems Laboratory, TR-93.

Reiter, R., 1979. On closed world data bases, *Logic and Data Bases* (H. Gallaire and J. Minker, eds.), New York: Plenum Press, 1979.

Reiter, R., 1980. A logic for default reasoning, *Artificial Intelligence* **13**, 81-132.

Reiter, R., 1988. On integrity constraints, *Proc. Second Conf. on Theoretical Aspects of Reasoning about Knowledge* (M. Vardi, ed.), Los Altos: Morgan Kaufmann.

Russell, S. J., and Grosof, B. N., 1987. A declarative approach to bias in concept learning, *Proc. Sixth Nat. Conf. on Artificial Intelligence*, 505-510.

Shoham, Y., 1987. Nonmonotonic logics: meaning and utility, *Proc. Tenth Int. Joint Conf. on Artificial Intelligence*, 388-393.

Smith, D. E., 1985. Controlling inference, Stanford: Department of Computer Science, Stanford University, Ph.D. thesis.

Thomason, R. H., 1987. The context-sensitivity of belief and desire, *Reasoning about Actions and Plans* (M. P. Georgeff and A. L. Lansky, eds.), Los Altos: Morgan Kaufmann, 341-360.

Tideman, J. N., and Tullock, G., 1976. A new and superior process for making social choices, *J. Political Economy* **84**, 1145-1159.

Touretzky, D. S., 1986. *The Mathematics of Inheritance Systems*, London: Pitman.

Touretzky, D., Horty, J., and Thomason, R., 1987. A clash of intuitions: the current state of nonmonotonic multiple inheritance systems, *Ninth International Joint Conference on Artificial Intelligence*, 476-482.

Tukey, J. W., 1960. Conclusions vs decisions, *Technometrics 2*, 423-433.

Wellman, M. P., 1986. Consistent preferences in a mind society, MIT 6.868 term project report, unpublished.