Preferences as Conclusions

Richmond H. Thomason
Artificial Intelligence Laboratory
University of Michigan
130 ATL Building
1101 Beal Ave.
Ann Arbor, MI 48109-2110
rich@thomason.org

Abstract

This paper is an informal and preliminary investigation of how in a logical framework we might account for how we come to have a large, complex system of conditional desires. I try to show how work that has been done in the formalization of causal reasoning might be adapted to show how this is possible.

Introduction

This paper is highly appropriate for a workshop, since it represents preliminary and tentative thoughts on what seems to me to be a difficult and important problem. The preferences that operate in realistic examples of practical reasoning are frequently novel and highly circumstantial. (Consider, for instance, my preference, in planning a trip that I have never made before, to rent a car rather than to use public transportation, and my revision of this preference after a discussion with the rental agency about available cars and prices.) Where do these preferences come from? How can they be concluded from beliefs and from preferences that are more stable?

I don't yet have a satisfactory solution to this problem. The purpose of this paper is merely to describe the problem, to relate it to ideas in the literature, to indicate some constraints on possible solutions, and to solicit help from the readers of this paper.

This question arises out of my recent attempts to advance the approach that is taken in (Thomason 2000) to practical reasoning based on defeasible desires. I'll begin by summarizing the ideas of that paper.

The distinction between *prima facie* and all-things-considered attitudes is crucial. I model *prima facie* beliefs and desires as defaults, using the approach to nonmonotonic logic known as Default Logic. All-things-considered beliefs (representing actual epistemic commitments) and all-things-considered desires (representing goals) are selected by choosing an extension that is allowed by the logic; this is done in terms of a prioritized formulation of Default Logic. In designing the system, it is crucial to avoid *fallacies of wishful thinking*, in which assumptions are justified merely by a desire for them to be true.

The approach only makes sense in the context of a more general planning formalism, in which the desires and beliefs relate to alternatives that can be reached by performing series of actions. (Thomason 2000) develops such a formalism.

I'll illustrate the ideas with a blocks-world example. The artificiality of the example is compensated by the fact that in this familiar domain it won't be necessary to explain the part of the domain having to do with preconditions and effects of actions. Suppose there are three blocks, a, b, and c, and that the agent believes that a is on b and that c is on the table. Putting a on the table costs $2. The agent would like to have b on c. The agent would not like to spend money. In this case, the theory generates two extensions, if plan length is limited to at most two steps. In one extension, the agent does nothing. In the second, the agent puts a on the table and then puts b on c. This logic is purely qualitative—it does not calculate tradeoffs, but generates alternatives that could be passed on to a tradeoff reasoner. The use of extensions in combination with default reasoning produces an effect that resembles dominance reasoning; for instance, the plan of putting a on c in this case does not correspond to an extension because it is dominated by the plan of doing nothing.

Assuming that this work at least provides a plausible specification of the practical conclusion sets that can validly be drawn from a set of conditional desires and a set of beliefs (some of which may also be default conditionals), and shows how to relate the consequence relation to the traditional planning paradigm, two fundamental problems remain:

(1) What would algorithms capable of performing this reasoning be like?

(2) Where do the conditional desires come from?

I will concentrate here on the second question, which seems to be more fundamental: we already have an idea of how to address nonmonotonic reasoning algorithms.[1] And certainly, no algorithm for doing the required nonmonotonic reasoning would be useful without a rich source of default desires.

Conditional Desires

The theory of (Thomason 2000) requires a large number of statements of *conditional desire*. For instance, the example

[1] See, for instance, (Lifschitz 1987; Ginsberg 1989; Przymusinski 1989; Niemelä 1995; Beneliyahu & Dechter 1996; Cholewiński *et al.* 1999).

that is elaborated in the firstsections of the paper requires the following premises.

Example 0.1. *Desires and beliefs on beginning a hike.*

1. I think it will rain.
2. If it rains, I'll get wet.
3. (Even) if it rains, I wouldn't like to get wet.
4. If I get wet, I'd like to change into dry clothes.
5. If I change into dry clothes, I'll have to walk home.
6. If I walk home, I'll have to walk an extra two hours.
7. I wouldn't like to walk an extra two hours.

These premises are then formalized as follows, where $\overset{B}{\hookrightarrow}$ represents default belief (these are simply rules, as in ordinary Default Logic, so they are not first-class logical citizens) and $\overset{D}{\hookrightarrow}$ represents default desire.

Example 0.1, continued. *Formalizing the Premises*

1. $\top \overset{D}{\hookrightarrow} \texttt{Rain}$
2. $\texttt{Rain} \overset{B}{\hookrightarrow} \texttt{Wet}$
3. $\top \overset{D}{\hookrightarrow} \neg\texttt{Wet}$
4. $\texttt{Wet} \overset{B}{\hookrightarrow} \texttt{Change-Clothes}$
5. $\texttt{Change-Clothes} \overset{B}{\hookrightarrow} \texttt{Home}$
6. $\texttt{Home} \overset{B}{\hookrightarrow} \texttt{Walk-Two-Hours}$
7. $\top \overset{B}{\hookrightarrow} \neg\texttt{Walk-Two-Hours}$

Our abilities to create such conditionals are astonishingly productive and flexible. Obviously, we don't remember all of them, and we don't simply make them up arbitrarily. The challenge, then, is to describe reasoning mechanisms that are capable of deriving a wide range of conditionals of this sort as conclusions.

Philosophical Background

One of the oldest philosophical traditions in ethics and practical reasoning deals with the origins of appropriate maxims for action. But this tradition is mainly concerned with whether there is an ultimate human good, and—if there is such a good—with how to resolve competing claims for what it is. These high-level issues do not begin to provide an adequate account of how we find appropriate desires to guide our conduct in specific circumstances. Take, for example, the utilitarian solution to the problem of the ultimate human good—it is whatever induces the greatest utility for a given population. This has the advantage of being less vague than many of its competitors. But still, it provides almost no specific guidance.

To further refine the recommendation, let's suppose it is interpreted decision-theoretically, say, using the model of (Jeffrey 1965). Even then it can't be applied, because we can never be in a position to calculate the relevant utilities and probabilities. These methods can be applied with some success in policy decisions where welfare can be modeled and probabilities estimated. But we don't have, as far as I know, a satisfactory account of how this modeling and estimation can be done; and it is highly unlikely that it can be applied to every practical problem-solving situation.

Deontological work in ethics, such as (Ross 1930), exhibits a rich, detailed system of fairly high-level conditional imperatives. But it is less helpful about how such a system is derived, or about how to derive the conditional imperatives that actually apply to novel, detailed practical circumstances.

The conclusion appears to be that the ethical tradition in philosophy doesn't provide useful clues to the solution to our problem.

Logical Background

The history of the problem of deriving conditionals has taken several turns. In mathematics, and with the traditional logical theory of 'if', deriving a conditional is unproblematic. Because conditionals are left-monotonic, any argument from information contained in the antecedent to the consequent will suffice.

But when the logic of "counterfactual" or left-nonmonotonic conditionals was introduced over thirty years ago (Stalnaker 1968; Lewis 1973), the problem was reintroduced, since $A \rightarrow B$ does not allow $[A \wedge C] \rightarrow B$ to be inferred in these logics. Thus, if we are in a practical situation in which we know A holds, we can't in general be sure that B follows from a background theory containing the conditional $A \rightarrow B$, because, where C compiles all the conditions other than A that obtain in the given situation, we can't derive $[A \wedge C] \rightarrow B$.

In fact, the recent logical tradition has been far more successful with the ontological question of under what circumstances a conditional is true than with the epistemological question of what reasons we can have for believing conditionals. The school of thought that relates conditionals to conditional probabilities ((Adams 1975) and the work deriving from it, for instance) does address the epistemological issue. But the ideas are at best suggestive, because the relations between conditionals and conditional probabilities are problematic, and because it isn't easy to say what reasons we can have for probability judgements in specific circumstances—especially, in novel circumstances where we do not have a good statistical model. Nevertheless, I believe that recent ideas about how we can infer probabilities provide the most promising way to approach the corresponding problems about conditional imperatives.

More Recent Work

Although, as far as I know, the problem of how to derive useful conditionals was not recognized as a serious problem until computational logicians began thinking about conditionals, it was there from the first. There has been some exploration of nonmonotonic logics involving left-nonmonotonic conditionals in which $[A \wedge C] \rightarrow B$ can be inferred by default from $A \rightarrow B$; see, for instance, (Asher & Morreau 1991), but these logics are complicated to a problematic degree, and until they are better understood, I prefer to explore alternative approaches.

There is a school of thought according to which, as far as reasoning is concerned, common sense conditionals are a more or less hopeless mess; (Goodman 1955) is still the most persuasive presentation of this point of view. I believe that advances in formalizing logics of counterfactual conditionals deals well with the problem of the nonconditional consequences of conditionals. From a contemporary perspective, I think that the main remaining problems raised by Goodman concern how to reason *to* conditionals, and can be divided into the following two questions:

(1) Under what conditions can we infer a conditional from other conditionals? In particular, when are we justified in inferring $[A \wedge C] \rightarrow B$ from $A \rightarrow B$?

(2) Under what conditions can we infer a conditional from other evidence that may be available in the reasoning situation?

Although I am using intuitions from conditional logics, in what follows I am concentrating on a formalization of conditionals as default rules. Here, the form of conditionals is greatly simplified (there can be no nested conditionals), and conditional inference will be a matter of inferring a default rule from premisses that may include monotonic formulas as well as default rules. In other words, we are thinking of inferring defaults from default theories. Ultimately, a formalization of conditional inference in this format would take the form of a strengthened definition of the extensions of a default logic. In the simplest form of such a strengthened definition, the extensions are formulated using the closure of the actual defaults under the inferred defaults; but there are reasons to consider more complex definitions. I will not go into this matter here.

To a large extent, much of the foundational logical work in logical AI and common sense reasoning has been concerned with these problems—though often, the work is not presented in the form of an explicit solution to them. I want to argue that two strands of work are particularly important in this respect.

Work on formalizing independence

It is very plausible to say that $[A \wedge B] \rightarrow C$ follows from $A \rightarrow B$ when (1) it is possible that $B \wedge C$ and (2) C is independent of B. For instance, suppose that I believe that if it rains I'll get wet, that (1) it's possible that I have a dime in my pocket and I'll get wet and that (2) whether I have a dime in my pocket is independent of whether it will rain. Then I should believe that if it rains and I have a dime in my pocket, it will rain.

Here, we can exploit work in AI that deals with inferring causal independence. The primary goal of (Pearl 2000) is to convince the statistical community of the value of reasoning explicitly about causal notions. An important part of the case Pearl makes for this conclusion is an extended investigation of how causal independence can be represented in diagrammatic form and used in statistical inference. Pearl makes a very convincing case that methods used by many scientific communities can be used to model causal independence in familiar cases. I believe that we can use this independence more generally, to infer conditionals by means of rules like the one I just described. This includes not only plain declarative conditionals, but conditional preferences and conditional desires.

The other ingredient in the above inference, possibility, can also be approached in familiar ways. In many reasoning environments, a failure to derive a contradiction from $B \wedge C$ or the construction of a model that satisfies $B \wedge C$ can be taken to establish the possibility of $B \wedge C$ for the purposes of this conditional inference is concerned.

CP-Networks

CP-networks, a graphical representation and reasoning system described in (Boutilier *et al.* 1997), exploit graphically represented independence information in much the same way as Bayesian nets to permit the derivation of conditional preferences. CP-networks support useful algorithms for inferring conditionals; there are more recent complexity results in (Domschlak & Brafman 2002).

However, there are important differences between the way I think about the role of desires in practical reasoning and the CP-network formalism that make it impossible to apply the framework directly to the problem as I have formulated it. I think of all-things-considered preferences (or wants) as derived by a process of tradeoff resolution from desires which, like defaults, can compete. Preferences in CP-networks are consistent, however, and can't compete. Because they can't compete, they can't be thought of as desires.

It may be possible to develop a framework similar to CP-networks that provides for the derivation of defeasible, competing desires. Such a framework would be a first step towards a solution to the problem of inferring desires. At present, I have not tried to work out the details, and—although this approach seems promising—I don't know how it will work out.

Conditionals and Frame Reasoning

There are also some useful analogies between inferring a declarative description of a future state, and the role of frame-based reasoning in such inferences, and inferring the value of a future state. And the analogies extend to conditionals.

In some of his recent work, John McCarthy has suggested that frame reasoning is a way of inferring conditionals; see, for instance, (McCarthy & Costello 1998). Frame reasoning certainly has something to do with conditionals; a frame axiom like the one in the Yale Shooting Problem essentially says that since the pistol is loaded now, it will be loaded even if I wait; and 'even if' conditionals are conditionals. But I don't see how to use this idea very generally, since frame axioms are connected to actions, and actions do not take us from arbitrary counterfactual situations to counterfactual situations. However, it is useful, I think, to use a situation calculus-like formalism in formalizing inference to conditionals. The resulting ideas are reminiscent, in a general sort of way, with frame reasoning.

Let's generalize the familiar situation-calculus formalism to one that uses *cases*. (I'll use this term because I don't want to use terms like 'context' or "possible world', and to indicate that we are trying to formulate some aspects of case-based reasoning.) Like situations, cases are primitives, and there is a `Holds` relation between contexts and fluents.

Let me illustrate how causal reasoning could be used to produce standing conditional desires with the hiking example. Suppose that I can learn about my likes and dislikes by experiencing my reactions to instances. I am hiking, for instance, it rains, and I get wet. I catch a cold and experience a great deal of discomfort, which I attribute to getting wet, to the absence of equipment that would let me get dry, and to cold weather. It is important to note that here I am assuming that the reasoning behind the causal attribution is causal. The attribution takes the form of a causal explanation. The causes are divided into standing conditions 'I am hiking', 'the weather is cold', 'I have no means of getting dry' that are natural language statives, and a triggering event, expressed by 'it rains'.

(i) c = hiking case, April 1996.
$\mathtt{Holds}(c, \text{I have no rain gear})$.
$\mathtt{Holds}(c, \text{I am hiking})$.
$\mathtt{Causes}(c, \text{It rains}, \text{I get wet})$.

(ii) c = hiking case, April 1996.
$\mathtt{Holds}(c, \text{Weather is cold})$.
$\mathtt{Holds}(c, \text{I am hiking})$.
$\mathtt{Holds}(c, \text{I have no means of getting dry})$.
$\mathtt{Causes}(c, \text{I get wet}, \text{I catch cold})$.

This experience is limited to one case. But causal explanations are intended to be general. I will suppose that they give rise to standing conditionals. The advantage of formalizing things this way is that we can look at the conditionals as default generalizations over cases. (I use the rather vague term 'subsequently' to avoid getting into complexities of temporal reasoning that I think are irrelevant here.)

(i′) $[\mathtt{Holds}(c, \text{I have no rain gear})$
$\wedge\, \mathtt{Holds}(c, \text{I am hiking})$
$\wedge\, \mathtt{Holds}(c, \text{It rains}] \overset{\mathrm{B}}{\hookrightarrow}$
$\mathtt{Holds}(\mathtt{subsequently}(c), \text{I get wet})$

(ii′) $[\mathtt{Holds}(c, \text{Weather is cold})$
$\wedge\, \mathtt{Holds}(c, \text{I am hiking})$
$\wedge\, \mathtt{Holds}(c, \text{I have no means of getting dry})$
$\wedge\, \mathtt{Holds}(\mathtt{subsequently}(c), \text{I get wet}\,] \overset{\mathrm{B}}{\hookrightarrow}$
$\mathtt{Holds}(\mathtt{subsequently}(c), \text{I catch cold})$

Suppose, now, that the hiker is in a different context which is similar to the other one.

(iii) c = Current hiking case.
$\mathtt{Holds}(c, \text{Weather is cold})$.
$\mathtt{Holds}(c, \text{I have no rain gear})$.
$\mathtt{Holds}(c, \text{I have no means of getting dry})$.
$\mathtt{Holds}(c, \text{I am hiking})$.

Now, in general we can't infer $A \rightarrow B$ from B and $[A \wedge B] \rightarrow C$ in conditional logic. But this too is one of the inferences we can make if A and B are causally independent. So, on the assumption that the three premisses of (iii) are causally independent, we can infer the following about the current hiking case. (Note again that the standing conditional (ii') applies in the current case because it is a generalization over cases.)

(iv) c = Current hiking case.
$\mathtt{Holds}(c, \text{I get wet}) \overset{\mathrm{B}}{\hookrightarrow}$
$\mathtt{Holds}(\mathtt{subsequently}(c), \text{I catch cold})$

The experience of the cold was unpleasant; again, there is a causal relation between the cold and the unpleasantness. This, finally, will induce the inference of a desire (which, under the circumstances, is unconditional).

(iv) c = Current hiking case.
$\top \overset{\mathrm{D}}{\hookrightarrow} \mathtt{Holds}(\text{I don't get wet})$

This is the place where direct, emotional reactions to circumstances produces a desire. This could be treated as a logical inference by introducing a special constant V for "undesirable", like the constant that is sometimes used for similar purposes in deontic logic. But at the moment, I prefer to regard it as a direct mechanism that creates standing desires out of conditional beliefs together with conclusions that are somehow labled as directly undesirable.

Note that this particular inference is simplified by the fact that, since conditional desires are treated as *prima facie* attitudes, we expect them, in general, to conflict. When we infer

$$A \overset{\mathrm{D}}{\hookrightarrow} B,$$

because of a negative reaction to a causal consequence of B in circumstances A, there is no need for consistency checks; we may very well infer

$$A \overset{\mathrm{D}}{\hookrightarrow} \neg B$$

from a negative reaction to some other causal consequence of B in circumstances A.

Note that the final default desire that is inferred here is the premise $\top \overset{\mathrm{D}}{\hookrightarrow} \neg\mathtt{Wet}$ that I needed in the earlier paper.

This example involves a direct experience, but we are able to enlarge the conditional desires we can learn in this way by imagination: we construct hypothetical situations, and arrive at conditional desires by considering the vicarious reaction we have to the hypothetical circumstances. This is not as vivid as real experience, but is far safer.

Conclusion

The main answer that I am currently exploring to the question of how we seem to be able to infer such a large and flexible set of conditional desires is (like so many current approaches) causal. We are constructed to recognize the causal features of our experiences. This structures them in a way that supports general conditional beliefs, applicable to a wide variety of circumstances. We have direct reactions to many simple propositions; we can use these to form conditional desires from conditional beliefs.

I realize that, presented in this form, the ideas are abstract and sketchy. But the abstractions I am using actually arose out of a consideration of many reasoning examples, and I believe that they do provide useful suggestions about how to carry out the reasoning. The problem, of course, is to provide more detail about the reasoning mechanisms that could realize this sort of reasoning, with the ultimate goal of producing a testable reasoning architecture. This is not a short-range task, but I am working on some of the simpler reasoning mechanisms, and I hope to have something specific to say about at least some of the details of the reasoning architecture very soon.

References

Adams, E. W. 1975. *The Logic of Conditionals*. Dordrecht: D. Reidel Publishing Co.

Asher, N., and Morreau, M. 1991. Commonsense entailment: a modal theory of nonmonotonic reasoning. In Mylopoulos, J., and Reiter, R., eds., *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, 387–392. Los Altos, California: Morgan Kaufmann.

Beneliyahu, R., and Dechter, R. 1996. Default reasoning using classical logic. *Artificial Intelligence* 84(1–2):113–150.

Boutilier, C.; Brafman, R.; Geib, C.; and Poole, D. 1997. A constraint-based approach to preference elicitation and decision making. In Doyle, J., and Thomason, R. H., eds., *Working Papers of the AAAI Spring Symposium on Qualitative Preferences in Deliberation and Practical Reasoning*, 19–28. Menlo Park, California: American Association for Artificial Intelligence.

Cholewiński, P.; Mikitiuk, A.; Marek, W.; and Truszczyński, M. 1999. Computing with default logic. *Artificial Intelligence* 112:105–146.

Domschlak, C., and Brafman, R. I. 2002. CP-nets—reasoning and consistency checking. In Fensel, D.; Giunchiglia, F.; McGuinness, D.; and Williams, M.-A., eds., *KR2002: Principles of Knowledge Representation and Reasoning*. San Francisco, California: Morgan Kaufmann. 121–132.

Ginsberg, M. L. 1989. A circumscriptive theorem prover. *Artificial Intelligence* 39(2):209–230.

Goodman, N. 1955. *Fact, Fiction and Forecast*. Harvard University Press.

Jeffrey, R. C. 1965. *The Logic of Decision*. New York: McGraw-Hill, 1 edition.

Lewis, D. K. 1973. *Counterfactuals*. Cambridge, Massachusetts: Harvard University Press.

Lifschitz, V. 1987. Computing circumscription. In Ginsberg, M. L., ed., *Readings in Nonmonotonic Reasoning*. Los Altos, California: Morgan Kaufmann. 167–173.

McCarthy, J., and Costello, T. 1998. Useful counterfactuals and approximate theories. In Ortiz, Jr., C. L., ed., *Working Notes of the AAAI Spring Symposium on Prospects for a Commonsense Theory of Causation*, 44–51. Menlo Park, CA: American Association for Artificial Intelligence.

Niemelä, I. 1995. Towards efficient default reasoning. In Mellish, C., ed., *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 312–318. San Francisco: Morgan Kaufmann.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge, England: Cambridge University Press.

Przymusinski, T. C. 1989. An algorithm to compute circumscription. *Artificial Intelligence* 38(1):49–73.

Ross, W. 1930. *The Right and the Good*. Oxford: Oxford University Press.

Stalnaker, R. C. 1968. A theory of conditionals. In Rescher, N., ed., *Studies in Logical Theory*. Oxford: Basil Blackwell Publishers. 98–112.

Thomason, R. H. 2000. Desires and defaults: A framework for planning with inferred goals. In Cohn, A. G.; Giunchiglia, F.; and Selman, B., eds., *KR2000: Principles of Knowledge Representation and Reasoning*, 702–713. San Francisco: Morgan Kaufmann.