Towards a Logical Theory of Practical Reasoning

Richmond H. Thomason

Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA 15260
U.S.A.

1. Introduction

From the very beginning, logicians have counted *practical* (or action-oriented) reasoning as well as *theoretical* (or belief-oriented) reasoning as part of their subject. However, despite a tradition that continues to the present, logicians have not produced formalisms that could be considered of any use in designing an agent that needs to act intelligently, or in helping an intelligent agent to evaluate its reasoning about action. In contrast, the decision-theoretic paradigm that grew out of the economic tradition is widely applied in many areas, including AI, and has dominated recent philosophical thinking about practical reasoning, without serious competition from the logical tradition.

This lack of progress is largely due to the unavailability of qualitative mechanisms for dealing with the appropriate inference procedures. To handle even the simplest cases of practical reasoning, it is essential to deliver a reasoning mechanism that allows practical conclusions to be nonmonotonic in the agent's beliefs. (If I have decided to drive my car to work rather than to ride my bicycle, I may well want to to withdraw this conclusion on learning that the car has a flat tire.) And, until recently, the only way to formalize inference procedures with these characteristics has been to use probability functions. Therefore, quantitative utility theory has remained about the only game in town.

2. The expected-utilitarian paradigm

In other presentations of these ideas, I have discussed shortcomings of the Bayesian paradigm, according to which practical reasoning is devoted to the calculation of expected utilities, and to the maintenance of the probability functions that enable this calculation. In this version I'll take for granted the need for a more qualitative account of practical reasoning and will point out some general features of the Bayesian approach that I would consider to be essential to any successful approach to practical reasoning.

Any adequate account of practical reasoning has to model beliefs and desires, and their interaction in decision making. Regarding belief, any theory, even if it relaxes the quantitative commitments of decision theory, should:

1. Provide for belief kinematics, in allowing a belief update function to be defined.
2. Allow realistic domains to be axiomatized. The information needed to support and update beliefs should be acquirable in some practicable way in a form that the reasoning system can use. For an artificial reasoning system, this information should either be machine learnable or knowledge engineerable.

Regarding desire, an adequate theory should:

1. Distinguish between desires that are immediate (below, I will call these immediate desires *wishes*), and reasoned desires that depend on wishes and beliefs (below, I will call these *wants*). Reasoned desires should approximate intentions, and be more or less directly connected to actions.
2. Provide an account of the central inferential processes in practical reasoning that infer considered from immediate desires (i.e., that turn wishes into wants).
3. Provide mechanisms for maintaining (i.e., for creating, canceling, and reprioritizing) wants or considered desires in light of changing beliefs (or, more generally, in light of changing beliefs and immediate desires).

Moreover, I think it is reasonable to expect that a logical theory of practical reasoning should be compatible with or even accommodate and subsume quantitative decision theory in those cases where we have good, applicable decision theoretic models—where relatively few variables are present, and we have good probabilities and quantitative estimates of desirability for states made up using these variables.

However, an adequate generalization of decision theory should definitely *not* treat the outcome of practical reasoning as unique. The theory should allow for the possibility that agents with the same beliefs and wishes could reach different conclusions, even while conforming to whatever principles of rationality are available.

The modification that I will explore here retains the utilitarian idea that the source of immediate desires or wishes is external to practical reasoning, so that such desires do not originate in, nor are they maintained by, an inferential process. I think this is a reasonable simplifying assumption, even if it doesn't seem to fit fully human practical reasoning.[1]

[1] For discussion of some of the relevant issues, see [De Sousa 87].

The model that I will present also will ignore complications in the relation of wants to intentions. In particular, I will not try to take weakness of will into account.

3. A conditional-based approach to belief-desire maintenance

An analogy to truth maintenance should help to motivate the general approach that I will advocate. You can think of truth maintenance as a qualitative approach to belief maintenance that uses nonmonotonic instead of probabilistic reasoning. Practical reasoning, however, involves desires as well as beliefs—and beliefs enter crucially into reasoned desires. If I am choosing whether to drive or ride my bicycle to work, and have an immediate preference to bike, this preference may well change if I look outside and learn that it is raining. Though some beliefs may seem to be purely theoretical, the point of an elaborate, carefully maintained system of beliefs—in human and in simulated agents—is most often practical and action-oriented. This suggests the project of generalizing pure belief maintenance to a framework capable of supporting belief-desire maintenance. Thus, a primary technical part of the project consists in articulating ways in which belief-oriented nonmonotonic accounts reasoning can be extended to deal with action-oriented reasoning.

In general, *conditional* attitudes or modalites are more informative and more useful in reasoning than absolute attitudes. A basic idea of the solution I want to recommend for BD maintenance is to conditionalize desires as well as beliefs.

The basic elements of the approach on which I'm working are these:

1. There are conditional wishes $W(\psi/\phi)$ as well as categorical wants $I(\phi)$. ('I' is for 'intention'; unfortunately, 'want' begins with a 'w'.)[2]
2. Wants are derived from wishes and beliefs by a process of nonmonotonic reasoning that aims at creating grounded extensions that somehow maximize wishes. The idea is to achieve reasoned desires that satisfy as many as possible of the immediate desires. In simple cases, we can imagine these wants to be derived from 'axioms' representing the agent's wishes and beliefs. In more complex cases, an agent's BD state is updated in light of new beliefs and desires. This updated state may depend not only on prior beliefs and wishes, but on the prior wants, because drastic and global replanning may be expensive for an agent whose reasoning resources are limited.[3]
3. Many of the problems that Bayesian decision theory classifies as deliberation under uncertainty reappear in this framework as cases in which conflicting wishes need to resolved. Often, such conflicts give rise to multiple extensions. As in the nonmonotonic theory of pure belief, to choose among such extensions we will need to represent priorities among conflicting conditionals in a way that will allow us to resolve at least some conflicts. In the case of practical reasoning, however, this is complicated by Dutch book arguments that seem to require a quantitative approach to both beliefs and desires in order to secure optimal solutions to certain sorts of classical decision theoretic problems. To the extent that we can introduce local quantitative mechanisms into the framework to guide conflict resolution in such cases, we may be able to approximate the results of Bayesian decision theory.

On the approach I'm advocating, conflict resolution becomes an essential mechanism in reasoning cases in which a Bayesian analysis is appropriate. This provides a welcome connection between the theory I wish to develop here and Wellman's recent attempt in [Wellman 1988] to develop qualitative foundations for decision making in medical domains. Wellman's term, 'tradeoff', is merely another way of talking about conflict resolution. Wellman's work, which is integrated into the AI planning tradition, in fact gives me some reason to hope that it may be possible to provide the sort of local and relaxed quantitative information that is needed to model conflict resolution in cases where risk is important. The scale of this problem, which calls for a long-range research program involving many people, is not particularly disconcerting, when one considers that (1) the need for this research can be well motivated independently, and (2) some excellent work (such as Wellman's) has already been done in this direction. Relying on this work to create a bridge to domains like medical decision making will enable me to concentrate here on less regimented types of reasoning.

4. Characterizing the logic

The crucial logical problem in developing this approach is to define the admissible extensions of a theory involving the three conditional operators mentioned above: B, W, and I.

We begin with the sublanguage involving B only, which is better understood. A basis for the definition can be found in recent work on conditional logic, such as [Bell 1991], [Geffner & Pearl 1992 19], and [Lehmann & Magidor, forthcoming]. However, for reasons that are motivated in part from research on inheritance and in part from a desire to make the axiomatization of realistic domains feasible, I believe that this basis has to be modified to provide for general patterns of default reasoning that are specific to the conditional operators.

The shortcomings of the more traditional conditional-based approach are clearly presented in at least two recent works: [Asher & Morreau 1991] and [Horty 1991]. Both of them argue for an approach that blends ideas from nonmonotonic logic, inheritance, and conditional logic, and that extends a basic conditional logic by (1) adding a principle of specificity that automatically allows more specific conditionals to preclude less spe-

[2]The theory will eventually need to be generalized to provide for conditional wants or intentions; these are clearly necessary for contingency planning. But in preliminary versions of the theory, it seems an appropriate simplification to disallow conditional wants.

[3]Thus, the theory I am aiming towards would be broadly compatible with Bratman's work on the value of plans; see [Bratman 1987].

cific ones, without having to state exceptions in the antecedent, and (2) construing conditionals to be implicitly generalizable, within the constraints provided by (1), so that $B(\psi/\phi)$ defeasibly implies $B(\psi/(\chi \wedge \phi))$. Horty motivates these conditions well and provides enlightening connections to both deontic and to default logic.

Asher and Morreau, however, go further in providing an explicit semantics,[4] and it is this work that I am currently relying on as a basis for the conditional constructions of BD maintenance. One consequence of the generalization of this sort of logic of belief to operators expressing desire is that dominance (in the decision theoretic sense) comes to play a role similar to that played by specificity in the inheritance tradition.[5]

In this paper, I omit the technical definitions. The idea is to define extensions for the B operator using Asher and Morreau's semantics, which evaluates conditionals in a world where total ignorance is assumed, and which uses the Henkin construction to guarantee that there are enough worlds to make conditionals generalize. The key idea is a notion of normalization that converts as many wishes as possible to wants, relative to the normalized beliefs. That is, the set of wants $I(\phi)$ such that $W(\phi/\mathbf{T})$ also holds is maximal. Normalization is used in the Asher-Morreau semantics in a way that parallels the more familiar use of preferred worlds in defining defeasible entailment.

It will become clear that further preferences will need to be added to deal with interesting cases of conflict resolution, but even at this stage we have a quite interesting formalism, that I think can qualify as a promising logical framework for practical reasoning. I will try to support this claim by illustrating how the formalism deals with some specimen arguments.

5. Examples illustrating the formalization of arguments

5.1. Example 1

Natural language version.

I'd like to fly to San Francisco.
If I fly to San Francisco I'll have to buy an air ticket to San Francisco.
So I want to buy an air ticket to San Francisco.

Formalized version, with additional steps.

1. $W(p/\mathbf{T})$
2. $B(q/p)$
3. $I(p)$
4. $I(q)$

According to the logic I am proposing, 3 follows (defeasibly) from 1, and 4 from 2 and 3. (I have in mind an interpretation that assigns truth values to the conditional B and W statements along the lines of theories like [Asher & Morreau 1991], and that assigns truth values to the I statements—which are unconditional in these examples—relative to a limited set of extensions. This set is simply part of the model.) In fact, the theory containing 1 and 2 will have only one extension, which will contain p and q. The motivation for the inference in step 4 is that wants are reasoned desires, and so their consequences have to be weighed in the balance when they are concluded. Thus, the principle of "desiring the means" is built into wants on this account, though it is definitely not part of the logic of wishes, which are not reasoned and do not have to be practical. Note that the phrase 'have to' in premiss 2 is taken to have no content that affects formalization. This is a simplification; we could add a branching time structure, for instance, and interpret this construction as an indication of necessity over futures. At this stage, though, I am ignoring complexities due to time.

5.2. Example 2

Natural language version.

I'd like to fly to San Francisco.
If I fly to San Francisco I'll have to buy an air ticket to San Francisco.
If buy an air ticket to San Francisco I'll have to spend some money.
I wouldn't like to spend some money.
(Even) so, I want to fly to San Francisco.
So, I want to spend some money.

Formalized version, with additional steps.

1. $W(p/\mathbf{T})$
2. $B(q/p)$
3. $B(r/q)$
4. $W(\neg r/\mathbf{T})$
5. $I(p)$
6. $I(r)$

Here, the conflict between the desire to travel and the desire to be frugal creates two extensions of the theory consisting only of premisses 1–4: one in which I want to fly to San Francisco and another in which I want to stay home. (Since the logic is "credulous" with respect to wants, there is no extension in which I am indifferent.) Thus, the inference is not valid in any familiar sense, but it is certainly a specimen of a reasonable practical argument, in that it holds relative to some set of reasonable extensions of the premisses. (In fact, this set is chosen in the course of the argument.)

The logic itself makes no recommendation about which extension an agent should choose (no matter how much or how little a ticket costs, or how urgent the need to fly to San Francisco is); and this intuition seems right, as far as logic is concerned. If mechanisms of the sort discussed in [Wellman 1988], however, are incorporated into the definition of preference, I believe that we could extend the logic of practical reasoning that I

[4] [Asher & Morreau 1991] is explicit only about the solution to (1), though they believe that they have a solution to (2). The details are complicated, and I am not entirely convinced that an adequate solution has been worked out yet. Given the attention these issues are now receiving, though, I believe that we can expect solutions to appear in the near future.

[5] There are also some affinities between the semantics I would like to supply for the theory presented here, and the one presented in [Wellman & Doyle 1991].

have described here to a theory that, like decision theory, provides a more robust guide to decisions when some quantitative information is available.

In example 1 there is a choice of the "spendthrift" extension, in which the desire to fly to San Francisco is practicalized at the expense of the desire to save money. In this extension, according to the logic, the agent must want to spend some money; otherwise, he cannot want to fly to San Francisco. This illustrates the contextualization of 'want'. In isolation, it may sound peculiar for this agent to say he wants to fly to San Francisco. But after concluding it is best to do so, it would be appropriate to say 'I guess I do want to spend some money, after all'.

Note that there is a dynamic aspect to the reasoning in these arguments; in the course of the argument an extension is chosen, and so the reasoning context is shifted. (This context shift is particularly noticeable in Example 2, in which one of two extensions is chosen at Step 5.) Thus, even though BD maintenance allows intentions to be derived from beliefs and immediate desires by a process of reasoning, the derivation process does not yield a reduction, since the part of the reasoning process is a nondeterministic choice.[6]

6. Examples illustrating BD maintenance reasoning

The preceding examples were meant to illustrate the adequacy of the formalism for characterizing at least some simple types of discursive practical reasoning. The following four examples are chosen to illustrate some simple features of the associated reasoning; in this version of the paper, I will give only very compressed explanations of them.

6.1. Example 3

Natural language version.

If I drink coffee, I'll have to drink decaf.
I'll drink coffee.
So, I'll drink decaf.

Formalized version.

1. $B(q/p)$
2. $B(p/\mathbf{T})$
3. $B(q/\mathbf{T})$

This is simply *modus ponens* for conditional belief. Though it is not logically valid in the theories that I propose to use, it is important for this inference to hold as a default. The theory associated with this argument, then, has one extension: $\{p, q\}$.

6.2. Example 4

Natural language version.

If I drink coffee, I'd like to drink decaf.
I want to drink coffee.
So, I'd like to drink decaf.

Formalized version.

1. $W(q/p)$
2. $W(p/\mathbf{T})$
3. $W(q/\mathbf{T})$

Example 4 is the analog of Example 3 in which desire replaces belief.[7] I assume that the desire to drink coffee is an immediate one; i.e., that it isn't obtained by a chain of reasoning from beliefs and other desires. I count such desires as wants, and would be willing to assume as a first approximation that at least these desires are consistent. Like unopposed instances of *modus ponens* for belief, instances of *modus ponens* for desire hold as defaults; the idea is that in practical reasoning we try to satisfy our immediate desires unless there is some reason not to do so. Thus, there will again be only one extension: $\{p, q\}$.

6.3. Example 5

Natural language version.

If they have coffee here, I'd like to order coffee.
I believe they have coffee here.
If they don't have decaf here, I wouldn't like to order coffee.
I believe they don't have decaf here.

Formalized version.

1. $W(q/p)$
2. $W(p/\mathbf{T})$
3. $W(\neg q/\neg r)$
4. $B(\neg r/\mathbf{T})$

This case of conflict between two wishes is just like a conflict between defaults in a belief-oriented nonmonotonic system. There are two extensions: $\{p, \neg r, q\}$ and $\{p, \neg r, \neg q\}$. Example 5 does not differ in important respects from Example 2.

6.4. Example 6

Natural language version.

If I go outside, I'll have to get wet.
I'd like not to get wet if I go outside.
I'll have to go outside.
So, I'll get wet.

Formalized version.

1. $B(q/p)$
2. $W(\neg q/p)$
3. $B(p/\mathbf{T})$
4. $B(q/\mathbf{T})$

Though (in view of Step 1) there is a conflict here between 2 and 3, and both are defaults, it would simply be wishful thinking to allow the desire to compete on a par with the belief. This illustrates a general principle of practical reasoning as it appears in this framework:

[6] In the current formalizations of the theory, I have left the dynamic aspects of the reasoning implicit; of course, it will be important eventually to explore formalizations in which the dynamicism is made more explicit.

[7] Note that natural language is often ambiguous between the two; 'If I drink coffee, I'll drink decaf' could be interpreted in either way. BD maintenance provides a partial explanation of the conflation; both operators play a similar role in extension construction. (Though, as we'll see, one operator dominates the other.)

beliefs override desires, when both are interpreted as defaults that may compete with one another. Thus, there is only one extension here: $\{p, q\}$.

At the moment, the details of the formalization that I have described here are rather volatile. I am uncertain about many things: for instance, how to represent ability, and whether there is a need for an independent kind of operator that *forces* the construction of extensions (as, for instance, our knowledge of methods of travel would force the construction of a bus and an airplane extension for a trip from Pittsburgh to Washington, DC, but would not force the bus extension to be constructed for a trip from Pittsburgh to Seattle.) I expect that details of the theory will change as the semantics is fleshed out. I hope that this level of volatility is acceptable in a working paper like this one, and that the material that I have presented here will at least suffice to motivate the approach.

7. Foundations of planning

In the rest of this paper, I will try to avoid some defects that previous presentations of the ideas have suffered from, by concentrating on foundational issues and contrasting the views with some of the so-called "BDI" approaches. It is easy to correlate

Belief–Desire–Intention

with

Belief–Wish–Want

and to understand the suggestion as nothing more than a change in terminology.

This correlation is actually not bad as a first approximation, and even in the last analysis you could understand the position I'm trying to develop here as a BDI theory. But the motivation of the BD maintenance approach that I'm trying to develop here, and the implementations that the approach suggests are sufficiently different from the current BDI theories and architectures that it is helpful to contrast the two.

First, the origins of the two approaches are different: Bratman's work, for instance, is a reaction to earlier philosophical views in the theory of action that attempted to reduce intention to desire and belief; of course, Bratman was also influenced by interactions with AI planning theorists. The views I am advocating grew out of the logical tradition in practical reasoning. I have been interested for a long time in trying to improve the logical theories of practical reasoning, and the current project is an attempt to use the theories of nonmonotonic reasoning that have developed in AI to improve the models in [Thomason 1981a] and [Thomason 1981b]. I was also influenced by conversations over several years with Jon Doyle, whose views, I think, are broadly compatible with the approach that is described here.

This work uses the methodology of philosophical logic; that is, one looks at specimens of common sense reasoning, and then seeks to build logical systems that allow the reasoning to be formalized in a way that provides an enlightening distinction between correct and incorrect reasoning. To a certain extent, these techniques have been imported into AI—particularly, in the area of formalizing common sense reasoning. But there is no guarantee, of course, that a theory of practical reasoning created by this logical methodology will be of value in the design of planning systems. Though I hope that the ideas will be useful in this way, I have no evidence yet to offer for this claim.

Second, I mean to suggest a very different area of emphasis than "belief-desire-intention" seems to suggest at the moment in the AI literature. In fact, the philosophical project of [Bratman 1987] (to refute reductions of intention to desire and belief) should correspond to an emphasis not only on plan synthesis, but on plan maintenance, since Bratman's main argument for the necessity of separate mechanisms of long-term intention is that plans are useful to resource-limited agents. He acknowledges in many places that to be useful in this way, plans will need to be adapted to circumstances, or even partially retracted or repaired. But—perhaps because a closer examination of the need for such mechanisms would weaken the contrast he wishes to make between his position and one that makes intentions derivative—he does not place much stress on the process of plan retraction and adjustment.

This emphasis is heightened in the AI planning community, which has tended to concentrate on means-end reasoning at the expense of other aspects of practical reasoning. Thus, the "BDI architectures" that I have seen, though they may allow for reasoned change of belief, for the synthesis of plans from goals, and even for some form of plan maintenance, do not focus on mechanisms that might help to make plan maintenance (including plan retraction) a process as reasoned as the maintenance of beliefs. For instance, [Rao & Georgeff 1992], which is designed as a general architecture for implementing rational BDI agents, assumes that desires must be consistent, and there is no provision at all for conditional desires.

In fact, however, conflicting desires are much more common and pervasive than conflicting beliefs—common enough to be typical of conscious decision-making. When most people are considering making a purchase, for instance, a desire for the purchase will compete with a desire to avoid parting with its price.

Without a way to distinguish a plan that is constructed by resolving competing desires from one that is constructed by arbitrarily choosing one of two indifferent alternatives, we will lack information that helps to guide plan revision. When I have decided arbitrarily in advance to take one of two equally good routes to the airport, it is not unreasonable for me to change my plan on the spot when I encounter a slight difficulty on the way, like unexpectedly heavy traffic. When I have debated whether to cancel my classes or to fly across the country to visit my sick mother, and have decided finally to do the latter, it is less reasonable to change my decision because of a small inconvenience like a busy phone line.[8]

But if we allow desires to conflict, we must have a mechanism for resolving conflict. A mechanism for re-

[8] To make the comparison fair, you must imagine that I have not announced the decision; after I decide, my first action is to try to call my mother, and the phone is busy.

solving conflicts among desires leads from immediate to reasoned desires. And, if we ignore problems having to do with weakness of the will (which in any case we can safely ignore in a first approximation model of reasoned agency), there is no good reason to distinguish reasoned desires from intentions. Thus, you can think of my "Belief–Wish–Want" trio as an attempt to generalize the BDI model to accommodate conflicting desires in a way that is reasonably faithful to common sense examples, and that does not unduly proliferate the attitudes needed in the account.

Conditional desires, too, are absolutely necessary in any moderately general theory that seeks to account for plan maintenance, since it is hard to imagine a realistic desire that cannot be undermined by changing circumstances—in which case, of course, we may need to reconsider the plans that were generated from these desires and that depend in part on them.

Though the model on which I am working does, like utility theory, construct reasoned from immediate desires, it does not represent an attempt to revive a philosophical reduction of intention to belief and desire. (Remember the indeterminism that is involved in passing from beliefs and immediate desires to intentions.) In fact, most of Bratman's arguments against the older reductions remain good on the view that I am advocating.[9]

8. Wanting the consequences

The model that I am proposing also differs in its treatment of the relevant attitudes in a number of details; I'll illustrate these with the one that is perhaps the most striking.

Bratman emphatically denies that logically rational intentions are closed under logical consequence. All of the AI formalizations of intention seem to have followed him in this, despite the fact that this makes it more difficult to apply possible worlds semantics to intention. No doubt, the reason for the general acceptance of non-closure is that it sounds crazy to say that I intend (or that I want) to suffer when I go to the dentist. But I am assuming this closure principle in my model.

The purpose of this section is to convince you that this is not as crazy as it sounds.

8.1. The problem restated

Any adequate account of deliberation has to allow for the need to compare the merits of the relevant alternatives. In deciding whether to buy a house or to continue renting, I need to consider the consequences of alternative courses of action. If I buy, I will be able to arrange the house to suit me, but I will have to pay real estate taxes. If I rent I will be free of responsibility for major maintenance, but I will have to pay rent.

[9]My view does make the efficacy of future-based intentions something more of a puzzle than I think Bratman wants it to be; but in fact I am not bothered by this, because I think that this is a difficult philosophical puzzle. The human phenomena regarding future intentions are very complex, and it is not easy to do justice to them. I don't think that either Bratman or BD maintenance begins to solve them.

Some of these consequences are desired, and some are not. (Note the natural use of 'but' to mark this distinction.) But if I make my decision thoroughly and well, I will need to take all relevant consequences into account, without reference to whether these consequences are welcome or not. Both optimism (ignoring or discounting undesired consequences) and pessimism (ignoring or discounting desired consequences) are liable to lead me to a suboptimal conclusion.

When, after comparison of such alternatives, one of them is chosen, intentions will generated. The act of choosing to buy the house generates a primary intention, together with many subsidiary intentions—for instance, the intention to obtain a mortgage.

Like quantitative decision theory, this theory of practical reasoning relies on comparison of alternative scenarios (or extensions). Thus, on this theory, the natural object of choice (or reflective desires) is an extension.[10] (Of course, an extension is often best characterized in a reasoning context by means of the distinctive hypotheses that generate it.) In general, as we have seen, some features of extensions will be desired, and others will not. But both desired and undesired features of an extension will be chosen along with the extension, despite the fact that the extension will be chosen *because of* what is desirable in it, and *in spite of* what is undesirable in it. The comparison is a package deal: it is the pros and cons of the extension as a whole that must be compared with those of other extensions.

We are now close to saying that undesired consequences are intended. Perhaps this conclusion could still be avoided by trying to establish a relevant distinction between choice and intention. But, since choice and intention are so close, such a distinction may be difficult to motivate. Rather than seeking to do that, I think it will be more rewarding to tentatively accept the notion that undesired choices are intended, and to treat this result as a test of the plausibility of the theory. If it can't be solved, the theory is in trouble.

8.2. The solution strategy

Theories that allow expressions to be context sensitive provide an alternative, pragmatic way of accounting for invalidities. The technique is used convincingly on indicative conditionals in [Stalnaker 1975] and has been plausibly applied in many other cases.[11] Essentially, the idea is to argue that the apparent invalidity of an inference

ϕ; therefore ψ

is due to not to semantic differences involving truth conditions in a fixed context, but to the fact that the context has been changed in passing to ψ. To make the argument plausible, of course, one has to show that some crucial constructions of the inference are context sensitive, and that the assertion of the conclusion is likely for pragmatic reasons to change the context in a way that will make the conclusion false. I will now proceed to argue that the model I'm proposing does this

[10]More generally, the object of choice would be a set of extensions.

[11]See, for instance, [Lewis 1979].

for wanting the consequences of what one wants. If I am correct, there are no reasons (other than the usual doubts that would apply to any attitude such as belief) for withholding closure under logical consequence from intentions.

8.3. Reorientation

Michael Bratman has argued in his book and in several articles for the need to distinguish intended from unintended consequences of intentions, and has provided intuitive evidence for the distinction.[12] Bratman is trying to undermine a theory of rational agency according to which the appropriate rational intentions must be calculated *ab initio* from current raw desires and probabilities. So he is concerned to draw attention to the fact that, though plans represent *calculated* choices that certainly are not immediate desires, they can nevertheless be used as the basis for deliberation in light of changing information. Bratman stresses that plans are useful not only as scores from which future behavior can be played by rote, but can serve as inspiration for more flexible and creative reactions to contingencies. Even in cases where a direct, utility-theoretic calculation from basic desires and probabilities may be feasible in principle, a resource-bounded agent may not have time for such a calculation, and it may be more economical to act on an existing plan. This emphasis on the dynamic importance of plans in the deliberation of resource-limited agents is shared by many members of the AI community.

8.4. Bratman's examples

For Bratman (and for me, though perhaps not for all members of the AI planning community) it is important for the intentions of a theory of deliberation to have robust connections with the intentions of common sense psychology. Thus, the theory is influenced by common sense intuitions about the role of deliberation in various examples. As J.L. Austin pointed out in [Austin 1957], we are used to making a large number of subtle distinctions in judging matters of deliberation, and these distinctions and the judgments they support are highly sensitive to context. Therefore, we will need to consider a number of examples, in order to probe the issues even superficially. For reference, the examples will be numbered and named.

Example 1: Bratman's bombers.
A military strategist decides to bomb a munitions factory, knowing that this will kill the children in a nearby school, and accepting this consequence.

Case 1. Strategic bomber.
The strategist accepts this consequence because of the importance of winning the war, though he deplores killing the children.

Case 2. Terror bomber.
The strategist values this consequence, because he believes that the terror caused by the children's death will weaken the enemy's will to fight.

[12]See [Bratman 1987].

Some readers may find the heightened emotional content of Bratman's example distracting. Dramatic consequences are in fact irrelevant to the distinction, and it may make it easier to focus on the theoretical issues to substitute a parallel example, of Kurt Konolige's and Martha Pollack.[13]

Example 2: Kids will be kids.
A child decides to play with his sister's toy fire engine, knowing that this will make his sister unhappy, and accepting this consequence.

Case 1. Selfish brother.
The child accepts this consequence because of the importance he attaches to playing with the toy, though he regrets it.

Case 2. Spiteful brother.
The child values this consequence, because he enjoys making his sister unhappy.

In both of these examples, Bratman would say that the side effect is intended in the second case, and not intended in the first. The reasons for this conclusion lie in the differences between the way that the side effect figures in replanning in the two cases.

Applying Bratman's suggested tests to Kids will be Kids, we can test whether the side effect is intended by examining whether, in his subsequent planning, the brother will protect the outcome of making his sister unhappy. For instance, if the sister says she is tired of the fire engine and doesn't like it any more, her brother (a) may be indifferent to this or even think it a reason in favor of playing with the engine. Or (b) he may think it counts in favor of substituting some other activity that is more likely to make her unhappy. The former behavior is what we would expect of the selfish brother, who wants to play with the engine despite the effects on the sister; the latter is what we would expect of the spiteful brother, who wants to play with it in part because of these effects.

The difference to which Bratman refers here is familiar, and is important both in practical reasoning and judging the conduct of others. But the extent to which it has to do with the presence or absence of an intention is far from clear; in both Cases 1 and 2 of Example 1, it seems accurate and natural to say, if the children are killed, that the bomber killed them intentionally and that he chose to kill them or meant to kill them, though (to me and to some others at least) it does seem misleading to say that the strategic bomber had an intention to kill them.

Bratman's work rightly draws attention to the use of intentions (or plans) as resources in practical reasoning; intentions do constrain our future choices. Since the paradigm within which most philosophers have worked recently has ignored this fact, it is worth stressing. But the constraints that intentions impose on future deliberation are flexible and dependent on context. So, when a contemplated action is withdrawn in the course of replanning, this need not mean that it must not have been intended—even when the action is gladly retracted.

[13]In an as-yet unpublished manuscript.

make room for a great many kinds of intention. The models of reasoning that support the distinctions will need to be fairly complex.

8.7. Context selection

A context for practical reasoning will involve, among other things, assumptions about the possibilities for action that are open; in specimens of practical reasoning, this context may be shifted as we entertain various hypotheses. It is not difficult to see, using principles of the sort that govern implicature generally, that an assertion (in a deliberative context) to the effect that I want an eventuality to occur will in general implicate that I have in mind a context in which it is an open possibility both that the eventuality may occur, and that it may not occur.

For example, when a nurse comes up to me with an unpleasant-looking needle and says

"Do you want it in your right or your left arm?"

she is inviting me to accept a context in which it is ruled out as a possibility that I will not get the needle in either arm. It is perfectly appropriate here for me to reply that I want it in my left arm. It is difficult to see how I would want this without wanting it in one of my arms, but (because, I would claim, of the inappropriate context that it implicates) it would be inappropriate for me to say at this point that I want the needle in one of my arms, since that implicates that this matter is open for deliberation. The utterance is incoherent because it suggests that I have not accepted the context I was invited to accept; and this incoherence persists even though the sentence says something that is true in that context.[14]

With this point about implicature, my case for closure under logical consequence is completed.

8.8. Summing up

The reason why Bratman's examples may seem compelling to a planning theorist is that there is a genuine difference between the efficacy of desired and undesired intentions in replanning: the spiteful brother may withdraw his plan on learning that his sister doesn't care if he plays with the toy, while the selfish brother will not.

However we develop a planning architecture that delivers a genuine mechanism for replanning, we should be able to use it to provide for cases like that of the Weekend Worker, in which genuine intentions are retracted in the light of new information that undermines the conditional desires on which they rest. Such a mechanism will automatically yield a way of retracting undesired elements of plans, that in no way requires us to say that these things must have been unintended in the first place.[15]

One of the problems with a monotonic system (like, for instance, the Bayesian account of full belief as subjective probability 1) is that it forces us to treat cases in which a belief is retracted as cases in which the belief was never actually present in the first place.[16] In a genuinely nonmonotonic model, there is no such need.

Thus, a BD maintenance system will provide for the retraction of intentions without any need to suppose that they were never present in the first place. Since this treatment seems to provide an equally good (pragmatic) account of the relevant examples, allows for simpler logical models of the relevant operators, and the mechanism that enables it seems to be required on independent grounds, it seems preferable to denying that these consequences are in no way intended.

9. Conclusion

This paper attempts to bring together several strands in recent logical work (much of it related to AI), along with work in limited rationality, and to use them to revive the idea of a true logic of practical reasoning. Demonstrating that these proposals really work is a large project. But by concentrating on the basic theoretical ideas and illustrating them with simple examples, I hope to show in this preliminary work that the ideas have some promise.

Bibliography

[Asher & Morreau 1991] Asher, N., and M. Morreau, "Commonsense entailment: a model of nonmonotonic reasoning." *IJCAI 1991: Proceedings of the Twelfth International Conference on Artificial Intelligence,* J. Mylopoulos and R. Reiter, eds., Morgan Kaufmann, San Mateo CA, 1991, pp. 387–392.

[Austin 1957] Austin, J.L., "A plea for excuses." Reprinted in J. Urmson amd G. Warnock, eds., *Philosophical papers by the late J.L. Austin,* Oxford University Press, 1961, pp. 123–152. (Originally published in *Proceedings of the Aristotelian Society 57* (1956-57).)

[Bell 1991] Bell, J. "Pragmatic logics." *KR91: Principles of knowledge representation and reasoning,* J. Allen, R. Fikes, and E. Sandewall, eds., Morgan Kaufmann, San Mateo CA, 1991, pp. 50–60.

[Bratman 1987] Bratman, M., *Intention, plans, and practical reason.* Harvard University Press, 1987.

[14] See [Thomason 1981a] and [Thomason 1981b] for background on this matter. The account that is offered here is similar to the one of indicatives in [Stalnaker 1975]; that, in fact, is not surprising, given the close connection between indicative conditionals and deliberation.

[15] It also seems to me that the work deriving from [Cohen & Levesque 1990], while it sheds light on some aspects of intention, is incompatible with models of plan maintenance that would be able to accommodate retraction of intentions in light of new information which (as in the case of the Weekend Worker) undermines the intention without rendering it impossible to achieve. Space limitations prevent me from discussing these "persistence" theories of intention in detail here.

[16] What seems to be a retracted belief, on the standard Bayesian account, has to be treated as the readjustment of a probability that was very close to 1. See, for instance, [Harper 1975].

8.5. An analogy

Consider the analogy to belief.

Example 3: Beliefs about the porch light.
I believe that the porch light is off.

Case 1. Guesswork.
I believe that it's off because I asked my daughter to turn it off, and she went off to do it.

Case 2. Seeing.
I believe that it's off because I go to the window and see that it's off.

Like an intention, my belief that the light is off acts as a constraint on future deliberation; for instance, if it is dark outside, and I decide to look for something on the porch, I will decide to turn the light on before looking. However, the ways in which this belief is maintained and affects choices will depend on the reason for the belief.

Suppose that I look at the light switch, now, and see that it's on. In Case 1, I will probably withdraw the belief that the light is off; after the belief is withdrawn, it of course will no longer play a role in guiding my choices. In Case 2, on the other hand, the belief that the light is off will persist, and I will start to look for trouble with the light mechanism; most likely, I will set out to change the bulb.

The fact that the belief is retracted and loses its efficacy in Case (1), however, does not mean that it was never present.

8.6. An example of undermined desire

Now, take a similar case, where a desire is undermined.

Example 4: The weekend worker.
I do not like to work on weekends, and try to avoid doing so.

But now I have to meet a deadline, and come to believe that I have to work on a particular weekend. I decide (perhaps because I have shopping to do in stores that aren't open on Sunday) that I want to work on Sunday.

In some examples of this kind, my original intention to work on Sunday at all might be questionable. But also there are many cases in which we do have clear, unquestionable intentions that work against our inclinations—and this is meant to be such a case. So, assume that I definitely intend to work on Sunday. In this case, my attitude towards working on Sunday has all the earmarks of any other intention. (1) It is natural to describe what has happened by saying that I intend to work on Sunday. (2) I chose to work on Sunday. There were no abnormalities in the choice situation; the choice was perfectly everyday and ordinary. (3) I can properly be blamed for this choice. (For instance, by a radical Christian who feels that it is irreligious to work on Sunday.) (4) In planning other actions, I will avoid alternatives that conflict with working on Sunday. (For instance, if I decide that I want to play tennis on the same weekend, I'll try to arrange to play on Saturday.) (5) I will treat working on Sunday as a goal when I replan in light of new information. (For instance, if I discover that I can't work at home and will have to do my work at the office, and if I have to drive to work, I will arrange with my wife to use the car.)

However, if this is an intention, the fact that it is opposed by a relatively strong desire (I really don't like to work on weekends) makes it particularly fragile. I will gladly discard it, given an opportunity. If my office deadline is unexpectedly extended, the intention will be retracted and will lose its efficacy in deliberation. Moreover, my desire not to work on weekends is not entirely uninhibited, even after I have chosen to work on Sunday; there is nothing irrational about my choosing to act in a way that facilitates the possibility of retracting the need to work on the weekend. (If I'm offered an extended deadline, I will gladly take it.) Surely, this is no reason to think that the intention to work on Sunday was never present.

Practical reasoning makes available a rich and complex system of distinctions for judging choices. There are in fact differences between the Weekend Worker example, and the Selfish Brother case of the Kids will be Kids example: the most important is that in the first case the undesired consequence forced is by a belief (I believe that I have to work on the weekend), while in the other it is the side effect or believed consequence of a desire (the selfish brother desires to play with the fire engine and believes that doing so will make his sister unhappy). And indeed, though it seems correct to criticize Bratman's strategist by saying either

> "You shouldn't have chosen to hurt the children,"

or

> "How could you mean to hurt the children?"

it does seem as accurate to say

> "You shouldn't intend to hurt the children"

and definitely inaccurate to say

> "You shouldn't mean to hurt the children."

Perhaps these differences could be exploited to produce an improved argument for Bratman's claim that unwanted, but chosen consequences are not intended. But the example of the Weekend Worker shows that the claim also has to be qualified, since some unwanted consequences are intended (or rather, consequences that are unwanted given some beliefs are wanted given other beliefs). Moreover, such examples show that Bratman's criteria for intentions are too coarse as stated, and if applied without qualification would exclude clear cases of intention.

Even though we may grant that intentions are needed to constrain future practical reasoning in various ways, they would not serve this purpose well if they strait jacketed agents. In fact, intentions appear to be defeasible and context sensitive, and to differ in ways that affect their range of applicability in future deliberations. An accurate account of practical reasoning will have to

[Cohen & Levesque 1990] "Intention is choice with commitment." *Artificial intelligence 42* (1990), pp. 213–261.

[De Sousa 87] De Sousa, R., *The rationality of emotions.* MIT Press, 1987.

[Geffner & Pearl 1992 19] Geffner, H., and J. Pearl, "Conditional entailment." *Artificial intelligence 53* (1992), pp. 209–244.

[Harper 1975] "Rational belief change, Popper functions and counterfactuals." *Synthese 30* (1975), pp. 221–262.

[Horty 1991] Horty, J., "Moral dilemmas and nonmonotonic logic." Manuscript, UMIACS, University of Maryland, 1991.

[Lehmann & Magidor, forthcoming] Lehmann, D. and M. Magidor, "What does a conditional knowledge base entail?" *Artificial intelligence,* forthcoming.

[Lewis 1979] Lewis, D., "Scorekeeping in a language game." *Journal of philosophical logic* 8, pp. 339–359.

[Rao & Georgeff 1992] Rao, A., and M. Georgeff, "An abstract architecture for rational agents." In *"KR92: Principles of knowledge representation and reasoning,* B. Nebel, C. Rich, and W. Swartout, eds., Morgan Kaufmann, San Mateo CA, 1992, pp. 439–449.

[Stalnaker 1975] Stalnaker, R., "Indicative conditionals." *Philosophia 5*, pp. 269–286.

[Thomason 1981a] "Deontic logic as founded on tense logic." In *New Studies in Deontic Logic,* R. Hilpinen, ed., D. Reidel, Dordrecht, 1981, pp. 141–152.

[Thomason 1981b] Thomason, R., "Deontic logic and the role of freedom in moral obligation." In *New Studies in Deontic Logic,* R. Hilpinen, ed., D. Reidel, Dordrecht, 1981, pp. 153–162.

[Wellman 1988] Wellman, M., *Formulation of tradeoffs in planning under uncertainty.* MIT/LCS Technical Report TR-427, Massachusetts Institute of Technology, Cambridge MA, 1988.

[Wellman & Doyle 1991] "Preferential semantics for goals." In *AAAI-91: Proceedings of the ninth national conference on artificial intelligence,* T. Dean and K.McKeown, eds., MIT Press, Cambridge MA, 1991, pp. 698-703.