Report Number: WUCS-90-35

1990-10-01

Edited Transcription of the Workshop on Defeasible Reasoning with Specificity and Multiple Inheritance

Jennie Dorosh and Ronald P. Loui

EDITED TRANSCRIPTION OF THE WORKSHOP ON DEFEASIBLE REASONING WITH SPECIFICITY AND MULTIPLE INHERITANCE

Jennie Dorosh and Ronald P. Loui, eds.

WUCS-90-35

October 1990

Department of Computer Science
Washington University
Campus Box 1045
One Brookings Drive
Saint Louis, MO 63130-4899

Additional transcription by: Kevin Ballard, Stephen Bashuk, Jim Daues, Steve DuBach, Peter Friesen, Ginny Gent, Jordan Kimberg, David Knickmeyer, David Kocs, Judy Lewis, James Miller, Janette Mize, Jerome Plun, Chung Shan Tai, Matthew Thomas, James Washek.

Additional audio work by: Alo Hasselbring, David Mitchell III.

To appear in ACM SIGART 1990.

Edited Transcription of The Workshop on Defeasible Reasoning with Specificity and Multiple Inheritance, St. Louis, April 1989

Jennie Dorosh and Ronald P. Loui, eds.
Washington University
St. Louis, MO

Additional transcription by: Kevin Ballard, Stephen Bashuk, Jim Daues, Steve DuBach, Peter Friesen, Ginny Gent, Jordan Kimberg, David Knickmeyer, David Kocs, Judy Lewis, James Miller, Janette Mize, Jerome Plun, Chung Shan Tai, Matthew Thomas, James Washek.

Additional audio work by: Arlo Hasselbring, David Mitchell III.

Editors' Note: The dialogue has been liberally edited, especially in the interest of brevity and comprehensibility. Most of the deletions are of fragments, rhetoric, references to visual aids, and redundancy. Most insertions resolve anaphoric reference.

For various reasons, the speakers were not permitted to edit their own remarks. The editors invite checks of accuracy and integrity with the original tapes, which can be made available. Ultimately, this work is an interpretation of the proceedings and the editors are responsible for any misrepresentation, none of which is intentional.

We have tried to produce a document that captures the zeitgeist during a time of clashing intuitions, a document that makes public some of the discussions that do not belong in technical papers or have not appeared in print, but nevertheless should be accessible to students of the area.

The transcription must also be assumed to be imperfect. The original tapes truncated some of Brian Haugh's remarks and all of James Hawthorne's presentation. The remarks by Lin Padgham, Bart Selman, and Randy Goebel were generally too technical to be included here.

Friday Session I: What Is This Thing We're Trying To Formalize?

Speakers: David Etherington, Donald Nute, Judea Pearl.
Moderator: Ronald Loui.

Loui: We are here to talk about defeasible reasoning with specificity and multiple inheritance. Presumably everybody here agrees that it's a good idea that we have something called specificity defeat, subclass defeat, or subset defeat in our defeasible reasoning. What else should direct us in formalizing systems that exhibit other behavior?

Etherington: It occurred to me that maybe we should say that what we are trying to formalize is exactly what Hector Geffner's latest implementation achieves because Ron Loui tells me the system is perfect.

It's probably premature to award that prize. In '84, a few other people and I were ready to give that prize to Dave Touretzky but a few people have come up with reasons for not saying Dave Touretzky solved the problem with *TMOIS*.

At Catalina, Len Schubert wrote a paper about there not being any denotation to the term "semantic networks" anymore. Any defining characteristic for semantic networks also applied to almost every other knowledge representation system. He argued that it was time to retire the term "semantic network" or else replace it with something else.

A question this raises is whether "defeasible reasoning with specificity and multiple inheritance" has any denotation or whether each of us takes something from it that we like and leaves the rest. Also whether there's anything that distinguishes us from doing commonsense reasoning or just doing reasoning? With a history search we might see the underlying motivation for this work.

One idea that seemed to come from early work on inheritance systems is that we want an economical representation, something that gives tight control over inference. All we did was path following, but at least there was some psychological validity to it. Sometime after that, people decided to have exceptions to inheritance reasoning, like the shortest path heuristic, and we lost sight of our original goals. Later somebody discovered that the shortest path heuristic doesn't do exactly what we would like.

Loui: Did anyone ever believe the shortest path heuristic?

Etherington: A lot of people implemented it. A lot of people still implement it.

We've discovered inheritance reasoning isn't simple. It's not obvious. We've gotten to the point where we don't even have a nice local, efficient way of doing it.

It seems that we want economical representations where we only have to say once what isn't inherited to get proper inheritance. A lot of us have let go of the psychological validity question and control of inference. We've tried to work out a general theory which is undecidable or worse. We have this idea that if we can sort problems out, we'll narrow them down to inheritance and intractability problems will go away. I don't know whether that's a reasonable assumption, but it motivated a lot of my work.

What are we trying to formalize? Are we trying to restrict qualitative probability? Are we trying to get human commonsense reasoning? Or are we just trying to come up with a useful representation and reasoning mechanism? Are the links that we have typicality links, high probability links, causality links or all of those things and none of those things with a few other things put in along the way? Is the general case, general reasoning, a difference in kind or degree from inheritance reasoning?

One view of inheritance is that if A's inherit from B's then prefer the consequences associated with A's to those associated with B, which is sort of what we call property inheritance. A second view might be if A's entail B's then prefer A's consequences to B's. That is sort of the general logical defeasible reasoning question. There's a temptation to say that the second one is too hard.

If we're trying to build systems that reason, it seems that even the first view can be too hard. The nets will get too big.

There are a lot of questions, such as whether we should use on-path/off-path preemption, about what we should get out of networks. One question we're not addressing can be seen in the Royal Elephants example. Is being royal a cause for being not gray or do royal elephants happen not to be gray?

Suppose we say that most people do average or better on scores, college students are people, and people who are historically economically disadvantaged tend not to do as well on scores as other people. Then we learn that Fred is both a college student and a poor person – what do we believe? We're less willing to say he won't do well on scores because we believe that being a college student is somehow connected with having OK scores, despite the fact this is not represented in the network.

Loui: Is that fair since you didn't draw that in the network?

Etherington: That's the question. How much intensional information we can put in these networks and how much do we need before we can argue our intuitions on diagrams. If they contain causal information, they will mean different things than if they just contain typicality or frequency informations.

Pequeno: I'd like to make a point that was made at the workshop in Tuebingen. In these diagrams you shouldn't put labels on nodes because you should figure out what to conclude just from the configurations themselves. When you put labels, they influence your intuitions.

Etherington: I agree that putting labels on them reflects that we're trying to use the same network to express different things. Different relationships hold even though we try to use the same structure to represent both of them. The field needs to think about whether there is any meaning to these things.

Shastri: Bill Woods wrote this paper, "What's in a Link", and we seem to be forgetting all of the lessons we ought to have learned from it. Drawing graphs with links, which are sometimes about probabilities, sometimes about tendency, and sometimes about causality, is part of the problem.

Etherington: I think that's one thing we need to think of.

The last thing I want to say, in terms of formalization, is about this thing you can call the Nixon ladder. I've taken your suggestion and not put any labels on its nodes. With this network we have a lot of arguments about whether it is appropriate to be credulous, so you get a large number of possible sets of beliefs, ambiguity-blocking skeptical, where you get a parity on your sets of beliefs, or ambiguity-propagating skeptical.

Back to the proposed original question of psychological motivation. Maybe we should say, "Gee, I don't know."

Nute: First of all, what are we trying to formalize? We are not all trying to formalize the same thing I think. Whether we think of it as a simplified probabilistic reasoning or restricted probabilistic reasoning depends upon what we are looking at. An obvious thing that we are trying to formalize is inheritance.

Some cases of inheritance perhaps can be treated probabilistically or we can think of some of the links in inheritance networks as mostly A's or B's. Another case where probabilistic approaches might be reasonable is causation, where events of a certain kind tend most of the time to cause events of type B.

There are other cases, moral or ethical ones for example, which don't lend themselves to this. You have to keep your promises but if keeping your promise will do great harm to somebody, then you ought not keep it. You can deduct expenses from your income tax, but if your expenses were reimbursed, then you can't deduct them. I don't think we want to say that we can deduct most of the time. That's not the intention behind saying "you can deduct your expenses but not if they're reimbursed." In the case of temporal reasoning too, we have cases which can't be represented probabilistically. If we say things tend to stay the way they are, that doesn't mean most of the time they stay the way they are. We expect something and then it's defeated in particular instances. Few have done much work with normative reasoning in this context.

More importantly, normative reasoning, the patterns of normative reasoning, and the way rules are used in normative reasoning, defeasibly, probably had more influence on the way we reason defeasibly about non-normative situations than we realize. It's probably important to look at how we do defeasible normative reasoning to get some idea about how people do defeasible reasoning in normal contexts.

Besides saying we're not trying to formalize the same thing, another answer to your question "what is it we're trying to formalize?" is that maybe we're not trying to formalize anything. By that I mean there is not any preexisting phenomena that we're trying to capture in our formalizations.

We're trying to develop and implement correct defeasible reasoning but there are different criteria for correctness. One criterion for correctness would be an objective criterion of reliability for reasoning based on results independent of the system. We could see by a system's results whether most of the time it keeps us out of or gets us into trouble.

Etherington: If you have a notion of good results, aren't you trying to capture whatever defines that notion of goodness?

Nute: That's correct.

Loui: I don't see how you can tell whether it did poorly because you drew the network improperly, forgetting to put in that link that college students tend to do well.

How will you tell whether bad results are the fault of the reasoning system and not of the person who created the networks.

Nute: That's true. Still, we can have different kinds of criteria. One is an abstract objective criterion, like truth, but another is how well it models what people do. I think there is probably a middle position where we're headed. We look at what people do, actual practice, and we're interested in formalizations that improve upon that. What we're trying is both descriptive and normative.

The philosopher uses logic to describe the way people act or conceptualize certain issues. But cases philosophers are most likely to be interested in, are the ones where people contradict themselves, or relatively simple ones, where people can't arrive at conclusions because there are gaps.

Philosophers look at actual practice and try to describe it. They expect to abandon actual practice and prescribe. Their final analysis must fit well enough to be a solution to a problem of actual practice but at the same time eliminate contradictions or fill gaps in actual practice. We have a mixed criterion for correctness. I'm suggesting that there is not any one thing we are trying to formalize.

Correct analysis of defeasible reasoning might be different for different domains. We might require different analysis of defeasible reasoning for ethics, law, science, inheritance, and causation and temporal analysis. The criteria of correctness may be mixed between objective criteria (that's what we're trying to get out of formal semantics) and actual practice.

Thomason: Why do you think gaps and contradictions will go away if you restrict to a reasoning task? Take commonsense physics. In the history of Hellenistic philosophy people got into contradictions doing physical reasoning. You think redefining the problem and specializing it will do away with this or is there a more basic problem involved in formalizing commonsense reasoning which philosophers can tell us about?

Nute: I don't think it goes away just through formalization. I think when we make decisions we don't capture anything that preexists, we make decisions. We don't just capture intuition, we reform it. We don't just describe it as a conceptual scheme, we change it.

Thomason: You must sense that the commonsense physics concept is misguided. Either you work on it enough so it isn't commonsense or it is incoherent.

Nute: I'm not sure what I would say about commonsense physics specifically. I'm opting for pluralism.

Certainly all of my remarks won't apply to every situation. I'm trying to describe an attitude. Generally I'm trying to use logical techniques to attack philosophical problems. Start with a practice. That is where your initial data comes from. I expect the final analysis won't match that practice entirely.

Commonsense physics depends on how you set it up. You may restrict yourself in task too much. That may be your goal. Then you wouldn't be doing what I'm talking about.

Thomason: Maybe you can eventually find something coherent about fluid, space, and so forth, but have no way to put it together.

Fetzer: Why do we want to capture commonsense physics? We have a better physics. The only reason to capture commonsense physics is to represent a lot of people's beliefs that may have vague reasons and ambiguities but aren't adequate for explaining or understanding anything.

Nute: If you want to model the way certain cognitive agents behave, you shouldn't reform. But you have a point.

Thomason: Suppose we're building a tutorial system for people; presumably in teaching them the right thing you'll need to teach them to use the module.

Fetzer: You might just want to tidy up their faulty set of beliefs in the first place.

Nute: It depends upon your goals. The system may have inconsistencies or doubts. If you're modeling humans you know that's going to happen.

Audience: You might want to use it for natural language applications.

Nute: You can at least explain when they contradict themselves or when they are speechless.

Fetzer: In '49 Robert Lynd pointed out a bunch of cases where commonsense is inconsistent. For example, "absence makes the heart grow fonder" but "out of sight, out of mind." Commonsense beliefs are compatible with any behavior one displays.

Nute: I want to point out semantics that we might want for rules, like "truth conditions" or "justification conditions." (Note, they might not be the same. Similarly, justification conditions and assertability conditions are close but not necessarily the same in meaning. Truth conditions and assertability conditions are often not the same.) More kinds of conditions – particularly when we're looking at the norm and I think that a lot of these rules we use are much like norms rather than like probabilistic statements – are compliance conditions. Part of knowing what a rule means is knowing how to comply with the rule. Knowing what it means to comply with a rule isn't the same as knowing what you have to know to justify accepting a rule. Justification and compliance aren't the same, just as justification and truth conditions aren't the same. There may be other ways of approaching semantics of rules or links.

Etherington: Can you give us a specific example of the difference between justification and compliance conditions?

Grosof: First, what are they?

Nute: Suppose I gave you a rule of chess and I told you begin the game by moving king's pawn to king's pawn four. Do you know how to use that rule?

You take a piece and you move it there, okay? You understand how to comply with the rule. You have no idea whether the rule is justified, whether it is a good rule to adopt.

Grosof: Do you mean by compliance I have enough background information to interpret the instruction of the assertion of the rule?

Nute: You know how to follow the rule.

Grosof: Okay.

Nute: If we interpret "birds fly" as a policy for forming and revising beliefs then knowing how to use that policy, "birds fly," as part of a group of policies, and knowing how they interact could be quite different from knowing when to adopt the policy. The justification for adopting the policy "birds fly" and the justification for adopting the policy "if you have a business expense, deduct it on your income tax" might be quite different. But the compliance conditions, the way these rules or policies interact with other rules or policies in the same domain, might be quite similar.

Grosof: It seems you're making a definition of justification conditions in terms of why you would introduce a rule into a system of rules on some condition (which I'm still a bit unclear about) which is in the system. How does this relate to assertability and truth?

Nute: A good criterion for justification is: if a rule is true, adopt it. Norms are particularly clear cases. "If you make a promise, keep it." To say that this rule is true is peculiar. To say of any rule that it is true is peculiar. If you interpret things like "birds fly" as a policy for forming and revising your beliefs, then (as a rule) figuring out the truth conditions of these rules is

difficult. Do rules have truth values? If you follow the rule "if you know something is a bird and, having no evidence to the contrary, conclude it flies," it doesn't seem that the rule has a truth value.

Audience: How is an assertability condition distinct from a justification condition?

Nute: Assertability and justification conditions may be the same. Those circumstances where you are justified in making an assertion might be the same circumstances under which you are justified in adopting the rule.

Loui: Have we ever been able to write assertability conditions down and see what they look like?

For any rule, what does the language look like?

Nute: You write them down in English.

Fetzer: Basically, you're talking about rules of evidence and inductive inference.

Nute: Not the norms.

Fetzer: For ethical norms?

Nute: Sure.

Fetzer: You have general theoretical principles that follow from ethical norms?

Nute: Not necessarily inductively. You have other kinds of norms like policies and tax laws.

Fetzer: I thought your distinction between assertability and justificatory conditions had to do with assertability being an epistemic issue: under what conditions are you entitled to assert or should you accept the truth of some claim where the evidence may be inconclusive or partial? The traditional forms of justification or forms of assertability conditions assume the character of inductive and deductive principles of reasoning.

Nute: If it doesn't make sense to call rules "true" or "false" then that characterization of justification or assertability conditions wouldn't make sense either.

Fetzer: Because conditions of justification in the ethical context could be conditions on which you're justified to perform a certain action, they would not have to do with assertability at all.

Nute: I'm not going to try to say what the justification conditions for an ethical rule are.

Fetzer: I interpreted the question as an invitation for clarification of the general notion of assertability, justification, and so forth – which could apply to many contexts. Do you want to relate it only specifically to rules?

Nute: That's all I was talking about here. We're talking about semantics through these systems, particularly of rules. Looking at rules in the context of defeasible reasoning, like "if it's a bird, then it flies," as "true" or "false" is problematic. We can talk about whether we're justified in adopting a rule. This criterion is separate from what it means to follow the rule, or the compliance conditions.

I wasn't trying to make a big distinction between justification conditions and assertability conditions. If these links have truth conditions, they are probably different from their justification conditions, which they surely have. They will be different for different domains. And their compliance conditions, which they seem to have, might be similar when the justification conditions are different. Compliance conditions for normative reasoning and inheritance reasoning, for example, might be similar even though the justification conditions for adopting an ethical or a normative rule, and a link in an inheritance hierarchy might be quite different.

Pequeno: This discussion sounds like the discussion in the thirties between Wittgenstein and Russell. They discussed whether logical philosophers should seek the meaning of rules, in terms of their use, or their truth conditions. You talk now of compliance conditions.

Nute: I'm not favoring compliance conditions. I suggest that to understand a rule like "birds fly," you have to know the evidential issues. We know all birds don't fly, but we need evidence before we're justified in adopting the rule. We would have to know how to apply that rule in the context of other rules it interacts with in the appropriate domain. Complete understanding involves both.

Pearl: Matt Ginsberg gave me the Jerk Lawyer Example during our bus ride. It was interesting because I had a strong intuition about it a year ago and I lost it. That's typical. We deal with complex systems; once we have three, four, or five rules, we don't have any intuitions. Having disputes about things for which we lack strong intuitions is a sign of maturity. Perhaps we are close to consensus.

I had a conversation with Grosof about examples and we decided to make believe we're reaching maturity. Some people believe that the harder they kick a ball, the farther it goes. Imagine in the 14th Century, they were happy with this qualitative physics. Then Newton asks, "is that ball going in a parabola or hyperbola?" They say "Who needs accuracy? We reached a consensus: we know how to kick the ball." Who needs the precision of Newtonian physics when there is a naive physics consensus? Can we conceive of an era when the difference in AI will be drastic? I wonder. I don't believe.

Don Nute suggested pluralism. I think pluralism is a cop out. It says the human mind has different procedures for different domains. We haven't yet captured the things common to most reasoning activity. And there's more to capture. Before we go to pluralism, let's try to capture what is common to all rules. It's premature to see a pattern common to normative and epistemic reasoning; normative reasoning is less specific than probabilistic reasoning because utilities can be erratic.

Nute: Normative reasoning doesn't have to be utilitarian either.

Pearl: I could interpret any normative rule in terms of utility and probability. Let's agree that inheritance nets deal with beliefs. There's a lot to be done even if we limit ourselves to the interpretation of these rules as beliefs.

Kyburg: I took Don Nute's pluralism to mean that you ought to explore different kinds of logic, different rules for logic, different axiomizations, and maybe different procedures for defeasible reasoning in general. That is not to say that one is as good as the next. It's great to have a lot of possibilities but I hope we can narrow the field and not have to live in a pluralistic society where anything can be said by anybody. The issues are more serious than that would suggest.

Etherington: There's pluralism in inheritance too. We have no shortage of systems, no shortage of different logics and approaches. There's the question of whether we want to bring in other problems like normative reasoning and have a plethora of systems for normative inheritance reasoning.

Kyburg: The question is to what extent can you unify various domains with similar structures.

Fetzer: Surely, part of Don's appeal was for clarification of what we do when we develop such a system. I think Henry Kyburg's words were harmonious with that. At least give us a crude idea of what you are accounting for, explaining, or improving when you offer a system. Give the pragmatics of the enterprise. What are you trying to do?

A specific response to your suggestion that we should assume these networks represent beliefs, is to ask "the beliefs of whom?" Infants? Senile elderly? There are potential problems.

Pearl: My point is that even if we limit ourselves to beliefs of ordinary, observant, rational individuals, we still have problems.

Shastri: We might still have too many problems if you do that. Many people suggest we avoid the problems associated with assuming we are discussing beliefs. Be more specific: ask what's the nature of these beliefs then focus on what beliefs mean.

Pearl: Are you saying the problem will go away if we widen the scope to include other issues?

Shastri: No. Try to make distinctions between different sorts of beliefs. Are we talking about beliefs based on underlying probabilistic information, notions of causality that we have about the domain, or something else? What I mean by "refine what we mean by beliefs" is find the basis of our beliefs, and then try to formalize specific aspects of that.

Pearl: It's all factual knowledge. The problem is to define whether we mean high probability, very high probability, or something else.

Shastri: What do you mean "something other than probability?" Maybe you mean what I mean when I say the sorts of beliefs are different.

McCarty: Let me make a comment in support of the pluralistic position, (meaning that many different domains can include different types of defeasible reasoning). I suggest pluralism doesn't lead to an anarchic situation, but to a stable methodology of investigation. We have a general pattern of reasoning which is identified in a lot of examples, different domains of rules, exceptions, and exceptions of exceptions. We see these complex structures. The interesting cognitive question (which might focus research in particular domains) is "why would people choose to set up their knowledge structures that way?" If you look at the epistemic as opposed to normative questions, you might get different answers as to why people would set up their knowledge base the way they do.

There are simple quasi-legal examples. There are numerous examples of defeasible rules in legal systems

adopted for a great variety of reasons. We can usually be specific and precise about why particular rules were structured the way they were. One common reason is our burden of proof rule: presume something unless the opponent proves the contrary. A simple example from Kowalski's work on the British Nationality Act: if X is found abandoned as a newborn infant in the United Kingdom, we presume that X's parents were British citizens, unless shown otherwise. That's a perfect example of a default rule. If you can't find the parents, you have no way of knowing whether X is a British citizen. That's why you adopt that default in this case. Other normative rules, in the legal system, can serve as examples. These are examples of the justificatory conditions.

I think we'll get very different structures for rule exception in networks depending on what the justification is.

Nute: They may differ, but then again they may not. We're talking about what I call the compliance conditions for the rule. I am not an anarchist, but I suspect that there is tremendous similarity between the way we use rules in an epistemic and normative situation. This was apparently too well disguised an attack on probabilistic approaches.

In normative situations I don't think we can justify rules in a probabilistic fashion, or by induction. If we use rules and exceptions in similar ways, in the epistemic and normative case, then that would seem to be an argument against probabilistic accounts of epistemic reasoning. So, if you are fond of probabilistic semantics, or epistemic defeasible reasoning, then you might be forced to a position of pluralism. The reasoning in both situations is basically the same, and it seems like McCarty gets along without a probabilistic mechanism just fine.

Pearl: Can you show us an example?

Nute: Of what?

Pearl: Of a normative rule which is held by a different mechanism that couldn't be described probabilistically?

Nute: Sure, the one he just said: find an abandoned child in Great Britain, presume that his parents are British citizens.

Pearl: Yes, but isn't that justified probabilistically?

McCarty: No, not probabilistically. You might justify the rule that way, even if you were absolutely confident in a statistical survey which said the probability would go the other way.

Nute: Even if you found that most of the abandoned children in Britain were Pakistanis who were abandoned there so they'd have a better life.

Pearl: It's not only the statistics of people abandoning children, but also the damage to society of trying to find the exceptions, and the utilities involved.

McCarty: We have gone beyond just probability.

Pearl: The damage to society is based on past experience.

Grosof: There is a class of normative rules which runs counter to most probability and security rules. Like airport check: the presumption is that if people walk around the metal detector, they're terrorists. The probability is high that a person is spaced out but not a terrorist.

The problem is that for any probability model, you can find a utility model that would decide differently. There's no way to justify that the assertion is tied to probability in a way that distinguishes it from other explanations for such defeasible rules. In a defeasible system, we'll never be able to make distinctions about whether systems are suitable for probabilistic methods, or whether one system is better than another, unless we ground particular assertions of a representation's or rule's intuitiveness or usefulness in a particular application where we can root intuitions and say "the system was convenient, useful, elegant, computable and did the right things."

Methodologically, we should proceed by grounding in particular application categories. A paradox in this field is knowing how to define these application categories, except by saying one application is like another, thus they have a formal system in common. There's a methodological circularity I think we're all struggling with.

Pearl: Suppose we have a nice narrow domain and settle on reasonable beliefs. Nuclear disasters, terrorists, all of these are just one util, plus or minus one. Are we happy then? Can we build a reasonable reasoner?

Poole: Maybe we should require people who propose new systems to write a user's manual so we can give it to a graduate student and they can get the right answers; so people like Ron Loui can't say you should have put that arrow in. The user's manual will tell you when to put the arrow in. Somehow we need to say "this is how to use the system or how to put the utilities in."

Loui: What language will you express that in?

Nute: I thought Judea Pearl was going to say it doesn't matter what system you have as long as you can put in the right rules to get out the right answers.

Poole: That's what I thought I heard you say and that made me wonder how to do it.

Pearl: You still have problems even if you give a pure probabilistic interpretation of the rules. There are problems interpreting observations.

Nute: It's not that we still have problems. Some of us think we might not have them at all. We might be justifying rules the other way.

Pearl: I hope in two days we can come to a consensus on this.

Saturday Session I: Research Strategies: Individual Perspectives.
Speakers: David Poole, Fahiem Bacchus, James Delgrande, John Horty.
Moderator: Ben Grosof.

Poole: This session was intended to get more people who were doing different things to explain why their approach is interesting, why others might be interested in them, and why different approaches are interesting to different people.

I have a wire in the bottom of the drawer that cuts cheese well, but I never use it because I have a knife on the counter. The knife works pretty well for cutting cheese, carrots, and a lot of things. It's a tool I use every day. I don't use the specialty tools in the bottom of my drawer. Since computers are now home appliances, I want them to be useful, general tools which don't sit unused at the bottom of a drawer.

First, if you claim what your doing is science, it's important to conjecture there are some tools we have to use, and that they are all we need. It's important to show that this conjecture is wrong so we'll learn something.

The second idea is: if I give you a fancy new screwdriver, a screw, and the idea that they will solve all your building problems, and you use the screwdriver to hammer screws, you'll find it doesn't work well. When you give someone a tool you should give them the intended way of using it. Books about programming languages don't tell about the formal language but rather how to use the system. That's particularly important if someone says "I have Horn clauses that can compute everything."

I'm talking about how to get computers to do things. I want to talk about the minimal tools we need, and how to use them. Logic tells us the consequence of premises, but not what the premises should be. What should the premises about knowledge and logical argument be?

I'll use my knowledge to try to figure this out, but as I don't know everything, it won't be enough. But we make reasonable assumptions about the world. I made assumptions about where this room was. I happened to wind up at the right place. I claim that looking at properties of reasonable assumptions is a reasonable way to use logic. Reasonable means here that the assumptions are known not to be false. The facts together with our assumptions should be consistent.

Where do our assumptions come from? I think we must learn the answer by trying to build systems using assumption-based reasoning. Here's an idea: let the user provide the formula for an acceptable hypothesis. Then we can learn how to supply hypotheses. The THEORIST framework is just a simple way for the user to provide facts about the domain together with possible hypotheses. The "scenario" is the consistent set of premises we can use in explanation of an argument. An "extension" is a maximal description of a way the world could be based on our assumptions.

Why is this interesting? Because I'm not trying to introduce new logic, new rules of inference, new operators, or fancy semantics but rather a simple, sensible way to use logic. I'm also trying to conjecture that we don't need any new logic, new rules of inference, new operators, fancy semantics, or anything else. Even if I can show that I'm wrong, I will have made an advance. If I show that I need funny new rules of inference, then maybe I'll be prepared to accept them. But it seems to me that you want to start with something simple. Is this plausible?

Gelfond: This will be used for what?

Poole: The problems we want to solve.

Is this plausible? It seems a lot of different reasonings fit into defeasible reasoning: one hypothesis is that defaults be used in predictions. We can use abductive reasoning in the same framework. Finger and Genesereth use the same framework for doing design. Inductive learning is sort of the same thing, but there is the problem of coming up with a hypothesis in the same sort of framework. It seems to be implausible.

Shastri: How is coming up with hypothesis a solution to that problem?

Poole: I'm claiming this has nothing to do with coming up with a hypothesis.

Shastri: That may be the problem. What's the nature of these hypotheses? Why do we come up with them?

Poole: I'm trying to get experience coming up with hypotheses and how to use them before we start that project.

Let's look at some cases. Start with hypotheses before we argue how to generate them automatically. This is meant to be simple. We want to build up programming methodologies which say how to use this system to solve problems.

For example, there has been some work naming and canceling hypotheses or parameterized defaults and predictions. You can conjecture we can do everything with circumscription, we can do everything with Horn clauses. This is no use. We want to prove this conjecture is providing a methodology to go about solving problems. It's important to build implementations.

It seems the validity of the conjecture is empirical, not theoretical. You've got to try to solve the problems we want to solve to show you can't do it. We have to build systems, and find out the problems we don't want to solve. To prove the conjecture is false show there are cases we can't do very well.

One case is specificity. We know how to get specificity by canceling defaults, arriving by specific exceptions. It seems there's enough knowledge in the conjectures "birds fly," and "emus don't fly" to conclude "Edna doesn't fly." We now have a framework to talk about these things.

One of the reasons I prefer direct arguments, rather than indirect arguments, is that direct argument works essentially in terms of the hypotheses that go toward the conclusion.

One idea we can follow up is having "birds fly" be applicable even if we only knew "Edna is a bird." In which case, "emus don't fly" wouldn't be applicable.

Essentially if you don't like the conclusion, you criticize the premises.

Pequeno: That you should prefer direct arguments rather than indirect ones, could be debated.

Poole: I want to say why THEORIST is interesting. It's interesting because it's simple, because it's not trying to do things.

What's a non-problem of THEORIST? I think the non-problem with THEORIST is intractability. What does that mean? Well, intractable means powerful, not inefficient. In some sense THEORIST is a programming language. It's a language to state problems and knowledge about the domain. It's essentially the stating of algorithms. Anyone who has tried to build a system knows you need a methodology for programming the system. Programming is what you have to do. You want the ability to solve simple problems simply. We don't want to remove the ability to solve difficult, real, and interesting problems. If anyone asks me whether tractability's a problem, I say "No, I'm writing a programming language." We want tractability to be a property of a programming language.

It seems there are a few ways we can view reasoning: one is as arguments. The argument for this is that language is the window to the mind, and we look at languages in terms of arguments. This symbolic reasoning consists of arguments for and against things. We look at how people reason amongst themselves in terms of arguments. THEORIST is trying to be the simplest argument system.

The second way people view reasoning is by weighing evidence for and against things. The argument for this reasoning is that everything we do is decision making. If we bet, we should use probability theory.

Another group of people who are into imagery. They are going to build one model of the world that's not the solution.

THEORIST is trying to figure out what we do with a simple argument system, and then trying to do it in the simplest way we can. It seems that weighing evidence is reasonable. Maybe we would only use probabilities at the end of an argument if we have a conclusion. Maybe the logic is an antique phenomenon of probability. It seems the only way we can study this is to take one of these two things, weighing evidence or probability, seriously, and follow it through.

In conclusion, I claim we have a system which is simple, powerful, in the sense that it can solve a lot of problems, and reasonably easy to implement. (It ends up being like John McCarthy's system where you only have one step negation as failure.) It's useful in a lot of applications, and showing that we need more. We have not invented a new logic; we're just using logic in a sensible way. By showing how the conjecture is wrong, we will learn something. I claim it's just common sense. THEORIST attempts to get by with few tools.

Randy Goebel and his students have been working on this at Alberta and UBC. In some sense all that they've done is taken David Israel's suggestion that we need rational epistemic policy ideas. Jon Doyle has been talking about wanting assumption-based reasoning for a long time.

Etherington: I want to know the relation between the formal model and the world. From this point of view default logic and circumscription both have problems.

Konolige: So does deductive logic.

Bacchus: Not the same problem. I claim that the relation of the Tarskian models to the world is a much clearer relation than the relation between minimizing predicates and what that means in terms of properties of the world.

These semantics do not address the issues I am interested in. Defaults, for example, are specific warrants to make very specific defeasible inferences. How does an agent learn its defaults? Telling the agent all the defaults won't work well if you build complex agents.

From the point of view of learning, what things will the agents have experience with? Any field of defeasible inferences and any defeasible policies the agent makes must be based on its experience with its environment.

Goebel: Is this different from the dynamic view of starting from a collection of beliefs?

Bacchus: This relates to that view. In the case of new evidence, how do we generate policies?

Goebel: You seem to be saying something stronger, that we don't know where a default comes from, but it's equally mysterious where a set of beliefs come from.

Bacchus: Sure.

Etherington: You say that they must come from interaction with the environment. I hope you're using environment in a general form where reading a textbook that says "most birds fly" would be the environment.

Bacchus: The key issue is that defaults provide an operational specification about what to do in your system. This operational specification doesn't really have any relationship to the properties of the world. Why are defaults the way they are? Why is "Birds fly" the default? If our knowledge base can't be used, if it doesn't entail "fly Tweety" or "not fly Tweety," then that default is consistent if and only if its negation is consistent. Why have one default per system rather than some other default? I think some properties of the world make "birds fly" the reasonable default. These properties are not captured in the formalism.

I think circumscription suffers from similar problems. The operational specification of what to do is being pushed from the proof-theoretic level to the semantic level. We have specification of what to do with certain predicates. The relationship between these minimizations and properties of the world is not clear to me. Why does the world choose either to minimize or not?

McCarty: The world just is. The language tries to capture what the world is. I can't relate defeasible inferences to things in the world with this formal process because I'm not sure what it means. What does it mean to minimize the predicate?

Doyle: Do you include the agent's preferences as properties of the world? If you found someone near the scene of a crime, a judge might say the appropriate default is "innocent until proven guilty" but a detective on the scene would say the suspect is "guilty until proven innocent."

Bacchus: It will be impossible to construct complex defeasible reasoning systems which handle many competing defeasible inferences in an intuitively reasonable manner without relating them to properties of the world. Statistical properties are used to justify defeasible reasoning.

Why would one defeasibly infer that Tweety flies? Because most birds fly. Why should one defeasibly infer that one's car will still be in a parking lot at the end of the day? Because most days it is. Why should one not defeasibly infer that a dollar bill left on a bus will remain there? Because most of the time it isn't.

One idea of deriving probabilities from statistical information goes back to Reichenbach and Kyburg. The notion of reference classes is that statistical information over sets of individuals is used to compute a probability, a particular statement about a particular individual. Kyburg has arguments for why reference classes are applicable when statistics are available. The problem of where these numbers come from does not arise.

This type of defeasible reasoning allows the possibility that the agent can learn its own policies through its experience. Another nice feature is that you get a graded evaluation of how good an inference is because the deduced probability is essentially unique. You need a graded evaluation, a way of determining how good your defeasible conclusions are if you're guiding actions from your defeasible conclusions.

You still have the problem of choosing the reference class but I think it's isomorphic to the problem of dealing with interacting defeasible reasoning.

Pollock: Are you claiming that the only basis for defeasible reasoning is statistical?

Bacchus: No, although I think you can make arguments for that. In many cases you can reduce things

to consideration of utility if you include a wide enough class of things.

Pollock: To get statistical information, you have to be able to do some defeasible reasoning. For example, the acquisition of information through perception is defeasible.

Bacchus: Right.

Pollock: Presumably, you've got to perceive the world before you can accumulate data that you can use for statistical information.

Bacchus: But we accept our perceptions as more likely to be true when we have a built-in system.

Nute: You don't gather statistics on how often your perceptions are correct.

Kyburg: But you can use your perceptions and what you know about the world to infer, defeasibly, that you make errors a certain amount of the time.

Pollock: But you have to get started by perception.

Kyburg: You don't have to regard perceptions as defeasible to start.

Pollock: You can't learn something that isn't defeasible. If you think of perceptions as deductive, then there will never be anything to make you take them back.

Israel: You might think something is terrific and then learn it's not.

Pollock: That's different from saying something is defeasible. To say something is defeasible simply says it's possibly wrong.

Bacchus: You don't have to regard perceptions as reasoning (it may be a process). Is this process defeasible?

Konolige: What role does argumentation play in this? It seems bothersome to say that you believe your car will still be in a parking lot at the end of the day because over the past year you've noticed it always has been. I think there's reasoning you could use even without knowing anything about a particular parking lot. You might use general arguments about things tending to stay in one place.

Bacchus: Right. That qualifies as general statistical knowledge about objects and their properties. Those are reference properties.

Konolige: Then you combine these arguments in certain ways and they're all the reference classes.

Bacchus: Right, that's the problem of choosing a reference class.

Etherington: According to Ben Grosof's talk last night, choosing a reference class is a problem in nonmonotonic reasoning. You get a regression.

Bacchus: I feel you have policies for choosing a reference class. I don't know if they are defeasible.

Grosof: We'll have to talk about this in the probabilistic approaches panel. Our next speaker will be James Delgrande from Simon Fraser University of British Columbia.

Delgrande: I've been developing a formal investigation of a particular approach to default and defeasible reasoning. I'm talking briefly about a proposed logic for defaults where it is meaningful to talk about deriving defaults from each other or defaults being mutually inconsistent. I'm talking a little about inference, about properties and issues that arise.

My basic approach looks at assertions like "birds fly" and interprets them roughly as "birds fly." We don't worry about penguins, birds we roasted, or ones with broken wings. I've added an operator for conditional or default relations. A single shaft will be for strict implications, where the "A" condition implies "B." The intended interpretation is, all other things being equal, "if A then B."

All other things being equal, if something is a bird then it flies. The idea is that sentences of this form are intended to correspond intuitively or as a guide to statements in some naive scientific theory. It makes eminent sense to state that a particular fluid boils at a particular temperature although it doesn't because it's impure or the thermometer is inaccurate. Birds fly, although we know that all birds don't fly. Planets move in ellipses, although we know that no planet moves in a perfect ellipse.

First, this approach is claimed to be fundamentally independent of any probabilistic interpretation. It's conceivable that no bird on this earth flies and yet we accept "birds fly." Secondly, this approach is independent of consistency-based approaches such as Reiter's. Thirdly, it's an attempt to specify what we mean by these default sentences. I leave the implementation issues to David Poole and John Horty.

Pequeno: About that statement "similar to scientific theory," I think there's a difference. When doing science, we believe that what we do is true. When we use a statement like "birds fly," we know beforehand it has exceptions.

Delgrande: You work with a particular scientific theory under the assumption that "it is" and the question of "whether or not it is something else." I guess it's more of a mental device, a guide.

Goebel: You're willing to use a scientific theory when it's accurate enough?

Delgrande: Scratch the "naive scientific theory."

If we follow through with this conditional logic approach, we end up with a logic where we state things consistently, such as: "Ravens are birds, ravens are black, but ravens that are albino are non-black." We could consistently assert for example, birds fly, and yet Opus is a bird that doesn't fly. On the other hand, we do get some non-trivial relations. For example, the interpretation of this first sentence is if it's the case that ravens are normally black and not normally albino then you can conclude that ravens that are not albino are normally black. It makes sense to talk about one default following from another and about a set of defaults being inconsistent.

My conditional operator doesn't allow us to make default inferences concerning a particular individual, as we can't apply modus ponens. If birds normally fly, and Opus is a bird that doesn't fly, then the ability to apply modus ponens would lead to inconsistency.

Thomason: That's not modus ponens unless you've got instantiation, too.

Delgrande: Yes, with instantiation.

Nonetheless, if all one knows is, "Opus is a bird" and you have the default that "birds normally fly," then a reasonable default inference is that "Opus flies." We don't have modus ponens since we do have instantiation, why is this reasonable? I claim in such cases when "birds normally fly" and "Opus is a bird," the default inference relies on the assumption that no exceptional conditions hold. As we do know, they hold. So we assume that the world at hand is one of the least exceptional of those consistent with Opus being a bird. Then it has to follow that "Opus does fly." If we know "penguins are birds," "penguins don't fly," and "Opus is a penguin," then under this same assumption we would deduce "Opus does not fly."

Poole: If you need assumptions, then why say conditional logic in the first place?

Delgrande: The logic is of defaults. You have a conditional arrow that says "If A then normally B." It specifies what defaults hold from other ones. We have a binary operator.

Poole: We get by without this new binary operator and conditional logic if we add assumptions.

Delgrande: No, the idea is we have a logic of defaults, which tells you what defaults follow as theorems from other defaults. This isn't present in other systems. The only trouble under this particular approach is that it doesn't have modus ponens. Arguably for any default reasoning system, if we make a default inference, we must assume that the world is least exceptional with respect to that inference.

Etherington: Those are metatheoretical assumptions. They're not assumptions about birds flying; they're assumptions about how the logic should work.

Delgrande: Yes, they're metatheoretical assumptions. Having to bring in these assumptions is arguably a strength of the system.

Goebel: Rather than fabricated on the fly.

Delgrande: I'll reiterate some points I've made and form conclusions. First, the logic itself provides a basis for representing reasonable default statements. Arguably one strength of this approach is if something looks wrong you can argue about the semantics. The semantics are independent of sets of beliefs, believers, or consistency-based . Default reasoning systems involve metatheoretical assumptions. The propositional case is obviously decidable but the first interesting case has a complexity order N cubed, N being the number of defaults.

Delgrande: This system has been implemented. Arguably its main contribution is provision of some justification for other implemented systems of default or defeasible inferences.

Selman: How do the ideas in your system compare to Touretzky's notion that uses paths in the hierarchies? Do you have to build something on top of your logic to get the same results?

Delgrande: No. Inferential distance, comparing two default proofs, comes free in the logic.

Geffner: How do you chain given "A then B" and "B then C"?

Delgrande: If you have "A then B," "B then C," then C follows by default from A unless it's specifically blocked.

Etherington: Then you have specificity of some sort?

Delgrande: Yes.

Selman: Going back to the impression that this comes for free, does skeptical inheritance from Horty come for free too? If I can only mention one of the various proposals, which one does it match?

Delgrande: I don't know. How do dimensions from "The Clash of Intuitions" fit in with this? The present system gives you some stuff for free, such as the orderings, otherwise the system is skeptical. Although, again, by messing around with the logic you could have a credulous reasoner.

Goebel: How can you tell whether it's okay to make that assumption? Regardless of where the assumption is made, at the meta-level or not, you're making the assumption that the conditional sanctions the use of the material implication. Assuming that's okay, how do you know to make that assumption?

Delgrande: No, you're not doing that at all. You assume information is part of the world. The world being described is as unexceptional as possible with respect to your set of defaults. That assumption effectively gives you the grounded material conditional in place of the variable conditional.

Goebel: So the assumption that you're working with the least exceptional model identifies the situation in which you can use the material implication. How do you determine which is the least exceptional?

Delgrande: That comes from semantics.

Goebel: Isn't there anything you must do to ensure you don't use an implication incorrectly?

Delgrande: No.

Pollock: The semantics you give seem to be only for singular conditionals. You talk about B being true in worlds in which A is true. How do you get free variables in there?

If you have free variables you can't talk about them being true or false in a world. You can't say "ISA bird" is true in the world. You have to say "Tweety ISA bird."

Delgrande: No, for example, the default would be "For all X, bird(X) conditionally implies fly(X)."

Pollock: But the semantics you give assume that the antecedent and the consequent of the conditional have truth values in the world. Free variable conditionals don't have truth values.

Etherington: Could you take it over all ground instances?

Delgrande: Effectively, yes.

Etherington: So you do what default logic does: ground over a Herbrand universe?

Thomason: Except then there's a problem about reference in a world in which penguins fly.

Delgrande: According to semantics, in the simplest world in which penguins exist, they don't fly.

Thomason: Why? There's a default "birds fly." That means for all X, if X is a bird, X flies. You make penguins fly.

Delgrande: With semantics, in the simplest world in which there are birds, penguins don't exist.

Etherington: You assert that penguins don't fly?

Delgrande: Yes, you assert they don't fly.

Thomason: I suppose if we have a total simplest world then we have nothing.

Delgrande: A set of simplest worlds.

Etherington: You have axioms that say what exists.

Delgrande: No, that's not true at all. You could have a simplest set of worlds in which there are birds. If you have a simplest world in which there are penguins, they don't fly but are birds.

Horty: How is there a simplest world in which there are penguins, penguins are birds and they don't fly? Where are you putting in preferences?

Delgrande: We need a blackboard.

Grosof: We're not going to be able to go into this level of technical detail. We're running over already.

Our next speaker will be John Horty, from Carnegie-Mellon.

Horty: My assignment from David Poole was to justify work on path-based reasoning and to explain why it might be interesting to someone else. Look at a complicated net. Consider what conclusions it supports.

Shastri: Could you begin by explaining what those links mean.

Horty: These are strict links, "all humans are mammals." Single arrows are defeasible.

Shastri: Could you relate that to what Fahiem Bacchus talked about, properties of the world? What does it say about the world?

Horty: I'll say that later.

Shastri: You want us to tell you what conclusions to draw before telling us what the diagram means?

Horty: That often happens. Telling how to reason with diagrams is part of explaining their meaning.

Shastri: Maybe we should begin by asking what the information is, then ask what conclusions we can draw from the information.

Horty: OK, at the meta-level I'm not answering this question. I'm not allowed to give a technical talk; David Poole said. So I'm not going to tell you what conclusion to derive. I'm talking about strategies for answering that question.

I've noticed two of them: I call one the indirect or translational approach, which involves adding a network, a bunch of links, usually into your favorite nonmonotonic logic or something else and ...

Ginsberg: You mean map it into your favorite formalism as opposed to one that's well understood.

Horty: Yes.

I went down the list of attendees, and I noticed that various people were taking this approach. They use probability and conditional logics, deontic logics, default logics, and circumscription in various ways. The approach I want to justify is the direct approach, which involves reasoning just on the basis of information in the network in an unanalyzed way. Don't try to map it into another logic. That usually involves constructing cows and things.

I'm going to show you a picture of what these theories look like. You start with a net, like the one I showed, and construct extensions. Then those get mapped into supported sets of statements (which is what you really care about). Depending on how you do this, you can get various extensions or one extension. That's a picture of the approach that I'm justifying and arguing for the interest of.

There are various ways of doing this, not just one theory. This came as a surprise to Rich Thomason, Dave Touretzky and myself. When you construct these chains, it matters whether you construct them from the beginning forward or from the end backward. This construction takes place inductively and surprisingly there's a semantic difference depending on which way you go. Bart Selman learned that there's a computational difference, too. There is a third option you can take and this is sort of a hybrid of the other two options.

When you have unrelated conflicting paths as the Nixon diamond, you also have various options. You can pick a path and jump to one extension. You can decide what you want for your conclusion so you intersect the right extensions. Now it matters whether you're going to intersect the paths with the extensions, or the statements supported by the paths with the extensions. You get different conclusions depending on which way you do that. If you have conflicting paths you can do something Rich Thomason and I defined called "inheritance," (I'll put that in quotes because that's not really the correct word for it) which lets both paths die whenever they clash. Or you can do something I call "flexible inheritance" which mixes credulity and skepticism at various conflict points. You take different strategies and mix them together to get the result you want.

Goebel: Conservative and reckless.

Horty: You could call them that.

Konolige: Brave and cautious are words that have somehow fallen unfairly.

Horty: There are different definitions for preferring one path to another. There are a lot of options. We have 72 possible theories. Some are incoherent and some collapse but if you subtract those you still have a healthy range, and those are not all the options. I want to talk about why people who take the first approach should care about this, even if your life's work is figuring out the number of things you can do with circumscription.

One thing you want to do with circumscription is translational theory in these networks. I want to give you a picture of what such a theory would be like. You take a network, your favorite of those 72 theories. You have extensions associated with the network. Now you have this function that maps the net into your default logic theory. That has some extensions too.

Gelfond: Can you explain the picture and notation?

Horty: These pi's are extensions of the network path sets. Someone who wants to do a theory of nonmonotonic inheritance in default logic should define this function in such a way that it's modular, meaning when you translate a big network you translate it piece by piece. The translation should be sound so that each default logic extension corresponds to a network extension. You have to find the correspondents because they will be in different languages. The translation should be complete so that you get all network extensions to your default logic.

Pearl: Why shouldn't it be the other way around?

Horty: You mean why shouldn't I try to manipulate my paths?

Pearl: Yes. You have 72 ways to do things.

Horty: I'm not sure how to answer that.

Etherington: Something was brought up last night about when you're proposing a translation it has to match a certain amount so you can tell them where they were wrong. You should get a correspondence with soundness and completeness. But I might want to show you you've come to wrong conclusions along the way.

Goebel: Yes, the mapping is a debugging exercise.

Horty: You have all these options, and one of them must be right. Just as a methodological point, I've never seen somebody who seriously mapped a lot of networks into a more powerful logic. When you make logics, you do a few problems. You do Tweety, Nixon, and the Yale shooting problem. But I've never seen an exercise, that's what it would be, of translating a classic network.

Pearl: Jim Delgrande said his conditional logic treats networks as a special case of a wider framework.

Horty: But Jim also said that it was an interesting question as to which of these theories it would correspond to. I'm not saying this should be a criterion of adequacy. I'm saying it gives you a place to tie your theory in a simple formalism that is easy to understand.

Etherington: Which one is easy to understand?

Ginsberg: Let's say some path people figure out the right way to do this. Then translational people embed the path algorithm in one of the various nonmonotonic reasoners. The networks become a special case of the general purpose inference engine. At this point does the path stuff go away? Is the path algorithm a step en route that translational people believe they can bypass? Or does this work retain some fundamental importance?

Horty: Except for implementational concerns, I think paths would go away.

Thomason: I think it is more like the relationship between proof theory and model theory in logic. There are a lot of intuitions about proofs at a level of detail we never consider logically interesting.

Ginsberg: Let me strengthen my question.

Horty: Why don't you hold your question because I believe I'll talk about it later.

I have copied objections to path-based theories, written by friends of mine. All of them are here. One objection, which really isn't an objection, is that these theories use funny formalisms. I interpret that as saying the theory has no semantics and is proof-theoretic. There is nothing like truth in a model. We don't explain why "bird fly" is true. That's maybe what Matt Ginsberg was complaining about. But you can answer questions like "what follows from what? when is the net inconsistent?" I don't think it is a devastating objection

An objection I do think is serious is what I feel Matt was talking about. These are limited theories about particular forms of inference, which may be interesting, but are not useful unless generalizable. The response to this objection is that they are generalizable. This inheritance formalism has to do with paths and linear sequences. You can go from paths, which are sort of sequences to propositional things like proof trees. Once you jump from paths to proof trees, you open up the possibility of having trees. Then you can represent arguments through these trees. You can define the things needed to define inheritance when two trees conflict. If you have a set of trees, you define

when one is preempted in the context of a bush of arguments. It is complicated, but notational. Once you spend three weeks to do it, you can define things which are like inheritance theory, except they have proof trees in them. There are essentially no new ideas involved. It is all notational.

It seems that these inheritance theories are a historically distinguished fragment, the implicational fragment, of full defeasible logics. It is useful to limit yourself to the implicational fragment because you avoid a lot of mess, but it is important to see some reasoning can be generalized. I think they are fragments of the theories that Loui, Pollock, and Nute are working on. Questions?

Pearl: Why should there be anything like a one-to-one mapping between the arcs and the logical sentences?

Horty: You wrote a chapter in a book about this, so I do not understand your question.

Konolige: Given this mapping, are there paths you could approach in path-based theory that would be hard or almost impossible to approach in other theoretical basis without dragging in notions like path conflict? For instance, in path-based approaches you can compare chains based on the type of arguments you make. Types of links in the formalism deal with action: A result of doing the action is different from an inertial law. An argument based on one path should win over an argument based on another path. This reasoning is relatively hard in other theories without explicitly dragging in notions of paths.

Horty: This is special to extending the path-based approach to other areas.

Konolige: Yes, but with the path-based approach you have a lot more mechanisms to deal with. You can compare paths based on how they work through a network. Other approaches don't have objects like paths that you can talk about.

Horty: I agree.

Morris: Is modularity a reasonable constraint on translation into another formalism? Incremental or algorithmic translation seems sufficient.

Horty: Are you thinking that when you introduce a new link you may want to redo your translations? That would be routine.

Morris: It would be routine.

Horty: You could bound how much information you had to search through.

Morris: Perhaps, perhaps not.

Ginsberg: It's got to be something like linear time. I guess polynomial.

Horty: There might be weakenings of this. This is the strongest statement.

Saturday Session II: Conventionalism's Curse Or Philosophy Of The Future?

Speakers: James Fetzer, Henry Kyburg, Donald Nute, John Pollock, Richmond Thomason.
Moderator: David Israel.

Israel: Let me say how I thought about it then. Being very crude, I'll say there are two genera of reasoning: deductive and nondeductive. Indefeasible and defeasible might be plausible synonyms, but that makes nondeductive reasoning look like only defeasible reasoning. I actually think that that's a good way to think about it. But I'm inclined to think that there is defeasible reasoning and there is *defeasible reasoning.* There are kinds of reasoning which are nondeductive which shouldn't be called defeasible. One was belief revision. I think it's *gonzo* to think of belief revision as defeasible.

Defeasible reasoning is a not very clearly defined or even characterized family of kinds of reasoning which are nondeductive and in a certain sense more local than belief revision. One can ascertain some structural characterization of the differences. For instance in defeasible reasoning you start with some things that aren't up for grabs within a certain stretch of the reasoning. Though ultimately, over the life of the species, maybe they are up for grabs. You get what is called inclusion or reflectivity. You don't lose the things you start with. They are taken for granted. There is a separation between what is up for grabs and not.

What about inheritance? I do not know what I think now but I did think that the property inheritance or subclass-superclass inheritance was a simple but special case of defeasible reasoning. Inheritance had a life of its own. About four years ago, there was a lot of interesting and relatively simple work. There were just two nonlogical relations, two kinds of specific relations (or one depending on how you think about it): namely ISA and IS NOT A, or inclusion and preclusion.

Notice, that this has nothing to do with people shooting each other and living or not living. You could

arguably, with great strain and no good reason, represent various other problems in this framework. For instance, you can get an ISA situation followed by a situation in which somebody shoots somebody and you could, I thought, get something like good in virtue of form. That is, it would be what you have in deductive reasoning. This form of argument is good. Labels would be irrelevant.

I notice a lot of people have talked about more general forms of defeasible reasoning. One more thing to keep in mind is whether we should do things all at once or keep track of differences.

Fetzer: David is right. Defeasible reasoning is a species of inductive reasoning. In fact, defeasible reasoning is inductive reasoning when you don't know what you are doing.

I would like to talk not so much about the future as about the past with respect to a certain tradition that hasn't been represented here. It's the tradition that takes seriously issues in the area of explanation and the differences between explanations and predictions, in particular, explaining why something occurs and understanding that it might be true.

This tradition is represented by Carl Hempel's, Wesley Salmon's, and my own work. Whether or not these traditions are represented by Reichenbach, Kyburg, Pollock or others, in many ways they trace back to Reichenbach's rule: an individual instance is to be resolved by assigning it to the narrowest class for which reliable statistics can be applied, which he offered in *The Theory of Probability*, 1949. If you distinguish between explanation contexts and prediction contexts, however, then certain problems are generated when we pursue this strategy.

This concept is well illustrated by an individual saying "abra cadabra" to a supermarket door, the door opening, and him walking in. We have a narrow reference class, but we have not improved our explanation. This case illustrates defeasible reasoning because this might or might not be a normal door mechanism. An operator might stand out of sight waiting until a customer says "abra cadabra," and if they do in an allotted amount of time, the door opens. We need to be aware of all relevant factors at work in a given case to figure out what kind of case it is and what we should expect to occur.

Hempel originally distinguished between two species of explanations, deductive and inductive. A typical deductive example is: a crystal of rock salt put into a bunsen flame turns the flame yellow because it is a sodium salt and all sodium salts impart a yellow color to a bunsen flame. Correspondingly an inductive explanation is: John Jones is almost certain to recover from his streptococcus infection for he was given penicillin and almost all cases of streptococcus infection clear up quickly upon the administration of penicillin.

This approach, which is known as the covering law model of explanation, shows us what these explanations should look like in principle.

Hempel also articulated a characterization of an inductive form of explanation involving a statistical probabilistic law, instead of a universal general law. The truth of the premise doesn't guarantee the truth of the conclusion, but lend it a certain degree of support. Here's a conundrum of the theory: if the statistical law were ".9 F are G," then the degree of entailment should bear the same value, ".9." The value is a degree of nomic expectability which provides a measure of how strongly we could expect the event to occur under corresponding conditions in the future. It is the key link between explanations and predictions in this approach.

Wesley Salmon, especially in some of his earlier work on statistical explanation and statistical relevance, became intrigued with the model. I think Salmon had been a student of Reichenbach and he had to wrestle with the same problem as Hempel, because both he and Hempel had supposed that Reichenbach's approach was the right general approach in dealing with narrower classes. But Salmon came to the realization this was not quite right. I will describe and explain some examples he offered.

John Jones is almost certain to recover from his cold within a week because he took vitamin C, and almost all colds clear up within a week with the administration of vitamin C. John Jones experienced significant remission of his neurotic symptoms for he underwent extensive psychoanalytic treatment, and a substantial percentage of those that undergo extensive psychoanalytic treatment show significant remission of neurotic symptoms.

Salmon wanted to make the point that he disbelieved vitamin C was causally efficacious in preventing colds. He believed that most individuals are almost certain to clear up from cold whether or not vitamin C is administered. Similarly, he believed psychotherapy does not help individuals overcome their neurotic symptoms, but rather most individuals experience significant remission of their neurotic symptoms over a substantial period of time anyway.

These explanations are not as good as one might initially think. Salmon was uncertain which of these inductive cases had deductive correlates but Henry Kyburg showed him an instance that did: the dissolving spell. A sample of table salt dissolves in water because a dissolving spell had been cast upon it and all samples of table salt that had a dissolving spell cast upon them dissolved in water. Casting a dissolving spell narrows the class, but does not improve the adequacy of our explanation, nor would using cubes of

salt, or dissolving the salt on a Monday, Wednesday, Friday or a day divisible by nine.

Salmon offered another case of a similar kind: the birth control case. John Jones avoided becoming pregnant the past year for he has taken his wife's birth control pills regularly, and every man who regularly takes birth control pills avoids pregnancy. This case suggests the necessity of distinguishing between different types of cases.

It is essential to distinguish between statistical relevance relations, which are changes in statistical frequency from causal relevance relations, which means changing the tendency of something to be brought about. In the case of prediction, statistical relevance relations can be sufficient, but are not necessary. For explanation statistics, relevance is neither necessary nor sufficient. For prediction, nomic causal relevance is sufficient, but not necessary. For explanation, nomic causal relevance is both necessary and sufficient. In general we ought to prefer causal nomic relevance relations over mere correlations because they need not reflect nomic causal relations.

How can we know no one can win a nuclear war unless we try? If our knowledge must be only statistically based, we could not theorize and conjecture by extending our knowledge in a inductive fashion.

Kyburg: The theme I've taken, from Susan Haack's book *Philosophy of Logics*, is "prudence demands a reasonably radical stance on the epistemological status of logic," that is, it will not be easy to decide whether or in what sense the world dictates one form of logic or another. This question which applies to classical logics, such as many-valued logics, temporal logics, monotonic logics, deontic logics, intuitionistic logics, clausal logics, and so on.

There are logics which involve defeasible reasoning: the new ones such as non-monotonic logics and default logics, and the old ones such as inductive logic, scientific inference in general, and probabilistic logics (what David Israel calls real rules of inference).

Are they all in the same boat? Yes. Suppose you have a bunch of logics of either of these kinds. How do you choose among them? One way is to look for soundness and completeness. This clearly will not be the distinguishing mark. Nowadays everybody knows enough to devise clever models of their logics relative to which they are sound and complete. Conformity to human habit is something that people in cognitive science tend to want to focus on. This doesn't necessarily lead us to the things we want. We'd like to be able to improve on people. Intuitive persuasiveness: that seems to be a stock in trade around here. People devise systems that correspond with their particular intuitions on specific examples. I went through the list of examples we were handed, and I consulted my intuitions. I only had intuitions on 10% or 15% of the cases. I was simply a bit puzzled about them. If I were obliged to come up with some kind of conclusion in these cases, I would consult my intuitions concerning the relative frequencies of various constraints corresponding to premises and the conclusion.

If we focus on the intuitiveness of the principles that lead to solutions in various logics, rather than on examples, maybe we can make some useful distinctions.

In general, I say "not semantics, not really intuition, not really principles, not certainty, not conformity to the way the world is built, and not conformity to an idea (that is the notion that you can look at a few clear principles and everything will fall neatly out of them)."

So then what? It is a matter of convention which logic you take, like in classical logics, whether you want two or three-valued logic, infix, prefix, or postfix operators, first-order or higher-order.

Pearl: How does convention differ from habit?

Kyburg: One can adopt a convention, but not a habit. Somehow habit is more psychological than convention.

How are we going to take convention into account? Conventions don't have to be "arbitrary." We need principles for choosing among conventions, and some that have been talked about are principles of simplicity: the simplicity of the resulting system, its power, its familiarity (which is tied to our habits), and elegance. Surely we want something more. For example we could settle for practical certainty in the resulting set of propositions. That could be construed as resulting experience, and would prefer one convention to another. This a vague notion.

Pollock: What is practical certainty?

Kyburg: Practical certainty invokes probability. In general, that probability depends on statistics. I'll go further than Fahiem Bacchus; I say all defeasible reasoning is based ultimately on some knowledge of statistics, not the statistics with which nuclear war is won, but other statistics about physics, human character, and so on, from which we could easily infer something about the relative frequency with which nuclear wars can, in principle, be won.

There are other considerations. One is the choice of a language in which this will be expressed. I think that is a matter of convention, of decision.

Back to John Pollock's question about observation, I think there are epistemic shifts which occur on a basis other than probability. Again, I don't think they

are defeasible, because I don't think they are falsifiable. I don't think you could be persuaded to abandon them, but rather persuaded to abandon a certain epistemic framework.

I have two relatively complicated examples in which specificity fails. For simple examples, specificity corresponds to Reichenbach's principle of choosing the smallest reference class about which you have knowledge. In these two complicated cases, you don't want to go to the smallest reference class, you want to go to a different one. Is it more specific?

Some people wanted to generalize specificity to the logically strongest description. It's not clear what that comes to: our old-fashioned friend the principle of total evidence, take everything you know into account, doesn't solve the problems. Principles for choosing reference classes that go beyond the subclass principle do. I have a couple of examples which I claim, had I time, would show that other principles are needed.

The final claim is that logic, defeasible or otherwise, is conventional in the sense that you choose between them. There is a non-arbitrary basis for choosing one language over another, one logic over another. The choice depends on the interplay among facts of human condition and human cognition. I claim, and this is my extreme claim, that probability gives us a handle on all of defeasible reasoning just as it gives us a handle on induction and decision theory, scientific inference, are rational belief. What underlies that claim is the subclaim that where probabilistic considerations and other considerations conflict, probability had better win out. Specificity construed as the rule of subsets falls short of accomplishing defeasible reasoning. Specificity construed as logical strength goes no further than the principle of total evidence. And construed as something in between. I don't know how to apply specificity to these two illustrative examples that I have not given you.

Bacchus: Would the policies for choosing reference classes be defeasible or not?

Kyburg: They are part of the underlying logic, part of the convention we hold.

Pearl: Are these policies of specificity compelled?

Kyburg: No they are not.

Doyle: Do you mean more practical certainty is better? That leaves out questions of whether certainties are more important in one convention than another.

Kyburg: You might want to include something other than the mere cardinality of what you get among the pragmatic considerations. We're not even taking cardinality into account yet.

Loui: I wonder about those examples.

Kyburg: This is the familiar one. You are trying to figure out whether a bird flies. You have an aviary with a lot of healthy crows and twenty fat penguins. They are in ten cages. Here is some data about how they are distributed in the cages. The first line is an argument that Tweety can fly, because most of the birds in the cages can. You want to defeat that. You also want to defeat the intermediate argument, if you look at pairs of birds in cages, the proportion of those in which the second member can fly is relatively high. I see no way of saving this bottom inference against this top inference in virtue of specificity. It is really a Bayesian construction. Tweety certainly belongs to no subset of birds in the aviary in which they mostly can't fly.

Pollock: One way of construing AI research on reasoning, which includes inheritance hierarchies, is saying it's aimed at understanding correct reasoning and building systems to carry out the reasoning.

Another way is thinking of work on defeasible inheritance hierarchies in a more limited way: as looking for a good way to process information. That is not the way I am viewing it. If we were interested in understanding reasoning and rationality, the question arises: what's the criterion for correctness for a system that does the kind of reasoning we want? The only possible criterion is what humans regard as correct reasoning. We don't have any other grasp on rationality. To say "our only criterion for correctness is the human standard" is not to say "build systems that do exactly what humans do." As Henry Kyburg just said we would like to improve upon that.

The human standard really comes in when evaluating conclusions from a reasoning system.

If the reasoning system comes up with conclusions that we regard as unreasonable, then that is criticism of the system. If it fails to come up with other conclusions we think are reasonable, that's a criticism of the system.

Coming up with conclusions the same way we do doesn't seem to be necessary. We can use intuitions about rationality and the correctness of conclusions to try to set up logical systems that describe various kinds of reasoning. In the deductive case we have been rather successful. Using our intuitions about correctness led to first order logic, formal semantics, logic, and so forth. We have not been as successful in defeasible reasoning. There are formal theories regarding defeasible reasoning such as circumscription or default logic, but none adequate. I think we don't have

reasonable semantics for defeasible reasoning and it's probably a mistake to even be looking for semantics at this point. We should be trying to figure out how defeasible reasoning works. We can worry about semantics once we understand what the target is.

I was told that I should say something about skeptical and credulous systems. Ron Loui thinks I am talking about a strawman. Some claim that using skeptical or credulous defeasible reasoning is a matter of convention. They are wrong. We want a skeptical reasoner. Simple epistemological intuition legislates that if you have equally good reasons for believing P and for believing not-P then you shouldn't believe either.

A lot of people urged that it's important that an AI system tells us something rather than nothing. Jon Doyle gave an example about a system that plans the company picnic and must choose between two parks. One park is better if it isn't going to rain and the other is better if it is going to rain. There is a fifty percent chance of rain. What do we want the system to do? It's reasonable that the system pick one at random. That's right, but that's not what the skeptical/credulous controversy in defeasible reasoning is about.

We have to distinguish between what philosophers call theoretical reasoning and practical reasoning. Theoretical reasoning is reasoning about what you believe. Practical reasoning is reasoning about what you do. The logics of these two kinds of reasoning are not the same. During theoretical reasoning, we want a skeptical reasoner. For example, we don't want the picnic planning system to randomly conclude it won't rain.

We want the system to decide that it can't tell whether it will rain, then go to a second stage of reasoning, practical reasoning based upon the theoretical reasoning, and infer that because it can't decide whether it will rain, it should randomly choose one of the two parks. Reasoning about practical matters is a two-stage process. You have to do some theoretical reasoning to get facts to plug into the practical reasoner.

Practical reasoning is governed by how much is at stake. When a system trying to determine what disease a patient has can't tell whether the patient has disease A or disease B, the system has equally good reason for each diagnosis. The question is "How is this patient to be treated?" We don't want the system to decide randomly which disease the patient has. We want the system to tell us, "I can't tell which disease the person has."

Different possibilities arise. Suppose that treating for one disease would not make the other disease worse. Suppose that each disease is of equal severity. It seems reasonable to pick one randomly to treat because that way you maximize expectations. Suppose instead, that treating for the wrong disease would be fatal. Surely in that case you shouldn't treat either. Suppose that one disease is more serious than the other. Obviously treat the serious one.

We want our theoretical reasoner to be skeptical and our practical reasoner to look at the output of the theoretical reasoner, tell when the theoretical reasoner can't draw a conclusion, then base its practical reasoning on that avowed ignorance.

Padgham: You have to do credulous reasoning to get options. When you have a skeptical reasoner you don't get any of the possible diseases.

Pollock: You will if you have the right kind of skeptical reasoning. For example, mine. When you have equally good reason for A as for B, my system will reach the conclusion "A or B."

Bacchus: What about a reasoner that gives you a grade as to how much you believe something?

Pollock: This presumably is a matter of having equally good reason for two things.

Bacchus: Right, but it'll say the grade is the same. In the practical component, how about deciding among alternatives based on how strong the belief is?

Pollock: If you had an equally good reason for each, that won't do anything in the practical component.

Bacchus: It comes down to the situation.

Pollock: Yes.

Etherington: In the credulous approach you don't just have A or B, you have (A or B) and (C or D). Right? The credulous reasoner gives you a set of different sets where possibilities hold, whereas a skeptical reasoner gives you a disjunction of all the possibilities.

Pollock: Mine would. Mine would give you (A and B) or (C and D): all the possible disjunctions.

Doyle: I can see how the examples you gave make a good case that neither skepticism nor credulity is good as a uniform policy. But how does that in itself argue for separating theoretical reasoning into two types of reasoning.

Pollock: The logic of theoretical reasoning is simply different from the logic of practical reasoning. I haven't argued for this, but it could be argued.

McCarty: Methodologically, would it be reasonable to build a skeptical reasoner with whatever degree of credulity you want built in the rules and information you give the reasoner?

The reasoner is skeptical about conclusions it draws from the information it has. Would it be reasonable to put rules in that would allow it to draw riskier conclusions about that information?

Pollock: You don't want to tamper with the degree of ignorance that the skeptical reasoner exhibits rather than its conclusions. Practical considerations will dictate how the system should behave independently of state space.

Nute: The topic, "Philosophy of the Future or More of Conventionalism's Curse?," reminded me of another question that had been posed to me recently: "Do you want George Bush or Michael Dukakis for President?" I think the answer is no. I suspect the answer to today's question is no.

There are people in the professional philosophy community who consider all philosophy to be conventionalism. If they're correct, then anything we do having to do with philosophy will be more of conventionalisms's curse. Any starting points are unjustified. Criteria of rationality may be impossible.

That is not the way I act, whether I have good reasons for acting one way or not.

What is Artificial Intelligence and what is its relation to philosophy? Is it unsupported conventionalism? Are there criteria for deciding between them? There are criteria, although maybe not criteria for selecting criteria.

In work I have been doing in different domains, not in areas of defeasible logic, but in traditional logic, temporal logic, deontic logic, and other areas, the basic approach starts with a problem. The problem motivates formalization. Typically our problems are conceptual problems: concepts, constructions, maybe linguistic constructions or ways that we see concepts fitting together.

In deontic logic we worry about the concepts of right and wrong, permitted, or obliged. We have difficulty seeing what conclusions to draw. We draw opposing conclusions with what appears to be the same information, like the same assumptions and starting points. Perhaps we can be led into internal contradictions, not just conflict among individuals.

We might call using logic "conceptual impotence." We have a conceptual framework for dealing with certain matters and we have relatively simple cases – the information isn't hard to understand – but we can't figure out what to believe or do. The philosophical problems that I've attacked with formalization dealt with this kind of conceptual impotence. How do you solve the problem? We start by looking at what agents do and trying to model that, to prevent them from being conceptually impotent in certain areas where we want to draw conclusions but can't.

It was said that we don't just want to model people; we would like our machines to do better. That's correct and also incorrect. When the machine does better, we have to recognize that it's better. We need criteria for deciding what's better. If a machine does something absolutely better than any of us, it is doing something none of us can understand or justify. Why should we trust that it's doing better?

We start with agents that can't arrive at endpoints from fairly simple starting points because these agents have internal conflicts. We have to reform that. We want systems that avoid conflicts and give answers in the hard cases. Then we apply some additional criteria.

If we find a nice theoretical underpinning, whether it's probabilistic or other semantics, that's excellent. But, we don't need that to start. When Frege or Russell were fiddling around with truth tables working on what we now know as "God's logic" I wonder if they were concerned that they wouldn't come up with soundness and completeness results for a long time. Probably not.

It's been suggested that Artificial Intelligence is a subarea of Cognitive Science. If Cognitive Science means something entirely empirical, developing reasonable, empirically verifiable, or testable theories about human cognition, I don't think Artificial Intelligence ought to be about that. It ought to be about understanding what intelligence is.

Intelligence is a notion that is subject to philosophical analysis. Intelligence and human intelligence aren't necessarily the same thing. We probably will be reforming our notion of intelligence, rational belief, or belief revision, to make it consistent or to fill in gaps. Certainly what Haugeland calls "good, old-fashioned Artificial Intelligence," seems closer to philosophical analysis than to empirical science. That probably will not be a well-received comment.

"Philosophy of the Future or More of Conventionalism's Curse?," I guess the answer isn't "no"; the answer probably is "yes." What we're doing probably fits well with philosophy of the future and probably is more conventionalism, whether it's a curse or not.

Thomason: I decided that it would be useful if I tried to talk as a computer scientist who knows something about philosophy. One question when assuming that Artificial Intelligence is either a cognitive or computer science is, what is a healthy relationship between philosophy and any science?

Even though at first people in these areas were confused because there were no-holds-barred problems ev-

erywhere, they isolated the clear problems (on which they could make progress) from the foundational problems. Perhaps a small group remains engaged in foundational problems and in close contact with the philosophers. But the main discipline ignores those problems, except in prefaces to books about scientific methodology. Any good philosopher of science can make a quantum physicist contradict himself in two minutes about the Measurement Problem. The problems are real but we can do the scientific work.

Philosophical logic is traditionally a part of philosophy although I think they are quite different. Philosophical logic is a technical area funded by the National Science Foundation and has pretensions of being a science. It has belonged to the philosophy department partly for historical reasons and partly because there are connections with their methodology. Logic was taught as an important tool in doing philosophy. Logic has been taken over by mathematicians to a large extent. A number of interesting problems remain in developing new logics and keeping the connection with logic and general-purpose reasoning rather than mathematical reasoning.

Philosophical logic might be transferred into computer science. Most people who would have gone into subjects in philosophy departments ten or fifteen years ago are going into computer science.

With regard to the relationship between AI and philosophy, I suggest we develop AI as a science, isolate foundational problems, and keep philosophy at arm's distance. If you want to do scientific work, keep a healthy distance from the no-holds-barred all-ends philosophical problems.

Maybe we should develop a strong relationship between philosophical logic and scientific work. As a philosophical logician I'm interested in new ideas for logic. The subject has become over- mathematicized and ought to move in new directions. Some of the most important new directions will come from computer science.

A lot of my interest in this work is in what new ideas we can get for logic. That's why I'm disappointed when computer scientists feel that they do better if they use a familiar logic. That bores me.

Keeping your tires on the road is important. Philosophers talk about the curse of conventionalism and neglect technology, possibly the most important aspect of computer science. Trying to make systems perform real tasks in industry, while working under practical pressures, is important to the field.

At some level, work in AI has to connect with work that is technologically useful. The connection doesn't have to be direct. If you compare the connections between sciences and theoretical work and sciences and applied work, you'll find the comparison more problematic in AI than in most subjects. That doesn't mean we shouldn't be sensitive to the needs of technology. For instance if people can't write a manual for systems suggested by their work, and they are impossible to use, that counts against them.

Let's be aware of technological constraints and use them to limit our options. A lot of conventionalist issues are left behind once we try to engage in scientific problems. A philosopher will always be able to show you problems lurking at a deeper level, but they are there in any science. We shouldn't worry as long as we find some agreement among ourselves and sense that we're making progress.

It occurred to me while I heard Jeff Horty's presentation that I like inheritance-based reasoning because one solution to the problem of relating theories to technology is introducing levels between theory and technology, particularly in an area where it is hard to connect the high end of theory to actual practice. You need various kinds of intermediate levels. It seems that the path-based stuff has new logical ideas, and gives an intermediate stage. We need to look at more intermediate stages of that kind.

Even from an a priori perspective, you can't say logic is entirely conventional. It's peculiar to choose logic the way you can choose anything. Saul Kripke gave a talk about a person who believed contradictions could be true. How can you have a rational debate with such a person? You may show him by reasoning that contradictions can't be true. He would say "Yes! I believe you now," and then utter a contradiction. You'll say, "Didn't I convince you that contradictions can't be true?" and he'll say "Yes."

At some level, choice of logic and rational debate about it becomes impossible. Maybe the problem goes deeper. One thing Kripke was trying to show with his talk was that there was something peculiar about conventionalism as applied to logic. He had a point.

The other point is that we shouldn't be imperialistic in AI. There's no Nobel Prize in AI. There's no grand prize for somebody who finds the solution that solves everything. The problems are so difficult and varied that we'll have healthy research trends if we limit the problems we work on.

We don't need to claim we have a system or theory that handles every reasoning task. In particular, I grew up thinking well of decision theory. I would want to use it if I were working on a problem like medical diagnosis.

But, that kind of reasoning will not handle every task, particularly cases where you try to coordinate people in societies. The kind of reasoning that you do in speech act understanding or the kind that you do when expecting someone to stop at a red light is not probabilistic reasoning. If it were, we would drive quite differently. The conventions that we establish for communications are not probabilistic ei-

ther. If you tell some people that you bought a bird and you would like them buy a bird cage, they'll get one with a top unless you tell them that you bought a penguin. That's the kind of application probabilistic approaches won't handle well because of the need for creating mutual structures, like mutual beliefs in communication. Default reasoning plays an important part here.

Another reason we can't solve everything with probabilistic reasoning is that in practice probabilistic reasoning depends on the construction of micro-worlds. We can only apply that kind of reasoning when we have solved what Savage called the "small worlds" problem and when from our wealth of knowledge and evidence, we have attached probabilities and utilities to variables we isolated in a particular problem situation.

AI can't ignore that problem by saying it belongs to another discipline. AI has to be reasoning-task complete with any problem; AI has to deal with any problem in robotics and not shove it off to another discipline. We can't bootstrap using probabilistic reasoning. We have to fall back on something more basic, qualitative.

An interesting consequence of this is the need to put things together. I was pleased that Horty and I did not look at Brachman's challenge of handling definitions that could be defeasible as a controversy between two schools, but instead we tried to put them together in a bigger system. A lot of insights came from that. I would be interested in a project putting probabilistic and defeasible reasoning in a single system.

Fetzer: I want to distill some of my remarks on the underlying problems of defeasible reasoning and suggest consequences related to remarks made by John Pollock and Henry Kyburg. Henry said his intuitions weren't very good. He looked at the benchmark problems and he had a sense of what should or should not be inferred in only ten or fifteen percent of the cases. There is a reason for that; these problems are not well defined.

The benchmark problems lack essential elements of well defined problems for which there are definite answers. The phrase "tends to be" appears in many places in the examples. This phrase is vague and ambiguous. We need to be aware of whether we're talking about causal tendencies, which might be deterministic or indeterministic, or relative frequencies. We need to be aware of what we mean when we use qualitative rather than quantitative locutions, and how they are related. All this is involved in unpacking the notion of probability.

The fundamental differences between causal relations and standard probability is that standard probabilities are symmetrical. If you have a probability from A to B you have a probability from B to A. Causal relations are nonsymmetrical. If you have a causal relation from A to B, sometimes, usually, but not always, you have a causal relation from B to A.

Be aware of what the purpose of inquiry is. What are you trying to do, explain or predict? I suggested some reasons why rules for assigning single cases to reference classes in the case of explanation may be different from rules in the case of prediction. If you're unclear of the purpose, you cannot answer a question that asks you to draw a conclusion or make an inference.

The cases that we're talking about technically qualify as singular predictive inference. It's important, in cases of singular predictive inference, to distinguish between aspects of the problem, between degrees of support and rules for acceptance.

Even in a simple case: "People usually wake before noon, Tom usually wakes before noon," and the question is, "Does Tom wake in the afternoon?," we have the suppressed premise about what morning we're talking about. We might add that Tom is asleep today. We might therefore modify the conclusion. The question is "Does Tom wake this afternoon on this particular occasion?"

More importantly, suppose we drew a double line between the premises and the conclusion. Then we're acknowledging that we're dealing with an inductive inference. The premises might be true and the conclusion still false. We have to ask how much evidential support do the premises convey on the conclusion. This could be indicated by a bracketed number. It would stand for "usually," i.e. usually the conclusion would be true if the premises were true. Once we can establish a measure of evidential support, is the evidential support strong enough to justify accepting the conclusion as true, or do we merely have a degree of evidential support that leaves us uncertain as how to proceed because the support is insufficient?

This relates to a point John Pollock was making. Especially when we talk about rules and acceptance, we have to distinguish between situations of pure science (drawing an inference), and applied science (arriving at a decision). The situations are very different. When we draw an inference we usually have unlimited time and can revise our decision, acquire and conduct new observations and experiments, and revise our thinking. When we make a decision, we limit time and our decision is often irrevocable.

As Pollock hinted, the seriousness of making mistakes must be factored in. We might have overwhelming evidence that something is the case. Yet the seriousness of being wrong might be so great that we are reluctant to accept it for the sake of making a decision.

I would encourage you, when you pursue problems of defeasible reasoning, to clarify exactly the nature of

the enterprise and to see if you can formulate rules of evidential support and conditions under which acceptance or rejection would be warranted. Frankly, those problems underlie all defeasible reasoning issues you confront.

Saturday Session III: Probabilistic Approaches.
Speakers: Fahiem Bacchus, Hector Geffner, Eric Neufeld, James Hawthorne.
Moderator: Judea Pearl.

Pearl: I call this slide the "skeptical chemist" in memory or remembrance of the celebrated skeptical chemist, Robert Boyle, who in the 17th century gave a death knell to alchemy.

I want to show two approaches: the skeptical chemist approach versus the great alchemist approach.

Etherington: Who are you calling the alchemists here?

Pearl: I'm not mentioning names.

I strongly believe that we have a common goal, to formalize convention. The question is only how we go about it.

I liken it to a cookie factory where everybody likes the cookies. We'd all like to have cookies, sell them, or enjoy them. The question is how. In our cookie shop we have a display window where the cookies named "reasoning conventions" are offered for customers. These cookies are partitioned: belief, action, and other things. You buy a cookie and enjoy it so much that you want to duplicate it.

Now you have an option. You can go into the alchemist's workshop and try to duplicate a cookie with what you know best. You know that it had saw dust and you know all about saw dust; it glues together so you know it has glue; it looks brown so you put in some brown paint. You're hoping to build your cookie with the things that you know best. That's one way of duplicating the convention: look at it, see how people communicate, try to duplicate it with a familiar mechanical facility. Let me propose a modest approach ...

Ginsberg: Wouldn't it also be fair to say that people who seem to take any sort of reasoning and recast in probabilistic terms ...?

Pearl: Who talked about probabilistics? I'm talking about cookies.

Ginsberg: People who recast things in probabilistic terms aren't falling into the same trap of using the things they're familiar with?

Pearl: Oh, of course.

Ginsberg: OK, so they're just as alchemical as the rest.

Pearl: Sawdust makes the wheat flour. Can we make a conjecture about why we chose one convention and not the other? We go back to basic experience. Basic experience mixed with water, which is statistical summarization, gives you dough, wheat flour. You can study all the chemistry or physics of wheat flours and not get a cookie mix.

Ginsberg: I want to know how you know to add statistical summarization as opposed to some other.

Pearl: That comes later.

The way we perceive the world is biased and contaminated with our linguistic heritage and culture. In the process of turning dough into something resembling the cookie, there are constraints we must obey. The chemistry of the flour is only one constraint. You have additional constraints. The most important one in my opinion is computational resources. Inherent in the problem is an additional biological constraint like a slow, short-term memory which results in local perception of the world. That's another mechanism which shapes our reasoning.

To form the dough into a sheet, reason qualitatively instead of numerically. Then comes the cookie cutter, which is a qualitative idealization of the probabilistic theory. Hopefully you will get cookies. I don't pretend that by this process alone you will get the cookie without additional baking facility. But at least you start with dough and not saw dust.

Etherington: This theory is only half-baked.

Pearl: Our distinguished panel will demonstrate a few attempts to shape this cookie cutter from continuous probability into a discrete cookie.

In qualitative physics, we abstract continuous quantities into landmark values. In probability theory, you only have three distinguished landmarked values: zero, one-half (neutral), and one. Zero and one are predicates for propositional logic. One-half gives you a neutral state of ignorance. Fahiem Bacchus will talk about semantics of probability where one-half, or, in general, greater than alpha or smaller than one minus alpha is the distinguished landmark value upon which

you abstract and reason. Ben Grosof will discuss reference classes. Eric Neufeld will show a tradition in qualitative physics, where you talk about the size of the derivatives of the critical quantities. He'll talk about the changes that occur in probabilities of propositions and the logic of those changes, like when the probability of B goes up or down if you know A. People in naive physics use an order of magnitude analysis: analysis of proximity to the landmark values. Hector Geffner will start with that topic. Hawthorne will talk about how qualitative probability can be derived from non-probabilistic assertions or axioms.

Grosof: Some of you have been hearing about reference classes. I'll be talking about an inheritance network which arises in proper probabilistic reasoning, as opposed to the approximation of probabilistic reasoning by zero-one probabilistic reasoning.

Here's a basic problem: Suppose the probability is ten percent that a car made by the Neptune Corporation is a lemon. What probability should we adopt, for whether a particular car, like a Triton station wagon, is a lemon in a specific situation? The axioms of probability leave this question open. The probability could be anywhere between 0 and 1. People in the probabilistic community use a traditional principle which says that to determine the probability of whether the car is a lemon use information from the class for which we have information, in this case the Neptune class. The answer would be .10. We can view this as a default conditional independence of the two events: the car being a lemon and the car being a specific case, given the event Neptune.

This conditional independence plays an important role in probabilistic reasoning. I propose that we view it as defeasible reasoning at the level of statements about probabilities as opposed to statements about zero-one truth values. Suppose we learn that the probability for a Neptune Triton model being a lemon is 5 percent better than for the general class of Neptunes. A refinement of this principle of looking for reference classes is that you should look for the most specific reference class among the potential reference classes you might use. Here Neptune and Triton is strictly more specific that Neptune, so we adopt the value from there. We can see that as having a second conditional independence of the lemon and the situation propositions given the Neptune and Triton proposition.

We need something more, namely, a principle of preference, like a priority. We can formulate this in terms of circumscription. Preference is prioritization in circumscription and makes things go neatly for this sort of case. Look at the blue arrows in this diagram. This is similar in topology to the case of inheritance that everyone agrees about, namely where the subsumption between the classes which makes for specificity is a strict link. In the car situation we have a strict link from the two classes. The more specific path, .05, preempts, or dominates.

This is a simple case that everybody agrees about, just as there is a simple case that everybody agrees about in our community.

Things can get more complicated. I bring you more problems than solutions.

Suppose we have information about cars made on Mondays. Those of you familiar with the auto industry know that absenteeism and alcoholism are rank and that cars made on Mondays tend not be made as well as cars in the middle of the week. We might have a value for that, maybe .20 are lemons out of all cars made in the United States. The question is, should we inherit from the Neptune class, the Neptune/Triton class, or the Monday class? There is no strict subsumption or domination. We have multiple inheritance or multiple reference classes.

How should we combine or regard domination and conflict when we have bounds on probabilities?

Should there be priority principles similar to specificity for reconciling multiple inheritance issues? Should we have rules about doing convolution or compromise operations in such circumstances, which, given the fact that we're working with numbers, makes more sense than in categorical situations?

Suppose that a link is probabilistic as opposed to certain. That gets you into more problems.

It seems there are similarities between inheritance and probabilistic reasoning, and inheritance and zero-one reasoning. The problems posed by inheritance in the probabilistic realm offer pause to those who look to the realm of probability as a bedrock of underlying semantics, as a scheme of approximation of zero-one logic.

How should we interleave probabilistic reasoning at a level of finer representation of information and reasoning than at zero-one level? Utility information or computational heuristic principles might lead us to compile the same propositions from the realm of probabilities to the realm of zero-one. At what point should we jump? Delay commitment? Maybe that's too expensive. It seems that the relationship between inheritance and underlying probabilistic information has to be dealt with.

Poole: Don't some tell where general knowledge is more applicable than specific knowledge? We don't always use the more specific. Is there some way of doing that?

Grosof: Henry Kyburg, and Ron Loui in his thesis, discuss principles for dealing with complex cases and

the kind of meta-information which furnishes principles for defeasibly selecting reference classes. What is the statistical basis for the probabilistic information you use? Typically, a bound is imposed on numbers. Were these bounds based on a lot of evidence? If you had a subclass, was it based on a subsample of the the general class or was it a different sample entirely? Was there overlap? Did you have a small sample regardless of its intersection with the other sample you considered? These problems need to be wrestled with.

Shastri: There's no strictly right answer. One way has been maximum entropy: combine information from different sources of evidence. It was applied in a system I worked on.

Grosof: Maximum entropy is, in effect, another non-monotonic principle of augmenting monotonic implications of your probabilistic information. It has different properties. In many cases it produces default conditional independences, but also imposes a uniformity assumption. You end up with point numbers, not bounds. It offers provocative competition, but I am dubious that it offers panacea.

Bacchus: This inheritance system will be presented at IJCAI this year. It takes a statistical look at defeasible inference. In this system you have two types of links: strict and statistical. We take a particular interpretation of the defeasible links: statistical majority. The statistical majority can be anything as long as it's more than 50 percent. You could have a higher threshold value. With this we have the subset preference which here is its own system. You can solve a number of different problems with it but I won't bother to go into that now.

With this system, you have restricted chaining. A path in this net can only contain one defeasible link because statistical links don't chain no matter how high your threshold is. You can chain as many strict links as you want.

This system is not a general system for doing all defeasible inference. The information in the system has clear semantics. We take a definite interpretation of the defeasible links and see what can we generate from such a limited system. In this way we can justify what we can generate. This addresses a question raised by Ron Loui last night: How do we know if we've drawn the net correctly? Should we put another link in? This is the question of meaning. We need to know what our nets mean to know whether a piece of information is present in the net. This limited approach gives us that information since we have a particular interpretation.

This statistical approach is being applied in some areas. Recent work by Weber and Tenenberg deals with the frame problem, in particular the qualification problem.

In this case we have some representation of knowledge of statistics about what happens when you perform certain actions; actions which are not deterministic. You have statistics about them and, in this case, the statistics are essentially the time points. You can represent this in the manner where X represents a time point. For all of the time points where you have an ignition advance, a certain proportion of them cause the car to start in the time point X+1, using discrete time. You can make assertions like: Given that you just had an ignition advance among the class of all ignition advances, the proportion of car starts in the time right after is greater than the proportion in the reference class of ignition advances where it's cold that day. You can make assertions about different relationships. Essentially these conditionals correspond to qualifying your actions on more conditions. A particular starting instance can be a member of many different reference classes which corresponds to the qualification problem. How do you know what to take into consideration? Now, if this value is very high then you've already taken into account the things which could cause the car not to start. In other words, you don't have to consider the potato-in-the-tailpipe problem. You don't have to qualify on that because the probability is very low.

My thesis developed a general logic for representing and reasoning about statistical information. It's an extension of first order logic where we have a probability distribution over the domain of discourse. Denotations of open formulas represent the set of satisfied instances. That's a common interpretation. Bird(X) would be the set of birds. In this logic you can take open formulas and form numeric terms which represent the probability of sets because you have a probability distribution over the domain of discourse. The probability of the set of birds doesn't make much sense, but when you conditionalize it you get interesting information. You could have the proportion of birds that can fly. You can represent a lot of different information in this way.

It has a sound and complete proof theory relative to Tarskian models with a probability distribution. It's not an esoteric model. You have an inductive way of relating particular sentences about particular individuals to underlying reference classes. If you want to know about Tweety flying and you know it's a bird, you can make it variable and relate it to a statistic in the logic about this term. You can reason in the logic about the value of this term. It can deal with N-place relations, so you can deal with things like inheritable relations.

Where do we stand as far as the statistical approach to defeasible reasoning? We have some formal tools

for representing and reasoning about these reference classes. We have a few applications. We still have important problems about the choice of reference class, although I don't agree with Ben Grosof that we need to regress to another formalism to decide on the feasibility of our policies choosing reference classes. We can come up with reasons for our policies that don't infinitely regress.

Geffner: You said you cannot chain inheritance networks' defeasible links. If I give you a strict link followed by a defeasible link, can you do that type of chaining?

Bacchus: Yes. If you have a defeasible link in the middle you can chain onwards.

Geffner: If you have a strict link and a defeasible link that follows, you need to make some conditional independence assumptions to chain.

Bacchus: If you have a strict link and then a defeasible link, you can go through any chain as long as you only have one defeasible link. The place of the defeasible link doesn't matter.

Geffner: From one point of view, doing that type of inference invokes a conditional independence assumption.

Bacchus: No, you invoke a different reference class.

Poole: How about "emu's are birds and birds fly"?

Bacchus: If you know that an individual is an emu, go via strict links to bird. Use bird, not emu, as the reference class.

Pearl: The problem is the subclass relationship. Penguins are birds. Birds fly. Can you attribute to penguin the property of bird in your system?

Bacchus: No.

Geffner: So why would that chain?

Bacchus: There's no defeasible reasoning about properties. You defeasibly reason about individuals and the properties these individuals have. You don't defeasibly reason about relationships between properties.

Etherington: If "Opus is a penguin, penguins are birds, and birds typically fly," you would get "Opus typically flies"?

Bacchus: Right, unless you have statistics about penguins not flying. That would form a more specific reference class.

Grosof: What about defeasible inference about the probability for a class? Suppose all you know is "robins are birds" and the probability of birds flying. Would you be able to infer, ceteris paribus, to use the same probability of flying for robins that you have for birds?

Bacchus: No. Essentially you choose the reference class that you have statistics for. If you know a particular bird is a robin, you don't have any statistics for that, so you have to go up to the next reference class.

Grosof: In other words, you can only get reasoning at the level of the individuals, not at the class "robin."

Bacchus: Right. But you do have some reasoning. For example if you know "most birds fly" and "penguins don't fly," you can infer that most birds are not penguins.

Thomason: I don't see how you can get relational reasoning on your system because such reasoning requires two ISA links which might be defeasible. That is if you have something like "X is related to Y," generically, but then you have "X1 is an X, Y1 is a Y," you essentially have two defeasible links. Even the simplest relational reasoning involve paths with two ISA links.

Bacchus: Are you talking about the inheritable relations?

Thomason: Yes.

Bacchus: The inheritable relationships are not done within this inheritance reasoner. That's done in the more general logic. You condition on more complicated reference classes. In the general logic you can condition on whatever you want because you have conjunction and other things. In inheritance nets you don't have conjunctions. You have a limited set that you can represent.

Thomason: What applications do you have in mind for this? I think there might be problems in applying it because you can't chain. How easy would it be to get people to use this system in real applications?

Bacchus: One application is work on discourse analysis. A Waterloo professor, Robin Cohen, has done this work. She doesn't want to make more than one defeasible inference in her application.

Thomason: What's the task?

Bacchus: The task is understanding some stories. You can get yourself into trouble if you do more than one inference. You're conservative, you do one. Then you check, because you have some feedback from the user as well, I understand.

Poole: Is it possible to have two sorts of information? If you have a high value for the input value threshold but half of your threshold for the output values, you can only do a few chains. In some sense, like epsilon semantics and the lottery paradox a lot of "nears" get to be "transitively near." You end saying two things are "near," and collapse into the lottery paradox.

Bacchus: You could chain more than once but that sort of chaining involves additional independence assumptions. If you have these, your probability becomes "C squared." That's a possible extension. Here we aim to be clear in what we're doing and see what we get.

Neufeld: I want to describe a modest probabilistic system that we can understand, compare with the world, and maybe verify. Out of the many epistemological issues, we can at least see that certain necessary conditions are satisfied.

I want to capture a certain fragment of reasoning. I believe that in artificial intelligence, all problems are about reasoning with incomplete information. Many years ago the claim was made that the theory of probability is epistemologically inadequate for reasoning under uncertainty. Having tried logic and not being satisfied with the results, I chose to use probability as a tool, with three simple ideas. I want to get by without numerical probability distributions (they seem hard to get) with a graphic formalism.

You can attach a number of different probabilistic meanings to these links. Judea Pearl outlined many different interpretations of these links, mine being change in probability. I have a strict link that says on learning A, B becomes certain. Furthermore, I assume that the probability strictly increases. Secondly, this is a defeasible link, which means the probability of B increases upon learning A. If I draw a cross through that link that means the probability of not-B increases, etc. This is a system that reasons about shifts in beliefs rather than which beliefs to accept.

The topology of the graph encodes the belief that there is an underlying probability distribution that we may not know. If it is factorable, in the same way influence diagrams are, that is you can state the probability as the product of the conditional probabilities of any node conditioned on its parents, then the topology gives us knowledge about independence, tons of knowledge about relevance. A redundant link doesn't exist in this system.

The third thing is: condition on all observations. My query to the system is: Is B more likely having observed A?

The result that made this work exciting and different from things I've seen, is that it allows modest chaining. If you have a link from A to B and another link from B to C and C is conditionally independent of A given any possible outcome of B, then A favors C. I chose "favors" for this relationship because it was uncontroversial, as opposed to "confirms" or "supports" which appear in other work on this topic.

I'll do the "birds fly example." What's interesting is not that it gets the correct answers. The default reasoning systems get the correct answers but they also get wrong answers. This system gives you that emus don't fly. It doesn't give you "emus fly." You can't chain from "emu" to "bird" to "fly" because "fly" is not conditionally independent of the outcome of bird, if you know emu. There are other things you can infer. "Emu" goes to "not flying." "Not flying" goes to "not airborne." I don't know of other systems that do. Furthermore, you can infer "emus have feathers," "feathered things fly" and so forth.

What's also nice is that the system is reversible. If "emu" conditions "bird" or increases belief in "bird," the reverse is true. Run the system backwards and a diagnosis pops out. Your observing feathers increases your belief in bird. If you relabel the graph with symptoms of a disease and with the name of the disease, you get a qualitative diagnosis. It seems to be a unifying formalism for reasoning in commonsense domains. It doesn't give you a different set of results or methodology when you're reasoning forward and reasoning backwards.

Furthermore, the topology of the graph can be verified. You can do a statistical experiment in the world that determines whether airborne is independent of bird once you know whether it flies. To the best of my knowledge nobody has defined any necessary or sufficient criteria for defining defaults, cancellation laws, abnormality predicates, or anything like that.

Ginsberg: It is non-intuitive to me that non-flying things should not be airborne.

Neufeld: It's probabilistically a fact of life. If "flying things tend to be airborne," you shift your belief

that way, then to be consistent with the laws of probability you have to believe that non-flying things shift your belief downwards in airborne.

Audience: What? Are you sure?

Pearl: We're talking about shifts.

Neufeld: We're talking about shifts in belief.

Ginsberg: But there are all sorts of constants, like the number of airborne objects in the universe. If airborne varies, then when I know that flying things tend to be airborne, non-flying things will be like they always were.

Neufeld: Anomalous cases arise where, for example, the belief shifts downwards from a high value to still a quite high value. But that shift downwards occurs.

Ginsberg: The shift downward only occurs because the probability of airborne is kept constant when I learn that flying things are airborne. In reality when I learn "flying things are airborne," the probability of random objects being airborne goes up just enough to leave probability that non-flying things are airborne alone. When I haven't learned anything about non-flying things, I shouldn't think I have because I will be in danger of concluding that ravens are black when I see a red rose.

Neufeld: I don't think this is the confirmation paradox. I'm saying the system allows us the contrapositive. If you see a non-flying thing you will conclude it's airborne, not a red raven.

Ginsberg: I believe in contraposition, but I think this property of taking inverses is anomalous.

Thomason: Contraposition is "if bird then flies," and then "not flies then not bird." You said "if flies then airborne" then you get "if not fly then not airborne." That's not contraposition.

Neufeld: You get both by the way.

Grosof: This is an argument against the conditional independence probability assumptions. Matt Ginsberg says his intuition is violated. I say that your intuition is violated by the conditional independence assumptions which underlie the factorability of this representation. What Eric Neufeld says is almost equivalent to saying that the likelihood ratio ...

Neufeld: Yes.

Grosof: ... When you learn "fly" is positive, if the likelihood ratio is greater than one, then when you learn that fly is negative in terms of its impact on airborne, the likelihood ratio has got to be less than one. A flaw must be in the independence assumptions which forced this mathematical result.

Ginsberg: This is an attempt to model my intuition in a mathematical way that has as a result a rule I find totally non-intuitive.

Pearl: Could you explain what you find non-intuitive?

Ginsberg: I find learning that no-flying things are not airborne non-intuitive. A rule says flying things are airborne. I learn "Tom cannot fly" and conclude from this "Tom is not airborne." That is non-intuitive.

Pearl: Does anybody share this conflict?

Etherington: It's the choice of words. If you had a rule which said "notebooks tend to be on tables" and you learn that the thing I'm talking about is not a notebook, then you lower your confidence that it's on the table?

Pearl: That's right. You lower the confidence by epsilon but don't conclude. Counterintuitive examples will be more revealing the smaller you make the epsilon. But epsilon is still there. That will not be shocking.

Neufeld: It's a mathematical result. You use information from the world to analyze your intuition.

Ginsberg: I don't question the mathematics.

Neufeld: How does the system handle paradoxes? The multiple extension paradox? It automatically handles situations with iterated exceptions. You can have a category of "flemus," flying emus, that fly because they're flemus. The system makes this inference on the basis of the narrowest reference class, by conditioning on knowledge as opposed to cancellation laws that have to be implemented in addition to your defaults, or whatever.

With the lottery paradox, in a default system you can reason from "bird" to "fly" and then, using the contrapositive, "not emu." That inference violates the independence knowledge in my system. I believe that if I see a bird it will increase my belief that it is an emu. (Once again I'm talking about shifts in belief) It seems reasonable to believe that a bird is some kind of

a bird. In this qualitative lottery paradox of defaults you believe that a bird is no kind of bird if every class is slightly exceptional.

In the Simpson paradox at first it irritated me that the system would not support, confirm or favor C if all it knew was A or B. But analysis reveals that you can consistently believe that A increases belief in C, B increases belief in C, but knowing only the disjunction does not increase belief in C.

Ginsberg: This goes against a general, well-accepted principle in non-monotonic reasoning.

Neufeld: Yes I know. But I'm not doing this for non-monotonic reasoning.

Pearl: I think when we discuss convention, Simpson's paradox is a good example to focus on because it conflicts with probability. Probability theory allows you to conclude "not-C" as well as "C." Non-monotonic logic compels you to believe "C." Most conventional discourse concludes "C." From a probabilistic viewpoint we have to build additional declarations on top of the basic probabilistic axioms to implement this convention.

Neufeld: This does not conflict with the axioms of probability theory. If A and B are mutually exclusive then it would. If A and B overlap and I rewrite the links as the probabilistic inequalities everything is consistent. I cite a paper written by Chung in 1942. I conclude that reasoning by cases is wrong in default logic.

Etherington: You can't reason by cases in default logic, right?

Neufeld: Depends on how you represent it. If I use Reiter's system and say "MA implies C, A implies C, MB implies C, B implies C," and I learn "A or B" then I can reason by cases and conclude "C." It's possible in Reiter's system to do "A : MC."

Etherington: That's the representation of the case you're talking about.

Neufeld: Reiter's system allows this representational distinction. If you have default logic, reasoning by cases won't work.

There are many examples of Simpson's paradox in sampling theory.

Audience: An intuitive example?

Neufeld: Suppose I randomly sample all people who have ever traveled on an airplane. The likelihood is that I pick somebody who flew on a holiday. But for each particular planeload of people the likelihood is that they are on a business trip. I take my sample. I get different conclusions if I take them from the disjunction (of planeloads) or from one of the disjuncts (a particular planeload).

Ginsberg: Give a compelling example with just two disjuncts.

Pearl: I can make it more compelling if you want.

Neufeld: With just two? I can't give you a compelling one for just two.

Pearl: I can. If you marry Ann you will be happy. If you marry Barbara you will be happy. But God forbid if you marry both!

Neufeld: That's horrible.

Pearl: That's the point.

Neufeld: I can't give a compelling one for A or B.

Pearl: That's another convention we should focus on.

Kyburg: The classical example comes from Nancy Cartwright. The worry is affirmative action at Berkeley. Each department reported that the chance of a woman being admitted was better than the chance of a man being admitted. The dean objected on the grounds that among all people who applied, the disjunction of the people in various departments, the chance of a woman being admitted was less than the chance of a man being admitted.

Pearl: You cannot give compelling examples with two disjuncts because it doesn't hold for strict probability. It's only for semantics of the majority.

Neufeld: I can get a majority/minority situation. But it's close to the midpoint.

Geffner: You present semantics that you call support or favoring. You draw conclusions by default. How do you see the semantics of defaults?

Neufeld: I see default reasoning as trying to capture this approximate inference. This system seems to do the kind of approximate inference that people want: to chain a little but not all the time because they don't want to make "emus fly." This seems to give a simple, statistically motivated reason for making those inferences you want.

Ginsberg: People do case analysis all the time.

Geffner: If Tweety is a bird, do you conclude it's a penguin?

Neufeld: I don't conclude it's a penguin. I increase my belief that it's a member of a subclass. I conclude slightly differently than the people using defaults.

Etherington: If the knowledge about the world is "penguins are birds, penguins don't fly, birds do fly and Tweety is a bird," you conclude "Tweety is a penguin" because you don't know of any other kind of bird?

Neufeld: I don't conclude that Tweety is a penguin. I increase my belief that Tweety may be a penguin. I know that Tweety is a member of a subclass of birds. This example is also artificial. I chose penguins to be sort of relevant.

Etherington: Is your highest belief Tweety flies or doesn't fly? Which belief is stronger?

Neufeld: It flies, obviously. You only know it's a bird.

Pearl: Only shift belief. Not accept.

Etherington: You cannot say which one is higher?

Neufeld: In this particular case, nothing is inconsistent about believing penguin and fly. You won't get into trouble building bird cages because you'll put tops on them. You're likely to be right.

I wanted to say something about probabilistic extensions, but it's too late.

Geffner: I'm going to talk about a relationship that can be established between probability and default reasoning. I'm trying to use probability to do what people who don't use probability try to do. Perhaps we can learn something from probability for non-probabilistic reasoners.

The approach is motivated by similarities found between inheritance networks and probabilistic networks that people use in evidential reasoning. If you impose the assumption "links encode high conditional probability statements" supplemented by conditional assumptions, you can capture the kind of inference that inheritance networks try to capture.

We need to divide the facts in the theory, T, into two different sets: the background set, denoted by K, and the evidence set, denoted by E. Usually the background set encodes generic knowledge about the domain in the form of defeasible and indefeasible rules, while the evidence set encodes facts of interest about the situation.

The semantics define the notion of probabilistic entailment in which we assert a proposition H from T: any probabilistic distribution that assigns unit probability to inferences in the background context, assigns a conditional probability of H given E (evidence that will be in T). E is as close to 1 as we want if the distribution makes the conditional probability associated with the defaults in K sufficiently high.

This is essentially the same type of semantics developed by Ernest Adams in 1966 to define his logic of conditionals. With this probabilistic semantic characterization you arrive at a sound and complete axiomatization; to evaluate whether you can assert H from T according to this criterion, you only have to test whether you can construct the derivation for H given E in the logic defined by these rules.

My rules two to five are essentially the same as those used in most conditional logics. Their object-level operators differ from the variable conditional that Jim Delgrande talked about; we use meta-level here. For convenience, we encode in something like McCarthy's abnormality style. That is, a default "A then B" can mean a member of A is likely to be a member of class B. Negative abnormalities are used as switches to turn the default on and off. The first variety of rules, rules 1 to 5, are validated by these semantics, this type of logic. If you forget the last rule you can get certain simple examples right, like the penguin triangle without appealing to notions like specificity.

In general, the logic from the first five rules is too conservative. It's not fully non-monotonic. It's non-monotonic in the evidence set: as you increase the evidence, you may lose some entailed consequences. It's monotonic in K, in the sentences in the background. Certain simple conclusions like "red birds fly" with the default "birds fly" are not sanctioned by these rules. You need assumptions of conditional independence; that's essentially the last rule, an attempt to supplement the rules.

The first rule permits one to conclude that a given default holds only if you have the antecedent. The last rule says the assumptions associated with a default hold if you can prove the additional evidence is irrelevant. If you show that "redness" is an irrelevant property regarding flying in "red birds fly," you sanc-

tion that "birds fly," and chain as usual.

Etherington: Are you able to show "red" is irrelevant to "fly?"

Geffner: Here is a simple example. For any argument that refutes the "irrelevant" assumption, you postulate a number of assumptions. That amounts to assuming certain abnormalities, permitting some defaults. Evidence is consistent with this set of assumptions. If you can prove it's abnormal, then to conclude it's still abnormal, you must show that this set of assumptions doesn't apply to this class. For every argument that shows the abnormality of the default, you can refute it, the assumptions are defeated. Then you can assume irrelevance.

What have we learned from this approach?

Ginsberg: I want a "yes" or "no" answer. Does this tell you that Nixon is politically motivated?

Geffner: It tells you by reasoning by cases. One of the rules that is criticized gets that Nixon is politically motivated.

I was happy because I thought it did not sanction the conclusion. But it does. That's a conclusion I would rather not have.

The main question was: if you use a probability interpretation of defaults, what can you tell to people who don't work with probabilities? This probabilistic interpretation provides a meaningful vocabulary to discuss problems that arise in default reasoning. Certain notions based on specificity can be explained in terms of these rules and the distinction between background and evidence.

These features are not particular to probabilistic interpretations. Many model-theoretic accounts, like Jim Delgrande's, Danny Lehmann's, Makinson's, and many other people's who are defining preferential types of entailment, validate the same core set of rules. But they take a different stance with respect to irrelevance.

That is the main problem that all these approaches have. The model-theoretic and probabilistic approaches that regard defaults as high-conditional probability assertions not only have these features, but the same problems.

The main problem is that a notion of irrelevance can be justified well. I presented the rule that makes sense with the test examples. Usually this rule gives reasonable results, but it lacks a clear foundation. I won't be able to convince somebody who doesn't share my intuitions (which also change with time) of the assumptions that underlie irrelevance. This is also a problem of Jim Delgrande's and Danny Lehmann's approach. They share the problem of justifying an extension of the core set of rules to use assumptions about conditional independence.

The second problem is the treatment of causality and contraposition. I want contraposition in this system. You use contraposition for refuting conclusions, but not for building proofs. It has a conservative stance.

Fetzer: You may want contraposition, but not for all kinds of relations; it won't do for causal relations. For example, striking a match causes a match to light, but not lighting a match doesn't cause not striking it.

Pearl: But you accept the belief.

Fetzer: It permits an inference, but it's not a causal relation.

Geffner: That's true. But in terms of acceptance, what is a legitimate conclusion? If the match doesn't light, you probably want to infer ...

Fetzer: You mean modus tollens is proper, but that's different from contraposition.

Geffner: I should say more.

Fetzer: The contraposed causal claim is false. The original is true as a causal claim. Modus ponens operates on that conditional the way it operates on other conditionals. But contraposing across causal contexts is fallacious.

Geffner: I'm not saying from one causal relation I get another causal relation. (This is an example I want to show you.) As in other accounts, in the account of irrelevance I use, there are problems. When I say "causality," I am talking about simple issues that arise in simple nets, not problems like the Yale Shooting. Two different inheritance nets can be encoded as logically equivalent; however applying different intuitions, you want to conclude different things. Logically they are the same: a link in one net is the contraposition of the corresponding link in the other. This is probably Horty's point.

The two issues that need to be addressed are irrelevance and causality. My belief is that probability will say little about the solution to these two problems. Maybe model-theoretic approaches can clarify these two issues. We use probability to get intuitions – we understand and can explain probabilities – but must look for a solution elsewhere.

Etherington: Is there something that probabilities can't make sense of?

Geffner: You want to hear an answer to that question?

Pearl: What's the question?

Etherington: He said probability would say little about something and I was confirming that he actually said that.

Geffner: Probability has little to say about the solution, but probability says something.

Pearl: Just as chemistry says little about the baking of cookies, you still have to know about the chemistry.

Geffner: The vocabulary of this type of interpretation of defaults is its main contribution. But the vocabulary is not a solution. At least you can identify the problems in terms of probability. In a different framework, you have a different space of problems.

Pearl: I would like to re-emphasize that starting from completely different viewpoints, Lehman, Makinson, and Maggidor rediscovered the same core of axioms, rules one to five.

Geffner: It's a way to encode defaults. Essentially by defining a preferential entailment relation and models, you get rules one to five and disjunction. Based on preferential models these rules will be valid. The problem is how to validate something like the irrelevance assumption we introduced. The problem is that we are proposing a solution that is ad hoc.

Ginsberg: What you say is true and it speaks well for epsilon semantics. Epsilon semantics gives the same basic semantics as Makinson's and Lehman's investigation of what non-monotonic inference is. It's also bad because Makinson and Lehmann realized that the current action in non-monotonic reasoning is in finding what to add.

Pearl: If rules one to five are the conservative core, I call this the brave ship.

Ginsberg: Okay. But how much will probability say about that addition?

Geffner: Actually, we are arriving at some interstting results, but in model-theoretic terms.

Ginsberg: But your approach seems to be profitable.

Geffner: Yes, it is a suitable language. Doing a model-theoretic approach, we get the same thing. We are not getting any new insights into meaning. Model-theoretic approach can validate the solution, can prove consistency, completeness, and other things, but we like epsilon semantics because it's a method we can understand.

Saturday Provocation: Specificity is The Wrong Generalization.
Provocateur: John Pollock.

Pollock: I want to argue that defeasible reasoning in defeasible inheritance hierarchies has a more complicated logical structure than has often been appreciated, and that specificity defeaters represent an incorrect generalization of that logical structure.

How do I think defeasible inheritance hierarchies work? In a general theory of defeasible reasoning, reasoning is viewed in terms of reasons and arguments. Some reasons are subject to defeat. There are two kinds of defeaters: rebutting defeaters, which attack the conclusions of a single inference step, and undercutting defeaters which attack the connection instead of the conclusion. This is roughly the difference between normal and non-normal defaults.

Here is an example of an undercutting defeater. "X looks red" is a prima facie reason, a defeasible reason, for "X is red". "X is illuminated by red light" is clearly a defeater. If you knew that X is illuminated by red light, you wouldn't draw the conclusion that it is red. But notice this doesn't give a reason for "X is red". So it's a defeater that is not a rebutting defeater, but is an undercutting defeater; it attacks the connection between looking red and being red.

I have wielded this general framework of prima facie reasoning and defeaters in epistemology for more years than I care to think about. The key to giving a correct account of a particular kind of defeasible reasoning in this framework lies in getting the undercutter defeaters right. That's basically what gives the flavor to any particular kind of defeasible reasoning.

I am going to look at reasoning that's involved in inheritance hierarchies. Here the reasoning proceeds in terms of this prima facie reason: "most A's are B's" and "this is an A" is a prima facie reason for "this is a B". This is a non-numerical version of statistical syllogism. It's often supposed that all defeasible reasoning is basically just an instance of this statistical syllogism. This is wrong. You can't do a lot of cases this way. For example, before you can get the statistical information needed for the statistical syllogism, you have to acquire some information about the world,

and often the only way to do that is perceptually; and perceptual reasoning is defeasible. I gave you an example right here. So you must have other kinds of defeasible reasons besides statistical syllogism.

Turning specifically to the statistical syllogism, the simplest kind of undercutting defeater is the subset defeater. Suppose we know "this is an A" and "most A's are B's." If you also know "this is a C as well as an A," and "most things that are both A and C are B's" is false, then that defeats the inference. Various people have noted these on various occasions.

It has frequently been asserted that a principle called specificity defeat is involved in defeasible reasoning in general. This tells you that if P is a prima facie reason for Q and if P and R is a prima facie reason for not-Q, then given P and R, we should infer not-Q rather than Q, allowing the more specific prima facie reason to take precedence.

This is a generalization of and subsumes subset defeaters.

This is false on two grounds. First, the first conjunct of the subset defeater is not "most A's and C's are non-B's" but rather that "most A's and C's are B's" is false. You don't need to know that most of them aren't C's, just that most of them are C's. Of course knowing that most of them are non-B's entails this also. That becomes a special case of subset defeater.

You need the more general defeater to reason correctly. The principle of specificity defeat doesn't subsume the general principle of subset defeat. If anything, it would only subsume the special case, when most A's and C's are non-B's.

Pearl: This is a narrow definition of specificity defeat. I thought Ron Loui would include that among his definition of specificity defeaters.

Loui: Let's focus on Don Nute's system. Nute's system represents A and C as an interferer for B rather than a supporting reason for not-B. Logical strength instead of subset is his notion of specificity. A lot of systems can do this.

Ginsberg: Any abnormality-based system can do this. You just call it abnormal.

Pollock: How do you get the abnormality? From the statistical information?

Ginsberg: The point is it's possible to say the inference is blocked. It's not even awkward.

Pollock: I'm observing that you can get this directly out of the statistical information.

Ginsberg: But the abnormality-based systems aren't interested in statistical information. They expect to be presented with the fact that it's not so.

Pollock: Secondly, you need the stronger version of subset defeateres to make statistical reasoning work right. This is not captured by specificity defeaters. The prima facie reason provided by the statistical syllogism has two conjuncts: "this is an A" (R1) and "most A's are B's" (R2). The special case of the defeater also has two conjuncts, that is "this is an A and C" (D1) and "most A's and C's are non-B's" (D2). Although one part entails its counterpart, the other doesn't. This is an essential ingredient in the reasoning.

If you are only interested in inheritance hierarchies, not in the reasoning per se, then you can leave that out, take it for granted.

Konolige: Aren't you talking about specificity of evidence?

Pollock: I am talking about both. In inheritance hierarchies, where they can suppress this, they get entailment. But when generalizing to other cases of defeasible reasoning that aren't necessarily a matter of statistical syllogism, they say they're generalizing what goes on in the statistical syllogism. I am pointing out that their formalizations of specificity defeat don't capture what goes on in the statistical syllogism.

Nute: Some of *them* make this generalization.

Pollock: I guess I'm attacking Ron Loui's view.

Etherington: Why would you expect D2 to entail R2?

Pollock: No, the conjunction of D1 and D2 doesn't entail R2.

Etherington: Why would you expect it to?

Pollock: You would have to if you want a specificity defeater. The conjunction of D1 and D2 needs to entail both R1 and R2.

Thomason: That's not what we mean by specificity.

Ginsberg: Right, that's not what we mean by specificity.

Pollock: That's what Ron Loui means by specificity.

Etherington: Ron Loui would mean that the antecedent of one entails the other antecedent and the consequent of one entails the negation of the other consequent.

Pollock: But the antecedent doesn't entail the other antecedent.

Etherington: "A and C" entails A.

Pollock: The reason is the whole conjunction.

Ginsberg: You're saying that the antecedent is the entirety (the statistics and the membership). But we don't need the statistics.

Pollock: Not everybody uses specificity defeaters to talk about reasoning in general. Some use specificity strictly in the context of inheritance hierarchies. Some do it more generally.

Poole: We know how to write rule defeaters and their conclusions by writing explicit cancellation axioms. I see specificity saying that there's more information than these explicit cancellations; the implicit cancellations are already here. If you represent it your way, your priority rules have to be explicit. You're not getting out any implicit knowledge. At your level, it's not specificity, not something you can say is derived.

Pollock: Are we agreeing or disagreeing?

Poole: I can see having one level with all of the cancellation rules. Then you can remove a few things that are implicit.

Pollock: How is that related?

Poole: You want to say that if you know A and C, that defeats the first rule.

Pollock: Yes, but you also have the conjunction of D1 (A and C) and D2 (Most A's and C's are non-B's) to get the defeat.

Poole: I can translate that into abnormality.

Pollock: I have some more offensive things I want to say. We can come back.

In inductive reasoning in philosophy, you need a projectibility constraint. Try to endorse a general principle of induction, which says a sample of A's, all which are B's, can be a prima facie reason for all A's are B's without any constraints on what A and B are. You get difficulties.

Audience: Nelson Goodman showed that a long time ago.

Pollock: I want to point out that the same thing happens for the statistical syllogism. I don't have time to go through the argument but I will show you the conclusion.

Assuming the following principle of collective defeat, basically I have a skeptical system. If you get prima facie reasons whose conclusions conflict, collectively defeat all of them – they're equally good reasons. Given any case of high probability less that 1, you can generate, in a mechanical way, another probability just as great so that with the statistical syllogism you can draw the exact opposite conclusion. That is, if you know that "this is an A" and that it is very probable that an A is a B, you can generate another probability, which is just as high, which gives you the conclusion that this is not a B.

Thomason: You need a big enough language.

Pollock: Yes, but I am not talking about language. I am talking about propositions.

Thomason: You need a big enough space of proposition, I guess.

Pollock: It's big.

I won't go through all the arguments. But the point is simply that to block this, you need a restriction on the A's and B's that you can use in the statistical syllogism. We can label them and say they have to be projectible. We need a theory of projectability which nobody has. A lot of cases of non-projectibility arise from disjunction, and in general, we want to say that projectibility is not closed under disjunction. So in using this statistical syllogism, you want to keep disjunctions out.

Loui: What would projectibility correspond to in an inheritance network?

Pollock: That's interesting.

Loui: You can't draw the link if you can't project one from the other. Even if there's a high statistical connection?

Pollock: That's right. This problem has not been encountered in work on defeasible inheritance hierarchies because people haven't looked at disjunctions. Ultimately we have to allow logically complicated nodes in hierarchies. This indicates that once we put

disjunctions in, and we want to, we'll have to be careful because if we put just any logical form in we will run into this problem.

Pearl: Can you demonstrate in more detail how this problem arises?

Pollock: Suppose we know that the vast majority of birds can fly. There are many more birds than giant tortoises. It follows that most things that are either birds or giant sea tortoises can fly. Herman is a giant sea tortoise. This gives us a reason for thinking that Herman can fly. We certainly don't want that. We want to make our inferences from the "not anything" or disjunction.

Fetzer: This doesn't affect what you are saying about disjunctions but the funny predicates are defined conjunctively: blue before the year 2000, green after, or green before 2000, blue after; just funny definitions of properties.

Pollock: Disjunctions don't solve the whole problem. There are other cases. In general it seems disjunctions will cause problems.

Bacchus: This is part of the problem of choosing a reference class.

Pollock: I guess it is.

Geffner: This doesn't happen in most default logics because you take the default to be primitive. There's a distinction between conclusions sanctioned and defaults that are given.

Pollock: Sure, but that's an ad hoc way. If you begin by listing a finite set of defaults, then of course you can leave out the non-projectible cases. But we'd like a general theory of what we should and shouldn't put in. Doing that, you will solve the projectibility problem.

Fetzer: I have a theory about that.

Pollock: I have one more point, which I learned from Kyburg. You need more defeaters than just subset defeaters to get the logic of the statistical syllogism. Paradoxically, the more defeaters acknowledged by the statistical syllogism, the more inferences get licensed by the theory. The probability calculus gives such a rich source of logical interconnections that it is easy to find conflicting probabilities and end up with collective defeat. The only way out of collective defeat is by defeating interfering inferences in some other way. If you adopt more defeaters, you actually save more at the end.

An example of another defeater seems to be required. Suppose that 98% of the people with enzyme B in their blood have Roderick's disease. We investigate other correlations, and discover that having red hair doesn't make any difference. We discover, however, that some other considerations disrupt our correlation. For instance, of the patients undergoing radiation therapy, only 32% of the ones with enzyme B in their blood have the syndrome. Because we found that hair color isn't relevant to the reliability of the general diagnostic technique, we typically wouldn't check specifically to see whether this holds for redheads. Now consider Jerome who is red-haired, undergoing radiation therapy, and found to have enzyme B in his blood. What should we conclude about whether he has the syndrome? Intuitively, we should not conclude that he has the syndrome because he's undergoing radiation therapy.

Can we get this out of our theory? Not with the subset defeaters that we have. We have statistical knowledge about three reference properties. Persons with enzyme B, redheads with enzyme B, and persons undergoing radiation therapy with enzyme B. We think being redheaded is irrelevant in this case. What rule could we adopt to capture this? I propose a notion of domination defeaters, which says: if A is projectible with respect to both B and C, then probability of (A on B) is the same as probability of (A on C) and C entails B. That's an undercutting defeater, for the inference from the stronger predicate: if you have this information, you are forced to make the inference on the basis of the weaker predicate rather than the stronger predicate.

Thomason: Where will you get knowledge about projectibility with respect to arbitrary conjunctions? You will conclude it from more basic projectibility or will be storing it directly?

Pollock: I assume the conjunctions are projectible. Projectibility is closed under conjunction. You can give a good argument for that.

This is only a defeasible defeater. Other considerations could defeat the defeat. For instance, if you know that in another predicate between B and C the probability is different, then that defeats the defeat.

Fetzer: Don't you want to say something like redhairedness is possibly statistically irrelevant but not causally irrelevant? You'd give preference to causal relevance over statistical relevance in this situation.

Pollock: Something like that seems to be happening.

Etherington: Why do you get the probability from "red-head" and not from "radiation-therapy?"

Pollock: In virtue of knowing that these two probabilities are the same ...

Etherington: They aren't the same.

Pollock: These two probabilities are the same. We presume that the additional strengthening (by red-head) is irrelevant, consequently anything defeating the inference from one (enzyme B) ought to defeat the inference from the other (enzyme B and red-head). That's the idea behind domination defeaters.

Etherington: But neither will be applicable to this case.

Pollock: You don't get a conclusion, just defeat. One (radiation therapy) will defeat the inference from either of the other two (enzyme B, enzyme B and red-head) but gives a low probability so you can't get a conclusion.

Loui: Can we go back to when you discussed me?

Pollock: All right.

Loui: I think we have to at least include Poole, Delgrande, and Nute. They used logical strength for defeat instead of subset defeat.

Pollock: I refer to these as subset defeaters, but I think of them as logical strength in what I'm doing.

Loui: You think of this as logical strength?

Pollock: It's logical. This is strength in part of the antecedent.

Loui: Right, it's only in part of the antecedent.

Now you only get grief because you unpack something like "bird is reason for flies" is reason for "most birds fly," and "Tweety is a bird" as "Tweety flies." Your reason relation is purely analytic, determinable from what's written. The way we write things, the default "birds fly" has empirical content.

Here's a spectrum: you, us (Nute, Delgrande, Poole and me), and then Konolige. You allow evidence to have empirical content – and why shouldn't you? You don't allow reasons to have empirical content; they're purely analytic. All of the defeat relations, the rules that decide one argument is better than another, are analytic. I, Poole, and Delgrande, allow statements about the world to be empirical, but our reason relations also have empirical content. Our rules for arguments are analytic. Konolige recently wrote that he wants all three (statements, reason relations, and rules) to have empirical content. He wants to alter our conception of defeat based on which domain we're in.

Why do you want to be extreme?

Pollock: Whether this makes a difference is not clear. If you want to do more than inheritance hierarchies, if you want to talk about the entire structure of your beliefs, you have to worry about from where you got something. You can't pull it out of the air.

Thomason: No. If someone had a system with probabilistic and default reasoning together, I think that the "most A's are B's" connections would fall under the probabilistic representations.

Pollock: Sure. That's what I'm doing.

Thomason: But the things we talk about in inheritance diagrams needn't be like that. They can be ...

Pollock: That's what I said. If we only do inheritance hierarchies, we suppress part of the antecedent. If we give a general theory of defeasible reasoning, as applies to all beliefs, then we have to worry about the rest of the antecedent of a reason because we believe other parts of the antecedent too. You believe them as a result of other reasoning.

Thomason: That's fine.

Fetzer: Those statements are just statistical summaries? They don't have any nomic sense?

Pollock: You can do it either way. Given certain assumptions, the statistical version follows from the nomic version.

Fetzer: Could you tell me what you mean by nomic probability?

Pollock: Heuristically, you can think of it as a measure of A's that could be B's.

Saturday Session IV: Issues and principles.
Speakers: Matt Ginsberg, John Horty, Ronald Loui, David Poole.
Moderator: Paul Morris.

Horty: John Pollock told us that trying to generalize specificity from inheritance networks to richer languages is not obvious. I'm here to say that comparing one argument to another in inheritance networks is not obvious.

I'll give you options among which I don't have any preferences.

Doyle: Seventy-two?

Horty: No, four or five to provoke discussion. Everybody agrees we should start with two conflicting unrelated arguments to which you have to do whatever you do. I'll tell about you two clear cases. The question is, "How do you state in general what's going on in these two clear cases?"

This is the vocabulary, notation, I use in answering this question – I'm sorry for this. Pi means an arbitrary path running from X, moving arbitrarily through sigma, to D, and then extending with a link to Y. I do this because I don't want to tell you anything about what the path looks like except the beginning and the endpoints.

In the context of arguments, when is an argument of this form preempted? When do you like another argument better? The strict-subsumption view, which says this argument is preempted if there is direct information to the contrary, will be uniform throughout these examples.

The question is when do you like one node better than another? When do you like arguments emanating from Z better than those from D. The first answer is when there's a strict path of any length from Z to D. I attribute this idea to Thomason, who has never written it down. If you like that, then you view this as an ambiguous network.

The second answer is you like one node better than another whenever there's a defeasible path, not just a strict one, running from the first node to the second.

If you like option number two, if you compute multiple extensions, specificity varies from extension to extension. So this is a double diamond, one nested inside another. With two diamonds you have four extensions, but this actually only has three because if you take a choice in the inside diamond, you're forced to prefer chaining off P to chaining off Q.

Loui: What did you call that?

Horty: "Defeasible subsumption" because the connection between the two nodes was defeasible. Defeasible subsumption is general subsumption since it allows for strict connections too.

Loui: That's what I called "strong specificity" but wished I'd called "broad specificity."

Nute: That's what I call "naive."

Pearl: Why do you claim it, if you will revise it later?

Horty: I don't believe in any of these. I'm skeptical because I don't know anything about it.

The first option I call "off-path," because the other's been called "on-path." You reason about X and prefer to chain off Z rather than D. They might go off to two conflicting things. if there's not only a path from Z through D, but in fact a path running from X through Z to D. For example, Matt Ginsberg has interesting labels for this net.

A path is permitted from A to P and there's a path from P to R, but there's not a path from A to P to R. According to this view, this is an ambiguous net; according to the previous view, it's not. Does anybody have any feelings about which of these views is right, or how you could resolve these questions?

Bacchus: Can you resolve these questions without a more definite meaning of the links? Perhaps under different interpretations of the links the preferred argument would be different.

Nute: You're doing a taxonomy.

I refer to this as "intact" defeasible specificity and the previous one "naive" because you treat the defeasible links as if they're strict.

Horty: I don't know if you treat them as if they're strict.

Nute: Look at the last one. That's what I was calling naive.

Horty: This seems to make sense. The other seems to be a very conservative view, which I called the true on-path preemption. Suppose when we assure ourselves that X through V is okay we consider extending V onto Y. The preferable nodes to V are only the ones that lie on the path you're concerned with. Another node, Q, which tells you not-Y doesn't bother you because it doesn't lie on the path you're concerned with; It might bother you in other options.

The problem is it's not stable; you can perturb the conclusions of the network by adding links that are supported by it already.

Now add a link that was there implicitly. That's true on-path. I attribute this to no one. It's like what Dave Touretzky calls "on-path." He considers the case where you perturb by adding additional links, and defines the set of links that can give you opposite conclusions.

One problem with constructing a definite theory of these things, is that you have to choose.

Selman: Touretzky calls that "on-path?"

Horty: Touretzky doesn't need the nodes to lie on the path, and so it's funny he calls it "on-path."

Let me say something about contraposition. Answer this question: Is object number twelve a bird? We know that birds fly; object number twelve doesn't. I guess everybody thinks that object isn't a bird; you'd like to conclude that, right? Is there a consensus?

Konolige: If your whole universe were birds ...

Goebel: Could be exceptional ...

Horty: This would be a defeasible conclusion, right?

Israel: If that's the whole network ...

Loui: Let's take a vote.

Horty: How many people feel strongly that they want this conclusion?

Ginsberg: I do.

Horty: How many feel strongly they don't want it?

Konolige: In general?

Poole: It's open.

Grosof: People want to be able to represent either way in the system.

Horty: That's something you'd want. Or do you feel that way?

Fetzer: Are we distinguishing clearly between contraposition and modus tollens?

Horty: Probably not.

Let me infer using whatever that is ... modus tollens. If you want that, I guess you also want that E's aren't B's; That's the same thing. You can't get this in inheritance networks because if you allow yourself reasoning across a defeasible negative length and then down positive chains coming from the opposite direction, Nixon diamonds spring up all over in place in different ways.

At the choice point (and you have two choices in this ordinary Nixon diamond) you think you get to choose whether X is a C, but if you allow contrapositive reasoning, this could be your choice point too. You get more extensions and when you intersect them, you lose your conclusions. I know no way of handling this problem.

Padgham: Why don't you prefer to not have the contrapositives, as you usually do, then add them in as extra. I have a paper in IJCAI about this. You add contrapositives at a lower priority to the stuff you get normally.

Horty: You do it at the end so it's not an input to the default reasoner?

Padgham: You don't get any conflicts on it or you throw out conflicts you already have.

Horty: That seems wonderful.

Morris: I think it's time for us to move on to our next speaker.

Ginsberg: I'll talk briefly about contraposition. First, whenever I say that I'm in favor of contraposition, somebody asks "what about the fact that people are not normally diabetic?" Then they cackle at me because they think I will say, "that means diabetics are normally not people." But, if you write this down (here is a blockheaded way to do this) "if X is a person and X is not abnormal then X is not diabetic." Indeed it follows from this that if X is diabetic and X is not abnormal then X is not a person, but the point is that diabetes is an abnormal condition by itself.

This particular argument against contraposition does not hold water. There's still the issue basically on which we voted. I will put up the simplest legitimate diagram I can think of, my example of whether things contrapose.

Fetzer: Some statements contrapose and preserve truth, and some don't.

Ginsberg: I'm talking only about "is not a" links. I don't want to get derailed in a big, long philosophical discussion on contraposition versus modus tollens. I think we have a good common-sense understanding of what it means to say "birds fly," and I don't care whether it's statistical or communication's convention. I think our intuitions on simple examples are clear.

Shastri: You're absolutely right that we have good intuitions about what "birds fly" means and I don't want complicated discussions. The trouble begins when you take those intuitions from the simple examples and generalize to another example.

Ginsberg: Let's keep going.

Tom votes for Reagan. Tom is a Quaker. Most people who voted for Reagan are Republicans. Most Quakers are pacifists; Most pacifists are not Republicans. Basically, I want to know how many extensions this diagram has. It seems to have three. Tom is surely a Reagan-voter and a Quaker because I told you that. If you just look at it this way, the choice point is here: he's certainly a pacifist, and he might be a Republican because he voted for Reagan or he might not. If you reverse the arc, you get a different extension where he's a Republican and not a pacifist.

It seems completely reasonable that Quakers are pacifists, Reagan-voters are Republicans, and pacifists are not Republicans. It seems that each of these three possible conclusions is reasonable.

Goebel and Grosof: What was the third conclusion?

Ginsberg: He voted for Reagan and is a Republican. He's not a pacifist. He's a normal Republican, a Hawk, and happened to be a Quaker who voted for Reagan.

Etherington: What's the second conclusion?

Ginsberg: He's both a pacifist and not a Republican, an abnormal Reagan-voter. Then he's a Quaker, a pacifist, and a Democrat. He can be a Quaker, a pacifist, and a Republican (this is an abnormal arc) or he can be a Reaganite, a Republican and a Hawk. That's contraposition.

I also wanted to talk about specificity being object-dependent. Intersecting paths you're prepared to sanction and looking at conclusions that follow from these paths, is not the same as intersecting extensions you get when you expand the paths to complete extensions.

Here is the Politically Motivated example: Nixon is a Quaker and a Republican. Quakers are pacifists. Republicans are Hawks, but both Hawks and pacifists are politically motivated. The question is, "Is Nixon politically motivated?" The answer you want is "yes."

The conclusion is supported by both paths, neither of which is okay by itself. When I introduced this, Horty gave a talk and complained that it had cyclicity. Makinson and Horty had come up with this acyclic diagram with the same property. Even if you're not prepared to think about cyclicity, looking at the paths that are sanctioned by all extensions is not the same as looking at the extensions themselves.

There are four things people seem to be doing. Some people invent special-purpose methods which is roughly what I think Jeff Horty said his work was. The idea is to get the special purpose to work, and hope eventually it will be subsumed in general methods and will go away.

Other people skip that part and directly show that their general purpose method is better than anybody else's. For example, they show that circumscription is wonderful because you can do inheritance reasoning. These methods are both reasonable.

The third thing people do, which I applaud Judea and Hector for, is work in epsilon semantics. Geffner uses a sharp description of a general approach; the intuitions for which are based on careful consideration of principles underlying a special case. They use probability to illuminate what goes on in general. I think people should not focus too sharply on particular examples, but keep track of the principles that underlie the examples. It doesn't matter whether your conclusion is that Nixon is politically motivated; what matters is the reason for that conclusion. Are you intersecting the conclusions rather than the paths?

The last thing people do, which I argue isn't a good idea, is describe inheritance reasoning in a God-awful way using general purpose methods. This produces a stretched purpose that can't be extended to more general stuff. They do it only to boost their confidence in their tool. Now that's not right.

Eventually we want to have general purpose principles, general principles that we extracted from special cases. We either have to understand the special cases or find a way to bypass them.

Pearl: I like the comments here.

With strong inference principles and weak inference principles can we keep tabs on arising conflicts? As far as contraposition is concerned, I don't think it is either strong or weak; it depends on the context.

If you find a flying object, I'd like to conclude it is not a penguin not only because we have a strong intuition, but also because it is a result of epsilon semantics. It's part of what many approaches to non-monotonic reasoning consider the conservative core. I avow that. Other kinds of contraposition do not have

this quality, being part of epsilon semantics or the conservative core.

Morris: Suppose you accept contraposition as a principle of default reasoning. Do you have an idea of the circumstances under which it gets blocked? You gave one example.

Ginsberg: My opinion is that when it gets blocked, there is explicit information in the database to that effect, as in the diabetic example. In the absence of explicit information, it runs. I'd like to see Lin Padgham's examples because it seems that in these diagrams you get multiple Nixon diamonds In all the examples I've come up with, reversing the arrows and getting a whole pile of extensions is pretty reasonable.

Goldszmitt: Suppose you have the example of penguin-bird. You add a node between them. If you allow contraposition with no discrimination, and you ask if the penguin is a bird, by contraposition, you get one path that's a bird and one that's not.

Ginsberg: That diagram is basically ...

Goldszmitt: A Nixon Diamond.

Ginsberg: A lot of examples have single arrows that are really double arrows In my examples the single arrows are really single arrows, so you get the contraposition propagating as far as you can. If single arrows are natural (which is where Pacifist-Republican-Reagan-voter came from) my intuition is that the contraposition propagates all the way.

Goldszmitt: Mine is also. But I always get puzzled by these ...

Ginsberg: That's because they're really hiding double arrows.

Goldszmitt: That's true.

Etherington: Isn't there psychological evidence that people have trouble with contraposition?

Ginsberg: It's something that people don't automatically do, but feel stupid if they miss it. They sanction it, but don't come up with it themselves. I don't know what that means we should do. You can know Tom wasn't at the store because he didn't come home with a can of coffee. And people say, "I should've figured that out." But they didn't.

Fetzer: Your last concern, Matt, was about individuals who use special case methods to defeat general-purpose. I'm concerned about your dismay.

Ginsberg: That wasn't what I meant to say.

Fetzer: What did you mean to say?

Ginsberg: They take general methods and turn them into special-purpose methods. When they're done, they have something you could never embed in a larger theory. There's nothing wrong with that. It helps us understand the special case. But ...

Fetzer: What about your fourth principle?

Ginsberg: They use the special case to emasculate general methods. What if I use circumscription to write axioms for inheritance nets. In circumscription, I conclude that pacifists and hawks are disjoint, not because I was given it as an arc but because I have a complex theory of political views. If my description of inheritance using circumscription is so screwy that it can't incorporate conclusions drawn from using first order logic that circumscription allows me to take advantage of, what's the point of my using circumscription?

If you emasculate a general theory to the point where you can't take advantage of something like the descriptive power, you might as well be working with a special-purpose method where it's clearer what the principles are.

Poole: Matt Ginsberg asked me if THEORIST doesn't allow me to do things because it's so weak.

It allows us to criticize the premises if we don't like the conclusions. In the example where you can explicitly write "birds fly," "emus don't fly," and a cancellation, we have reason not to like the premise "birds fly" when we know "Edna is an emu." By augmenting THEORIST with this, we can get the same results that other people got.

The idea, which is in the IJCAI paper, is "birds fly" is applicable even if you know "Edna is an emu," but "emus don't fly" isn't applicable if you only knew "Edna is a bird." It's not symmetric.

Gelfond: Using direct or indirect arguments?

Poole: This is an intuition: prefer direct arguments rather than circular indirect arguments.

We'd like logically equivalent formulations to give the same results, to express the same idea. This indicates a need to distinguish between the general knowledge "emus are birds" and the contingent. In path-based formalisms we distinguish the nodes from the

arcs when we translate into logic. Also in probabilistic formalisms, we distinguish what we put on the right hand side from probabilistic statements.

In the Nixon diamond, there's a symmetry as you take different nodes to be known. This tells me that only the things one step away from a node are distinguishable. Probabilistic analysis gives this context-sensitivity, the one-step default property. With a circular argument, you want to conclude what is one step away.

Etherington: You never want to chain?

Poole: When there is a loop, you want to use the one-step property. On symmetric grounds, there's nothing else you could do.

Etherington: That's complete contraposition.

Poole: Contraposition is a side-effect.

Ginsberg: Do you want this extreme view?

Poole: In one paper I said exactly what you said, "It may be that you don't want contraposition." I'm schizophrenic about contraposition.

Then I tried to put in a one-step default property in terms of a more general class. I messed up. I didn't have the notion that arguments contradicted each other.

THEORIST has a reasonably nice implementation of two agents having an argument. One agent finds explanations of the goal. One finds scenarios inconsistent with the explanations found by the other. One sequential protocol is: when X succeeds, Y tries a new explanation; when X fails, ...

Ginsberg: They play by different rules, right?

Poole: They both find explanations, but of different things. Y finds explanations of G, and X finds explanations of not-G.

Ginsberg: But X undercuts what Y says while Y refutes what X says.

Poole: They both find explanations.

Ginsberg: But Y could just repeat G.

Poole: With this protocol they prune each others' search space. Y knows it doesn't need to find an explanation that's inconsistent with any that X has found. X doesn't need to incrementally search. That's how it's implemented.

Ginsberg: I don't believe you.

Poole: What? It runs. In your circumscription theorem prover, I think, Y finishes all of its arguments then X starts.

Ginsberg: That's right. But Andy Baker and I have a paper in IJCAI this year about extending this to prioritized circumscription and it matters that X's job is just to mess Y up while Y has to utterly refute what X is saying. They play by different rules.

Poole: In this case they use the same rules. They both find explanations.

THEORIST tells us where to find arguments for and against things. Because we always try to defeat some goal, the notion of skepticism arises.

When I implement I get stuck on things like disjunction. Birds fly, emus don't fly, ostriches don't fly, but emus are birds and ostriches are birds, and Edna is an emu or an ostrich. My definition works but my implementation doesn't.

Grosof: About the one-step-away principle, do you imagine that in a knowledge base created by several people the left hand might not know what the right hand is doing? You might have someone put in as a one-step rule something an omniscient observer says corresponds to the compilation of other stuff already there, including conflicts with other rules, and therefore we should be skeptical.

Ginsberg: No, I think exactly the opposite happens. When people put in a direct rule, they are telling you to stop contraposing right there. They're being specific. This is a more general form of the degenerate case I talked about with the diabetics.

Etherington: So you don't hold to stability: adding things that are already there doesn't change ...

Thomason: We don't claim stability for general rules.

Ginsberg: I believe in stability for particular instances. But if you have "A's tend to be B's" and "B's tend to be C's" and then you add "A's tend to be C's," you may get something very different. Everyone believes that.

Morris: We're under time pressure.

Loui: The crime in reasoning by cases is not that if A1 is defeasible reason for C, A2 is defeasible reason for C, A100 is defeasible reason for C, and on evidence of the disjunction A1 through A100 you conclude C. That's not the crime even when you show Simpson's paradox, that C might not have very high probability. A lot of times you might have defeasible arguments where the probability of the conclusion is not high given the premises.

The real crime is that you may want to add that this disjunction is defeasible reason for not-C. When you reason by cases, in case 1, A1 is reason for C. A1 allows you to use the disjunction as reason for not-C, but the first reason is more specific. Therefore, you still conclude C. Case 2: the same thing happens. After you go through each case, you say "I've exhausted the cases, therefore C."

Goebel: Why is the first reason more specific?

Loui: Because A1 entails the disjunction.

Ginsberg: You should be getting a bad shivering feeling, right Judea?

Pearl: ... If you have machinery to realize you have an inconsistent system according to certain standards.

Loui: Right, it's inconsistent from the Pearl-Geffner point of view.

If you bring up the reasoning by cases dilemma, the problem is not that you can't choose between C and not-C, but that you get the wrong answer. You get C when by writing the disjunctive rule you were trying with this evidence get not-C. It's a little worse than Poole, Neufeld, and Pearl have reported. It's not that you can't get the right answer, it's that you get the wrong answer.

Pearl: If you ignore the inconsistency and proceed ...

Etherington: You also have another problem here. You assume going into cases preserves specificity.

Loui: Right. You might recast your specificity rule so that it doesn't occur in cases. You might say reasoning by cases in a specificity-based system is a lower-priority style of defeasible reasoning. Lin Padgham's suggested making contraposition a lower-priority style of defeasible reasoning.

Directness is a property systems should have. Garfield is a cat, he dislikes people in virtue of his being aloof, defeasibly. But cats defeasibly like people. This is the more direct argument, so it defeats the other. Most systems have this property. How do we write formally that a system has this property or not? John Horty tried with off-path preemption. He tried to write formally the condition to have this property. He said, about his symbols, "this is the vocabulary I use; I'm sorry for this."

Fetzer: This is just specificity. Directness is defeated by taking more properties of Garfield into account.

Loui: I don't think that's what's happening here.

How do you write in a language that applies to enough systems that you can see whether Poole, Geffner, or Delgrande has that property. The way I've written it, I can tell if my system has it, but I can't apply this condition to Horty's system.

Thomason: He's apologizing for his notation.

Loui: That's exactly what I'm saying. We're all apologizing for our notations. Can we figure out what language is adequate for expressing all systems?

I've separated all the systems I know well into two columns. Geffner, Delgrande, Neufeld and maybe Nute on one side. Pollock, Touretzky, Poole, Loui, and maybe Horty on the other.

I separate them this way because people in the first column are interested in something like irrelevance or independence, which allows you to do transitivity, or to enrich the antecedent. Their principal interest in defeasible entailment is the following. If I have something like "bird defeasibly entails flies," and all of my evidence is "Tweety is a bird and ugly," can I introduce "ugly" on the left of the defeasible entailment relation and keep "flies" on the right? If I can, then I can conclude "flies" on the basis of the total evidence. I introduce "ugly" on the left side by saying "ugly" is irrelevant to "flies" in the context of "bird." We can't do this with "penguin," because "penguin" is relevant to "flies" since it's an argument for "not flies."

The people on the right are interested in defeat, preemption, and preclusion. They're interested in a path from "bird" to "flies." When "Tweety is a bird and ugly," I'm still allowed to use this argument. "Tweety is a penguin and a bird," gives us a new argument or new path: "penguin therefore not flies." I look at defeat relations among these arguments, talk about these arguments as objects, and come up with relations among them.

Think of proof theories. You don't say there's a relation between those proofs.

I said I bifurcated on irrelevance and defeat, but that's only half true. I get the same bifurcation and categorization if I say which systems have axioms and inference rules, and which ones reason about prima

facie inference. I claim that people interested in irrelevance have axioms and inference rules while people interested in defeat reason about the paths because there are two paradigms. I wonder if you accept that.

Pearl: I accept that.

Loui: Why not three?

Pearl: If you are born inhibited, you have to work on gaining courage. If you are born promiscuous, you have to work on acquiring limits to your behavior. Automatically, if you have inference systems, you are semi-monotonic. So you have to work on giving yourself more power. The dichotomies coincide.

Loui: This is a more interesting dichotomy than just that. Look at a situation in which you want to compute defeasible entailments under limited computation. (Let me remind you how negation-as-failure works. If C is consistent, infer C and if C is not consistent, infer not-C.) Under limited computation I might focus on one rule and see if I can use it, which means attempting to prove not-C. I fail to prove not-C because I run out of time, so I say the rule is fine and conclude C.

It's irrevocable. Having decided you can't prove not-C, allow this rule to fire. Had you tried to use the other rule instead, you would have found that you can't prove C, and therefore should be allowed to conclude not-C. The control strategy I just described would focus only on one rule. If it tried to equivocate, to spend some time on each rule, then both rules fire and that's a problem.

In irrelevance systems the control strategy under limited computation appropriate for these systems would have: "if bird defeasibly flies" and I want to see if "ugly and bird" gives "flies," I need to show that "ugly" is irrelevant to "flies" in the context of "bird". What does it take to show irrelevance? It takes trying to prove "not flies" from "ugly." If I can't, then I'm allowed to say it's irrelevant. I don't have enough time to check all possible proofs for "not flies" from "ugly." I might say I couldn't find a proof, so it's irrelevant. Having made the irrelevance assumption, it's irrevocable. That means I'm allowed to put "ugly" on the left-hand side.

Defeat is irrevocable in systems on the right hand side. What may you have to call time on in argument systems? What in argument systems says "I may have determined this, but I didn't have enough time, so I assumed that it is or isn't there?" Lacking time, we might assume there's no relation of defeat between the two arguments.

Systems with defeat are interesting because we can make inferences even if we don't have defeat relations. We only need defeat relations to resolve a conflict when two arguments interfere. All the interesting cases require defeat, but we can infer without defeat. You can't infer in other systems without irrelevance.

Hawthorne: It seems your bifurcation is whether independence, irrelevance, or randomness is in the object or metalinguistic language.

Loui: In the defeat paradigm, in Poole's theory, and Touretsky's theory, what corresponds to irrelevance?

Goebel: ... Failure to find contradiction.

Konolige: ... Something like failure of conflict among arguments.

Poole: Yes.

Goebel: You said earlier that you were reifying derivations. That's what it is.

Grosof: In general, you have to fail to find in the argument systems, just as you have to fail to find relevance. Non-defeat is like irrelevance where you cut your search short. In some special cases, maybe there are syntactic checks like connectivity so that you don't need to look because you know the relation's there already. But those are special cases. I think there's symmetry.

Sunday Session I: Path, Argument, and Model-Based Approaches.

Speakers: Gerhard Brewka, Hector Geffner, Michael Gelfond, Brian Haugh.
Moderator: Guillermo Simari.

Brewka: I agree with what David Poole said yesterday with one exception. You cannot represent hierarchies with THEORIST. I'm motivated by the desire to represent hierarchies.

First, I'll show a general framework which subsumes THEORIST as a special case. Then I'll give an example which illustrates that there are some things that cannot be nicely represented by THEORIST. Then I'll show two generalizations.

With any standard default formalizations, you start with an inconsistent set of premises and extend the inference relation. In circumscription you use an additional schema. In modal logic, we have additional non-standard inference rules.

A default is a default because of our attitude towards it in cases of inconsistency. If we have conflict, we tend to give up the default, not the premise. Naturally, default reasoning is a special case of inconsistency handling. Instead of starting from consistent premises, start with inconsistent premises and try to handle the inconsistency.

We modify the inference relation so it has fewer inferences than classical logic. The standard way is by considering maximal size consistent subsets of premises. If you consider all of them, then there's no way of representing priority between them. You have to select some of them. This gives you two notions of provability, derivability in all maximal consistent subsets, and a weaker version, derivability in at least one maximal consistent subset.

How does provable fit in this framework? We have to define the preferred maximal consistent subsets. THEORIST has a consistent set of premises and a possible set of hypotheses. A notion of explainability corresponds to our weak provability if we define preferred maximal consistent subsets (I call them preferred subtheories) as subtheories containing the premises. This seems to be exactly the THEORIST system.

There is one problem with THEORIST. You can block the applicability of the defaults. David Poole gives names to the defaults so you can express "this holds" and "this default should not be applicable." But it's not possible to represent priorities. It's not possible to say if one default is applicable then another conflicting default should not be.

For example, usually we have to go project meetings, but there are exceptions. For instance you could be sick or on vacation. But there's an exception to this exception. If you only have a cold, then the rule does not apply. A cold is no excuse for not going to a meeting. If you want to represent this in THEORIST, you need these defaults: a default for going where if sick is true, this will not be applicable. Or if vacation is true, this will not be applicable. But since there's an exception to sickness being an excuse, you need another default. Maybe this is not so bad, but the repercussions to this other exception are. The introduction of the new default forces you to explicitly state that if you are on vacation, this is also an exception to the new default. This seems to show that it's not sufficient to block in certain situations, but it must be possible to represent priorities.

Geffner: You need the ability to represent exceptions to exceptions. Why do you need priorities?

Brewka: You need another formula to explicitly state being on vacation is an exception to the new default. This seems to show that it's not sufficient to block in certain situations, but it must be possible to represent priorities.

Priorities signify more than exceptions to exceptions.

The first generalization is represented graphically. In Poole's approach we have different levels of hypotheses with decreasing levels of reliability. A hypothesis in a lower level is only given up if it contributes to an inconsistency with a hypothesis at a higher level. To get the intended results within a framework of maximal consistent subsets you prefer a subtheory if it is a maximal consistent subset of something containing a maximal consistent subset of something else containing a maximal consistent subset of something else, and so on.

There are cases where this hierarchy is not general enough. You need formulae to be incomparable. A's are red, B's are blue, C's are green, and A is more specific than B. This hierarchy forces you to put C implies green somewhere. They must link, leading to unintended results: priorities where you don't want them.

The second generalization is a partial ordering among premises. S is a preferred subtheory of T if there exists a total ordering. S can be obtained by picking the premises according to this total ordering. Whenever you pick a new premise consistent with what you have already, you put it in the subtheory that has been used.

Pearl: In your condition you talked about premises. In your example you talked about defaults. How are they related?

Brewka: The defaults are schemata. If Poole represents birds fly, he has a schema and ground instances of the schema.

An advantage of this approach is that you have to solve the problem of handling inconsistencies anyway. In this approach, the solution is explicit.

Geffner: A number of definitions of the defeasible consequence relationship, in different languages and in terms of different meta-level constructions, encode the relationship found in inheritance nets. In some schemes, this information is encoded in the links and the different inferences defined; in other schemes, this information is encoded in first-order formulas and you reason in terms of models, probabilities, or default extensions. Still other approaches introduce a modal operator and define the defeasible entailment relation in terms of stable expansions.

In all cases you get non-monotonic behavior. Sometimes after tuning you get the intended behavior. The problem with the proliferation of notations, languages, and definitions, is that a common problem that the approaches are trying to capture is obscured. This prob-

lem has arisen in inheritance nets. As Matt Ginsberg pointed out yesterday, many proposed solutions are applicable or understandable in terms of the particular notation or particular language in which defeasible consequence is defined. There is nothing wrong with defining inheritance in terms of paths. The path-based approach is probably the most convenient. You're only interested in doing inferences in an inheritance net, and the language of paths is probably the simplest.

By uncovering the issues that underlie the inheritance problem you don't get anything by translating into a framework which doesn't attack central issues of the problem.

It is possible to construct path-based characterizations of inheritance derived from general principles, relevant to approaches, and not defined in terms of paths. We will start with an inductive definition of inheritance. The link in the net notation becomes conclude D from Y. The second part of the inductive schema is similar to that of Horty and Thomason, and can derive Y from X. Then you are able to derive the consequence of Y when the conflicting path from X to the complement of D can be ignored. We say that the conflicting paths can be ignored without using defeat. We avoid preemption. We use the term preemption to mean something different.

We have mapped the inheritance problem into this notion of defeat: under what conditions can conflicting paths be ignored? No special commitment comes this way. Almost any existing definition of inheritance can be mapped into an inductive definition of this form.

The question is how to have a good definition of defeat. This is usually where the clash of intuitions is found. Jeff Horty talked about the wide space of options yesterday. He mentioned there were at least 72. The problem is selecting a reasonable account of defeat. Most of the discussion has been centered around on-path versus off-path, different specificity considerations, skeptical versus credulous, and things like that. Most of this vocabulary is too embedded in the language of paths and the algorithms. It is possible that the space of options can be reduced if you can express the underlying assumptions of the choices we make.

The problem of inheritance is not so much a problem of design but of discovery. We won't arrive at a unique definition of inheritance, but I think the number of choices is not large if we represent the assumptions that underlie the choices.

I suggest two assumptions. First, when you have a default X then Y such that its antecedent presumes the proposition Z, then the default authorizes you to conclude Y from X and Z. This is not particular to default reasoning; this assumption is a weaker version of the commutativity rule I talked about yesterday in the probabilistic system and that many modal systems have today. I regard this assumption as a general principle for defaults.

Secondly, when you have defaults and can derive Z from X in a net with paths between X and the complement of Y, and you have a default from X to Y, what do you do with the path that connects X to the negation of Y? We assume that the paths connecting X to Y in this context are not applicable to members of the class X. These paths are defeated.

Ginsberg: You're taking the position that specificity is object-dependent, because it's possible to imagine A being an X and a Z, but not being a Z by virtue of being an X. If you have a node between X and Z that did not include A as a member, your assumption tells you it doesn't matter.

Geffner: It doesn't matter. I'm saying that what we do or how we do it is not a matter of what properties the individual has, but a property of the class.

Ginsberg: My jerky lawyer example is non-ambiguous according to this definition.

Geffner: According to this definition, not ambiguous.

Ginsberg: Not ambiguous.

Geffner: Yes, I think it is not ambiguous. If you have a triangle with something in the middle and you take the two extremes and negate the thing in the middle, the bottom link prevails. Yes, it prevails.

Ginsberg: People in the field have this opinion, but when they ask their friends, their friends say it's ambiguous. I don't know what that means.

Geffner: It remains a conflict of intuitions. From my point of view, you are arguing against a lesser assumption. The idea is that it's not so tied to the language of thought. The first is a weak assumption. The second is about how you read the default assumption.

Horty: You need paths to express the three-place relation that Matt is talking about. It's hard to say if you just have formulas. You need paths to have some way of "going through."

Geffner: When you mentioned that specificity is a three-place relation, what types of arguments go into the specificity relation? Propositions, classes, what?

Ginsberg: Two classes and an object. Class one is more specific than class two for object X.

Geffner: OK. This is not exactly the notion of specificity; I think the notion is more general. It's a two-place predicate.

Ginsberg: The question is, is it two-place or three-place?

Geffner: I think that's a better question to argue than whether you are on-path or off-path, things like that. I'm not saying this is the only reasonable assumption. I want to convince you it's reasonable. But whether it is reasonable or not, it's clear; it can be criticized.

Grosof: I think four-place is more natural. Individual, class one is more specific than class two, for the purpose of some goal class, G. As a naive inheriter, that's a more primary intuition of specificity.

Geffner: Okay. I will give a rationale that may be valid for why it depends on classes. The point is whether specificity depends on the individual, or is going to be the same for all individuals. We had thought that it's the same for all individuals.

Some implications of my definitions, for instance in the penguin triangle, where birds fly, penguins are birds, and so on: if you know that Tweety is a penguin, then bird is always irrelevant, or does not say anything about flying. For penguins, bird is irrelevant to flying. The clause, "birds fly," can be erased from the net. I want to say that for one individual, reasoning in the net after removing this path should get the same result. In a sense I accept the idea that the default "birds fly" is not applicable to the individual in this case.

What are the consequences of this account? I don't know. What are the consequences of these definitions? Networks are going to be consistent if every pair of competing links in the network has distinct antecedents which are not embraced by acyclic paths. We can deal with acyclic paths. Actually this has the same condition for consistency as epsilon semantics. This is not justified probabilistically. But it has the same type of sufficient and necessary conditions when you apply epsilon semantics to this type of network. This is only a sufficient condition, not a necessary condition. The complexity of this scheme is NP-complete. That is shown.

Summarizing my point of view, there is nothing wrong with doing inheritance in terms of paths, and doing so can uncover issues that underlie the inheritance problem. After all, we can make more progress and talk to each other in a more meaningful way by making clear the assumptions that underlie the approach. By "making assumptions clear," I don't mean a model theory, but instead that we can communicate without having to think in terms of algorithms. That's essential.

Horty: Besides being able to deal with acyclic nets does your scheme treat any well-known examples? Is there anything surprising about it?

Geffner: According to the definitions that Horty, Thomason, and Touretzky give, and according to the type of commutativity rules that we had, this scheme of inheritance is not stable. I've argued that probably stability is not a universal assumption that you want to make.

Horty: Non-stability follows from that issue we talked about earlier. Specificity is only two-place, right? If you had the three-place relation of specificity, it would be stable. Is this true?

Geffner: I depends on how you use it, probably. I don't know. Why is it not stable? We cannot say by these examples. How much time?

Simari: One minute.

Geffner: For instance in the net I have here you can conclude G from A, and can get F. What stability tells you, is that if you add a link from A to G, that is already conclusion from A, you should still be able to conclude F from A. When you add the link, you end up with a network in which skeptical inheritance will tell you A is an F. I don't conclude F from A. So my rules are not stable.

Horty: If you used a three-placed relation, in the net on the right, B doesn't give you more specific information with respect to A, than G does. It only gives it to you with the two-placed relation.

Geffner: B is more specific because, since A is a G, in some sense you get this path. I don't know whether you like it.

Simari: Can we postpone this?

Gelfond: I'm going to talk about some research of Halina Pryzmusinska, me, and a couple of us here. The main goal of this research was to classify autoepistemic logic, a formalization of commonsense reasoning. We had two goals. First we wanted to check if autoepistemic logic was applicable in practical

terms. We selected two areas of commonsense reasoning: temporal reasoning and inheritance hierarchies. We tried to see if autoepistemic logic works well in these areas according to our estimation. We checked on the conventional properties of autoepistemic logic. We also wanted to check that it should be expressive.

Now, why autoepistemic logic? As always it is difficult to answer at this time. At first, I was interested in circumscription primarily as a tool. Most formalisms are reducible to autoepistemic logic like negation as failure and logic programming. Circumscription can also be modeled in autoepistemic logic by a simple device. Recently, it was kind of shown that truth maintenance systems and abductive reasoning can be expressed in autoepistemic logic. I say "kind of" because I don't understand truth maintenance systems enough to know.

Why not try to work with a new logic? I worked for a couple of years as a programmer and my supervisor told me, "don't fix what's not broken." This leads us to methodology: we are suspicious of new formalisms and try to deal with mathematical properties of what is already known. Every time I look at a new formalism, I have a difficult time figuring out, even in a simple case, what holds and what doesn't hold from it.

I'm going to talk about inheritance hierarchies in a moment.

First, you might be surprised, we just have two types of defaults. Some can be easily expressed as: "birds typically fly", or "birds normally fly." When you look at sentences like this, one of the connotations is that you can afford to assume that birds fly, unless there is some reason to believe that they do not. Sentences of this form can be expressed in terms of autoepistemic logic, captured by Moore. It's easy to do using a simple device suggested by McCarthy. Just say that it's not true that you believe that birds have no reason not to fly.

As far as I understand the previous speaker's suggestion, you are saying that to understand inheritance hierarchies, you have to understand what paths to select, right? I suspect that there is one more problem, which is even more basic. I am interested primarily in what the link is. When you write a defeasible arrow from C1 to C2, I suspect you first have to decide what is meant. In attending this workshop, I noticed that to different people it means different things which may lead us to different notions of inheritance in hierarchies.

Geffner: I can answer that; I can say that all I am doing is just ...

Gelfond: I'm not trying to argue with what you did; I'm just choosing you to make my point.

In our interpretation of an inheritance hierarchy we have different types of arrows, classes, properties, positive strict links, defeasible links and negative links. These are all translated, and I believe the translation is natural.

Once this is written you can interpret what a link is. Since we are interested in non-monotonic reasoning, we think about communication conventions. I am not claiming that my interpretation is the only way, but you have to pick something. One axiom, which is written once and for all, explains what links are. You may need two axioms; one for positive links and one for negative links. This is entirely consistent with the general way of expressing normative statements in autoepistemic logic. Do the same for temporal reasoning, or any other type of reasoning.

McCarty: You don't have an the modal operator L in your consequent.

Gelfond: No, I don't have an L in the consequent. I don't have time to discuss it at the moment.

Doyle: Is it significant that you read the first two L's as "we believe," and the third L as "reason to believe?"

Gelfond: No. This is belief for all reasonable agents.

After you know what a link is, you have to look at what to use to select among different paths. We want to be able to express whatever we want for selecting. One thing most people agree upon is Touretzky's principle, which says subsets normally are more important than supersets. This is a normative statement. It's easy to write it as a normative statement. It works out in a natural way.

What kind of predicates do you need to express? Something like this: If one class is more specific than another then you need a "more specific" predicate. The real problem is multiplicity of approaches. What does the more specific predicate depend on? Does it depend on two classes? Does it depend on two classes and predicates? I suggest you can have four-predicate strategies.

You can do at least two of the approaches that were given earlier today. They seem to work as they should on examples commonly discussed. Someone asked if you get any surprises using this approach. You said you didn't have any surprises; you got exactly what you expected, because that was exactly what you put in. My definitions are more specific.

I've said two things. Some defaults can be expressed as normal defaults, but others seem completely different. They suggest that by default you select certain explanations. You heard about this from Poole

and Brewka. We didn't know how to express them at first, in terms of normative statements. Autoepistemic logic doesn't do it well. So what we did was expand autoepistemic logic to add some explanatory power. Later, we found some way to say it in classical autoepistemic logic.

Etherington: Did this exercise in translating inheritance hierarchies into autoepistemic logic tell you anything about inheritance or did it just show you that you had this power in autoepistemic logic?

Gelfond: It told me the following about inheritance. First, you have two problems: how to translate links and how to explain what reasoning principle to use. The path-approach doesn't explain to me what links are. As a result, I have no way to tell which approach is the one I like. You can derive two approaches, based on the way I understand "more specific." One way it doesn't depend on X; it's a property of properties. You can also decide that "more specific" depends on objects. So select on this basis. Do we agree on which one is good? I suspect it depends on your design.

Pryzmusinska: You can express different approaches.

Gelfond: Yes, you probably could, but I wouldn't say that. Unless one could explain how it is related to "more specific."

Goldszmitt: I'm confused. Specificity is a predicate you use at the object level? You have a predicate that says "penguins" is more specific than "birds"?

Gelfond: I can define two predicates: C1 and C2.

Goldszmitt: You define it as a predicate at the object level. The user will have to propose it? Or you?

Gelfond: No. The user has to tell me which he wants. It is incorporated in the system (you are not supposed to write anything about it).

Goldszmitt: Suppose a two-place predicate. You deduce the specificity from the relation?

Gelfond: Of course. The user is supposed to do what I showed in the beginning.

Pearl: In the version that allows you to deduce explanations, the user has to label the links and identify which link is explanatory and which is predictive?

Gelfond: He is only supposed to tell you the kind of specificity. In general, he is supposed to tell what explanations he prefers.

Nute: I'm not sure what you gain by using autoepistemic logic instead of first order logic. If you have a predicate for it, and there are different kinds of links, then you ought to be able to write in first order logic the specificity relation you are using in the net and compute that.

Gelfond: Of course not. These are defined as defeasible links. It says "if you have no reason to believe in something"; you are using something like negation as failure. You try to show that an object is exceptional; if you cannot show it or it is not true, then you have no reason to believe. That is captured by autoepistemic logic. You need some non-monotonic formalism to do it.

Ginsberg: The reason that doing specificity or inheritance reasoning in a general framework always seemed attractive to me is because you can learn general lessons from inheritance reasoning like arguments based on more premises are somehow to be preferred to arguments based on fewer premises. I'm concerned that this careful translation of the inheritance reasoning mechanism makes impossible learning the general lessons, which is the reason you are doing it in a general framework anyway.

Gelfond: In some sense, we differ completely. I'm primarily interested in what kind of specific questions we raise from doing translation. I don't believe, at the moment, that you can define general principles with this framework. I believe that you have different types of defaults and different types of axioms.

Ginsberg: You do not believe that an argument based on many premises would be preferred to an argument based on a few premises?

Gelfond: Not always. I believe it's domain-specific.

Poole: I have a question about the Nixon diamond. What if you write that he's a quaker and you don't know he's not a pacifist, then you write he's a pacifist? You're writing something that's not true.

Gelfond: False in what sense?

Poole: "If someone's a quaker and you don't know he's not a pacifist, then he's a pacifist." That statement is not true because there's a counterexample: Nixon.

Etherington: In autoepistemic logic, because you don't know those two things, you come to know one of them.

Gelfond: The answer I'll get is "I don't know."

Ginsberg: There are two extensions.

Etherington: Those axioms aren't inconsistent.

Gelfond: You have two defaults, the defaults conflict, and the answer I get is I don't know. I don't need to know anything about it.

Haugh: I call this "path-based inheritance reasoning in ordinary non-monotonic logic" and that gets me immediately into a dispute about what path-based reasoning is. I tried to clarify that in discussions with people. I believe that my conception varies a little. It seems that when you are using references to links, in the object language or the meta language, and you work to determine what the cancellations are, what the inheritance is, then essentially you are doing path-based reasoning.

You can do path-based reasoning in a formalism such as set-theory as Touretzky and Thomason have. You can do it in autoepistemic logic, or in circumscription. It seems that those distinctions are orthogonal to the path-based versus argument-based distinction. Argument-based reasoning as far as I can tell, is distinct simply because arguments are richer kinds of relations than paths. You get more general, stricter default relationships between antecedents and consequents.

The immediate problem that you encounter with model-based inheritance with specificity, as with any non-monotonic logic, is that you have to reason about specificity relationships between links or between arguments. You need some way of doing that. You have to be able to refer, either in the meta-language or in the object language, to links or to arguments. You have to be able to say that because the antecedent subsumes the antecedent of another argument, it's more specific. Or you give a higher or lower priority to circumscribing the appropriate abnormalities. You must have some way of reasoning about paths or arguments in general. You need to supplement your existing apparatus with the ability to talk about these things.

A lot of people have done it in the meta-language, so you don't put it right into the logic. You just say, if you have these kinds of axioms and these are the priorities, you apply circumscription, etc. But you can do the same thing in the object language. You just need to reason about paths and links, or arguments and statements.

Another problem that's peculiar to model-minimization theories, is that most of them are committed to a wide range of closed-world assumptions. They assume that penguins are abnormal birds with respect to flying. Effectively you assume that you know all the penguins, or at least all of the birds abnormal with respect to flying. Minimization of abnormalities entails that no others exist than those we can prove exist.

That is a clear violation of commonsense reasoning. We don't assume we know all the penguins. We don't assume we know all the abnormal entities in order to be able to make defeasible inference. This seems to be a general, serious problem in circumscription and model-minimization approaches.

To deal with the problem you have to move away from model-minimization. Autoepistemics have been successful in doing that. Introducing autoepistemic notions will allow you to address that problem.

There are other issues regarding predicate-based representations in model-based theories and non-monotonic logics, that you don't get derived generic relations.

Grosof: What is a generic relation?

Haugh: "Penguins are non-flyers," "penguins are non-winged," "flyers ordinarily have wings," etc. A generic default is one that applies between classes or between properties.

Grosof: That's not true in circumscription. If you have no known exceptions to the rule, you get the quantified rule. For example, if you don't know of any flying penguins, the system would conclude that penguins don't fly as a rule.

Gelfond: It depends upon the set of parameters used.

Haugh: With strict links, you get derived generic relations.

Grosof: No. Even if they are default links you get them.

Recordings truncated.

Sunday Session II: Complexity and Implementability.
Speakers: Randy Goebel, Lin Padgham, Bart Selman, Lynn Stein.
Moderator: Henry Kautz.

Not all speakers transcribed.

Kautz: A session on implementability at a computer science conference seems odd. It sends shivers up and down my spine. Surely, if AI contributes anything to the study of reasoning, it is appreciation of devising computational theories.

I hope you're not satisfied with coming up with post-hoc descriptions of how a piece of reasoning might be carried out, especially if that description appeals to ill-defined or non-computational processes. Instead, what AI would like is to perform that reasoning, to describe the procedure in such detail that it could be carried out by a Turing machine or any other universal computer in a reasonable amount of time. Why would we like to do that? Because AI doesn't take computation as a metaphor for thought, but rather takes thought as a specific kind of computation, which must, therefore, obey the laws of that discipline.

We worry about complexity because a procedure which requires an exponential amount of time to compute is little better than no procedure. Complexity results play two roles. An existing tractable algorithm tells us that people could reason in such a manner. This hypothesis can be tested, perhaps linguistically, perhaps psychometrically, perhaps by cutting people's brains apart. On the other hand, if we can prove that some class of problems are inherently intractable, then we can't expect people to solve those problems. We have either to look for a more tractable subclass of problems, or weaken our notion of a solution. In general, complexity and implementability are the key constraints that AI can bring to bear on problems of any kind of reasoning of defeasible reasoning specifically.

Goebel: I agree with almost everything that Henry Kautz said. Model theory is useful to me as long as it doesn't destroy my intuition about the thing I am trying to capture. Programming languages have gotten simpler because Dana Scott's P-Omega was too hard for us to understand. It didn't mean that the model theory wasn't useful. It did the right thing. It forced us to write simpler programming languages.

Doyle: Maybe the things you are interested in are not tractable by their nature. It's still worthwhile to get an accurate specification of them prior to finding how close you are to tractability.

Goebel: Yes, that's right. Software engineers assume a specification, and talk about the problem of meeting it. But in fact, we seem to work with a sequence of specifications. We need one specification to guide the construction of another. It's like the Galois connection algebraically, you always need one thing to compare to another, to see if the one is meeting the specification of the other.

Thomason: It's not true that we aren't doing that kind of optimization in engineering, if "we" means "people in AI." A third of Raj Reddy's presidential address was about how that affected speech recognition at CMU, and Masaro Tomita's context free algorithm is another case of that. The reason people doing inheritance are not doing that is that there aren't close connections between what we do and actual deliverables.

Goebel: Sometimes that's true. On the other hand, we've had systems that do hierarchical reasoning for a long time. For the most part, all of us would agree that what those systems did was wrong.

Thomason: That's right. There is a gap between the people delivering those things and the people doing the theory.

Goebel: That's right. I think we have to address that gap.

Stein: When I first saw the set of questions for this panel, my first reaction was: these questions don't make sense. The questions of implementability and tractability aren't hard to understand. There are two cases. Most of the people here have been talking about the first case, about problems with fairly straightforward answers. Yes, we want to build a program, and we want that program to be tractable. In that case the answers are relatively trivial. In the other case, the questions are irrelevant.

When doing mathematical reasoning, or using a classical logic, if you have a set of equations and an answer, there is an ulterior way of determining whether that answer is right. You have an exact specification of the problem. In AI this is typically not true because we do commonsense reasoning. I challenge you to give an exact definition of commonsense reasoning.

To show the two levels of specification, I'll use the Traveling Salesman Problem. We have specification for the right answer: to visit all the given cities at minimal cost. Unfortunately the problem is NP-hard. Heuristics can solve this problem. We even have exponential algorithms which use these heuristics, and when the heuristics fail, we can fall back on the exponential algorithm. The heuristics may not give us the

right answer in all cases. We have two formal specifications: the answers we want and what our programs do.

Typically people won't build the Traveling Salesman Program because this problem in isolation isn't interesting.

I claim that these levels of specification, of formalization, are both important. In the case of the "right answer to the problem" we don't need to concern ourselves with implementability and tractability. In AI, agreement on the right answer is rare.

Touretzky's definition of inheritance, which we now know to be NP-complete, was proposed as a right answer. I think most of us agree that Touretzky's contribution to the field is proposing a possible right answer, even if we think his definition is not right and know it is NP-hard. We learn something. Maybe we learn that inferential distance is more right than shortest path. We need to look at the cases where inferential distance doesn't work as well. We learn a lot about the principles underlying commonsense reasoning from something like Touretzky's work.

Definitional formalism in AI is trying to get the right answers. It's nice if our definition is tractable and implementable but more importantly it should be specific, arguable, and hopefully correct.

We also need semantics for our programs. We need to formalize what our systems do. These formalisms must be implementable. This is irrelevant if you're talking about definitions and want to know the right answers.

Kautz: If we haven't defined the problem or solution, rather than come up with a definition and then figure out if we can implement it, why don't we factor in the complexity constraints when we are formalizing the problem?

Stein: That's nice if we only want to create things in a circumscribed area.

Kautz: It should be one of the factors that goes in the pot.

Ginsberg: Circumscription has done the field an invaluable service. It's helped people with intuitions about model-based semantics. It may turn out to be misguided, but it's a good thing for AI.

Stein: I would use both definitional and specificational formalism.

Kyburg: One doesn't just want to look under the street light for one's lost keys.

Geffner: Why do you think an inheritance reasoner should depend on the laws of computer science?

Thomason: We are trying to do human reasoning on a computer.

Geffner: Reasoning and understanding are two different things. We can have a theory that allows you to understand inheritance reasoning. It's NP-complete. How are you going to reason about inheritances?

Grosof: I suggest we inherit this discussion of philosophical issues about complexity and tractability from the general AI level. Between experiment and mathematical formulation we'll see what we can implement and get intuitions about what we would like to implement.

Poole: It seems that linear isn't too bad, besides the number of facts. There is the assumption that the knowledge base is wide yet shallow. So we ought to be interested in complexity with respect to the depth rather than the whole thing.

Selman: We're looking at restrictions on the depth of our hierarchy, log N's, or balanced hierarchy.

Stein: Then it has to be not intuitive.

Thomason: People are misled by focusing narrowly on the nonrelational case because they think the depth is the longest number of ISA chains. If you can construct rules, you can construct ISA chains by going up and out through rules. You can easily get chains of forty or more in natural representations. In the relational case, you're compounding rule links and relational links to get ISA links. The idea that "it's just the height of the net" disappears in the more general case.

Etherington: The reasoning you're interested in may not be the reasoning in the whole of your knowledge base. You may be NP-complete except in a small part of the knowledge base. An exponential in a small number may not be a problem. Context and focus may be more important than the complexity of the algorithm.

Selman: Finding the right focus is often the hard problem. Looking at backward versus forward chaining and when to change direction, you often find that deciding upon the right search strategy is an intractable problem.

Etherington: If you narrow the focus, you're probably going to make some mistakes. That is the tradeoff.

Ginsberg: People have the fantastic ability to trade meta-level and base-level reasoning.

Selman: If we can be precise, we will succeed. If we just say "that's a problem for heuristics," or "this is an NP-complete algorithm, therefore just focus," then I think we're on the wrong track.

Ginsberg: Deciding where to focus and how to decide where to focus are both NP-complete problems.

Sunday Provocation: Some Impossibility Results.
Provocateur: Jon Doyle.

Doyle: Due to the unpleasantly huge variety of theories about non-monotonic reasoning, people have worked to unify some of them. There has been progress but there have also been doubts expressed: take the Yale Shooting Problem and Touretzy-Horty-Thomason's Clash of Intuitions. My assignment is to investigate whether we are able to get a unified logic that will tell us what assumptions to make in all circumstances.

We use Shoham's formalism for preferenced-based non-monotonic logic and translate default reasoning questions into questions about economics' social choice theory, which is about rational group decision making. We get some results for default reasoning from Kenneth Arrow's theorem. Shoham takes ordinary logic and adds a partial order over the set of models being considered. He gets varying intuitions about satisfaction and entailment. So he gets a variant logic. For example, circumscription and default logic can be phrased in terms of partial orders. You might also use Shoham's formalism to phrase probabilistic rules. In inheritance you might use it to formalize specificity or inferability.

Konolige: Does that work? I thought there were problems with default logic.

Doyle: We are going to use his formalism, not his theory. Parts of the formalism will be variable. In the theory the order is viewed as a preference order. You can think of non-monotonic inferences as inference to maximally preferred states of beliefs. That follows the ordinary motivation for defaults, which is to eliminate or reduce cost by making inferences or analyzing information.

This idea has a long history, mainly from Pascal's wager. If there were any probability that God existed, Pascal decided it would be better for him to believe than disbelieve. In mundane situations we often draw conclusions when we don't have enough information. But if we produce default rules by summarizing rules to make decisions when we don't have enough information, we get conflicts. For example, the Yale Shooting Problem has abnormalities for different conditions. The Nixon Diamond serves as an example where different dimensions of taxonomic specificity result in conflicts. I use Ron Loui's Met's Victory Problem to illustrate probabilistic maximization, which Ben Grosof talked about earlier in the conference. The most specific reference class gives the most precise but also the least reliable information because the more specific a piece of information is, the fewer the number of pieces there are. In the Met's Victory Problem, we require three conditions. There may be only one or two games (which fit all three conditions) that we can use to predict whether the Mets will be victorious today. Also you can have conflicts from different authorities. You have to decide which arguments to believe.

To approach this problem, we think of criteria for preferring one type of assumption over another as criteria to judge arguments. We're going to aggregate these criteria into a global order which tells us what's preferred overall. There are two main approaches towards this: the credulous approach (when there is a conflict, choose to satisfy one criterion over another), and the skeptical approach (when there's conflict avoid making a preference). It's not always rational to follow either of these prescriptions uniformly. A classical example where skepticism fails is the poor donkey who starves because he can't decide whether to go the hay bail on his left or on his right.

Similarly, credulity isn't always rational because there might be bad consequences, like wrong decisions or delays. We're going to look for a language in which we can express rules or policies about when to be credulous and when to be skeptical case by case.

We'd like ground rules for aggregating these criteria. The method of aggregation should be potentially mechanizable, so we don't have to solve all possible conflicts in advance, and modular, so we can plug in newly discovered criteria without having to rewrite everything. We would like the method to be non-arbitrary.

Let's view aggregation policies as functions which order sets of rules for preference over models. For example, in the rule regarding unanimous decision, the global order prefers one model to another, just in case all of the policies or all of the criteria agree. But

since there's little agreement, that makes for skepticism. Majority voting, another rule, doesn't always give you transitive preferences, or make this order a transitive relation. There's a classical result called the paradox of voting.

Economists have studied this question under the name social choice theory. They study things like candidates in an election. They think of the what we call "individual criteria" as the ordering of each member of the group that's making the decision. The global order is a social ranking, the result of group decision. The aggregation policy is the "social choice option." The mapping between these two fields fits well with what might be a *Society of Minds* view of AI. We want this aggregation policy, or social choice function to be justifiable when we put conditions on them. I'll give you five conditions.

One is collective rationality: global order is a function of the individual order which is a function of something else. So global order should just depend on what the members of the group say, or what the criteria we are considering say.

We'd also like the Pareto principle which is of version of unanimity. It basically says that when there is no dissension on a question then the group decision ought to agree.

Another condition is independence of irrelevant alternatives: in the final ordering, the relation between two alternatives depends only on how individual criteria rank the two. Their relation doesn't depend upon how they rank among other alternatives. You can view this as a syntax-independence condition; extra predicates in the description of your theory shouldn't change the outcome.

The fourth condition is non-dictatorship: no single criterion gives you the full answer. We want to view each criterion as having part of the truth or part of the preference order, not the complete order. In the human case, this is reasonable. In a case with computers, this just says we didn't work out all possible conflicts.

The last condition is conflict resolution, which you can think of as a way of doubting skepticism. It says that when a criterion orders two alternative models, then the global order must decide to agree or disagree with that criterion. The first four principles occur in Arrow's formulation of his problem. This last one doesn't because in social choice theory, economists look at total orders. With Shoham's theory we draw on partial orders, which allows criteria to be unrelated. They don't necessarily have to be in a greater than, less than, or ranked equivalent relationship.

A slight variant of Arrow's theorem states that no aggregation policy satisfies the first four conditions because anything which satisfies the first three, any policy mapping total orders to total orders, must have dictatorship. So much for Arrow's theorem. We're interested in defaults.

We'll write default rules in a modal logic just like in autoepistemic logic. We write the default "P without Q gives R," where if P is believed, and Q isn't believed, then we conclude R. We'll use Bob Moore's models of autoepistemic logic, which are a certain part of Kripke structures. We're aiming to give a preferential interpretation to default rules. This is the basic structure: if P and Q are mutually inconsistent, then write Q is preferred to P, meaning that every model of Q is preferred to every model of P, every model of Q is equivalent to any other model of Q in the ordering, and the same for P.

With this ordering, we can express the preferences of the approach we are looking for. We can express the preferences for skepticism about P by saying we prefer "not believing either P or not-P" to "believing either P or not-P." You get credulity by reversing that. You prefer believing either P or not-P to believing neither. An important property of this interpretation is that every model we consider is either less than, greater than, or indifferent to another. Given the total order, we can prove the following theorem: no policy for aggregating default rules into a global order satisfies the Pareto conditions in the sense of the irrelevant alternatives and non-dictatorship.

Another way to think about Arrow's proof, is that we must assure we can express any preference order among candidates with default rules. You couldn't do this if all default rules expressed an identical preference order. Our paper (with Mike Wellman), which will appear in the KR89 conference, shows this.

We would like to allow more generality in the orderings we put over interpretations to allow for partial orders. We have another result which says that no aggregation policy for the general case satisfies the five conditions. This is proved by the conflict resolution principle, which allows you to transform small cases of the general thing to get the same result. What does this say? Let me hold this question.

Are there ways around this result? Economists have looked without much success. They've expanded the language of preferences to say one criterion has greater priority over another. You might say specificity always rules over chronological considerations. Adding comparisons of that form (which I think we can do in our language) doesn't get around Arrow's theorem. Adding bigger criteria intensities instead, like not just saying that this model is preferred to another according to specificity, but this model is much more specific than another, also doesn't get around Arrow's theorem. If you add both, you get around the problem at a big cost. Namely, with inter-criteria comparisons and inter-criteria intensities, the theory turns out to be equivalent to having cardinal utility (or at least

intervals of it). That solution doesn't fit well with the original motivation for default reasoning, which is qualitative. We don't want to add information equivalent to putting numbers on everything.

Economists also try to get around Arrow's theorem by limiting the domain in which preferences occurred. For example, we could forbid certain types of defaults or preference relations. Economists' results along these lines don't look like they'll carry over well to AI. So this is an interesting area to look at.

Another way around this problem is weakening the conflict resolution condition or any of the conditions. We're trying to bound the degree of skepticism. One can think an axiom is too strong. On the other hand, if you express explicit preferences for when preferences should be taken or ignored, then you'll still have the potential for conflicts at the top of this hierarchy of meta-levels.

Finally, one might abandon the idea of getting globally rational agents and accept that the agent in one circumstance will draw one conclusion and in the next circumstance draw the opposite or a logically inconsistent conclusion.

This result tends to indicate that it's unlikely that a single logic of default reasoning or inheritance will satisfy all five criteria. Given that, I think there are a couple of things we need. There is a lot of interesting stuff in economics that appears to be relevant to the problems we're discussing. We barely scratched the surface in seeing what use we can make of it.

My final note is about resource-bounded rationality because people have talked about computational agents only having a small amount of time or resourcefulness to compute things. The limitation on rationality that we expect is independent of the resources we have. Limited rationality occurs in idealized economic agents that do not have any computational or mental limitations. I think if we face this limitation in AI, (I have no belief that we will) it will have nothing to do with computational ability.

Ginsberg: I agree in many ways, but I think you are beating a dead horse. David Poole said we are trying to do more than exploit preferences on models. Everyone appears to agree that there's something extra to which we have to appeal. In terms of your theorem, the idea that the global preference criterion is only a function of local preference criteria will probably disappear. It's a function of extra stuff, like what we write declaratively, as in "this circumstance prefers this criterion."

Doyle: I don't think that will be a barrier because in this paper we proved the theorem for models in which we only talk about provability and truth. The things we're really looking at when writing preferences are much larger structures, whole sets of arguments. I see these results working for larger structures just as well as for the limited models we talked about in the paper.

Ginsberg: I expect something dictatorial to basically be right. Figure out how to do it, code it up, and there won't be all these arguments.

Doyle: If you believe that we have a dictator, then this result doesn't have force.

Ginsberg: To the extent that we are capable of intuition, there is a right method.

Doyle: That's assuming there is just one intuition.

Ginsberg: But if there are many intuitions, all of which are equally good, I think we can be happy if our machines exhibited any.

Doyle: This result would indicate it is true. There will be a lot of intuitions about what's reasonable and a lot of reasons or questions about reasons.

Etherington: Your theorems, as I read them in the paper, require that if anyone has a preference, the resulting theory has a preference. This means you expressly exclude any skeptical research in the result.

Konolige: Hear! Hear! Which is it?

Doyle: We do, at the top level. We allow individual criteria to prefer skeptical conclusions.

Etherington: But skepticism would defeat your theorem. If two criteria disagree there aren't any criteria for deciding between them, so no decision is made.

Doyle: If you want a uniform skepticism, not bounded skepticism ...

Etherington: It doesn't even have to be uniform.

Horty: You're giving unbounded skepticism, not bounded skepticism.

Doyle: It's not saying you must decide all questions. It's saying you have to decide conflicts which arise.

Grosof: That seems to be the interesting case: composing an overall preference among agents when they can conflict. When there is something that they all throw up their hands about, why should you expect a global order to generate a preference? The interesting case is exactly the one that you mentioned.

Doyle: When they all throw up their hands, you're saying they all don't make a decision?

Grosof: If everybody says "I don't care about X," and this global thing comes up with nothing, you're permitting skepticism. That's the case where nobody asserts a preference. But that's the uninteresting case. The interesting case is when people have preferences.

Doyle: We have preferences about a particular case. We add that as a new criterion. That is the sort of thing that we're trying to capture by having a set of criteria. We take any set of those criteria and map them in. One criterion might be for a conflict between these other ones, that we want to remain skeptical in that case. Another criterion might say, in other cases, we might want to actually ...

Grosof: But it seems that your conflict resolution principle there ...

Doyle: That's the point only when nobody has ruled out or decided that conflict already.

Grosof: You have two defaults. One says pacifist the other says non-pacifist. You're saying that the theory has got to come down one way or the other.

Etherington: The global criterion has to be accepted. If an idiot has a preference, and nobody else does, then you have to go with the idiot.

Doyle: That's exactly right if no information is given to the system that the idiot shouldn't be taken seriously; wiser minds are puzzled.

Etherington: You aren't allowing for wiser minds to be puzzled. If two wiser minds disagree, then you can safely rule out the idiot.

Doyle: If two wiser minds disagreed, then one has expressed the things we will go with.

Etherington: If every wise mind says I can't reach a conclusion, then they don't have a preference. Nobody has a preference except the idiot, so the idiot's decision rules.

Doyle: That happens a lot of times.

The issue you're raising is whether you have inter-criteria comparisons: one fellow taken less seriously than another. You can add that to the theory and problems reappear at the higher level because those criteria conflict. If you haven't added enough, you have a tree structure or hierarchical order for which conflicts should be resolved over others. You'll wind up reproducing the conflicts.

One way to look at this result is not as an impossibility result. The title of this paper is "impediments." This might be an impossibility result, but it will be a lot of work to look at all conditions and see whether they're completely reasonable or whether we can get around them. You can have an intuition that there may be a dictatorial theory, or a full theory of everything. If there is, this is irrelevant, but it's reasonable to doubt that there is.

Horty: There are two kinds of dictatorship: absolute and lexical. My understanding is that with Arrow's theorem you might have a lexical dictatorship, which means if the dictator (criterion) is neutral then another criterion dictates, and if that criteria is neutral there's another dictator. That kind of dictatorship might be acceptable.

Doyle: That maps into a single figure called a dictator. There's still a single answer. I submit that we don't have any ordering for these conflicts now, but one type of criterion always rules over another. We don't even have a criterion which tells us whether the Nixon Diamond should be taken skeptically or credulously.

McCarty: This is an impossibility unless AI models the ideally rational case. But characteristics like intransitive preferences and dictatorial dominance are descriptions of real cognitive agents. In social choice theory, one can study a descriptive model of how society confronts the impossibility.

Doyle: I agree, but there are people who believe there's one true logic. There's reason to doubt them. I intuitively do.

Sunday Session III: Questions of Substance or Mere Clashes of Intuition?
Speakers: Jon Doyle, David Israel, Kurt Konolige, Rich Thomason.
Moderator: David Etherington.

Etherington: The last panel of the workshop is called "Questions of Substance or mere Clashes of Intuition?" How do we distinguish progress from motion? How do we decide between the 72 choices? Are any of the issues substantive? Can we come to a conclusion and say "This is the right approach" or "For different applications, and different situations you need different approaches"? Maybe sometimes we'll be skeptical and other times credulous. Is there any core to this work?

Konolige: This is the first time I've heard one of my projects called "dictatorial." They've been called many other things, including "funny," at this workshop.

Most of you probably saw the magician at dinner last night. He handed me a deck of cards, and told me to divide the cards into three piles, turn over the top card of one pile, show it to everyone, then fold the deck back together. Then he shuffles them and picks out the correct card. I didn't know he was planning to do that. He just handed me the deck and gave me directions.

When he gave me the deck, I noticed that one card was slightly bent, so I unbent it while his back was turned. He turned over cards, asked me if one of them was my card, and it wasn't. He never found out which card was. If I had been smarter, I would have kept the "crimp" in, put it somewhere else in the deck, and made him pick out the card that I wanted him to. I'll get to the point of the story later.

I have three general points. First, I think most people would agree that we need principles to guide defeasible reasoning. How else can we say the reasoning we do is worthwhile or useful? At first, people were happy to have these systems. They thought the defeasible reasoning problem was finally solved. The reasoning was more complex, interesting, and perhaps domain-dependent than most people were (or even now are) willing to admit.

General principles of default reasoning are useful, but the ones we agree to in every case are few. Perhaps one network will agree to our principle of direct versus indirect arguments, although I have qualms with that principle. Specificity is fairly well thought of in some respects. But people have clashing intuitions on contraposition, on-path/off-path preemption, etc.

I suggest we aren't specifying everything we can specify. People assume background information and they do so in different respects. We won't be able to find any more general principles, but rather we need crimps in our deck of theoretical cards. We need places in the deck to put cards.

When the magician crimped that card he essentially could make me pick out the particular card he had in mind. I suggest that our theories need more hooks where we can put in domain-dependent information.

What types of hooks do we need? Unfortunately, people haven't worried about this issue much. One type of hook is priorities among rules. (Perhaps we've concentrated too much on simple problems, like inheritance networks or simple action theories. Even in a very simple type of temporal projection problem, like the Yale Shooting Problem, we discover that our formal systems don't seem adequate to express the subtle reasoning going on.) I've done work trying to see what kind of domain-dependent information we need in these formalisms. The priorities among rules and flexible specification of priorities are two of the most important.

We've paid too little attention to the agent-world connection. I think a lot of the clashes of intuition develop here. In the first place, why would an agent attempt defeasible reasoning? Agents must accomplish tasks in the real world. They may only have partial information about the world and need better information than they have to act upon. The second reason might be that they have resource bounds, but this wasn't a question approached at this workshop. Both incomplete information and resource bounds affect what you want from defeasible reasoning. They argue for basing more of what we think on decision-theoretic ideas, as opposed to simple arguments based on probability. When we look at resource bounds, one of the things that I'd like to take time ...

Etherington: You just hit a resource bound.

Konolige: One final thing, the type of default theories we have are generally normative. That is, we think of them as an ideal agent given an arbitrary amount of time. This is curious given that default reasoning is an approach which might have resource bounds. I don't think people can approach the difficulties involved in doing something useful with normative theories under resource bounds. Perhaps the theories we're looking for aren't normative. Questions?

Thomason: If you want people to pay attention to a paper you write, get Dave Touretzky title it. He thought up "Clash of Intuitions."

I never thought this title would suggest crisis to people. Look at the article on modal logics in *The*

Handbook of Philosophical Logic. There are as many types of modal logics as there are reasoning alternatives here. People learn to live with it. Basically, you use theorems to compare systems with respect to power. Some theorems are better than others according to the theoretical criteria in them. You can classify modal logics as normal or pathological, and switch from one to another when you are working for theoretical purposes. We never thought of applying these things, but when a user community evolved in AI, modal systems became fairly restrictive. The conflict creates no tremendous scientific crisis.

It seems a lot of these issues have been discussed at recent meetings I've been to. This one is concentrating on issues of inheritance and default reasoning. To me the scientific progress in the area has been remarkable on the theoretical end. The rate that ideas diffuse into the research communities is frightening. It's amazing how fast people are to assimilate ideas.

A lot of results show that major frameworks for dealing with the problems of non-monotonicity, ranging from probability to preference over models, have a lot in common. We have principles for comparing them. We have a common fund of examples that we can talk about. We recognize what's important even if we don't agree exactly about what we should do with these examples. I think we agree about essentials and can communicate properly. I think the theoretical side is very healthy.

I think, and this was clear at the session for implementation, we need to work on getting good connections with the user community if we want to be complete computer scientists. At some point we can't just come up with specifications and hire programming engineers to implement them. The relations are too complex to be a workable model.

We have to have one science in which people are doing the theories. I said before that I'm hoping for this because I'd like to see new ideas for logic. I think the same goes for building systems. We can't separate those projects because there are too many feedback loops in the whole process. We might be missing essentials if we're not in touch with users. If you can't write a manual to tell people how to use your knowledge representation system, you won't be very effective.

It's hard for me to feel that the problem with scaling up is just efficiency because people have built new scaled up systems and found that their algorithms no longer work. I think there's a basic representational problem that makes building these things beyond 5000 frames or so impossible. If users are screaming because, for instance, your inheritance network can't use numbers, frames, or even procedures, Lord help us. Before you invest two or three years building a system, you ought to realize that you might have to rethink it to put a few other datatypes in and reason with them. I think we need healthy relationships with the users to find out what they need.

The problems that technology rubs your nose in are not terrible as long as you don't have a deadline. Deadlines are the difficulties with real-life problems, not the problem themselves. I would like find out from the users what they need and hopefully we can continue to have the luxury of working on long-term research problems without having to build a system under a deadline, so we can't think through the theories and specifications.

I'd like to see more hooks in making actions, too. This is the point that Kurt Konolige mentioned, so I won't say much more about it. I think a lot of problems about what kind of reasoning to do might disappear if we engaged the system in decision-making. We'll find that a lot of things aren't important or that some reasoning strategies are better for some purposes than others. A lot of these problems might go away if we put them in a larger context.

I'm making a plea for the unity of the subject of computer science and the development of useful technology. I feel that related subjects, such as psychology, economics, and logic, fit together. We have to do them all at once. We have a lot to gain by being ambitious and trying to put something together on that scale.

Selman: A quick remark about complexity. I agree that we cannot yet build these systems, so I guess it's impossible to get anyone to build a large network. I don't think complexity analysis will tell you that. If somebody would compose a large network, there might be no mechanism to draw conclusions from the network. We already have large networks laying around.

Thomason: They're fairly large but I think they're representations without algorithms. Ones at the center for machine translation are getting big.

Selman: That uses fairly straight-forward ...

Thomason: Yes, straight-forward AI techniques.

Selman: They don't use something like inferential distance. If they wanted to use that, they would want to know in advance how to use it efficiently. The complexity comes in because, for example, skeptical reasoning gives a straight-forward algorithm.

Horty: It's not clear how to do inferential distance except by an algorithm that lists all paths. I don't think they'd want to do that.

Thomason: I talked to them and they couldn't care less about complexity. I don't know if that speaks ...

Selman: That's because they're not trying to do what we are. In fact, they might not care about inferential distance either.

Thomason: I actually think that complexity is an important consideration. I waffled about how basic complexity is for AI. I don't know what I think of it myself, except that it's important. I want to know the complexity of the algorithms I have. It matters, although I'm not sure how much. The issues keeping us from scaling-up knowledge representation systems go beyond complexity.

Grosof: Another important question, as far as hooks go, is "what kind of integration do we want, or need between a large inheritance net and other kinds of default reasoning?" That's a challenge for the approaches which talk about embeddability. Many computational questions will be substantially modulated by how tight they are bound to other classes of default reasoning. Why haven't we seen really large inheritance nets? Was there ever a knowledge base big enough to talk tens of thousands of inheritance links? We'd want to do more general kinds of default reasoning at the same time. We won't want to stick to pure inheritance.

Ginsberg: This is what I was saying – people working on general methods make sure they can take advantage of the general methods of their logic.

Thomason: We're thinking of hybrid systems, expecting that the inheritance reasoning will be fast. Presumably in an expert system, in an inner loop, or in something of that sort, people don't use general-purpose methods if they can find fast special-purpose methods. You have to find fast special-purpose methods for those things. But, you might have other kinds of reasoning, things you can do off-line. I'm for a single system where you can do both special and general reasoning. If you have more complicated reasonings in your expert system, then it better be efficient. Even if it isn't inheritance reasoning, it better be efficient. I'm not sure how to define inheritance reasoning. I guess it's a visual representation at the low end of the complexity scale. Maybe you're talking about something I would call inheritance in these more general methods.

Grosof: Two of the few things that we agree on about inheritance are that acyclicity and implicit specification of priorities are important constraints. They are violated for much of the reasonings we want in our systems. That's maybe the hairier part of what could be called inheritance.

Thomason: I don't know about implicit priority. We might debate that. We couldn't agree more about cyclicity. But our need for cyclicity is fairly restrictive. We need it to define inheritance, but you can tell how inheritance ought to go in most examples of cyclic nets that you can write down. We would like to figure out how to relax the cyclicity condition and still be able to define inheritance. The mathematical problems are horrible. We haven't gotten far.

Etherington: Let's go to Jon Doyle.

Doyle: How do we tell where we're aiming? A lot of people have been trying to use their intuition to determine what properties systems ought to have. That's a wonderful idea because intuitiveness is an admirable property, but it has limits, particularly if we can't come to a consensus or if our intuitions are not stable over time. Judea said he had an intuition about one thing a year ago, and now he doesn't. The more we learn about things, the more we may change our intuitions.

I suggest in the future we investigate sources of intuitions. We want to think of this as probabilistic reasoning or autoepistemic reasoning. I suggest underlying intuitions represent different ways of economizing mental resources or the cost in reasoning. I agree with both Kurt Konolige and Rich Thomason in parts of what they said. The fundamental way of giving rise to intuitions is being interested in different architectures for reasoning. Each is represented in operations of various sorts. All have various costs and benefits.

We ought to look for rational or reasonable reasoning, meaning that the agent's reasoning is guided rationally by its preferences. The agent should take steps of reasoning it thinks are better rather than ones that it thinks are worse. Ideally, we'd like those choices to be optimal with respect to its preferences and beliefs, but that's not feasible.

You can expect many variations in intuitions. We might be interested in a lot of different architectures: human, machine, and reasoning architectures. Humans have different personalities. Some people miss inferences that others get. Do people use various architectures? Even with individuals, learning more about a subject can change the cost and benefits they get out of reasoning on particular things.

Even if you fix all those variations, there are a bunch of others that don't go away easily. One is that in different domains, you have different expectations about what types of reasoning are valuable. Even within

different purposes in a given domain, that is if you're trying to discover, diagnose, or explore something, you might have different expectations about what inferences are reasonable.

Even if you fix those, you might have multiple nondeterminisms. Just in the rational sense, that would give you multiple, equally reasonable, paths, sets of paths, or sets of conclusions. The multiplicity is multiplied. Suppose non-optimal control of reasoning. There are a lot of ways of limiting rationality. We can expect a lot of variation.

From the economic point of view, it's natural to view defaults as summaries of judgements of expected utility. That is, when you want A to inherit from B, or A to be B by default, or something like that, you make a statement like "this is a useful inference." That's the fundamental statement you're making. You may be grounding that statement in others like "it's likely" or "it's logically true."

The one thing you know is that you expect this to be useful to you. The reasoner makes up preferences and accumulates them over time. That's how you can think of defaults and other information we'd like to put in. They could be inconsistent because you select and formulate them in a limited purview. Our theories of interpretation must combine everything we might think of as revealed preferences into overall thinking.

There are 72 theories. Or, are there 72 patterns of probabilities and utilities about computational costs and benefits for different domains and different circumstances? Of course no one has really argued for 72.

Etherington: It's the lower bound.

Doyle: I was talking about that before. We will be stuck with only special-purpose theories of various forms. Each is good in a particular area. There's no general way to decide in advance which is best in particular circumstances. Can we formalize desiderata that inference systems should have, see whether they are jointly satisfiable, and ask whether there's a finite number of them, as in the Arrow's case. You can make up more cases just for social choices whether there's an unlimited or finite number.

Poole: One thing you mentioned was some way of jumping to conclusions. I've never seen a cost-benefit analysis of thinking and acting. It seems that you want to think more of the decision you make when you act.

Doyle: There's been a little work on that. People have recognized the problem. I.J. Good wrote about it several decades ago. Right now in AI, there's a fair bit of interest in it. One of the Spring Symposia was on "AI and the Limit of Rationality." Eric Horvitz, Stuart Russell, and a few others explicitly tried to assess the cost/benefit of thinking more. I think it depends a lot upon the architecture and the things you think of as costs and benefits. An essential problem is also deciding what things count as costs and what things count as benefits.

Thomason: Also, it takes a certain amount of time to figure out this stuff.

Doyle: If you have unlimited thinking ability ...

Konolige: One explanation for why plans get into the philosophical literature is that supposedly agents wouldn't need plans if they were not resource-bounded because at each moment they would evaluate their desires and beliefs and figure out what to do. That's hard to do on the fly. So follow your plans for what to do in a simplified matter without thinking much about it.

Israel: You believe things precisely to commit yourself because you want to be committed. Many of us probably will be. You want to be committed to block getting into alternative epistemic states. You hope your choices of blocking certain epistemic states and precisely committing to another are okay.

Konolige: That's an argument for credulity as opposed to skepticism.

Israel: It's an argument for the necessity, in a resource-limited agent, for beliefs as opposed to degrees of confidence. It's not an argument for credulity versus skepticism because it doesn't tell you that you should believe this or that. It just says that it's good to have all-or-nothing beliefs. If you want to build a resource-limited agent, you should stick that functionality into it.

Pollock: I don't see that this is an argument for all-or-nothing beliefs. It's an argument for belief but I don't see why the conclusion couldn't be probabilistic.

Israel: The conclusions distinguish between all-or-nothing beliefs.

Pollock: It's an argument for storing data, but I don't see why you couldn't store it with a number as opposed to all-or-nothing.

Israel: I'm not going to give this story because it's long. You want them to be strong so that other desires become ruled out. As opposed to just desiring something, you intend to bring it about.

Konolige: But your intentions have to be consistent.

Israel: Right, your intentions have to be consistent unlike your desires. They can be highly inconsistent. The idea is to channel your mind in a certain way. I think a similar argument can be made for beliefs, that is, for a state that is believed probabilistically. I think there is a similar argument for all-or-nothing beliefs, the contents of which can be probabilistic.

Grosof: In response to the question David raised (how is a decision-theoretic or use-oriented analysis a justification for having default rules?), there's literature about acceptance in philosophy and in decision theory.

I don't think a pressure to act removes us from a situation where we have all-or-nothing beliefs or degrees of beliefs. You can say we have to act, so we'll commit. By evaluating why it makes sense to have a default, we're missing a parameter about whether we should be credulous and commit under certain circumstances. To address this issue, we have to take the principles of compiling defaults, using utilities from probabilistic information or other defaults, or of compilation of knowledge into other representations. That's why we haven't seen systems learn defaults. We lack by having no principles to guide the learning of default rules.

Israel: With respect to the general area of defeasible reasoning, I urge people to look into special subjects of little inheritance networks. Even if you find that they're not important components of human reasoning, it might be a good thing to do with respect to research. In the general area we do not have a good specification of the problem. You can't ask the question yet. In what class is the problem? NP, P, Super-P? You don't know the problem yet. (I disagree with Henry Kautz on this.) So you shouldn't consider the class of the problem while deliberating about this specification. You shouldn't say the specification of the problem better be linear or P with low overhead. I think that's a mistake. (It's not a mistake to consider tractability important in AI.)

Imagine that complexity theory had arisen before 1928-1930, before the precipitation of background stuff such as first order logic, which is the theory of human deductive reasoning. My own view is that people reason deductively once a week for about a minute. Then competence theory with deductive ability came along. Consider what we'd have come up if tractability folded in earlier. We wouldn't have come up with relevance logic. It's intractable. Von Wright noted that relevance logic is the first natural propositional logic to be proved undecidable.

If you do inheritance, I think path-based or the argument-based stuff is good. If you look toward defeasible reasoning, argument-based stuff is nice, that is, forget for the moment about semantics. This is an odd thing for me to say, given the other things I've said about this. But forget about semantics for a while and look at arguments as concrete instructions. I hate to use that word because that's finite and has a "proof-theoretic" structure in a general sense.

The last thing to talk about are the "clashing intuitions." Rich mentioned the proliferation of modal logics. At a Symposium titled "Is there One Correct Modal Logic" or "Is there One True Modal Logic" at an Aristotelian society meeting, E.J. Lemmon and somebody else argued. This was pre-Kripke. Lemmon, one of the most reasonable people who ever lived, said there are different notions so you need different modal logics. That's kind of obvious today.

I'm not sure this is the case here. I think there might be more substantive clashes. People are looking to different notions and have different logics, but it's the same notion.

Etherington: When you say the clash is substantive, do you mean that it needs to be resolved?

Israel: Yes. It's resolvable in a way different from aggregating.

The payoff of reasoning is in action. Because a small set of weird species on a bizarre planet spend time looking for truth for its own sake is irrelevant. Reasoning is to guide action. As Rich said yesterday: "when the rubber hits the road ..." If you embed inheritance reasoners in larger acting systems, the intelligence of their actions could be judged. I think you may get good checks, further information to resolve clashes. My skepticism is that it may introduce so many complicating parameters that it will be hard to figure what's going on.

Thomason: I think you're right about the analogy to modal logics. I said, with my experience from modal logic, it doesn't bother me. That's just history, right? You can put many modal logics together in one big system and say "I'll use this for knowledge, this for obligation, and so forth." But we can't put all 72 varieties of inheritance in one system. Even if we could, I'm sure we wouldn't want to. It's more of a choice between alternatives. I take your point.