Reset reproduction of article published in *Artificial Intelligence*, Vol. 49, Nos. 1–3 (May 1991), pp. 97–128. Reprinted July 1994.

Impediments to Universal Preference-Based Default Theories*

Jon Doyle
Laboratory for Computer Science, Massachusetts Institute of Technology
545 Technology Square, Cambridge, Massachusetts 02139

Michael P. Wellman
Wright Laboratory AI Office, WL/AAA-1, Wright-Patterson AFB, Ohio 45433

Abstract

Research on nonmonotonic and default reasoning has identified several important criteria for preferring alternative default inferences. The theories of reasoning based on each of these criteria may uniformly be viewed as theories of rational inference, in which the reasoner selects maximally preferred states of belief. Though researchers have noted some cases of apparent conflict between the preferences supported by different theories, it has been hoped that these special theories of reasoning may be combined into a universal logic of nonmonotonic reasoning. We show that the different categories of preferences conflict more than has been realized, and adapt formal results from social choice theory to prove that every universal theory of default reasoning will violate at least one reasonable principle of rational reasoning. Our results can be interpreted as demonstrating that, within the preferential framework, we cannot expect much improvement on the rigid lexicographic priority mechanisms that have been proposed for conflict resolution.

1 Introduction

The proliferation of formalisms for nonmonotonic inference [16] attests to a diverse set of methods for reasoning by default. These include circumscriptive inference [29, 31, 35], which draws those conclusions valid in all minimal models of a set of axioms; autoepistemic inference [39, 42] and default logic [49], which permit rules of inference to refer to unprovable statements as well as provable ones; specificity-based taxonomic inference [60], which makes assumptions based on the most specific of the relevant prototypes; and chronologically ignorant inference [56], which draws conclusions based on the shortest or simplest possible histories of events. In addition to these generic patterns, there are often domain-dependent reasons for adopting default policies in particular problem situations. Unfortunately, none of these theories of default reasoning constitutes a comprehensive universal theory that indicates which assumptions are appropriate in every circumstance, all things considered.[1] All the known generic inferential patterns cover only some of the considerations relevant to drawing the best overall conclusions, and individual default rules concern only specific

*This paper is a revision and expansion of [12]. Authors listed alphabetically.

[1]Indeed, none even constitutes a universal representational formalism which can express the mechanisms or principles of all the known (and the undiscovered) theories of default reasoning. Whether there is a universal representational formalism is an interesting question, but not one that will be addressed in this paper.

propositions. This proliferation of formalisms is unsatisfying in the absence of an explanation for why it exists. Our purpose in this paper is to investigate the natural question of whether there is some deeper or more comprehensive theory which combines or unifies all patterns (those known and those awaiting discovery) of nonmonotonic inference.

Toward this end, some theories have been proposed as unifications or partial unifications of some of these ways of making assumptions [13, 25, 31, 32, 56]. At the same time, doubts about the existence of complete unifications have also been expressed because the different theories of nonmonotonic inference may indicate conflicting conclusions when the underlying default rules conflict. Early indications of difficulty appeared with Hanks and McDermott's [19] so-called "Yale shooting problem". Subsequently, Touretzky et al. [61] argued that the gross differences between competing theories of inheritance stem from disparate underlying intuitions about how to make assumptions. As they put it, the differing theories reflect a "clash of intuitions." Recently, Poole [47] displayed another fundamental clash among intuitive properties of default inference.

Unification in the face of these conflicts is possible only if we can divide responsibility among the different methods so that each theory of particular cases provides the right criterion of correctness for inferential problems appearing in its domain. Making this division requires identification of the various conflicts and the sets of cases in which they arise. But it is not feasible to detect all potential conflicts in advance, and there are simply too many to resolve manually. Instead we must seek some way of detecting and resolving them automatically as they arise in reasoning. Of course, we do not wish conflicts to be resolved in arbitrary ways, so our question becomes whether there is a universal theory that automatically combines all particular correctness criteria in a rational manner.

The answer is no: *any universal theory of default inference based on combining correctness criteria must sometimes produce irrational conclusions* (with respect to a very weak standard of rationality) unless one criterion is a comprehensive theory by itself. Put differently, the only way to guarantee rational conclusions is to rationally resolve the problematic conflicts in advance and then use this resolution as the universal theory. Since prior manual resolution seems infeasible (even if we make the rather dubious assumption that people can resolve every specific conflict correctly), some degree of irrationality seems inevitable, whether because of an imperfect automatic conflict resolution method or a manually constructed theory that does not resolve all conflicts. To support this conclusion, we use Shoham's formalism [4, 56] to translate questions about nonmonotonic inference into the context of rational decision making. This translation allows us to adapt Arrow's [1, 2] celebrated results about the impossibility of universal social choice rules to the case of nonmonotonic inference. We also draw on the literature of social choice to consider some possible ways around these results.

2 Preferential theories of default reasoning

The initial theories of default, circumscriptive, autoepistemic, chronologically ignorant, and specificity-based taxonomic inference had very different appearances. Despite their diversity, Shoham [56] has shown how to cast a number of these theories in similar form, as sound inference with respect to models maximal in some partial order. In the more general form of his construction [4], a nonmonotonic logic is characterized by a partial preorder (that is, a reflexive and transitive relation) $\sqsubseteq$ over a set $\mathcal{M}$ of *interpretations*, which, depending on the base logical language $\mathcal{L}$, represent truth assignments, models, Kripke structures, or similar objects. We write $\sqsubset$ to mean the strict part of $\sqsubseteq$, so that $M \sqsubset M'$ iff $M \sqsubseteq M'$ but $M' \not\sqsubseteq M$, and write $\sim$ to mean the reflexive part of $\sqsubseteq$, so that $M \sim M'$ iff $M \sqsubseteq M'$ and $M' \sqsubseteq M$. The meaning of a nonmonotonic theory in these logics is then obtained by modifying the usual notions of satisfaction and entailment to take the model ordering into account. A model M $\sqsubseteq$-*satisfies* a formula P, written $M \models_{\sqsubseteq} P$, iff $M \models P$ and there is no model M' such that $M' \models P$ and $M \sqsubset M'$. A formula P $\sqsubseteq$-*entails* a formula Q, written $P \models_{\sqsubseteq} Q$, iff $M \models Q$ whenever $M \models_{\sqsubseteq} P$. Substitution of these variants for the usual satisfaction and entailment concepts yields a complete description of the nonmonotonic logic $\mathcal{L}_{\sqsubseteq}$.

Shoham illustrates the construction by providing orders corresponding to circumscription [29], the minimal knowledge logic of Halpern and Moses [18], his own chronological ignorance [56], and a few others. In circumscription, for example, models are ranked by minimality according to subset relations among extensions of specific predicates designated as abnormalities. That is, $M_1 \sqsubseteq M_2$ if the extension of the circumscribed predicate P in M_1 contains its extension in M_2 and the two interpretations agree on all other functions and predicates. To capture chronological ignorance, models are ordered according to amount known about histories. These and other theories thus have the same formal structure, differing from each other only in how they order different models.

More generally, the theory may be formulated so that maximization is based on arguments or other epistemic notions as well as truth or belief. For example, Touretzky's [60] theory of inheritance with exceptions compares alternative resolutions by means of an "inferential distance" ordering based on paths or arguments for conclusions, in addition to the conclusions themselves. Such criteria may be captured in a simple variant of Shoham's framework in which the notion of satisfaction in a model is replaced by satisfaction in a mental state, where mental states may include information (e.g., paths or arguments) in addition to the beliefs of the reasoner. In this framework, a nonmonotonic logic is characterized by a preorder $\sqsubseteq$ over a set Σ of possible mental states and a "satisfaction" relation $\models$ between states and sentences such that P is a belief in S iff $S \models P$. We then modify the earlier definition of $\sqsubseteq$-satisfaction to say that S $\sqsubseteq$-satisfies P, written $S \models_{\sqsubseteq} P$ iff $S \models P$ and there is no state $S' \in \Sigma$ such that $S' \models P$, and $S \sqsubset S'$. We redefine $\sqsubseteq$-entailment accordingly, with $P \models_{\sqsubseteq} Q$ meaning that $S \models Q$ whenever S $\sqsubseteq$-satisfies P. We observe without proof that all of our results apply equally well to orders over entire mental states as long as all epistemic states are consistent, that is, as long as either $S \not\models P$ or $S \not\models \neg P$ for every state S and proposition P.[2]

One natural interpretation of inference in the preferential framework is as *rational* selection of maximally preferred states of belief, or of those conclusions that hold in all maximally preferred states. Shoham's terminology is in accordance with this interpretation, as he calls $\sqsubseteq$ a *preference* order, and the corresponding logical notions *preferential* satisfaction and

[2] For an illustration of this point, see the treatment of rational belief revision presented in [10]. Belief revision and default reasoning are closely related, as belief revision concerns how beliefs change nonmonotonically with increasing time, while default reasoning concerns how conclusions change nonmonotonically with increasing knowledge.

	God exists	doesn't
Believe	$+\infty$	$-\epsilon$
Doubt	$-\infty$	$+\epsilon$

Table 1: Pascal's utility assessments of the possible consequences of his decision about belief in God.

entailment.[3] In fact, this view of nonmonotonic inference is more than just an interpretation: it provides a justification for the formal structures of the various nonmonotonic logics. The original theories provided precise formal concepts, but motivated explanations of why these concepts were interesting appeared only later, when Doyle [6, 9], Shoham [55], and others [26, 28] justified default rules by an appeal to decision-theoretic rationality, saying that an agent should adopt a default conclusion or default rule if the expected inferential utility of holding it exceeds that of not holding it. Default rules and other assumption-making mechanisms are ordinarily not presented in terms of rational choice, and their mechanizations usually involve no decision-theoretic calculations. But they are used when the information needed in deliberation about actions and their consequences may be guessed with reasonable accuracy and when mistaken guesses do not lead to serious consequences. In such cases, guessing avoids the costs of acquiring and analyzing the needed information, and so represents a rational response to computational problems involving incomplete information.[4]

In fact, the notion of rationally adopted beliefs is quite an old idea, traceable at least back to the seventeenth century in the form of "Pascal's wager." Pascal [45] framed his problem of belief in God in the following way: he can either believe or doubt the existence of God, and God may or may 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 summarize Pascal's evaluations in the decision matrix shown in Table 1, where ϵ represents the finite amount of pleasure enjoyed or foregone due to belief during his life. Of course, these same quantities modify the first column as well, but finite modifications to infinite quantities are negligible. Since Pascal did not judge God's existence impossible, the expected utility of belief is $+\infty$, dominating the expected utility of doubt, $-\infty$. This convinced Pascal that doubt was not a viable alternative for him. Rational assumptions also play a large role in William James' [22] theory of the "will to believe". James argued that cases of rational belief are ubiquitous in mundane reasoning, an assessment corroborated by the pervasiveness of default reasoning in artificial intelligence.

[3] In the earlier version of this paper [12], we criticized Shoham's definition of $\sqsubset$ in [56] as opposite in sense to the usual notion of preference. That criticism was wrong, based on a misreading of Shoham's definition. Sorry.

[4] See [11] for more discussion of the roles of decision-theoretic and economic notions of rationality in artificial intelligence.

3 Resolving conflicting preferences about defaults

Each of the existing formalisms for nonmonotonic reasoning is either the direct expression of a single criterion for preference among competing interpretations, such as taxonomic specificity, or a means to specify a class of preference criteria, such as default rules (see Section 4.3). Since each can be viewed as a special theory of rational inference, many have hoped or expected that with careful analysis one could combine the choices made by the different theories of nonmonotonic reasoning into a single rational choice, yielding in effect a universal theory of default reasoning. Unfortunately, the potential for conflict among these criteria impedes integration attempts.

3.1 Examples of conflicts

The famous "Yale shooting problem" of Hanks and McDermott [19] illustrates that basic nonmonotonic logics are too weak to arbitrate conflicts among abnormality minimization of different properties. Initially, in their example, a gun is loaded and Fred is alive. After an interval of waiting, the gun is fired at Fred. Fred's survival is a problem for default reasoning because loaded guns normally stay loaded during waits, and living people normally remain alive after actions. Which violation is more abnormal? In this view, the normality of loadedness after waiting and life after shooting are two conflicting criteria. Defenders of nonmonotonic logics have responded by proposing a third criterion—such as chronological minimization [56] or some causality theory [30]—to resolve the issue. However, as Hanks and McDermott point out, in some contexts other criteria (perhaps even chronological *max*imization for diagnostic reasoning) may be compelling, leading to further unresolvable conflicts. It seems a good bet that enterprising researchers will always be able to generate problems that fall through the cracks of fixed configurations of criteria.

In fact, numerous examples suggest that conflicts are unavoidable. The most widely-known conflicts occur in inference from the most specific prototypes, where multiple dimensions of specificity within a taxonomic lattice can result in conflicting preferences between conclusions. An example is the famous "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? Since neither default is more specific than the other, one cannot tell simply from the information given in the taxonomic lattice. Moreover, though one might resolve the question of Nixon's pacifism empirically, such a resolution need not generalize to correctly resolve all formally similar but substantively dissimilar conflicts among other taxonomic defaults.

Conflicts are also possible between pairs of more global preference criteria. For example, ordering assumptions according to their statistical predictivity can conflict with specificity orders. A case in point is Loui's [34] "Mets victory problem," which asks for the probability that the Mets will win today. Statistics are available for three conditions: the game is at home, Dwight Gooden pitches, and Keith Hernandez plays. All three hold for today's game. The difficulty is that the most specific reference class of events, that in which all three conditions hold, may contain so few games that the resulting prediction is much less reliable than a prediction made from one of the more general reference classes in which only one or two of these conditions holds. This problem is a very practical concern for actuaries, who must estimate probabilities for various classifications of events. In their terminology, the *credibility* of a sample of events conflicts with its specificity. As Longley-Cook puts it, when they try to slice the cake too many ways at once they are "left with a useless collection of crumbs" [33].

Similarly, conflicts may occur within reasoners that have multiple informants or refer to multiple authorities to obtain their information. For instance, Milton Friedman presents arguments for free trade, while Lester Thurow presents arguments for controlled trade. These arguments seem individually coherent, but are mutually contradictory. Which conclusion should one believe? Most practical artificial intelligence systems are designed to incorporate all the available knowledge about the relevant subjects by combining expertise from multiple sources. In the simplest approach, one might encode 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 proffering them. In this case, as Thomason [59] 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, many 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 contra-indicating the others.

Yet another class of conflicts arises when criteria of social, ethical, or moral acceptability of conclusions rule out the conclusions indicated by statistical criteria. To use Levesque's [28] example, it may be statistically compelling but socially unacceptable to conclude that a given computer scientist is male. Or for a more consequential conflict, consider "redlining", the practice of not lending money to anyone in neighborhoods deemed to be bad credit risks. Redlining may be justified on statistical grounds, but is often prohibited because it may impede economic recovery of the neighborhood and discriminate against ethnic or racial groups.

These examples suggest that conflicts between preferential criteria among beliefs are unavoidable. While one might view this situation as a reflection of the limits of current epistemology, perhaps a better view is that these conflicts reflect the more general problem of irreconcilable conflicts among values examined by Van Frassen [63].

3.2 Skeptical and credulous conflict resolution

Any comprehensive mechanism for nonmonotonic reasoning must embody some way of handling the conflicts that arise among the different patterns of inference. Some theories provide explicit criteria for resolving unproblematic conflicts. Inheritance theories, for example, resolve conflicts between more specific and less specific information in favor of the former. But as noted above, this rule does not help when neither conflicting preference is more specific than the other. There are two major approaches taken to resolve such problematic conflicts: to choose to satisfy one preference instead of another, and to refuse to satisfy

	Result
Eat nearby hay	Live
Eat distant apples	Live
Refuse to decide	Starve

Table 2: Consequences of actions for the decision faced by the hungry donkey.

any of the conflicting preferences. Each of the different theories proposed for nonmonotonic reasoning takes one of these approaches. For example, nonmonotonic logic, autoepistemic logic, default logic, and "credulous" inheritance [61] describe how a single set of axioms and rules may yield several different, often incompatible sets of conclusions closed under inference. In these theories, problematic conflicts between specific defaults are resolved in every possible way, with each different set of conclusions representing a maximal consistent set of preferences. In contrast, Shoham's logics, circumscription, closed-world reasoning [48], Pollock's defeasible inference [46], and so-called "skeptical" inheritance [21, 57] resemble ordinary logic in that they describe how a set of axioms or rules yields a single set of conclusions closed under inference. These theories handle conflicts either by failing to draw any conclusions involving dissonant problematic preferences or by drawing every conclusion from them (explicit inconsistency).

Both approaches have their defenders. Pollock [46], for example, advocates skepticism in the face of problematic conflicts on the grounds that belief should be based on epistemically defensible positions. But neither skepticism nor credulity is always rational. The agent cannot always rationally choose to remain skeptical about questions very important to its prosperity, whether the skepticism stems from too much information (conflicting preferences) or from too little information (incomplete beliefs). In either case, it may be better to adopt a stance on some issue and risk error than to take no stance at all and risk paralysis. Nor can the agent always rationally choose to be credulous, particularly in situations involving serious consequences of error.

For example, the following elaboration of the classical example of Buridan's ass presents a case where skepticism fails as a rational inference policy. A hungry donkey has to choose whether to eat a nearby bale of hay or a more distant bucket of apples. The donkey prefers nearby food to distant fodder, but also prefers apples to hay. If the skeptical approach is followed, the donkey should refrain from choosing to eat either the hay or the apples. But this is irrational, since eating keeps the donkey alive while not eating makes the donkey starve (see Table 2). Correspondingly, credulity is not necessarily rational for a parent who finds two children fighting, each of whom claims the other started the fight. Because the need for skepticism and credulity may vary with circumstances, we seek a language for expressing when to be skeptical and when to be credulous.

4 Social choice and nonmonotonic logics

Any acceptable universal theory of default reasoning must provide a rationale for its treatment of conflicts, whether credulous, skeptical, or sometimes one or the other. It should also be potentially mechanizable. As noted earlier, placing responsibility for resolving potential conflicts on the human designer is infeasible because for large sets of criteria it is difficult to anticipate all of the potential conflicts and all of the varying circumstances that may influence how the conflicts should be resolved. Furthermore, introduction of new criteria may necessitate completely restructuring the global preference order. A more satisfactory solution would exploit the concept of *modularity* to base conflict resolution mechanisms on general rules of combination that could be applied either manually or automatically as the need arises, so that the same solution still suffices when new criteria are discovered. As is widely recognized, modular design is critical to the successful construction of complex structures, and large commonsense knowledge bases certainly count as such.

4.1 Aggregation policies

To investigate this approach formally, we say that an *aggregation policy* is a function that specifies the global preorder corresponding to any given set of individual preorders. Let the set I index the set of preference orders that are to be combined, so that if $i \in I$, $\sqsubseteq_i$ denotes the preference order corresponding to the ith pattern of inference to be included in the unified logic.[5] The multicriteria nonmonotonic logic problem is then to aggregate the set of preorders $\{\sqsubseteq_i \mid i \in I\}$ into a global preference preorder $\sqsubseteq$.

In this framework, the preferences based on a single criterion, such as predicate minimization, specificity, or chronological ignorance, might be represented by an individual order $\sqsubseteq_i$. Alternatively, individual orders might represent more narrow criteria corresponding to the separate predicates to minimize, the respective dimensions of specificity, or individual default rules (as in Section 4.3). In any case, each $\sqsubseteq_i$ reflects a distinct attribute, encoding the local preferences over interpretations according to its dictates. Modularity or generality of the aggregation method may be ensured by including a large number of vacuous preference orders (trivial preorders such that $M \sqsubseteq_i M'$ if and only if $M = M'$) to be replaced by more substantive orders as new criteria are discovered.

For example, one simple aggregation function is unanimous decision: $M_1 \sqsubseteq M_2$ iff $M_1 \sqsubseteq_i M_2$ for all $\sqsubseteq_i$ that rank the two. This policy, of course, is extremely skeptical as it fails to resolve any conflicts whatsoever.

Another aggregation function comes from applying a voting scheme, for example, majority rule among the criteria: $M_1 \sqsubseteq M_2$ iff

$$|\{i \in I \mid M_1 \sqsubseteq_i M_2\}| \geq |\{i \in I \mid M_2 \sqsubseteq_i M_1\}|.$$

Technically, however, simple majority rule is not a legal aggregation policy because the resulting global order $\sqsubseteq$ is not guaranteed to be transitive when there are more than two models to be ranked. (The intransitivity of majority rule is also known as "Condorcet's voting paradox", after the eighteenth century social scientist who discovered it [50].)

[5] Use of an ordered index set (e.g., $I = \{0, \ldots, n\}$) does not generally reflect any prioritization of these criteria. See Section 4.5 for a discussion of mechanisms where the ordering is significant.

Other aggregation functions organize the criteria in a hierarchy and delegate authority to each criterion according to its place in the hierarchy. We discuss this class of priority-based mechanisms extensively in Section 4.5.

An alternate formalization would be to take the aggregation policy to be a function from individual orders to a set of globally maximal elements, rather than to a global preference order. This would allow for voting schemes that selected a winning candidate without necessarily producing a ranking among the also-rans. Adopting this framework, although slightly more flexible in some respects, would not significantly affect the results of our analysis. We return to this point in Section 5.3.

This formalization covers the result of preference aggregation but abstracts from the process by which aggregation occurs. In particular, it does not seek to characterize processes (such as some forms of negotiation, persuasion, or intimidation) in which the preference orders themselves may change during aggregation. The preference aggregation framework merely describes the functional relationship between individual preferences at the start and group decisions at the end of some unspecified aggregation process.

The group decision-making analogy can be taken quite literally. The problem of designing aggregation policies has been studied extensively in economics, under the heading *social choice theory*. In the language of social choice theory, the ranked interpretations $M_1, M_2, \ldots$ are *candidates*, the $\sqsubseteq_i$ are *individual orders*, and the global order is the *social ranking*. The aggregation policy itself is called a *social choice function*. The main result of social choice theory is a startling theorem due to Arrow [1] that establishes the impossibility of social choice functions possessing several specific desirable and apparently reasonable properties. In Sections 4.3 and 4.4, we show that slightly modified versions of this result apply to preferential nonmonotonic logics, with important implications for the potential construction of universal default formalisms. We first discuss the hypotheses underlying these results.

4.2 Aggregation principles

The principled design of an aggregation policy for multicriteria preferences begins with a consideration of properties we think a reasonable policy should exhibit. The properties we propose are analogs of Arrow's desiderata for social choice. We first present the proposed properties, and then discuss their desirability.[6]

1. *Collective rationality.* The global preorder $\sqsubseteq$ is a function of the individual orders $\sqsubseteq_i$, which are unrestricted, possibly partial, preorders. That is, if Π denotes the set of all preorders over $\mathcal{M}$, an aggregation policy for criteria indexed by I is a function from Π^I to Π.

2. *Pareto principle (unanimity).* If $M_1 \sqsubset_i M_2$ for some $i \in I$ and for no $j \in I$ does $M_2 \sqsubset_j M_1$, then $M_1 \sqsubset M_2$. In other words, the global order agrees with uncontested strict preferences.

3. *Independence of irrelevant alternatives (IIA).* The relation of M_1 and M_2 according to the global order depends only on how the individual orders rank those two candidates.

[6] Consult sources on social choice theory [2, 50] for somewhat more rigorous versions of these desiderata, though for the case of total preorders.

That is, the global order restricted to a subset of candidates is equivalent to an aggregation of the individual orders restricted to that subset.

4. *Nondictatorship (noncomprehensive criteria).* There is no $i \in I$ such that for every M_1 and M_2, $M_1 \sqsubseteq M_2$ whenever $M_1 \sqsubseteq_i M_2$, regardless of the $\sqsubseteq_j$ for $j \neq i$. That is, there is no "dictator" whose preferences automatically determine the group's, independent of the other individual orderings. This principle reflects our presumption that each criterion provides only one consideration of limited scope, that no criterion is itself the universal theory.

5. *Conflict resolution.* If $M_1 \sqsubseteq_i M_2$ for some i, then $M_1 \sqsubseteq M_2$ or $M_2 \sqsubseteq M_1$. That is, if two candidates are comparable in an individual order, then they are comparable in the global order.[7]

Technically, these desiderata are a bit more general than Arrow's, as his framework requires the preferences to be total rather than partial preorders. That is, while social choice theory uses total orders in which, for each x and y, either $x \sqsubset y$, $y \sqsubset x$, or $x \sim y$, preferential nonmonotonic logic allows the additional possibility that x and y are unrelated. Our divergence from Arrow's problem is most apparent in the conflict resolution principle, which for Arrow is implicit in the requirement that the global order be total.

Collective rationality is just a statement of the aggregation framework in preferential nonmonotonic logics. It stipulates that aggregation policies define general methods for combining multiple preference criteria that yield answers no matter what preferential criteria are employed. In particular, it ensures modularity of the aggregation method by requiring that aggregation succeeds even when vacuous criteria are replaced by nontrivial new criteria.

The Pareto principle is clearly a desirable property of aggregation functions; reversing an uncontested preference would be difficult to justify.

IIA has been perhaps the most controversial condition among social choice theorists. In the logical context, however, it corresponds closely to the expected property of model preference that if M is maximal among a set of models, it is maximal in any subset including M. Adding an axiom that rules out only nonmaximal models of P should have no effect on the preferential entailments of P. For example, suppose $M \sqsubset_1 M' \sqsubset_1 M''$ and $M \sqsubset_2 M'' \sqsubset_2 M'$. If these are the only two criteria and the aggregate order makes M' the maximum element of $\{M', M''\}$, then IIA and the Pareto principle require that the aggregate order also makes M' the maximum element of $\{M, M', M''\}$.

Moreover, the independence condition is necessary in a precise sense for the existence of a satisfactory semantics of individual preference criteria. An aggregation function violating IIA cannot be "strategy proof" [15]; that is, it will be susceptible to strategic voting, in which an individual might best realize its own preferences by misrepresenting them to the aggregation procedure. For example, faced with a bully who "aggregates" his preferences with those of his victims by doing the opposite of what the victim wants, Br'er Rabbit professes a great aversion to being thrown into the briar patch even though that is what he actually desires. In such cases of strategic voting, the preferential interpretation of

[7] This weakens slightly the conflict resolution condition stated in our earlier paper [12], which required the global order to provide strict resolutions whenever one of the criteria expressed a strict preference. Such strictness is not necessary.

individual criteria does not reflect their true impact on the global order. Indeed, because of consequences such as these, no one has proposed default theories violating IIA. On the other hand, computational mechanisms implementing nonmonotonic reasoning commonly violate this property by employing rules in which preferences depend on the set of explicitly represented alternatives rather than on the (perhaps hard to compute) set of implicitly represented alternatives. One may view some processes of human negotiation (especially advertising) similarly, as cases in which the negotiation does not merely seek to determine the relevant existing preferences, but instead seeks to change preferences through repetition and association of different alternatives.

The condition ruling out dictators has two independent justifications, corresponding to its descriptive and normative readings. In the first of these, the condition simply states the problem faced by theorists of nonmonotonic reasoning at this time: namely, that all known (and foreseeable) preference criteria to be aggregated provide at best only single considerations to be weighed in making assumptions, and that each of them is prone to override in the face of enough opposition by other criteria. In this descriptive reading, the condition merely rules out the trivial solution to the aggregation problem; it says we cannot simply assume we possess some universal criterion that we in actuality lack. The second justification for the condition is that the existence of a sovereign authority undermines the decentralized representation of preferences motivating aggregation by obviating the need for combination of criteria. It is easy to see why decentralization is a normative ideal in the social choice context, but in the case of reasoning, the normative justification is less obvious. One can justify decentralization in terms of good programming practice, reflecting the limits of human theorists and designers to fully analyze complex structures. Alternatively, decentralization might be justified as reflecting the limitations of scope exhibited by humans and other sources of available information. But there is no objection in principle to monolithic solutions when the problem can be understood fully. In such cases, decentralization must be judged pragmatically. We defer discussion of the practical consequences of dictatorship to Section 4.5.

The conflict resolution condition rules out complete skepticism about conflicting preferences by mandating that the global order commit to some relationship, even if only indifference, whenever the individual orders express a preference. That is, it permits the global order to be skeptical about conflicting strict preferences between two alternatives only by explicitly considering them to be equally desirable. This does not rule out skepticism about conclusions, however, since in preferential nonmonotonic logics the conclusions drawn from equally preferable interpretations are just those conclusions holding in each of the interpretations. For example, if the order is indifferent between interpretations in which Nixon is pacifist and interpretations in which Nixon is not pacifist, then neither pacifism nor nonpacifism will be conclusions of the logic. (Credulity may be achieved simply by linearly ordering the incompatible interpretations so that, for example, all interpretations in which Nixon is pacifist are preferred to those in which Nixon is not pacifist.)

That skepticism about preferences is no panacea becomes apparent when we consider languages that permit explicit expression of preferences about skepticism about belief. The preferential framework applies directly to modal logics of belief, and in such a language we might express a preference for skepticism about a proposition P (as we exhibit in Section 4.3), that is, a preference to believe neither P nor $\neg P$. This preference for skepticism could conflict with a preference for credulity (believing P or $\neg P$) or for a particular stance on P (e.g., believing P). We cannot decide to be skeptical about whether to be skeptical about P, since each of P and $\neg P$ must either be a belief or not. Since there is no recourse to higher-level skepticism about belief, conflict resolution at this level is a requirement, not merely an axiom.

As noted earlier, the conflict resolution principle is relevant only when the aggregate order may be partial. Our first theorem concerns the special case in which all orders are total, for which it states that the desirable and apparently reasonable properties enumerated above are not simultaneously satisfiable by any aggregation policy for preferences expressed by total preorders. (We return to the case of partial preorders in Section 4.4.)

Theorem 1 (Arrow) *If the set of possible interpretations includes more than two models, no aggregation policy mapping total individual preorders to a total global preorder satisfies the properties 1-4 above.*

Proof: With the restriction to total preorders, this is exactly Arrow's theorem applied to choices among models. For a proof of the original result see Arrow [2], Roberts [50, Chapter 7], or any book on social choice theory.□

There is no problem finding good aggregation policies for choices among only two alternatives. But for the case of default reasoning there are always many possible candidates to choose from (for example, all possible models); hereafter we take it for granted that there are at least three.

4.3 Default rules

Arrow's theorem as expressed above need not rule out good aggregation policies for nonmonotonic reasoning, as the preferences occurring in this context may be of a special form which permits satisfactory aggregation. To investigate whether this is the case, let us consider aggregating a set of default rules in the sense of Reiter [49]. A default rule $P : Q_1, \ldots, Q_n/R$ specifies that R should be concluded if P is believed and $\neg Q_k$ is not believed, for each k, $1 \leq k \leq n$.[8]

In order to express preferences about when to be skeptical and when to commit to belief, we require models which describe belief states as well as the contents of beliefs. For this purpose, we employ Moore's models for autoepistemic logic [41].[9] As Konolige [25] shows, default theories correspond naturally to autoepistemic theories in which each default is rewritten in the form $LP \wedge \neg L\neg Q_1 \wedge \ldots \wedge \neg L\neg Q_n \supset R$, where we read LP as "P is believed" and $\neg LP$ as "P is not believed." Each Moore model M is a pair $M = (K, V)$ of an ordinary valuation V and a Kripke structure K. A Kripke structure contains a set of possible worlds and an "accessibility" relation on these worlds. The truth of a formula is evaluated with respect to each world, and a formula of the form LP is true in a world W just in case P is true in every world accessible from W. In Moore's semantics, each K is required to be a complete structure for the modal logic S5, that is, an equivalence relation in which every possible world is accessible from every possible world. Moore proves

[8] Actually, Reiter wrote defaults as $P : MQ_1, \ldots, MQ_n/R$, but we omit the M markers.

[9] One can also formalize these preferences using "situations" to describe belief states, as in Levesque's logic of explicit belief [27], or use the belief states directly, as discussed in Section 2.

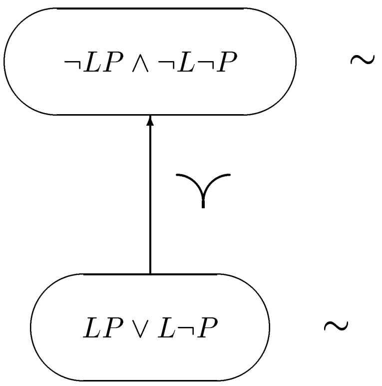

Figure 1: The order expressing preference for skepticism about P.

that such models are in exact correspondence with stable autoepistemic theories, that is, deductively closed sets of sentences which contain LP whenever they contain P and contain $\neg LP$ otherwise.

With such interpretations, we may express default rules as preferences in a natural way. Let us first introduce a bit of helpful notation. If p and q are mutually inconsistent sentences, then they are satisfied by disjoint sets of models, and we write $p \prec q$ (q preferred to p) to mean that $M \sqsubset M'$ iff $M \models p$ and $M' \models q$, and that $M \sim M'$ for all models M, M' of p and all models M, M' of q. In other words, the models of p (respectively q) are all equally preferable (the agent is indifferent among them), but all models of q are preferred to all models of p.

We may then express a preference for skepticism about P by

$$LP \vee L\neg P \quad \prec \quad \neg LP \wedge \neg L \neg P,$$

which says that believing neither P nor its negation is preferred to believing either, and depict this relationship as in Fig. 1. A preference for credulity about P is expressed by the opposite order.

Similarly, a default rule $P : Q_1, \ldots, Q_n / R$ may be expressed by the preferences $\sigma \prec \sigma' \prec \sigma''$ (read transitively), where

$$\begin{aligned}
\sigma &= LP \wedge \neg L \neg Q_1 \wedge \ldots \wedge \neg L \neg Q_n \wedge \neg LR, \\
\sigma' &= LP \wedge (L \neg Q_1 \vee \ldots \vee L \neg Q_n), \text{ and} \\
\sigma'' &= \neg LP \vee (LP \wedge \neg L \neg Q_1 \wedge \ldots \wedge \neg L \neg Q_n \wedge LR).
\end{aligned}$$

That is, if P is believed, the default rule $P : Q_1, \ldots, Q_n/R$ prefers believing R to believing any $\neg Q_k$, and prefers believing either R or some $\neg Q_k$ to believing none of these. As all models satisfy one of σ, σ', or σ'', it never happens that two of them are incomparable. Thus these preferences induce a total preorder (shown in Fig. 2): any two models are related by $\sim$ or $\sqsubset$, and hence by $\sqsubseteq$.

While there may be other motivated ways of interpreting default rules as preference orders over states of belief (see, for example, [13, 20, 56]), the interpretation above seems a natural one, and is corroborated by previous results of Doyle [6, 7] which showed that the

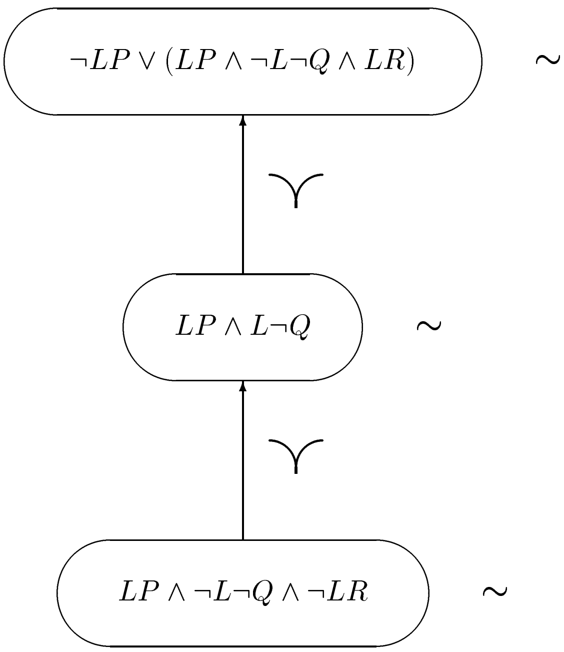

Figure 2: The total preorder expressed by the default rule $P : Q/R$.

extensions of default theories are Pareto-optimal choices, that is, correspond to maximal consistent sets of default-rule preferences.[10]

We now improve on Theorem 1 by showing that restricting the individual orderings to those arising from default rules is not sufficient to forestall the impossibility theorem.

Theorem 2 *No policy for aggregating every set of default rules into a total global preorder satisfies the properties 2-4 above.*

Proof: Since Theorem 1 applies whenever the class of possible individual orders contains every possible ordering of at least three alternatives (even if the usual case involves aggregating only a subset of these), it suffices to show that each pattern of preferences among three alternatives can be expressed by some default rule. Any three alternatives will do, as the set of candidate models can be arbitrarily constrained to these by nondefeasible axioms. In fact, we show that even normal defaults (those of the form $P : R/R$) can express all patterns among the alternatives chosen.

Let $R_1, \ldots, R_6$ be six logically independent propositions, and $: R_k/R_k$ defaults for $1 \leq k \leq 6$. Each default $: R_k/R_k$ expresses the preferences

$$\phi_k \prec \phi'_k \prec \phi''_k,$$

[10] Not all Pareto-optimal choices are extensions, however, which is one of the points of divergence of preferential nonmonotonic logics from default logic. For example, while our preferential interpretations of default rules express preferences satisfied by default logic, the nonmonotonic logic that results from considering *all* maximally preferred models can yield contrapositive conclusions not reached by default logic. (That is, $\neg Q$ will be a maximally preferred conclusion from axioms P, $\neg R$, and the sole default rule $P : Q/R$.)

Default	Preferences
$: R_1/R_1$	$\sigma_1 \prec \sigma_2 \prec \sigma_3$
$: R_2/R_2$	$\sigma_1 \prec \sigma_3 \prec \sigma_2$
$: R_3/R_3$	$\sigma_2 \prec \sigma_3 \prec \sigma_1$
$: R_4/R_4$	$\sigma_2 \prec \sigma_1 \prec \sigma_3$
$: R_5/R_5$	$\sigma_3 \prec \sigma_2 \prec \sigma_1$
$: R_6/R_6$	$\sigma_3 \prec \sigma_1 \prec \sigma_2$

Table 3: Six default rules which result in all possible preference orders over $\sigma_1, \sigma_2, \sigma_3$.

where

$$
\begin{aligned}
\phi_k &= \neg L \neg R_k \wedge \neg L R_k, \\
\phi'_k &= L \neg R_k, \\
\phi''_k &= \neg L \neg R_k \wedge L R_k.
\end{aligned}
$$

Consider now three sets of models constructed by conjoining all permutations of the individual default conditions:

$$
\begin{aligned}
\sigma_1 &= \phi_1 \wedge \phi_2 \wedge \phi''_3 \wedge \phi'_4 \wedge \phi''_5 \wedge \phi'_6, \\
\sigma_2 &= \phi'_1 \wedge \phi''_2 \wedge \phi_3 \wedge \phi_4 \wedge \phi'_5 \wedge \phi''_6, \\
\sigma_3 &= \phi''_1 \wedge \phi'_2 \wedge \phi'_3 \wedge \phi''_4 \wedge \phi_5 \wedge \phi_6.
\end{aligned}
$$

As Table 3 indicates, the six defaults express every strict preference order among elements from these three sets of models. Thus, the individual preferences for these elements are effectively unrestricted by the form of default rules. Because these elements could constitute the entire candidate set (given, for instance, a nondefeasible axiom of the form $\sigma_1 \vee \sigma_2 \vee \sigma_3$), Theorem 1 applies directly. □

4.4 A general impossibility theorem

The special case of default rules by itself does much damage to hopes for a unified theory of nonmonotonic reasoning since a general theory should cover at least these. But one might still escape this limitation by dropping the restriction that the preorders be total. The following theorem shows that the impossibility result recurs if we require global orders to be as complete as the individual orders.

Theorem 3 *No aggregation policy satisfies the properties 1-5 above.*

Proof: (sketch) The only difference between the multicriteria aggregation problem considered here and the classic social choice setup is that the individual and global orders can be partial whereas individual and social rankings are taken to be total. Partiality is constrained, however, by the conflict resolution condition's restriction that the global order be at least as complete as the constituent orders. An examination of Arrow's proof of Theorem 1 [2] reveals that it does not depend on totality of the individual orders, and presumes totality of the global orders only for cases where an individual expresses a preference. The validity of this presumption is enforced by our conflict resolution condition; hence the result generalizes to our situation.□

The conflict resolution principle is an extreme condition in that it rules out all skepticism about conflicting preferences. However, the negative implications of Theorem 3 do not hinge on this extremity. Dictatorial consequences are inevitable to whatever extent conflicts are resolved. The following corollary expresses this formally.

Corollary 4 *Suppose an aggregation policy satisfies properties 1-3. Then for any subset of candidates for which conflict resolution is satisfied, there exists a dictator whose preferences among this subset are adopted globally regardless of the preferences of the other criteria.*

Proof: By the independence of irrelevant alternatives, the global preference for a subset of candidates depends only on individual preferences among that subset. Therefore, the projection of the aggregation function onto this restricted domain must obey identical properties. Since conflict resolution is satisfied by hypothesis, Theorem 3 entails the impossibility of nondictatorship. □

Corollary 4 highlights the central tradeoff between skepticism and appeal to central authority in the design of preference aggregation schemes. The conflict resolution strategies admitted by the dictatorship condition form a class we call *lexicographic priority mechanisms.* Because this is the only kind of conflict resolution consistent with the other conditions, it is worth exploring the nature of these strategies in some detail.

4.5 Lexicographic priority mechanisms

In a "dictatorial" conflict resolution strategy, there exists a criterion which always gets its way regardless of what other criteria oppose it. In schemes where the dictator necessarily expresses a preference over all candidates (as in the usual social choice framework), this leads to very uninteresting choice mechanisms. Where the criteria may be expressed by partial preorders, a dictatorial aggregation policy may be slightly more complicated. In this case, there is a secondary dictator deciding conflicts among the candidates for which this criterion expresses no preference. This cascade of dictators continues until all conflicts are resolved.

As an example, we exhibit the preference aggregation function corresponding to the *stratified logic* scheme of Brown and Shoham [4]. Modifying their notation slightly, the problem is to determine the global order $\sqsubseteq$ from an ordered set of individual orders $\sqsubseteq_i$, $0 \leq i \leq n$. According to their definition, $M \sqsubseteq M'$ iff $M \sqsubseteq_j M'$ for some $j \leq n$ and $M \sim_i M'$ for all $i < j$. In other words, criterion 1 serves as dictator: the global preference order $\sqsubseteq$ is equivalent to $\sqsubseteq_1$ except where the latter is indifferent. In that case, $\sqsubseteq_2$ decides, regardless of the $\sqsubseteq_i$ for $i > 2$, unless it too is indifferent. Each underling criterion comes into play only when all of its superiors express indifference.

Choice rules of this form are called *lexicographic* because they resemble the method for ordering words alphabetically: compare the first letters; if tied compare the second, and so on. The implication of Corollary 4 is that every conflict resolution method satisfying collective rationality, unanimity, and IIA is necessarily equivalent to a lexicographic priority mechanism for some fixed sequence of criteria.

Note, however, that in stratified logics the dictator reigns even over unrelatedness; that is, if M and M' are not related by $\sqsubseteq_1$, they are not related by $\sqsubseteq$. According to the definition of Section 4.2, dictators are authoritative over indifference but not unrelatedness. In fact, overruling an explicit preference with unrelatedness sometimes violates the Pareto principle and always violates the conflict resolution principle. Thus, despite their lexicographic flavor, stratified logics do not satisfy our aggregation principles.

In McCarthy's prioritized circumscription [36], minimizing the abnormality of each predicate is a distinct preference criterion, and the global minimization is based on a lexicographic decision procedure using a predefined order on the predicates. Grosof [17] generalizes this to arbitrary partial preorders on arbitrary model-preference criteria, but the prioritization resolves conflicts only to the extent that the orders are *stratified* in the lexicographic sense illustrated above.

An example of lexicographic choice in everyday societal decision making is the judicial hierarchy of courts. The supreme court is the dictator, and in general the outcome of a dispute is decided by the highest court that addresses the case. Because each case is heard only along an ancestral line of courts, conflicts never arise among jurisdictions that are not strictly ordered by authority. As long as this constraint is satisfied, any order-preserving linearization of the structure will yield the same result under lexicographic choice as the original hierarchy.

Konolige's hierarchic autoepistemic theories [24] work much the same way for nonmonotonic reasoning. A fixed hierarchy is established initially to decide which criterion will be authoritative in case of conflict. Although the ordering of sub-theories is partial, conflicts are resolved only among theories that are linearly ordered in the hierarchy.

The main problem with lexicographic choice methods is that they are too rigid in how they combine criteria of limited expertise. Once assigned a position in the sequence, a criterion cannot be overridden by those below it, even by unanimous opposition. Thus, any criterion that is not absolute must be placed below those that potentially outweigh it, with the consequence that these other criteria *always* hold sway when in conflict with the original. If different criteria should properly be authoritative in varying situations, no lexicographic priority mechanism will be adequate for the conflict resolution task. For example, there exists no lexicographic method implementing majority rule for three criteria choosing among two candidates.

Another way of expressing this inflexibility of lexicographic orderings is to observe that they do not specify very many implicit preferences. We distinguish here the explicitly specified preferences from those implicitly determined by the aggregation function. Lexicographic choice determines relatively few implicit preferences, namely those defined by the transitive closure of the explicitly specified pairwise priorities. In contrast, more flexible aggregation rules like majority voting specify few (if any) explicit preferences beyond the individual preferences, but determine large numbers of implicit preferences. Since explicit preferences must be specified in advance, the designer of a lexicographic default reasoning system must essentially anticipate all potential conflicts in the process of specifying the criteria and their priorities. This means that if a new preference criterion is discovered, it cannot be considered an incremental addition to the set of criteria. Instead, it must be fit into the existing ordering, perhaps by replacement of the existing criteria by new versions that incorporate the new considerations, or by invention of further “higher-level” criteria to resolve conflicts arising within the expanded set. This certainly goes against the spirit of modular preference aggregation, in which the general combination rule accommodates new criteria without the need to explicitly reconsider or revise existing preference criteria.

Though fixed lexicographic orderings are too inflexible to capture all desirable choice rules, one can always find a lexicographic rule to achieve a particular outcome by adding a new criterion that dictates over the existing set. This, in fact, is the way many have proposed to resolve the Yale shooting problem and other conflicts: by expanding the set of criteria to include such factors as causality or chronological minimality. But while we may expect manual “patching” to be a useful way of incrementally improving any aggregation method, it seems imprudent to rely on it as the sole mechanism for resolving conflicts. As argued earlier, resolving all conflicts manually in advance is simply not feasible. Even if the effort is spread out over time, there are simply too many conflicts potentially requiring resolution. Worse still, the new special-purpose criteria may conflict with existing criteria in unanticipated ways; these conflicts will be hidden from view and so never explicitly considered because the new criteria are installed in a dictatorial position. Finally, there are no grounds to assume that we will always be able to find a simple criterion suitable for selection as the next dictator, nor does it seem reasonable to suppose that external (specifically, human) resolvers will be available at the time an important conflict happens to arise to make the necessary decisions, or that they will be able to make these decisions as quickly as needed.[11] Instead, prudence dictates that some mechanism be in place for resolving conflicts automatically between the time they are discovered and the time (if ever) the “right” resolution is found. Any conflict may indicate the need to seek out new principles or preferences. But sole reliance on timely human discovery of effective new criteria seems as unjustified as reliance on some *deus ex machina*.

4.6 Petty dictators

The existence of dictators might not be so distasteful if the identity of the dictator could vary depending on the candidates being ranked. For example, the path through a hierarchy or the structure of the hierarchy itself might depend on the choice involved. Conceptually, each dictator would wield authority only over some designated subdomain. This arrangement, however, does not satisfy the definition of “dictator”, and therefore the theorems above indicate its incompatibility with the other aggregation principles listed in Section 4.2. Nevertheless, Corollary 4 suggests that by satisfying conflict resolution only partially we might limit the scope of dictatorships to specialized choice contexts.

The concept of limited realms of absolute authority has sometimes been called *liberalism* in social choice theory, by analogy to the idea that an individual’s preferences should be the sole factor in choice among matters pertaining peculiarly to that individual (such as whether one sleeps on one’s back or on one’s stomach). Remarkably, the notion of liberalism is inconsistent with even the Pareto principle, as shown by Sen [53]. To see this, suppose that $\sqsubseteq_i$ is authoritative with respect to M_1 over M_2, and that $\sqsubseteq_j$ is authoritative about

[11] McDermott [37] rhetorically asks whether Lifschitz will always be there to bail us out when existing inference mechanisms prove inadequate.

M_3 over M_4. In addition, suppose all criteria (including $\sqsubseteq_i$ and $\sqsubseteq_j$) prefer M_2 over M_3 and M_4 over M_1. There is no global order consistent with unanimous preference as well as the individual authorities, even though the subdomains of dictatorship are disjoint.

5 Paths toward possibility results

The impact of the impossibility result is proportional to the judged importance of conforming to the premise conditions, as well as the degree to which they need be relaxed in order to achieve "possibility". For social choice, Theorem 1 has had great force due to the apparent reasonableness of the conditions and its demonstrated robustness despite countless mathematician-years spent laboriously tweaking axioms. For nonmonotonic logics, the reasonability of the desiderata is more in question, and further scrutiny is needed to determine the robustness of our results.

Paralleling the investigations made in social choice, one can identify three primary options for dealing with impossibility. The first and most direct way out is to restrict or expand the specification of preferences and the basic construction of nonmonotonic logics from them. The second approach attempts to find compromises among the conflicting desiderata, and to analyze the tradeoffs involved in different compromises. A third option is to investigate modifications of the output required of preference aggregation policies, requiring only that aggregation indicate maximal models, not full orders. We discuss these paths toward possibility in turn.

5.1 Modifying the expressiveness of preferences

The impossibility result is fundamentally a statement about the relation between the expressive power of a preferential nonmonotonic logic and the difficulty of combining multiple criteria. To accept the aggregation principles and yet avoid the implications of Theorem 3, the language for representing preferences needs to be more or less expressive than the framework presented above.

For example, the impossibility result can be circumvented by expanding the language of preferences to include some expression of intensity of preference.[12] More specifically, the ordinal expressiveness of the individual preferences $\sqsubseteq_i$ can be strengthened in two ways. The first is to allow *intercriteria comparisons*, permitting statements of the form "criterion i likes M_1 more than criterion j likes M_2." A circumscriptive example would be a comparison of the degrees of abnormality of two predicates in different situations, perhaps by counting their abnormal instances. The second enhancement introduces *intracriteria intensities*, where i's degree of preference for M_1 over M_2 can be compared to its preferences for M_3 over M_4. For example, the degree of chronological minimization might be measured by the temporal distance between events. Taken alone, intercriteria comparison only opens the door a crack, leading to aggregation policies that are almost-but-not-quite dictatorial (in a precise sense described, for example, by Roberts [51]). And incorporating only intracriterial intensity comparisons does not help at all. Together, however, the two measures induce a fully cardinal description of preferences (that is, a numeric measure of degree of preference),

[12] Strictly speaking, this only circumvents the theorem if the intensity information is *mandatory*.

which leads immediately to satisfactory aggregation functions of the sort recommended by multiattribute utility theory [23]. In this case, each criterion may be represented by a real-valued function over interpretations and a global utility function may be constructed by a weighted combination of these. Numeric comparison of utility values then defines the global preference order.

Although it solves the preference aggregation problem, we suspect that designers of nonmonotonic logics will not be eager to require in effect that numeric utility measures be assigned to every interpretation. Numeric representations are typically avoided because they are excessively precise and present an intolerable specification burden on the source of default assertions. To make this approach palatable, one would have to find some qualitative (direct) expression of the available preference information (going beyond purely ordinal comparisons) from which the numerical measures could be automatically constructed. Unfortunately, this global comparative information is just what seems to be lacking in our intuition in many cases, as indicated in Section 3. Nevertheless, it may be possible to learn pragmatically useful numerical measures through experience.

Similarly, limiting the expressive power of preferences by restricting the form of the individual partial preorders that are handled by the aggregation policy can lead to acceptable policies operating over the smaller domain. Theorem 3 declares the impossibility of a completely general aggregation policy but does not rule out satisfactory aggregation in special cases. Social choice theorists have explored this route in depth, but the special cases they consider (such as single-peakedness, a condition that the candidates be orderable according to one global dimension) do not appear to be viable for the multicriteria preference problem. On the other hand, one might discover aggregable special cases particularly well-suited for nonmonotonic reasoning, whether or not they make sense in the original social choice context. Theorem 2 demonstrates that the particular case of normal default rules is not special enough to avoid the difficulties of aggregation; other candidates, however, have yet to be investigated.

5.2 Modifying the aggregation principles

If we insist on maintaining the ordinality of constituent preferences and the universality of the aggregation policy, we must consider which of the desiderata may be abandoned or relaxed.

At the extreme, we could simply give up on global rationality, permitting $\sqsubseteq$ to be intransitive or inconsistent. The effect of intransitivity would be to make more models maximal as compared with those maximal under the transitive closure of the relation, and thus make the nonmonotonic consequences of a theory less complete. Rychlik [52] proposes such an intransitive scheme, and motivates it with an example where chronological ignorance loses force over long durations. However, it is our view that this situation and similar ones are better modeled by recognizing competing preference criteria (e.g., decay of persistence) rather than weakening the underlying concept of preference. Transitivity appears to be a minimal requirement for the interpretation of $\sqsubseteq$ as a "preference relation" in the spirit of rational choice. If it is necessary to abandon transitivity (and ideal rationality of choice along with it), we need not make a virtue of it. Instead, we might abandon transitivity on the simple pragmatic grounds that it is too costly or impossible to enforce in computationally convenient conflict resolution schemes (such as majority rule or random choice).

The effect of inconsistency depends on whether one uses the strict or nonstrict order in defining nonmonotonic entailment. The resulting conclusions will be either inconsistent (models exist, but none are maximal) or skeptical (the relevant models are indifferent), respectively.

Similarly, any of the other aggregation principles may be relaxed.

- We could choose to live with a bit of dependence on irrelevant alternatives. This would mean that the preferential semantics imperfectly describe the effects of individual orders, as criteria might better achieve their objectives by misrepresenting their preferences.
- We could accept skepticism in some cases in which credulity would be better.
- We could accept the rule of imperfect dictators which correctly resolve many conflicts but which wrongly resolve others.

Intelligent compromise on these principles requires a much better understanding of the tradeoffs we face. One task here is to obtain a deeper analysis of the sources of impossibility. If we can characterize a subclass of preference profiles that fully account for the pessimistic conclusions of Theorem 3, we can limit our desiderata violations to that class. This step is simply a less drastic version of the suggestion above that we restrict the expressive power of the language to exclude the problematic cases. For example, while we argued above that skepticism as a response to all conflicts is irrational, it would be less objectionable to suspend commitment when the conflict is further classified as one of the particularly difficult instances.

To justify this approximation approach, however, we need some way of judging the expected utility of different aggregation procedures when the costs and consequences of inferences are taken into account. This means estimating the likelihoods with which different conflicts appear and the risks and benefits that different forms of irrationality entail in each of these cases. For example, information about the reasoning process in which the conflicts arise might be used to determine the cases in which suspending judgment is rational because determining the proper resolution would take too long, or in which errors would occur infrequently enough to ignore. Or alternatively, this information might make it possible to compare the expected utility of imperfect dictators (based on the expected probability and consequences of their errors) with the expected effort of revising the dictatorial priorities in a way that improves their performance. In the context of voting, Tullock [62] argues that systems such as majority decision will produce satisfactory, approximately rational results given large enough numbers of voters. Although the conditions and assumptions underlying his conclusions are not clearly applicable to the default reasoning case, investigations of this sort suggest that regularities in preference structures may mitigate the undesirable consequences of the impossibility results.

5.3 Modifying the output of aggregation

Approaches that modify the expressiveness of preferences can be viewed as variations on the *input* of aggregation policies. In a dual manner, we can attempt to escape the impossibility result by modifying the *output* produced by the aggregation process. Social choice theorists have investigated one variation in which maximal elements are selected directly without constructing the entire global preference order. Formally, the aggregation returns a *choice function* mapping candidate sets to subsets of maximal elements. This particular modification seems appropriate for nonmonotonic logic because the concepts of preferential satisfaction and entailment distinguish only between maximal and nonmaximal elements of the preference order.

However, this apparent weakening of the aggregation framework does not offer much improvement in the prospects for obtaining satisfactory aggregation functions. Sen [54] points out that proofs of the various impossibility results typically apply directly to the case of choice of maximal elements. The key observation is that the choice function induces a relation R (roughly speaking, $M_1\ R\ M_2$ means "M_1 is chosen over M_2 in the context of some candidate set") that fulfills the role of the global preference order $\sqsubseteq$ in derivations of the original theorems.

6 Applications to mental societies

Reasoning has been viewed in social terms in artificial intelligence by several authors. The most prominent example is Minsky [40], who explicitly models thinking as the aggregate activity of many small mental agents. In the context of nonmonotonic reasoning, Borgida and Imilienski [3] appeal to committee decision-making as a metaphor for default inference, and Doyle [5, 7, 8] presents nonmonotonic reasoning from a group decision-theoretic perspective. Related views of thinking can be found in economics, philosophy, and psychology [43, 44, 58, 59].

The central tenet of Minsky's *Society of Mind* [40] is the rejection of the single-self viewpoint in favor of a mind made up of many, largely autonomous, agencies. If this idea is to be taken seriously, then analyses of behavior which take as primitive such single-self concepts as beliefs, preferences, and goals should be regarded with some degree of skepticism. More precisely, the presumption of consistency among these objects must be carefully scrutinized. Social choice theory is well-suited for this kind of study because it aims to characterize global properties of aggregate behavior without requiring mechanistic descriptions of the individual components.

It appears at first glance that the impossibility results discussed above should also explain inconsistencies in belief and preferences for minds—even under the most optimistic hypothesis about the rationality of its component agencies. Representing the individual preferences by $\sqsubseteq_i$ and taking the global, single-self preferences to be the output of the aggregation function, the impossibility theorems apply directly.

Indeed, the results have immediate consequences for the society-of-mind model proposed by Doyle [6]. This is not surprising, as the agents in this model correspond to a form of default rule that can be interpreted preferentially as described in Section 4.3. Theorem 2 implies that either IIA is violated in constructing the admissible extensions or the set of agents contains a dictator.

Unfortunately, few of the other society-of-mind theories are concrete enough to be analyzed in this fashion.[13] For the more general framework, we need to ask the obvious questions about whether mental agencies can have some way of circumventing the conditions. To some extent, the case against consistent preferences in a mental society shares a common basis with the arguments we offer against universal default theories. However, differences in the domain (mental agencies versus preferential criteria) lead to important conceptual distinctions in our interpretation of the results.

The case against dictatorship is, if anything, stronger in the mental society context, as decentralization is an explicit and fundamental attribute of these theories. Even if not absolute, dictatorial individuals represent an uncomfortable concentration of power that is more characteristic of bureaucratic structures than societies of autonomous agents.

The dictum against cardinal representations of intensity of preference, however, is more difficult to defend for mental agents, especially since several architectures for reasoning involve numeric representations for degrees of belief or other states (for example, neural networks). Without a theory of the ultimate source of preferences, it is hard to justify specific constraints on their form. Finally, the universality standard we imposed on preferential default theories may not apply as forcefully to mind societies. Mental societies may manage to function satisfactorily despite an inability to incorporate or obey particular preference criteria.

These trap doors prevent us from offering any sweeping conclusions about the possibility of globally rational agents built from societies of autonomous individuals. For analyzing specific society-of-mind architectures, however, the tools of social choice theory are likely to prove quite useful.

7 Conclusions

We proved in Section 4 that any universal theory of default inference based on combining noncomprehensive preference criteria must sometimes produce irrational conclusions, or alternatively, the only way to guarantee rational conclusions is to manually resolve all conflicts in advance. Our argument may be summarized as follows:

1. It is natural to formalize nonmonotonic logics as theories of preferential or rational inference. From this viewpoint, defaults express preferences about what to believe (or more generally, about what states of mind to inhabit), and the theories of different nonmonotonic inferences embody different criteria about how to identify the most preferred conclusions.

2. Unfortunately, these theories are very incomplete. Individual defaults only concern specific propositions, and all known theoretical inferential criteria cover only some of the considerations relevant to choosing conclusions rationally.

3. These theories are also somewhat incompatible. Individual defaults may express conflicting preferences, and different criteria may indicate conflicting conclusions.

[13] Fagin and Halpern's logic of local reasoning [14] might be one of these few.

4. It is not feasible to resolve these conflicts in advance. Therefore, any universal theory must rely on some method that takes the available set of preference criteria and combines them into a global criterion.

5. If conflicts are to be resolved rationally, the preferences resulting from the resolution must agree with the original criteria when those criteria agree and must result in the same ordering for pairs of alternatives independent of what other alternatives are under consideration.

6. But we prove, with respect to this weak sense of rationality, that there are no rational methods for aggregating criteria represented by preference orders unless the resulting order is simply one of the criteria being aggregated. Therefore, the only way to achieve rationality is to impose a priority ranking on the criteria, and to revise the ranking whenever new conflicts become important.

More fundamentally, resolving even simple conflicts requires empirical information about which resolution is best (or is of maximal expected utility to the reasoner). While designers might try to supply artificial reasoners with some of this information, most of it must be left unsaid as too hard to foresee or too voluminous to state explicitly. One cannot expect a purely theoretical combination method to possess this empirical information, so a purely theoretical solution to conflicting defaults seems unlikely (cf. [38]).

Of course, our results apply only to the case in which none of the available criteria is a comprehensive universal theory by itself, and would be irrelevant if someone were to discover a good comprehensive theory. But at present, all criteria are clearly limited in scope, and our strong expectation is that all theories discovered in the future will be similarly limited.

The impossibility results presented above expose previously unarticulated difficulties in the quest toward universal default mechanisms. We do not believe that these results constitute an indictment of the preferential framework. Instead, translating questions about nonmonotonic reasoning into the language of rational inference and social choice provides a rational justification for the nondeductive structure of some nonmonotonic logics and yields valuable insights into their design. Moreover, the problem is not attributable to the use of logical or mathematical formalisms for describing or mechanizing reasoning, nor is it due to limitations on the computational resources available for carrying out reasoning. Instead, our results delimit the nature of feasible forms of rationality for an agent that integrates preferences from multiple sources, independent of its representational structure, computational power, or extent of knowledge.

To address the problems posed by our results, we must continue to investigate special theories of reasoning and the conditions under which each of these is to be preferred or to be avoided. We expect that further analysis from the social choice perspective will suggest promising approaches, both because it provides the vocabulary for expressing concepts related to aggregation policies, and because it allows artificial intelligence studies to draw on a large literature of detailed investigations of social choice questions.

Acknowledgments

We thank Daniel Bobrow, Ruy Cardoso, Gerald de Jong, David Etherington, Benjamin Grosof, Ronald Loui, Joseph Schatz, Yoav Shoham, Mark Stefik, and Peter Szolovits for valuable discussions and probing questions. The comments of the anonymous referees helped identify some minor errors and suggested clarifications of several issues. Jon Doyle's work on this paper was supported by National Institutes of Health Grant No. R01 LM04493 from the National Library of Medicine.

References

[1] Arrow, K. J. *Social Choice and Individual Values.* (Yale University Press, second edition, 1963).

[2] Arrow, K. J. Values and collective decision-making. in: Laslett, P. and Runciman, W. G. (Eds.), *Philosophy, Politics and Society (third series)*, pages 215–232. (Basil Blackwell Oxford, 1967).

[3] Borgida, A. and Imilienski, T. Decision making in committees—A framework for dealing with inconsistency and non-monotonicity. in: *Non-Monotonic Reasoning Workshop*, pages 21–32, New Paltz, NY, 1984. American Association for Artificial Intelligence.

[4] Brown, Jr., A. L. and Shoham, Y. New results on semantical nonmonotonic reasoning. in: Reinfrank, M., de Kleer, J., Ginsberg, M. L., et al. (Eds.), *Non-Monotonic Reasoning*, pages 19–26. (Springer-Verlag, 1989).

[5] Doyle, J. A society of mind: Multiple perspectives, reasoned assumptions, and virtual copies. in: *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pages 309–314, 1983.

[6] Doyle, J. Some theories of reasoned assumptions: an essay in rational psychology. Technical Report 83-125, Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1983.

[7] Doyle, J. Reasoned assumptions and Pareto optimality. in: *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 87–90, 1985.

[8] Doyle, J. Artificial intelligence and rational self-government. Technical Report CS-88-124, Carnegie-Mellon University Computer Science Department, 1988.

[9] Doyle, J. Constructive belief and rational representation. *Computational Intelligence* **5** (1989) 1–11.

[10] Doyle, J. Rational belief revision. Presented at the Third International Workshop on Nonmonotonic Reasoning, Stanford Sierra Camp, CA, June 1990.

[11] Doyle, J. Rationality and its roles in reasoning (extended abstract). in: *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 1093–1100, Menlo Park, CA, 1990. AAAI, AAAI Press.

[12] Doyle, J. and Wellman, M. P. Impediments to universal preference-based default theories. in: Brachman, R. J., Levesque, H. J., and Reiter, R. (Eds.), *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 94–102, San Mateo, CA, May 1989. Morgan Kaufmann.

[13] Etherington, D. W. *Reasoning with Incomplete Information.* (Pitman, London, 1988).

[14] Fagin, R. and Halpern, J. Y. Belief, awareness, and limited reasoning: Preliminary report. in: *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 491–501, 1985.

[15] Gibbard, A. Manipulation of voting schemes: A general result. *Econometrica* **41** (1973) 587–601.

[16] Ginsberg, M. L. (Ed.). *Readings in Nonmonotonic Reasoning.* (Morgan Kaufmann, Los Altos, CA, 1987).

[17] Grosof, B. N. Generalizing prioritization. Manuscript, February 1990.

[18] Halpern, J. Y. and Moses, Y. Towards a theory of knowledge and ignorance: Preliminary report. in: *Non-Monotonic Reasoning Workshop*, pages 125–143, New Paltz, NY, 1984. American Association for Artificial Intelligence.

[19] Hanks, S. and McDermott, D. Nonmonotonic logic and temporal projection. *Artificial Intelligence* **33** (1987) 379–412.

[20] Horty, J. F. and Thomason, R. H. Deontic foundations for nonmonotonic reasoning. Presented at the Third International Workshop on Nonmonotonic Reasoning, Stanford Sierra Camp, CA. To appear., June 1990.

[21] Horty, J. F., Thomason, R. H., and Touretzky, D. S. A skeptical theory of inheritance in nonmonotonic semantic networks. *Artificial Intelligence* **42** (1990) 311–348.

[22] James, W. *The Will to Believe and Other Essays in Popular Philosophy.* (Longmans, Green, and Co., New York, 1897).

[23] Keeney, R. L. and Raiffa, H. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs.* (John Wiley and Sons, New York, 1976).

[24] Konolige, K. Hierarchic autoepistemic theories for nonmonotonic reasoning. in: *Proceedings of the National Conference on Artificial Intelligence*, pages 439–443. American Association for Artificial Intelligence, 1988.

[25] Konolige, K. On the relation between default and autoepistemic logic. *Artificial Intelligence* **35** (1988) 343–382. See also errata, **41**(1): 115.

[26] Langlotz, C. P. and Shortliffe, E. H. Logical and decision-theoretic methods for planning under uncertainty. *AI Magazine* **10** (1989) 39–47.

[27] Levesque, H. J. A logic of implicit and explicit belief. in: *Proceedings of the National Conference on Artificial Intelligence*, pages 198–202. American Association for Artificial Intelligence, 1984.

[28] Levesque, H. J. Making believers out of computers. *Artificial Intelligence* **30** (1986) 81–108.

[29] Lifschitz, V. Pointwise circumscription: Preliminary report. in: *Proceedings of the National Conference on Artificial Intelligence*, volume 1, pages 406–410, 1986.

[30] Lifschitz, V. Formal theories of action. in: Brown, F. M. (Ed.), *The Frame Problem in Artificial Intelligence: Proceedings of the 1987 Workshop*, pages 35–57, 1987.

[31] Lifschitz, V. Between circumscription and autoepistemic logic. in: *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 235–244, 1989.

[32] Lin, F. and Shoham, Y. Argument systems: A uniform basis for nonmonotonic reasoning. in: *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 245–255, 1989.

[33] Longley-Cook, L. H. An introduction to credibility theory. *Proceedings of the Casualty Actuarial Society* **49** (1962) 194.

[34] Loui, R. P. Computing reference classes. in: Lemmer, J. F. and Kanal, L. N. (Eds.), *Uncertainty in Artificial Intelligence 2*, pages 273–289. (North-Holland, 1988).

[35] McCarthy, J. Circumscription — a form of non-monotonic reasoning. *Artificial Intelligence* **13** (1980) 27–38.

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

[37] McDermott, D. AI, logic, and the frame problem. in: Brown, F. M. (Ed.), *The Frame Problem in Artificial Intelligence: Proceedings of the 1987 Workshop*, pages 105–118, 1987.

[38] McDermott, D. A critique of pure reason. *Computational Intelligence* **3** (1987) 151–160.

[39] McDermott, D. and Doyle, J. Non-monotonic logic—I. *Artificial Intelligence* **13** (1980) 41–72.

[40] Minsky, M. *The Society of Mind*. (Simon and Schuster, New York, 1986).

[41] Moore, R. C. Possible world semantics for autoepistemic logic. in: *Non-Monotonic Reasoning Workshop*, pages 21–32, New Paltz, NY, 1984. American Association for Artificial Intelligence.

[42] Moore, R. C. Semantical considerations on nonmonotonic logic. *Artificial Intelligence* **25** (1985) 75–94.

[43] Mundell, R. A. *Man* and *Economics.* (McGraw-Hill, New York, 1968).

[44] Nowakowska, M. *Language of Motivation and Language of Actions.* (Mouton & Co., The Hague, 1973).

[45] Pascal, B. *Pensées sur la religion et sur quelques autres sujets.* (Harvill, London, 1962). Translated by M. Turnell, originally published 1662.

[46] Pollock, J. L. Defeasible reasoning. *Cognitive Science* **11** (1987) 481–518.

[47] Poole, D. What the lottery paradox tells us about default reasoning. in: *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 333–340, 1989.

[48] Reiter, R. On closed world data bases. in: Gallaire, H. and Minker, J. (Eds.), *Logic and Data Bases*, pages 55–76. (Plenum Press, 1978).

[49] Reiter, R. A logic for default reasoning. *Artificial Intelligence* **13** (1980) 81–132.

[50] Roberts, F. S. *Discrete Mathematical Models.* (Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1976).

[51] Roberts, K. W. S. Possibility theorems with interpersonally comparable welfare levels. *Review of Economic Studies* **47** (1980) 409–420.

[52] Rychlik, P. The generalized theory of model preference (preliminary report). in: *Proceedings of the National Conference on Artificial Intelligence*, pages 615–620. American Association for Artificial Intelligence, 1990.

[53] Sen, A. The impossibility of a Paretian liberal. *Journal of Political Economy* **78** (1970) 152–157.

[54] Sen, A. Social choice theory: A re-examination. *Econometrica* **45** (1977) 53–89.

[55] Shoham, Y. Nonmonotonic logics: Meaning and utility. in: *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 388–393, 1987.

[56] Shoham, Y. *Reasoning about Change: Time and Causation from the Standpoint of Artificial Intelligence.* (MIT Press, Cambridge, MA, 1988).

[57] Stein, L. A. Skeptical inheritance: Computing the intersection of credulous extensions. in: *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1153–1158, 1989.

[58] Sternberg, R. J. Intelligence is mental self-government. in: Sternberg, R. J. and Detterman, D. K. (Eds.), *What is Intelligence? Contemporary Viewpoints on its Nature and Definition*, pages 141–148. (Ablex, Norwood, New Jersey, 1986).

[59] Thomason, R. H. The context-sensitivity of belief and desire. in: Georgeff, M. P. and Lansky, A. L. (Eds.), *Reasoning about Actions and Plans: Proceedings of the 1986 Workshop*, pages 341–360. Morgan Kaufmann, 1986.

[60] Touretzky, D. S. *The Mathematics of Inheritance Systems.* (Morgan Kaufman, Los Altos, CA, 1986).

[61] Touretzky, D. S., Horty, J. F., and Thomason, R. H. A clash of intuitions: The current state of nonmonotonic multiple inheritance systems. in: *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 476–482, 1987.

[62] Tullock, G. The general irrelevance of the general impossibility theorem. *Quarterly Journal of Economics* **81** (1967) 256–270.

[63] Van Frassen, B. C. Values and the heart's command. *Journal of Philosophy* **LXX** (1973) 5–19.

Contents

1 Introduction **1**

2 Preferential theories of default reasoning **2**

3 Resolving conflicting preferences about defaults **5**
3.1 Examples of conflicts 5
3.2 Skeptical and credulous conflict resolution 6

4 Social choice and nonmonotonic logics **8**
4.1 Aggregation policies 8
4.2 Aggregation principles 9
4.3 Default rules 12
4.4 A general impossibility theorem 15
4.5 Lexicographic priority mechanisms 16
4.6 Petty dictators 18

5 Paths toward possibility results **19**
5.1 Modifying the expressiveness of preferences 19
5.2 Modifying the aggregation principles 20
5.3 Modifying the output of aggregation 21

6 Applications to mental societies **22**

7 Conclusions **23**