The Logical Foundations of Goal-Regression Planning in Autonomous Agents

John L. Pollock
Department of Philosophy
University of Arizona
Tucson, Arizona 85721
(e-mail: pollock@arizona.edu)

Abstract

This paper addresses the logical foundations of goal-regression planning in autonomous rational agents. It focuses mainly on three problems. The first is that goals and subgoals will often be conjunctions, and to apply goal-regression planning to a conjunction we usually have to plan separately for the conjuncts and then combine the resulting subplans. A logical problem arises from the fact that the subplans may destructively interfere with each other. This problem has been partially solved in the AI literature (e.g., in SNLP and UCPOP), but the solutions proposed there work only when a restrictive assumption is satisfied. This assumption pertains to the computability of threats. It is argued that this assumption may fail for an autonomous rational agent operating in a complex environment. Relaxing this assumption leads to a theory of defeasible planning. The theory is formulated precisely and an implementation in the OSCAR architecture is discussed.

The second problem is that goal-regression planning proceeds in terms of reasoning that runs afoul of the Frame Problem. It is argued that a previously proposed solution to the Frame Problem legitimizes goal-regression planning, but also has the consequence that some restrictions must be imposed on the logical form of goals and subgoals amenable to such planning. These restrictions have to do with temporal-projectibility.

The third problem is that the theory of goal-regression planning found in the AI literature imposes restrictive syntactical constraints on goals and subgoals and on the relation of logical consequence. Relaxing these restrictions leads to a generalization of the notion of a threat, related to collective defeat in defeasible reasoning. Relaxing the restrictions also has the consequence that the previously adequate definition of "expectable-result" no longer guarantees closure under logical consequence, and must be revised accordingly. That in turn leads to the need for an additional rule for goal-regression planning. Roughly, the rule allows us to plan for the achievement of a goal by searching for plans that will achieve states that "cause" the goal. Such a rule was not previously necessary, but becomes necessary when the syntactical constraints are relaxed.

The final result is a general semantics for goal-regression planning and a set of procedures that is provably sound and complete. It is shown that this semantics can easily handle concurrent actions, quantified preconditions and effects, creation and destruction of objects, and causal connections embodying complex temporal relationships.

1. Introduction

This paper addresses some logical problems that arise in the course of formulating a theory of plan construction for autonomous rational agents operating in realistically complex environments. My ultimate objective is to understand how truly intelligent agents can get around in the real world. I approach the theory of plan construction as part of an attempt to construct a general theory of rational cognition in such agents. Within rational cognition we can

---

Artificial Intelligence **106** (1998), 267-335.

This work was supported by NSF grant no. IRI-9634106.

distinguish between epistemic cognition and practical cognition. Epistemic cognition is concerned with what to believe, and practical cognition is concerned with what to do. We can think of practical cognition as dividing roughly into four parts: (1) goal-adoption, (2) plan-construction, (3) plan-adoption, (4) plan-execution. Viewing plan construction from this somewhat broader perspective turns out to impose constraints not satisfied by standard planning systems. These constraints generate logical problems for a theory of plan construction, and the purpose of this paper is to formulate some of those problems precisely and propose solutions to them. The focus of this paper will be theoretical, however the ultimate intent is to implement the theory in the OSCAR architecture for rational agents.[1] This has been partially accomplished to date. I will say more about implementation as the theory develops.

In this paper, I am not particularly interested in constructing a theory of human cognition. However, humans are the most successful autonomous rational agents that we know about, and so it will be occasionally useful to reflect upon how humans solve some of the problems that face all autonomous rational agents. Human plan-construction is generally based on *goal-regression planning*. The basic idea is a simple one, going back at least to Aristotle. To achieve a goal, we consider an action that would achieve it under some specified circumstances, and then try to find a way of putting ourselves in those circumstances in order to achieve the goal by performing the action. Putting ourselves in those circumstances becomes a *subgoal*. The idea is to work backwards from the goal through subgoals until we arrive at subgoals that are already achieved. The resulting sequence of actions constitutes a *plan* for achieving the goal. My ultimate objective in this paper is to provide precise logical foundations for goal-regression planning.

Much work in AI has been directed at the task of formalizing and automating goal-regression planning. This forms the basis of a large part of AI planning theory, and the result is what I will refer to as the "conventional" theory of goal-regression planning. In section two I will give a precise formulation of the conventional theory. The ideas developed in that section will mostly be familiar, although there may be some novelty in the way in which I have combined familiar ideas and developed them into a logically precise theory. In section three I will argue that, assuming the basic correctness of the conventional theory, rational agents situated in a complex environment cannot in general perform goal-regression planning in quite the way the conventional theory proposes to implement it. In a sense to be explained, planning must be done defeasibly rather than by running a semi-decidable algorithm. Section four will describe how that can be done. In section five I will argue that, even given the modifications of section four, the conventional theory turns upon an indefensible assumption. In effect, the conventional theory runs afoul of the Frame Problem. In section six I will show how the conventional theory must be modified in light of section five. In sections seven and eight I will suggest further modifications aimed at relaxing the syntactical constraints the conventional theory imposes on goals and subgoals and on the relation of logical consequence. The final result is a general semantics for goal-regression planning and a set of procedures that is provably sound and complete. The short closing sections illustrate the power of this semantics by showing that it can easily handle concurrent actions, quantified preconditions and effects, creation and destruction of objects, and causal connections embodying complex temporal relationships and metric time.

---

[1] Pollock (1995).

2. The Conventional Theory of Goal-Regression Planning

Goal-regression planning is based upon conditionals to the effect that if an action *A* is performed under circumstances *C*, the goal *G* will be achieved. I will write such a conditional as "$(A/C) \blacktriangleright G$". I will refer to *C* as the *precondition* of the conditional, *A* as the *action*, and *G* as the *goal*. For the time being, I will not attempt to be more precise about the logical form of these conditionals. That is a topic to which I will return in section five. I will call these *planning-conditionals*. In goal-regression planning, human beings make explicit appeal to planning-conditionals. By contrast, most work in AI planning theory follows the lead of STRIPS (Fikes and Nilsson 1971) in building these conditionals into the actions from which the plans are constructed (the "plan operators") instead of storing them in a separate database of background information from which a planner reasons explicitly. Allowing multiple planning-conditionals concerning the same action is equivalent to employing plan operators with conditional effects.[2]

Goal-regression planning aims to construct a plan for achieving a goal. But what exactly is a plan? Plans are constructed out of *plan-steps*, which prescribe actions. Plan-steps cannot be identified with the actions they prescribe, because the same action may be prescribed by more than one step in a single plan. The plan-steps must be executed in a proper order, so I will take a plan to include both the set of plan-steps and the ordering of the plan-steps.

In constructing a plan, we must keep track of the purpose of each plan-step. A plan-step is included in a plan for the purpose of achieving some particular subgoal (or the ultimate goal of the plan). Subgoals, in turn, are adopted for the purpose of achieving other subgoals (or the ultimate goal) *by* performing a specific action (executing a specific plan-step). *Causal-links*, introduced by McAllester and Rosenblitt (1991), provide a convenient mechanism for recording these purposes. In the simplest case, I will take a causal-link to have the form "$n_1 \rightarrow subgoal \rightarrow n_2 \rightarrow goal$" where n_1 and n_2 are plan-steps. This causal-link encodes the information that step n_1 is intended to achieve *subgoal*, whose purpose is to (partially) enable step n_2 to achieve *goal*. The function of the set of causal-links in a plan is essentially explanatory. It keeps track of why the plan was built in the way it was. This explanatory structure is useful both in constructing the plan in the first place and in repairing the plan if, in the course of plan execution, things do not go as planned, i.e., a plan-step fails to achieve its objective. In the general case, I will allow causal-links to have the form $n_1 \rightarrow subgoal_1 \rightarrow \ldots \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow \ldots \rightarrow goal_m$, where $subgoal_1, \ldots, subgoal_n$ and $goal_1, \ldots, goal_m$ are finite sequences of subgoals and goals. A sublink of the form $goal_1 \rightarrow goal_2$ in a causal-link signifies that $goal_1$ participates in establishing $goal_2$ without a further action being required. For now there will be just one way in which this happens. $goal_2$ can be a conjunction and $goal_1$ one of its conjuncts. In section seven, other possibilities will be introduced. These causal-links are more complex than those employed in familiar planners like SNLP (McAllester and Rosenblitt 1991) or UCPOP (Penberthy and Weld 1992), which just have the form $n_1 \rightarrow subgoal_1 \rightarrow n_2$. The additional complexity is unnecessary for their use in constructing plans, but it is convenient to include it for the purpose of proving theorems about plans.

In order to use causal-links to record the purposes of plan-steps and subgoals, two special cases must be accommodated. A subgoal might already be true, in which case it is not established by any step of the plan. In order to record this with a causal-link, it is convenient to add a "dummy step" **start**, which precedes all other steps in the plan and does not itself prescribe an action. We can then have causal-links of the form $*start* \rightarrow subgoal_1 \rightarrow \ldots \rightarrow subgoal_n \rightarrow n_1 \rightarrow goal_1 \rightarrow \ldots \rightarrow goal_m$. The other special case occurs when the subgoal is the ultimate goal of the plan. In

[2] Conditional effects are discussed by Pednault (1987), and were first implemented in UCPOP (Penberthy and Weld 1992).

that case there is no step n_2 for use in a causal-link. To enable ourselves to use a causal-link to record the purpose of n_1, we add another dummy step **finish** and then add a causal-link $n_1 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m \rightarrow$ **finish** $\rightarrow goal_m$. For technical reasons that will be discussed below, we do not require **finish** to succeed all other plan-steps. A plan will be allowed to contain "extra" plan-steps that do not participate in the achievement of its goal.

In light of the preceding considerations, it is convenient to represent a plan as an ordered quadruple ⟨*plan-steps*,*causal-links*,*ordering*,*goal*⟩. Useful plans can be formulated using this representation, although this may be too simple a representation to accommodate some of our most sophisticated planning. I will discuss this in the final section. In this paper I will confine my attention to the logical structure of goal-regression planning aimed at constructing plans that can be represented in this way.

Goal-regression planning proceeds by performing several different kinds of operations. Describing these operations constitutes a description of the logical structure of goal-regression planning.

Null Plans

The simplest case of goal-regression planning is the degenerate case in which the goal to be achieved is already true, and hence nothing needs to be done to achieve it. A *null-plan* for the goal *goal* is a plan with no plan-steps and the single causal-link **start** → *goal* → **finish** → *goal*. The degenerate case of goal-regression planning can be regarded as proceeding in accordance with the following operation:

PROPOSE-NULL-PLAN

Given an interest in finding a plan for achieving *goal*, if *goal* is already true, propose a null-plan for achieving *goal*.

Note that the action prescribed by this operation consists of *proposing* a plan rather than *adopting* a plan. To adopt a plan is to form the intention of executing it. To propose a plan is simply to make it a candidate for adoption. Not all candidates are adopted. Multiple plans may be found for a single goal, and some may be better than others. Typically only the best plan found is adopted.

Goal-Regression

The core of goal-regression planning consists of an operation that I will call "goal-regression". Regarding this as an operation that proposes a plan, it can be formulated as follows:

GOAL-REGRESSION

Given an interest in finding a plan for achieving *G*, adopt interest in finding planning-conditionals $(A/C) \Rightarrow G$ having *G* as their consequent. Given such a conditional, adopt an interest in finding a plan for achieving *C*. If a plan *subplan* is proposed for achieving *C*, construct a plan by (1) adding a new step to the end of *subplan* where the new step prescribes the action *A*, (2) ordering the new step after all steps of *subplan*, and (3) adjusting the causal-links appropriately. Propose the new plan as a plan for achieving *G*.

Splitting Conjunctive Goals

The operations PROPOSE-NULL-PLAN and GOAL-REGRESSION do not by themselves constitute a complete description of goal-regression planning. The subgoals generated by GOAL-REGRESSION will usually be conjunctions. For example, if my goal is to light a fire, I may observe that I could do so by lighting a match provided I have a match and I have tinder. GOAL-REGRESSION will thus generate the conjunctive subgoal *I have a match and I have tinder*.

We will generally be unable to make further progress in our plan construction by applying GOAL-REGRESSION once more to such a conjunctive subgoal (SG_1 & SG_2). To do so would require our having a planning-conditional of the form $(A/C) \blacktriangleright (SG_1 \text{ \& } SG_2)$. But it is rare that we will have a single planning-conditional like this that will achieve both conjuncts of a conjunctive subgoal. The most we can generally hope for is to have two separate planning-conditionals $(A_1/C_1) \blacktriangleright SG_1$ and $(A_2/C_2) \blacktriangleright SG_2$, which will allow us to construct separate subplans for the individual conjuncts. Given subplans for achieving each conjunct, we can then attempt to construct a plan for achieving the conjunction by merging the plans for the conjuncts. Given two plans $plan_1$ and $plan_2$, let $plan_1 + plan_2$ be the plan that results from combining the plan-steps, causal-links, and ordering-constraints of each, with the following exception. Where G_1 is the goal for $plan_1$ and G_2 is the goal for $plan_2$, for each causal-link $n_1 \rightarrow goal_1 \rightarrow ... \rightarrow G_1 \rightarrow {}^*finish^* \rightarrow G_1$ of $plan_1$ and causal-link $n^*_1 \rightarrow goal_1^* \rightarrow ... \rightarrow G_2 \rightarrow {}^*finish^* \rightarrow G_2$ of $plan_2$, instead of including these causal-links in $plan_1 + plan_2$, we add causal-links $n_1 \rightarrow goal_1 \rightarrow ... \rightarrow G_1 \rightarrow (G_1 \text{ \& } G_2) \rightarrow {}^*finish^* \rightarrow (G_1 \text{ \& } G_2)$ and $n^*_1 \rightarrow goal_1^* ... \rightarrow G_2 \rightarrow (G_1 \text{ \& } G_2) \rightarrow {}^*finish^* \rightarrow (G_1 \text{ \& } G_2)$ Then we can plan for conjunctive goals by using the following operation:

SPLIT-CONJUNCTIVE-GOAL

Given an interest in finding a plan for achieving a conjunctive goal (G_1 & G_2), adopt interest in finding plans $plan_1$ for G_1 and $plan_2$ for G_2. If such plans are proposed, propose $plan_1 + plan_2$ as a plan for (G_1 & G_2).

There is, however, a well-recognized logical problem for planning for conjunctive goals using SPLIT-CONJUNCTIVE-GOAL. The difficulty is that planning separately for the individual conjuncts can produce plans that destructively interfere with each other, in the sense that although the separate subplans can each be expected to achieve their goals in isolation, the merged plan cannot be expected to achieve both goals (Chapman 1987). I will explore the nature of such destructive interference shortly, but before doing that let us consider the consequences that the possibility of destructive interference has for goal-regression planning about conjunctive goals.

The fact that $plan_1 + plan_2$ cannot automatically be expected to achieve (G_1 & G_2) suggests that the operation that should actually be employed in planning for conjunctive goals is not SPLIT-CONJUNCTIVE-GOAL but rather:

SPLIT-CONJUNCTIVE-GOAL-SAFELY

Given an interest in finding a plan for achieving a conjunctive goal (G_1 & G_2), adopt interest in finding plans $plan_1$ for G_1 and $plan_2$ for G_2. If such plans are proposed *and do not destructively interfere with each other*, propose $plan_1 + plan_2$ as a plan for (G_1 & G_2).

Conventional AI planning algorithms work in this way, ruling out the possibility of internal defects before proposing plans. I will argue below that this conventional AI approach cannot work for general-purpose goal-regression planning in autonomous rational agents, but before doing that I must lay some additional groundwork.

The difference between SPLIT-CONJUNCTIVE-GOAL and SPLIT-CONJUNCTIVE-GOAL-SAFELY is that the former must be viewed as a *defeasible* rule of practical reasoning. That is, if a plan is proposed on the basis of SPLIT-CONJUNCTIVE-GOAL, the planning-agent must be prepared to withdraw the proposal if destructive interference is subsequently discovered. SPLIT-CONJUNCTIVE-GOAL must then be supplemented with additional principles aimed at proposing new plans constructed on the basis of $plan_1 + plan_2$ but avoiding the interference. If instead it is proposed to base goal-regression planning on SPLIT-CONJUNCTIVE-GOAL-SAFELY, that principle must be made more complicated by building in the additional principles that aim to repair the destructive interference, producing something like the following:

SPLIT-CONJUNCTIVE-GOAL-SAFELY

Given an interest in finding a plan for achieving a conjunctive goal (G_1 & G_2), adopt interest in finding plans $plan_1$ for G_1 and $plan_2$ for G_2. If such plans are proposed and do not destructively interfere with each other, propose $plan_1 + plan_2$ as a plan for (G_1 & G_2). If the plans do destructively interfere with each other, then search for a way of repairing $plan_1 + plan_2$ so as to avoid the interference and propose that instead.

This is vague about how to repair plans exhibiting destructive interference, but we will see below how to make that precise.

In discussing the logical structure of plan repair, I will formulate the principles as supplements to SPLIT-CONJUNCTIVE-GOAL rather than addenda to SPLIT-CONJUNCTIVE-GOAL-SAFELY, because the principles can be stated more simply that way. It should be clear that the principles to be described can be used in either way. I will argue in section three, however, that general-purpose goal-regression planning must be done defeasibly, using SPLIT-CONJUNCTIVE-GOAL, rather than nondefeasibly, using SPLIT-CONJUNCTIVE-GOAL-SAFELY. A planner using SPLIT-CONJUNCTIVE-GOAL to reason defeasibly about plans has been constructed using the OSCAR system of defeasible reasoning. I will refer to it as the *OSCAR planner*.[3]

Partial-Order Plans

Plans produced by the exclusive use of PROPOSE-NULL-PLAN, and GOAL-REGRESSION will automatically order their plan-steps linearly, because when a plan-step is added by GOAL-REGRESSION it is ordered after all the previously constructed plan-steps. Such a plan is a *linear plan*. However, when SPLIT-CONJUNCTIVE-GOAL is used to merge independent plans, the resulting plan $plan_1 + plan_2$ simply combines the ordering-constraints of $plan_1$ and $plan_2$. This can leave plan-steps from one of the subplans unordered with respect to plan-steps from the other. Such a plan is called a *partial-order plan*.

It might be supposed that a partial-order plan is not yet a complete plan. Before we can execute a plan, we must decide in what order to execute the plan-steps, so it seems that an executable plan must be linear. But there are two reasons why it is useful to produce partial-order plans in the course of goal-regression planning. First, as is generally recognized in AI planning theory, planning is made more efficient by allowing planners to produce partial-order plans.[4] If a planner is required to produce linear plans, then when SPLIT-CONJUNCTIVE-GOAL is employed to merge $plan_1$ and $plan_2$, additional ordering-constraints would have to be added arbitrarily. There will not in general be any reason to choose one set of additional ordering-constraints over another. However, as planning proceeds and the merged plan (with the arbitrary additional ordering-constraints) is extended by GOAL-REGRESSION, it may then have to be merged with other plans by additional applications of SPLIT-CONJUNCTIVE-GOAL. At that point, the arbitrarily chosen ordering-constraints may cause destructive interference between merged subplans, while a different choice of ordering-constraints would have avoided that. The planner will thus have to backtrack and try other ordering-constraints. It is more efficient to wait and not impose additional ordering-constraints until we have to. This has become known as "least-commitment planning" (Weld 1994).

There is another, less familiar reason partial-order plans are to be preferred over linear plans produced by adding arbitrary ordering-constraints. This has to do with plan execution

---

[3] An experimental version of the OSCAR planner can be downloaded from http://www.u.arizona.edu/~pollock.

[4] For a discussion of partial order planning, see Weld (1994). The matter of efficiency is addressed in Barrett and Weld (1994).

rather than plan construction. If it is inessential to the structure of a plan in what order the plan-steps are executed, then it may be best to wait until the time of execution to decide which step to execute first. The cost of executing the steps in one order rather than another may depend upon factors not known at the time of plan construction. For example, a plan for baking bread may call for turning on the oven and for retrieving the flour from the kitchen cabinet, but leave it undetermined which to do first. If we find ourselves standing next to the oven, it may best to turn on the oven first, but if we find ourselves standing next to the cabinet it may instead be best to retrieve the flour first. Thus we may lower the cost of plan-execution by adopting partial-order plans rather than linear plans.

Achieving Goals

Thus far I have relied on little more than common sense and introspection in describing the structure of goal-regression planning. To make further progress, and to make the notion of destructive interference precise, we must consider what the objective of goal-regression planning is supposed to be. If we can make that precise, we can use it to evaluate proposals for how to perform goal-regression planning.

Presumably, the objective of goal-regression planning is to produce plans that will achieve their goals. Under what circumstances will a plan achieve its goal? Contemporary AI planning theory is based upon a particular answer to this question. First consider partial-order plans. A *linearization* of a partial-order plan is a linear plan that results from adding additional ordering constraints sufficient to linearly order the plan-steps preceding **finish**. To adopt a partial-order plan is to be indifferent between its various linearizations. Accordingly, we should define:

A partial-order plan will achieve its goal iff every linearization of it will achieve its goal.

Under what circumstances will a linear plan achieve its goal? The standard answer[5] to this question proceeds in terms of the technical notion of a "result of an action-sequence". To make the standard answer work, we must also make some assumptions. The assumptions are (1) that goals are always literals[6] or conjunctions of literals, and (2) that in a planning-conditional $(A/C) \blacktriangleright P$, C and P are either literals or finite conjunctions of literals. The relaxation of these assumptions will be discussed in section six. Where P is atomic, it will be convenient to identify $\sim\sim P$ with P, so that the negation of a literal is a literal.

Let us take an *action-sequence* to be a linear sequence of actions. Given an action-sequence $\langle A_1,...,A_n\rangle$, define:

(R1) Where *start-state* is a state of affairs and *conditionals* is a set of planning-conditionals, P is a **result** of $\langle A_1,...,A_n\rangle$ relative to *start-state* and *conditionals* iff either:
- (i) $n = 0$ and P is true in *start-state*; or
- (ii) $n > 0$ and *conditionals* contains a conditional $(A_n/C) \blacktriangleright P$ such that C is a result of $\langle A_1,...,A_{n-1}\rangle$;[7] or
- (iii) $n > 0$, P is a result of $\langle A_1,...,A_{n-1}\rangle$, and *conditionals* does not contain a conditional of the form $(A_n/C) \blacktriangleright \sim Q$ such that Q is either P or a conjunct of P and C is a result of

---

[5] This is implicit in both the situation calculus (McCarthy and Hayes 1969) and ADL (Pednault 1987).

[6] A literal is either an atomic formula or the negation of an atomic formula.

[7] I assume here that if $(A/C) \blacktriangleright (P\&Q)$ is in the set of conditionals, so are $(A/C) \blacktriangleright P$ and $(A/C) \blacktriangleright Q$. This assumption is automatically satisfied when we are talking about the set of all true planning conditionals. Without this assumption, one should add a fifth condition to the effect that conjuncts of a conjunctive result are results.

$\langle A_1,...,A_{n-1}\rangle$; or

(iv) $n > 0$ and P is a conjunction whose conjuncts are results of $\langle A_1,...,A_n\rangle$.

In other words, P is a result of an action-sequence iff either P is made true by the final step of the action-sequence in accordance with a planning-conditional whose precondition is a result of the preceding subsequence of the action-sequence, or an initial segment of the action-sequence makes P true and subsequent actions in the sequence do not reverse that. Let us define:

> A plan is **sound** relative to a state *start-state* and a set of planning-conditionals iff for every linearization of the plan its goal is a result of the sequence of actions prescribed by all of its plan-steps between **start** and **finish**, relative to *start-state* and the set of planning-conditionals.

Conventional AI planning theory makes the following assumption:

Soundness Assumption

> A linear plan will achieve its goal relative to a state *start-state* iff it is sound relative to *start-state* and the set of all true planning-conditionals.

For the moment, let us follow AI planning theory in assuming this. I will return to the evaluation of the Soundness Assumption in section four.

The Soundness Assumption provides the mathematical basis for a complete theory of goal-regression planning. As I will now show, it enables us to prove the correctness of a recursive characterization of plans that will achieve their goals. The steps of the recursion are formulated to correspond to procedures of plan-construction used in goal-regression planning. The end result is a proof that when goal-regression planning is performed in accordance with certain rules, the plans it produces will achieve their goals, and if there is a plan that will achieve a particular goal, some such plan will be found by following these rules of goal-regression planning. So this will be a kind of soundness and completeness proof for goal-regression planning.

Let us define:

> A plan is **presumptively-sound relative to** a set *conditionals* of planning-conditionals and a state *start-state* iff
>
> (1) where *goal* is the goal of the plan, the plan contains a causal-link $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow goal \rightarrow {}^*finish^* \rightarrow goal$, and
>
> (2) for every causal-link $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 ... \rightarrow goal_m$ of the plan:
>
> (a) if $n \neq 1$, then for each i such that $1 \leq i < n$, $subgoal_{i+1}$ is a conjunction, $subgoal_i$ is one of its conjuncts, and the plan also contains a causal-link $n_1{}^* \rightarrow subgoal_1{}^* \rightarrow ... \rightarrow subgoal_j{}^* \rightarrow subgoal_{i+1} \rightarrow ... \rightarrow subgoal_n \rightarrow n_2{}^* \rightarrow goal_1 ... \rightarrow goal_m$ where $subgoal_j{}^*$ is the other conjunct of $subgoal_{i+1}$;
>
> (b) if $m \neq 1$, then for each i such that $1 \leq i < m$, $goal_{i+1}$ is a conjunction, $goal_i$ is one of its conjuncts, and the plan also contains a causal-link $n_1{}^* \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2{}^* \rightarrow goal_1{}^* ... \rightarrow ... goal_j{}^* \rightarrow goal_{i+1} \rightarrow ... \rightarrow goal_m$ where $goal_j{}^*$ is the other conjunct of $goal_{i+1}$;
>
> (c) if $n_1 = {}^*start^*$ then $subgoal_1$ is true in *start-state*;
>
> (d) if $n_2 \neq {}^*finish^*$ then if A is the action of n_2, "$(A/subgoal_n) \blacktriangleright goal_1$" is a member of *conditionals*;
>
> (e) if $n_1 \neq {}^*start^*$ then the plan contains a causal-link $n_3 \rightarrow subgoal_1{}^* \rightarrow ... \rightarrow subgoal_{n^*}{}^* \rightarrow n_4 \rightarrow goal_1{}^* ... \rightarrow goal_{m^*}{}^* \rightarrow subgoal_1$; and
>
> (f) n_1 is ordered before n_2 by the ordering-constraints of the plan.

To say that a plan is presumptively-sound is just to say that its causal-links encode a causal structure derived from the set of planning-conditionals and the start-state. If a plan is presumptively-sound then every linearization of the plan satisfies conditions (i), (ii), and (iv) of the definition of "result". However, the plan may fail to achieve its goal because condition (iii) may not be satisfied. In other words, there may be a linearization of the plan in which, for some causal-link $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$, $subgoal_n$ is a result of the sequence of actions prescribed by the plan-steps up through n_1, but some step n occurs between n_1 and n_2 which makes $subgoal_n$ false again before it can be "used" by n_2 to achieve $goal_1$. Let us define:

> A plan-step n of a plan **undermines a causal-link** $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$ **relative to** the plan iff there is a linearization of the plan in which n occurs between n_1 and n_2 and either $\sim subgoal_1$ or the negation of a conjunct of $subgoal_1$ is a result of the sequence of actions prescribed by the plan-steps $*start*,...,n$ in the linearization relative to the set of all true planning-conditionals.

I will also say that the plan itself undermines one of its causal-links if one of its plan-steps does so relative to that plan. Let us define:

> A plan is **causally-sound** iff it is presumptively-sound relative to a set of true planning-conditionals and the plan does not undermine any of its own causal-links.

It is possible to prove the following theorems:

Theorem 1:
If a plan is causally-sound then it is sound.

Theorem 2:
If a goal is a result of an action-sequence $\langle A_1,...,A_n \rangle$, then there is a causally-sound plan for that goal some linearization of which prescribes the actions $A_1,...,A_n$ in that order.

The proofs are in the appendix.

It is easily verified that plans produced by reasoning in accordance with PROPOSE-NULL-PLAN, GOAL-REGRESSION, and SPLIT-CONJUNCTIVE-GOAL are presumptively-sound. However, if SPLIT-CONJUNCTIVE-GOAL is used the plans may fail to be sound because they may undermine some of their own causal-links. Thus for reasoning about plans, we must add some procedures that find and, if possible, repair underminings.

Searching for Underminings and Repairing Them

There are essentially two ways to go about finding and repairing underminings. The most straightforward is to search for underminings in essentially the same way we search for plans. A plan is undermined by a subplan, consisting of a subset of the plan-steps of the plan ordered in a manner consistent with the plan. The subplan undermines the plan by achieving a goal $\sim g$ which is the negation of some subgoal, and doing so between the time g is achieved and the time it is used. The subplan must be sound, i.e., it must really achieve $\sim g$. But more is required. It must achieve $\sim g$ in the context of the larger plan. That is, the larger plan cannot, in turn undermine the subplan. This is captured by adding the rest of the steps of the original plan into the subplan, even though they are not used, and requiring that the resulting plan still achieves $\sim g$. The result of adding the unused steps to the subplan is an *embellishment* of the original plan,

which is defined as follows:

> $plan_0$ is an **embellishment** of *plan* iff (1) the plan-steps of $plan_0$ are the same as the plan-steps of *plan*, and (2) any ordering imposed by *plan* on plan-steps of $plan_0$ other than **finish** is also imposed by $plan_0$.

The intent of this definition is that **finish** need not occur at the end of $plan_0$. If it does not, then the definition of soundness given above ignores all plan-steps succeeding **finish**. The *penultimate steps* of a plan are those occurring immediately before **finish** in some linearization of the plan. It can then be proven:

Theorem 3:
A plan-step n of a *plan* undermines a causal-link $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$ of *plan* iff there is a presumptively-sound embellishment $plan_0$ of *plan* whose goal is either $\sim subgoal_1$ or the negation of a conjunct of $subgoal_1$, and (1) n is the single penultimate plan-step of $plan_0$, (2) there is a linearization of *plan* consistent with the ordering imposed by $plan_0$ in which n occurs between n_1 and n_2, and (3) $plan_0$ is sound.

An immediate corollary of theorems 1, 2, and 3 is:

Theorem 4:
A plan-step n of a *plan* undermines a causal-link $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$ of *plan* iff there is a presumptively-sound embellishment $plan_0$ of *plan* whose goal is either $\sim subgoal_1$ or the negation of a conjunct of $subgoal_1$, and (1) n is the single penultimate plan-step of $plan_0$, (2) there is a linearization of *plan* consistent with the ordering imposed by $plan_0$ in which n occurs between n_1 and n_2, and (3) $plan_0$ does not undermine any of its own causal-links.

The only plan-steps of $plan_0$ that are relevant to undermining the causal-link are those not succeeding n_2. We can always make $plan_0$ linear, so theorem 4 constitutes a recursive characterization of undermining.

Underminings are produced by embellishments, and we can search for embellishments in essentially the same way we search for plans—using PROPOSE-NULL-PLAN, SPLIT-CONJUNCTIVE-GOAL, and an analogue of GOAL-REGRESSION that takes plan-steps from *plan* rather than building new plan-steps. I will postpone the discussion of this until section four.

Once an undermined link is found, we know that the plan will not achieve its goal in accordance with the causal-structure encoded in its causal-links.[8] But it may be possible to modify the plan so as to avoid the undermining. The current planning literature recognizes two ways of doing this. The simplest is by adding ordering-constraints (MacAllester and Rosenblitt, 1991). If a plan-step n of *plan* undermines the causal-link $n_1 \rightarrow subgoal \rightarrow n_2 \rightarrow goal$, but it is consistent with the ordering-constraints of *plan* that n not occur between n_1 and n_2, then the undermining can be avoided by adding to *plan* the ordering-constraint that n not occur between n_1 and n_2. Most AI planning systems do this by "promoting" or "demoting" n, i.e., ordering it either before n_1 or after n_2. However, it is also possible to simply impose the constraint that n not occur between n_1 and n_2, without specifying whether it occurs before n_1 or after n_2. The latter decision can be left to be determined either by subsequent planning reasoning or during plan-

[8] The plan might still achieve its goal fortuitously, because of other planning-conditionals that might relate its plan-steps appropriately without being encoded in the causal-links.

execution.[9] This suggests adopting the following reasoning-schema:

ADD-ORDERING-CONSTRAINT

Given an interest in finding a plan for achieving a conjunctive goal (g_1 & g_2), and plans $plan_1$ for g_1 and $plan_2$ for g_2, if *plan&* is a putative plan for (g_1 & g_2) constructed by merging plans $plan_1$ and $plan_2$ (and possibly other plans), but a plan-step n of *plan&* undermines one of its own causal-links $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$, construct a plan *plan+* by adding the ordering-constraint that n not occur between n_1 and n_2 (if this can be done consistently) and propose *plan+* as a plan for (g_1 & g_2).

ADD-ORDERING-CONSTRAINT must be taken to make a defeasible proposal, because *plan&* might undermine more than one of its causal-links and *plan+* repairs just one of these. The other underminings would also have to be repaired, one at a time.

The other recognized way of repairing underminings is called "confrontation", and is due to Penberthy and Weld (1992).[10] This consists of adding plan-steps to *plan* in such a way that the embellishment responsible for the undermining itself becomes undermined:

CONFRONTATION

Given an interest in finding a plan for achieving a conjunctive goal (g_1 & g_2), and plans $plan_1$ for g_1 and $plan_2$ for g_2, if *plan&* is a putative plan for (g_1 & g_2) constructed by merging plans $plan_1$ and $plan_2$ (and possibly other plans), but a plan-step n of *plan&* undermines one of its own causal-links $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$ by virtue of there being an embellishment $plan_0$ that achieves P where P is either $\sim subgoal_1$ or the negation of a conjunct of $subgoal_1$, then for each causal-link $n_0 \rightarrow G_1 \rightarrow ... \rightarrow G_m \rightarrow n \rightarrow P$ of $plan_0$, adopt interest in finding a plan for achieving $\sim G_1$ (or if G_1 is a conjunction, for each conjunct of G_1, adopt interest in finding a plan for achieving its negation). If a plan *repair-plan* is proposed for achieving $\sim G_1$ or the negation of one of its conjuncts, construct a new plan *plan+* by adding to *plan&* the plan-steps, ordering-constraints, and causal-links of *repair-plan*, with the following exception. Replace each causal-link of the form $n^* \rightarrow SG_1 \rightarrow ... \rightarrow SG_n \rightarrow$ **finish** $\rightarrow SG_n$ in *repair-plan* by the causal-link $n^* \rightarrow SG_1 \rightarrow ... \rightarrow SG_n \rightarrow n_2 \rightarrow SG_n$ and order n^* before n. If this ordering is consistent, propose *plan+* as a plan for achieving (g_1 & g_2).[11]

Just as for ADD-ORDERING-CONSTRAINT, CONFRONTATION must be taken to make a defeasible proposal.

I have formulated ADD-ORDERING-CONSTRAINT and CONFRONTATION as supplements to SPLIT-CONJUNCTIVE-GOAL, but they could instead be built into SPLIT-CONJUNCTIVE-GOAL-SAFELY by rewriting the latter roughly as follows:

SPLIT-CONJUNCTIVE-GOAL-SAFELY

---

[9] The OSCAR planner works this way. Adding the constraint that one node not occur between two others is equivalent to adding the disjunctive constraint that the node occur either before the earlier node or after the later node, and adding that to a partial ordering is equivalent to adopting a disjunction of partial orderings. Another planner that works this way is DESCARTES (Joslin and Pollack 1995).

[10] They actually called the technique "separation", but it was renamed "confrontation" in Weld (1994).

[11] This rule of confrontation differs from that of Penberthy and Weld in that theirs is formulated in terms of threats rather than undermining. See below.

Given an interest in finding a plan for achieving a conjunctive goal (G_1 & G_2), adopt interest in finding plans $plan_1$ for G_1 and $plan_2$ for G_2. If such plans are proposed and $plan_1 + plan_2$ does not undermine any of its causal-links, propose $plan_1 + plan_2$ as a plan for (G_1 & G_2). If the $plan_1 + plan_2$ does undermine some of its causal-links, then search for a way of repairing $plan_1 + plan_2$ by adding ordering-constraints and/or using confrontation to avoid the interference, and propose the resulting plan instead.

Searching for Threats and Resolving Them

The OSCAR planner works in the manner just described, but most AI planning systems work somewhat differently. Instead of searching for underminings, they search for "threats". Let us define:

> A plan-step s of a plan *plan* **threatens** a causal-link $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$ of *plan* **relative to** a set *conditionals* of planning-conditionals iff (1) there is a linearization of *plan* in which s occurs between n_1 and n_2, and (2) there is a conditional "$(A/C) \blacktriangleright \sim P$" in *conditionals* where A is the action of s and P is either $subgoal_1$ or a conjunct of $subgoal_1$.

Threats are "potential underminings". A threat is "real" only if there is a linearization of *plan* that constitutes an embellishment making C true at the time s is executed. The OSCAR planner ignores threats unless they are real. By contrast, most AI planning systems, like SNLP, UCPOP, or PRODIGY, take all threats seriously and try to resolve them by either adding ordering-constraints or using confrontation. In this connection, it makes a difference whether threats result from conditional or unconditional effects of actions. If a threat is produced by an unconditional effect, then it is guaranteed to be real, because the only way the action can get into the plan is by having the preconditional for the threatening effect already established. However, if a threat is produced by a conditional effect, then the action can get into the plan by having the precondition for some other effect of that same action established, and then the threat may not turn be real. In SNLP, actions have only unconditional effects, so all threats are real, but in UCPOP and PRODIGY that is not true. Parenthetically, it seems likely that in a domain of real-world complexity, all effects of actions are conditional effects. That is, the actions can, under some circumstances, be performed without having those effects.

If a threat that is not real is resolved by adding ordering-constraints, this can lead to unnecessary ordering-constraints, and that in turn can cause trouble later in the planning process and can also make plan execution more costly than it need be. In other words, this violates the spirit of least-commitment planning.

It may seem that resolving a threat that is not real by using confrontation will add unnecessary steps to a plan and will be computationally costly, but in fact that need not be true. If the threat is not real then *plan* never makes the precondition G_1 true, in which case a null-plan suffices for achieving $\sim G_1$. The process of finding the null-plan and resolving threats to it by constructing further null-plans and resolving threats to them, etc., is exactly as costly as searching for an undermining and failing to find it. So in fact, if the searches are done optimally, resolving threats by confrontation and repairing underminings by confrontation are equally costly. The former may add extra causal-links to a plan, but it will not add extra plan-steps.

Completeness

A planning system is *complete* if, given all true relevant planning-conditionals, it can always find a plan for achieving a goal if there is one. More precisely:

A planning system is **complete** iff for every goal and start-state, if there is a plan that will achieve the goal relative to the start-state, then when given all the relevant true planning-conditionals the planning system will find some such plan.

It can be proven that a planning system that searches for plans using PROPOSE-NULL-PLAN, GOAL-REGRESSION, and SPLIT-CONJUNCTIVE-GOAL, then searches for either underminings or threats and in response to finding them modifies plans by adding ordering-constraints or using confrontation, is complete.[12] In other words, this constitutes a complete basis for goal-regression planning. It should be noted, however, that this result presupposes the syntactical constraints on goals and planning-conditionals, and is based upon the Soundness Assumption. These assumptions will be examined critically in section four.

3. R.E. Planning and Defeasible Planning

Contemporary AI planning theory is based upon what I will call *r.e. planners*. Given a planning problem, an r.e. planner runs a program that systematically searches the space of possible plans until it returns one that purports to solve the problem. What is important about such a planner is that it executes an effective computation. Defining this precisely:

A planner is **r.e.** iff the set of pairs ⟨problem,solution⟩ that characterize the planner is recursively enumerable.

In effect, contemporary goal-regression planners are based upon the three operations PROPOSE-NULL-PLAN, GOAL-REGRESSION, and SPLIT-CONJUNCTIVE-GOAL-SAFELY. There is, however, an insuperable logical problem for attempting to perform general-purpose goal-regression planning in an autonomous rational agent by running such an algorithm. The difficulty derives from the fact that any such algorithm must use some variant of SPLIT-CONJUNCTIVE-GOAL-SAFELY to handle conjunctive goals. Recall that SPLIT-CONJUNCTIVE-GOAL-SAFELY was formulated as follows:

SPLIT-CONJUNCTIVE-GOAL-SAFELY

Given an interest in finding a plan for achieving a conjunctive goal (G_1 & G_2), adopt interest in finding plans $plan_1$ for G_1 and $plan_2$ for G_2. If such plans are proposed and do not destructively interfere with each other, propose $plan_1 + plan_2$ as a plan for (G_1 & G_2). If the plans do destructively interfere with each other, then search for a way of repairing $plan_1 + plan_2$ by adding ordering-constraints and/or using confrontation to avoid the interference, and propose the resulting plan instead.

Destructive interference was cashed out in terms of either underminings or threats.

Assuming that a planner is a goal-regression planner that works as above by splitting conjunctive goals into their conjuncts and merging the plans for the conjuncts, an r.e. planner will only be possible if the set of destructive interferences is effectively computable, i.e., recursive. If the set of destructive interferences is not effectively computable, the planner will not be able to use SPLIT-CONJUNCTIVE-GOAL-SAFELY to determine whether two plans can be merged or, when there is destructive interference, whether it can be repaired.

---

[12] The proof is essentially the same as the proof of the completeness of UCPOP given in Penberthy and Weld (1992). Alternatively, see section four.

In order for destructive interference to be computable, it must be computable whether a particular condition (the negation of a precondition of one of the plan steps) is a consequence of an action under specifiable circumstances. Standard AI planning systems accomplish this by assuming that all relevant planning-conditionals are contained in a database at the time planning begins, and hence the consequences of actions can be determined by simply looking them up in a table (using unification). Such planners do no reasoning or very little reasoning about the consequences of actions, relying instead on precompiled knowledge built into the plan operators.[13]

It is useful to make a distinction between applied planning systems and planning systems that are intended to formalize and automate the planning of an autonomous rational agent (e.g., a human being). AI planning theory has had a number of practical applications, and is one of the success stories of AI. However, practical applications of AI planning theory have been largely confined to well behaved domains in which the goals are fixed and all the relevant information can be precompiled and supplied to the planner. The planner then runs a program that searches the space of possible plans (relative to the given information) until it finds a plan whose execution is guaranteed to achieve the goals. In such "applied planning", a planner is a tool used by a human being, and in order to use the tool effectively the human must prepare the ground very carefully, being sure to give the planner all the knowledge needed to solve the planning problem.

One of the ideals to which AI aspires is the construction of autonomous rational agents capable of maneuvering through a complex, variable, and often uncooperative environment. Planning will be an essential ingredient in any such agent. However, the planning problem faced by such an agent contrasts in important ways with the kind of applied planning problem that is solved by current AI planning technology. The most obvious difference is that, in sharp contrast to applied planning, it cannot be assumed that a planning agent has exactly the knowledge it needs to solve a planning problem. An autonomous agent must build its own knowledge base. The system designer can get things started by providing background knowledge, but the agent must be provided with cognitive machinery enabling its knowledge base to grow and evolve as it gains experience of its environment, senses its immediate surroundings, and reasons about the consequences of beliefs it already holds. The more complex the environment, the more the autonomous agent will have to be self-sufficient for knowledge acquisition. I have distinguished between practical cognition and epistemic cognition. The principal function of epistemic cognition in an autonomous agent is to provide the information needed for practical cognition. As such, the course of epistemic cognition is driven by practical interests. Rather than coming to the planning problem equipped with all the knowledge required for its solution, the planning problem itself directs epistemic cognition, focusing epistemic endeavors on the pursuit of information that will be helpful in solving current planning problems.

Paramount among this information is knowledge about what will happen if certain actions are taken under certain circumstances. Sometimes the agent already knows what will happen, but often it has to figure it out. At the very least this will require reasoning from current knowledge. In many cases it will require the empirical acquisition of new knowledge that cannot be obtained just by reasoning from what is already known. For example, in order to construct a plan the planning agent may have to find out what time it is, and it may be able to do that only by examining the world in some way (e.g., it may have to go into the next room and look at the clock). In general, such empirical investigations are carried out by performing actions (not just by reasoning). Figuring out what actions to perform is a matter of engaging in further planning. The agent acquires the epistemic goal of acquiring certain information, and

---

[13] This originated with STRIPS (Fikes and Nilsson 1971), which built the requisite planning-conditionals into the plan operators themselves. Subsequent AI planners have followed suit.

then plans for how to accomplish that. So planning drives epistemic investigation which may in turn drive further planning. It follows that an essential characteristic of planning agents is that planning and epistemic cognition are interleaved. Unlike applied planning, it is impossible to require of a planning agent capable of functioning in realistically complex environments that it acquire all the requisite knowledge before beginning the plan search.

Now let us apply this to the question whether human beings (and other rational agents) can perform their planning by implementing planning algorithms of the sort described in section two. As I have argued, that is only possible if destructive interference is computable, which in turn requires that the consequences of actions be computable. As we have seen, autonomous planning agents cannot rely on precompiled knowledge. They must engage in genuine reasoning about the consequences of actions, and we should not expect that reasoning to be any simpler than general epistemic reasoning. Realistically, epistemic reasoning must be defeasible, which makes the set of conclusions at best Δ_2.[14] But even if we could construct an agent that did only first-order deductive reasoning, the set of conclusions is not effectively computable—it is recursively enumerable. Even for such an unrealistically oversimplified planner, destructive interference will not be computable—the set of destructive interferences will be only r.e. This means that when the planning algorithm computes plans for the conjuncts of a conjunctive goal and then considers whether they can be merged without destructive interference, the reasoning required to find any particular destructive interference may take indefinitely long, and if there is no destructive interference, there will be no point at which the planner can draw the conclusion that there is none simply on the grounds that none has been found. Thus the planning algorithm will bog down at this point and will never be able to produce the merged plan for the conjunctive goal.[15]

If destructive interference is not computable, how can a planner get away with dividing conjunctive goals into separate conjuncts and planning for each conjunct separately? The key to this problem emerges from considering how human beings solve it. Humans assume defeasibly that the separate plans do not destructively interfere with one another, and so infer defeasibly that the merged plan is a good plan for the conjunctive goal. In other words, human goal-regression planning is based on SPLIT-CONJUNCTIVE-GOAL rather than SPLIT-CONJUNCTIVE-GOAL-SAFELY. Having made this defeasible inference, human planners then look for destructive interference that would defeat it, but they do not regard it as essential to establish that there is no destructive interference before they make the inference. And if, at the time plan execution is to begin, no destructive interference has been discovered, then we humans go ahead and execute the plan despite the fact that we have not *proven conclusively* that there is no destructive interference.

One may be tempted to suppose that human beings are making an unreasonable leap of faith here, and that a more rational agent would postpone plan execution until it has been established that there is no destructive interference. However, the logic of the epistemic search for destructive interference makes that logically impossible. Given a logically complex knowledge base, there will not, in general, be a point at which an agent can conclude with certainty that there is no destructive interference within a plan, so an agent that required such certainty would

---

[14] Δ_2 sets are sometimes called "trial and error sets". R.e. sets can be "approximated from below" by an algorithm that systematically adds members without ever having to take any out of the set. By contrast, Δ_2 sets can only be approximated "from above and below simultaneously", by an algorithm that systematically adds members, but doesn't always get them right and may have to remove members later. For further discussion of this, see Pollock (1995), chapter three.

[15] Notice that a similar problem arises in applying PROPOSE-NULL-PLAN, which may require an indeterminate amount of reasoning to determine that the subgoal is already true. If the requisite reasoning is not at least r.e., then the planning cannot be r.e..

be unable to complete and execute any of its plans.

The problem I have raised is specifically a problem for goal-regression planners that employ SPLIT-CONJUNCTIVE-GOAL-SAFELY to reason about conjunctive goals. Might we be able to circumvent this problem by employing some other kind of r.e. planner? Erol, Nau, and Subrahmanian (1975) prove that for a wide variety of STRIPS planning domains, the problem of finding a plan is at least semi-decidable, and hence r.e. planners are possible in such domains. However, this result assumes a fixed (finite) set of STRIPS operators (or equivalently, planning-conditionals). My point concerns the quite different situation in which we may have to discover new planning-conditionals in order to solve the planning problem. In fact, the following simple theorem shows there is no way around the problem I have posed for r.e. planners:

Theorem 5:

If the set of planning-conditionals is r.e. but not recursive, then the set of sound solution-pairs ⟨problem,solution⟩ is not r.e.

The upshot of this is that a rational agent operating in a realistically complex environment must make defeasible assumptions in the course of its planning, and then be prepared to change its planning decisions later if subsequent epistemic reasoning defeats those defeasible assumptions. In other words, the reasoning involved in planning must be a species of defeasible reasoning. *Planning by autonomous agents in complex environments cannot be done by an r.e. planner.*[16]

4. Reasoning Defeasibly about Plans

The general way goal-regression planning must work in autonomous rational agents is by performing goal regression, splitting conjunctive goals into their conjuncts and planning for them separately, and then merging the plans for the individual conjuncts into a combined plan for the conjunctive goal. The planning agent will infer defeasibly that the merged plan is a solution to the planning problem. A defeater for this defeasible inference consists of discovering that the plan contains destructive interference. Whenever a defeasible reasoner makes a defeasible inference, it must adopt interest in finding defeaters, so in this case the agent will adopt interest in finding destructive interference. Finding such interference should lead the agent to try various ways of repairing the plan to eliminate the interference, and then lead to a defeasible inference that the repaired plan is a solution to the planning problem. The tentative conclusion being adopted is that the plan will achieve its goal. Goal-regression planning becomes a form of epistemic reasoning to the effect that if a plan is executed (in any way consistent with the ordering) then it is defeasibly reasonable to expect the goal to be achieved. This turns PROPOSE-NULL-PLAN, GOAL-REGRESSION, SPLIT-CONJUNCTIVE-GOAL, ADD-ORDERING-CONSTRAINTS, and CONFRONTATION, into epistemic operations leading to epistemic conclusions. In other words, they are epistemic inference-schemes. GOAL-REGRESSION differs from the other rules in that it produces conclusions that (given the Soundness Assumption) follow deductively from the

---

[16] Ferguson and Allen (1994) describe a different use of defeasible reasoning in planning. They use defeasible reasoning as a way of avoiding the ramification problem. I propose my own solution to the ramification problem in my (1998a). Ginsberg (1995) explores an idea related to the defeasible approach. He considers planners that find plans that "almost always work", and shows that under certain circumstances plans for the individual conjuncts can be merged to form plans for conjunctions that "almost always work". This accommodates incomplete planning, and is done in the interest of planning efficiency. By contrast, the defeasible approach described here is intended to accommodate incomplete knowledge.

premises to which it appeals. The other rules produce conclusions that follow only defeasibly. For instance, in the absence of any reason for thinking that $plan_1 + plan_2$ is internally defective, SPLIT-CONJUNCTIVE-GOAL makes it reasonable to conclude that $plan_1 + plan_2$ will achieve (G_1 & G_2), but if it is subsequently discovered that $plan_1 + plan_2$ is internally defective, then the conclusion that it will achieve (G_1 & G_2) should be withdrawn. ADD-ORDERING-CONSTRAINTS and CONFRONTATION work similarly.

This way of understanding goal-regression planning contrasts sharply with conventional AI planning theory, which attempts to proceed non-defeasibly by employing a variant of SPLIT-CONJUNCTIVE-GOAL-SAFELY instead of SPLIT-CONJUNCTIVE-GOAL, but as I have argued, such an approach to planning cannot work as a general theory of goal-regression planning. It can only work in narrowly circumscribed contexts in which the planner can be given all relevant knowledge from the beginning and does not have to engage in epistemic reasoning during the course of the planning.

I will assume the general theory of defeasible reasoning embodied in OSCAR (Pollock 1995), and the implementation discussed below will be based upon the implemented OSCAR architecture. Reasoning in OSCAR consists of the construction of natural-deduction-style arguments, using both deductive inference rules and defeasible reason-schemas. Premises are input to the reasoner (either as background knowledge or as new percepts), and queries are passed to the reasoner. OSCAR performs bidirectional reasoning. The reasoner reasons forwards from the premises and backwards from the queries. The queries are "epistemic interests", and backwards reasoning can be viewed as deriving interests from interests.

The complete set of inference-schemes required for this approach to goal-regression planning can be formulated as follows:

PROPOSE-NULL-PLAN

Given an interest in finding a plan for achieving *goal*, if *goal* is already true, infer non-defeasibly that a null-plan will achieve *goal*.

GOAL-REGRESSION

Given an interest in finding a plan for achieving G, adopt interest in finding planning-conditionals $(A/C) \Rightarrow G$ having G as their consequent. Given such a conditional, adopt an interest in finding a plan for achieving C. If it is concluded that a plan *subplan* will achieve C, construct a plan by (1) adding a new step to the end of *subplan* where the new step prescribes the action A, (2) ordering the new step after all steps of *subplan*, and (3) adjusting the causal-links appropriately. Infer nondefeasibly that the new plan will achieve G.

SPLIT-CONJUNCTIVE-GOAL

Given an interest in finding a plan for achieving a conjunctive goal (G_1 & G_2), adopt interest in finding plans $plan_1$ for G_1 and $plan_2$ for G_2. If such plans are proposed, infer defeasibly that $plan_1 + plan_2$ will achieve (G_1 & G_2).

ADD-ORDERING-CONSTRAINT

Given an interest in finding a plan for achieving a conjunctive goal (g_1 & g_2), and plans $plan_1$ for g_1 and $plan_2$ for g_2, if *plan&* is a putative plan for (g_1 & g_2) constructed by merging plans $plan_1$ and $plan_2$ (and possibly other plans), but a plan-step n of *plan&* undermines one of its own causal-links $n_1 \rightarrow subgoal_1 \rightarrow \ldots \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow \ldots \rightarrow goal_m$, construct a plan *plan+* by adding the ordering-constraint that n not occur between n_1 and n_2 (if this can be done consistently) and infer defeasibly that *plan+* will achieve (g_1 & g_2).

CONFRONTATION

Given an interest in finding a plan for achieving a conjunctive goal (g_1 & g_2), and plans $plan_1$ for g_1 and $plan_2$ for g_2, if *plan&* is a putative plan for (g_1 & g_2) constructed by merging plans $plan_1$ and $plan_2$ (and possibly other plans), but a plan-step n of *plan&* undermines one of its own causal-links $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$ by virtue of there being an embellishment $plan_0$ that achieves P where P is either $\sim subgoal_1$ or the negation of a conjunct of $subgoal_1$, then for each causal-link $n_0 \rightarrow G_1 \rightarrow ... \rightarrow G_m \rightarrow n \rightarrow P$ of $plan_0$, adopt interest in finding a plan for achieving $\sim G_1$ (or if G_1 is a conjunction, for each conjunct of G_1, adopt interest in finding a plan for achieving its negation). If a plan *repair-plan* is proposed for achieving $\sim G_1$ or the negation of one of its conjuncts, construct a new plan *plan+* by adding to *plan&* the plan-steps, ordering-constraints, and causal-links of *repair-plan*, with the following exception. Replace each causal-link of the form $n^* \rightarrow SG_1 \rightarrow ... \rightarrow SG_n \rightarrow {}^*finish^* \rightarrow SG_n$ in *repair-plan* by the causal-link $n^* \rightarrow SG_1 \rightarrow ... \rightarrow SG_n \rightarrow n \rightarrow SG_n$ and order n^* between n_0 and n. If this ordering is consistent, infer defeasibly that *plan+* will achieve (g_1 & g_2).

UNDERMINE-CAUSAL-LINKS

Given an inference in accordance with SPLIT-CONJUNCTIVE-GOAL, ADD-ORDERING-CONSTRAINT, or CONFRONTATION to the conclusion that *plan&* will achieve (G_1 & G_2), adopt interest in establishing that *plan&* undermines one of its own causal-links. If it is determined that it does undermine one of its own causal-links, take the inference to the conclusion that *plan&* will achieve (G_1 & G_2) to be defeated.

Let us say that a plan *achieves* its goal *between* two plan-steps iff it achieves its goal and all its penultimate steps are ordered between the two plan-steps. Similarly, a plan achieves its goal *before* a plan-step iff it achieves its goal and all its penultimate steps are ordered before the plan-step.

UNDERMINE-CAUSAL-LINK

Given an interest in establishing that *plan&* undermines one of its own causal-links, for each causal-link $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$ of *plan&*, adopt interest in finding an embellishment $plan_0$ of *plan&* that achieves $\sim g$ between n_1 and n_2 consistent with the ordering-constraints of *plan&*, where g is either $subgoal_1$ or a conjunct of $subgoal_1$. Given $plan_0$, infer nondefeasibly that *plan&* undermines one of its own causal-links.

The search for embellishments can be performed using analogues of the inference-schemes used in searching for plans in the first place, with the difference that the analogues use only the plan-steps of *plan&*. They start with *plan&* stripped of its causal-links, and simply add causal-links and ordering-constraints.

EMBEDDED-GOAL-REGRESSION

Given an interest in finding an embellishment of *plan* that achieves G before n_2 (and optionally after n_1) consistent with a set of ordering-constraints *order*, adopt interest in finding planning-conditionals $(A/C) \Rightarrow G$ having G as their consequent for which there is a plan-step n of *plan* such that (1) the action prescribed by n is A, and (2) it is consistent with *order* that n occur before n_2 (and optionally after n_1). Given such a conditional and plan-step, let *order+* be the result of adding to *order* the constraint that n occur before n_2 (and optionally after n_1). Adopt an interest in finding an embellishment of *plan* that achieves C before n consistent with *order+*. If an embellishment $plan_0$ is proposed for achieving C before n consistent with *order+*, construct an embellishment

plan+ by adding a causal-ink to record the achievement of *G* by *n* and adjusting the ordering-constraints accordingly. Infer defeasibly that *plan+* is an embellishment of *plan* that achieves *G* before n_2 (and optionally after n_1) consistent with *order*.

EMBEDDED-GOAL-REGRESSION, unlike GOAL-REGRESSION, is defeasible. This is because in EMBEDDED-GOAL-REGRESSION, the causal-link achieving *G* may be undermined by other plan-steps in *plan* that can be consistently ordered between the causal-link root and the causal-link target. In GOAL-REGRESSION, on the other hand, there are no plan-steps in the plan so far constructed that can be consistently ordered between the causal-link root and the causal-link target. The defeat of EMBEDDED-GOAL-REGRESSION is accomplished by the following variant of UNDERMINE-CAUSAL-LINKS:

UNDERMINE-EMBEDDED-CAUSAL-LINKS

Given an inference in accordance with EMBEDDED-GOAL-REGRESSION, ADD-EMBEDDED-ORDERING-CONSTRAINT, or EMBEDDED-CONFRONTATION to the conclusion that *plan+* is an embellishment of *plan* that achieves *G* before n_2 (and optionally after n_1) consistent with *order*, adopt interest in establishing that *plan+* undermines one of its own causal-links. If it is determined that it does undermine one of its own causal-links, take the inference to the conclusion that *plan+* is an embellishment of *plan* that achieves *G* before n_2 (and optionally after n_1) consistent with *order* to be defeated.

Embedded goal-regression terminates with goals that are already established:

EMBEDDED-NULL-PLAN

Given an interest in finding an embellishment of *plan* that will achieve *goal* before plan-step *n* consistent with *order*, if *goal* is already true, construct $plan_0$ by (1) letting its plan-steps be the plan-steps of *plan*, (2) letting the ordering-constraints of $plan_0$ be *order*, and (3) taking the only causal-link to be **start** → *goal* → **finish** → *goal*. From the truth of *goal* infer nondefeasibly that $plan_0$ is an embellishment of *plan* that will achieve *goal* before plan-step *n* consistent with *order*.

SPLIT-EMBEDDED-CONJUNCTIVE-GOAL

Given an interest in finding an embellishment of *plan* that will achieve a conjunctive goal (G_1 & G_2) before plan-step *n* consistent with *order*, adopt interest in finding embellishments $plan_1$ and $plan_2$ that will achieve G_1 and G_2 respectively before plan-step *n* consistent with *order*. If such embellishments are proposed, infer nondefeasibly that $plan_1 + plan_2$ is an embellishment of *plan* that will achieve a conjunctive goal (G_1 & G_2) before plan-step *n* consistent with *order*.

Note that unlike SPLIT-CONJUNCTIVE-GOAL, SPLIT-EMBEDDED-CONJUNCTIVE-GOAL is not defeasible. This is because $plan_1 + plan_2$ has the same plan-steps as both $plan_1$ both $plan_2$, so if a causal-link of either $plan_1$ or $plan_2$ is undermined by $plan_1 + plan_2$, it will already be undermined in $plan_1$ or $plan_2$ itself when it is constructed by EMBEDDED-GOAL-REGRESSION.

We can try to repair embellishments that undermine their own causal-links by either adding ordering-constraints or confrontation:

ADD-EMBEDDED-ORDERING-CONSTRAINT

Given an interest in finding an embellishment of *plan* that achieves *G* before n_2 (and optionally after n_1) consistent with a set of ordering-constraints *order*, if *plan+* is a putative such embellishment but a plan-step *n* of *plan+* undermines one of its own

causal-links $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$, construct a plan *plan*++ by adding the ordering-constraint that n not occur between n_1 and n_2 (if this can be done consistently) and infer defeasibly that *plan*++ is an embellishment of *plan* that achieves G before n_2 (and optionally after n_1) consistent with a set of ordering-constraints *order*.

EMBEDDED-CONFRONTATION

Given an interest in finding an embellishment of *plan* that achieves G before n_2 (and optionally after n_1) consistent with a set of ordering-constraints *order*, if *plan*+ is a putative such embellishment but a plan-step n of *plan*+ undermines one of its own causal-links $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$ by virtue of there being an embellishment $plan_0$ of *plan*+ that achieves P where P is either $\sim subgoal_1$ or the negation of a conjunct of $subgoal_1$, then for each causal-link $n_0 \rightarrow G_1 \rightarrow ... \rightarrow G_m \rightarrow n \rightarrow P$ of $plan_0$, adopt interest in finding an embellishment plan++ of plan+ that achieves $\sim G_1$ (or if G_1 is a conjunction, achieves the negation of one conjunct of G_1) between n_0 and n. If such an embellishment *repair-plan* is is found, construct a new embellishment *plan*++ of *plan* by adding to *plan*+ the ordering-constraints and causal-links of *repair-plan*, with the following exception. Replace each causal-link of the form $n^* \rightarrow SG_1 \rightarrow ... \rightarrow SG_n \rightarrow$ **finish** $\rightarrow SG_n$ in *repair-plan* by the causal-link $n^* \rightarrow SG_1 \rightarrow ... \rightarrow SG_n \rightarrow n \rightarrow SG_n$ and order n^* between n_0 and n. If this ordering is consistent, infer defeasibly that *plan*++ is an embellishment of *plan* that achieves G before n_2 (and optionally after n_1) consistent with a set of ordering-constraints *order*.

As before, ADD-EMBEDDED-ORDERING-CONSTRAINT and EMBEDDED-CONFRONTATION repair underminings one at a time. Further underminings may remain, so the inferences in accordance with these two inference-schemes are defeasible in accordance with UNDERMINE-EMBEDDED-CAUSAL-LINKS.

Evaluating a defeasible planner

An r.e. planner is evaluated by asking whether it is sound and complete. It is sound if every plan it proposes for achieving a goal is a sound plan, and it is complete if it finds a sound plan for achieving a goal whenever one exists. But how can we evaluate a defeasible planner? It will inevitably find unsound plans. Hopefully, it will retract them later.

A distinction can be made between the conclusions that a defeasible reasoner is *justified* in holding, at any given stage of its reasoning, and the *warranted conclusions* that it will be justified in holding *at the limit*, when all possible relevant reasoning has been performed.[17] What we want of a defeasible planner is that it will eventually draw warranted conclusions that constitute sound solutions to planning-problems. Let us call the plans endorsed by warranted conclusions *warranted plans*. A first pass at a criterion of adequacy would require that warranted plans are always sound, and whenever there is a sound plan for achieving a goal there will be a sound warranted plan. However, this criterion of adequacy is still too strong. The reasoner might be warranted in taking an unsound plan to be sound simply because the reasoner is unable to draw the conclusion that some relevant fact about the world is a fact or that some relevant consequence of an action is a consequence of that action. For the same reason it may be unable to find some sound plan.

We can usefully separate the plan-reasoning from the reasoning aimed at finding factual knowledge of use in the planning. The reason-schemas used in planning may be beyond reproach,

[17] This is made more precise in chapter three of Pollock (1995).

but the reasoner may still find unsound plans and fail to find sound ones because its factual reasoning is inadequate. This separation can be achieved by noting that the concept of a sound plan was defined to be relative to a start-state and a set of planning-conditionals. A plan is **sound** relative to a start-state and set of planning-conditionals iff for every linearization of the plan its goal is a result of the sequence of actions prescribed by its plan-steps relative to the start-state and the set of planning-conditionals. In defeasible planning, the relevant start-state consists of the set of all warranted conclusions, and the relevant set of planning-conditionals is the set of all warranted planning-conditionals. We can then define defeasible planner to be sound iff all its warranted plans are sound relative to the set of warranted conclusions and warranted planning-conditionals, and it is complete iff whenever there is a plan for achieving a goal that is sound relative to the set of its warranted conclusions and warranted planning-conditionals, the planner is able to find some such plan.

We can evaluate the soundness and completeness of a defeasible planner by simply *giving* the reasoner all the factual knowledge (including knowledge of planning-conditionals) that is relevant to solving the problem, and then asking whether under those circumstances all its warranted plans are sound and whether it is always able to find a sound plan for achieving a goal when there is a one. Giving the reasoner all the relevant factual knowledge has the effect of turning the defeasible planner into an r.e. planner. Search for planning conditionals or goals true in the start state will terminate after a single step, so for each plan there will be a determinate point at which there is no more relevant reasoning to be done. We can take the planner to "return" the plan iff at that point it is justified in concluding that the plan will achieve its goal. Because all the relevant reasoning has been done, the plan will be warranted iff the reasoner is justified in drawing that conclusion at that point.

The r.e. planner that is generated in this way by the defeasible plan reasoning described above can be shown to be sound and complete by appealing to theorems 1 and 4. For this purpose we need the assumption that the control structure for the defeasible reasoner searches the space of potential inferences systematically. It then follows that the inference-schemes have the effect of systematically expanding the recursive characterization of undermining provided by theorem 4. The result is:

Theorem 6:

If OSCAR searches the space of potential inferences systematically then the OSCAR planner is sound.

Theorem 7:

If OSCAR searches the space of potential inferences systematically then the OSCAR planner is complete.

5. Planning and the Frame Problem

I have presented a tentative account of the logical structure of goal-regression planning in autonomous rational agents. This account differs in some important ways from conventional AI planning theory, but it also makes heavy reliance on certain aspects of the conventional theory. In particular, it turns on the Soundness Assumption, according to which a linear plan achieves a goal relative to a start-state iff the goal is a result of the sequence of actions prescribed by the plan-steps relative to the start-state and the set of all true planning-conditionals, where "result" is a technical concept that was defined by (R1). To evaluate the Soundness Assumption we must consider more carefully what it means. It seems to say the following:

> Necessarily, a linear plan will achieve a goal *G* when executed from a start-state iff *G* is a result of the sequence of actions prescribed by its plan-steps relative to the start-state and the set of all true planning-conditionals

But so interpreted, the Soundness Assumption is obviously false (as, I think, everyone in AI agrees). The difficulty concerns clause (iii) of (R1). Clause (iii) asserts that once a subgoal has been established, it will remain true unless some later step of the plan makes it false. The world is a dynamic, continually changing place. It is certainly not a necessary truth that subgoals established by earlier steps of a plan will not be made false by events extraneous to the plan before the subgoals can be used in establishing further goals.

In AI it is often claimed that goal-regression planning relies upon the so-called "STRIPS assumption" according to which nothing changes in the world unless it does so as a result of executing a step of the plan.[18] But such an assumption is obviously silly. We engage in goal-regression planning all the time without believing the STRIPS assumption, so the STRIPS assumption cannot provide the logical basis for our planning.

We do not expect that nothing will change in the world unless we change it, but we do expect our plan to work. This means that we have a *limited* expectation, not that nothing will change, but that the particular subgoals established by initial steps of the plan will not change unless executing later steps of the plan causes them to change. We certainly do not believe that plans will *never* be disrupted by extraneous events, but we do expect that not to happen in any particular case. We are, however, always prepared to be proven wrong. In other words, our expectation is defeasible. We know that things change, but there is a presumption against it in any particular case.

Providing the logical foundations for such a defeasible expectation is just the *Frame Problem*. Early attempts in AI to give a logical reconstruction of reasoning about the consequences of actions tried doing so by axiomatizing the domain and then reasoning about it deductively. It quickly became apparent that such an approach required not only axioms describing how things change, but also a much larger set of "frame axioms" describing when things don't change.[19] Getting such axioms right in a complex domain seems to be a practical impossibility, and even if we had such axioms the deductive reasoning task would be made infeasible by the huge number of frame axioms required. The Frame Problem became the problem of finding some feasible way of reasoning about both change and non-change.[20] That is precisely the problem facing us here. In goal-regression planning we want to be able to assume defeasibly that the truth values of our subgoals will not change until we do something to change them, and we want to use that assumption in reasoning about what will change as a result of executing the plan-steps for which the subgoals are the preconditions.

AI researchers quickly gave up the attempt to solve the Frame Problem deductively, and proposed instead that there is a defeasible presumption that things don't change. The thinking was that given such a defeasible presumption, the only substantive principles we need are causal principles overriding the defeasible presumption in specific cases.[21] I have recently explored ways of making this reasoning precise (and implementing it) within the OSCAR system

---

[18] See Allen (1987) and Lifschitz (1987a).

[19] McCarthy and Hayes (1969).

[20] There is a lot of disagreement about what the Frame Problem really is. For historical substantiation of my interpretation of it, see my (1998a).

[21] McCarthy and Hayes (1969).

of defeasible reasoning, and I will summarize my results here.[22]

Temporal-Projection

As a first approximation, we can formulate a defeasible presumption against change as follows:

(1) If $t_0 < t_1$, believing P-*at*-t_0 is a defeasible reason for the agent to believe P-*at*-t_1, the strength of the reason being a monotonic decreasing function of $t_1 - t_0$.[23]

Principle (1) is a principle of *temporal-projection*. It amounts to a presumption that P's being true is a stable property of a time. A stable property is one such that if it holds at one time, the probability is high that it will continue to hold at a later time. Some such principle seems to be presupposed by much of our reasoning about the world.[24] However, as formulated, principle (1) is too strong. A constraint must be imposed on P. This is best demonstrated with an example, diagrammed in figure 1.[25] Let P and Q be unrelated propositions. Suppose we know that P is true at t_0, and false at the later time t_1. Consider a third time t_2 later than t_1. P-at-t_0 gives us a defeasible reason for expecting P-at-t_2, but ~P-at-t_1 gives us a stronger reason for expecting ~P-at-t_2, because $(t_2 - t_1) < (t_2 - t_0)$. Thus an inference to P-at-t_2 is defeated, but an inference to ~P-at-t_2 is undefeated. This is as it should be. However, from P-at-t_0 we can deductively infer $(P \vee Q)$-at-t_0. Without any restrictions on the proposition variable in temporal-projection, $(P \vee Q)$-at-t_0 gives us a defeasible reason for expecting $(P \vee Q)$-at-t_2. Given the inference to ~P-at-t_2, we can then infer Q-at-t_2. In diagramming these inferences in figure 1, the solid arrows symbolize deductive inferences, and bars connecting arrows indicate that the inference is from multiple premises. The "fuzzy" arrow symbolizes a defeat relation. In this inference-graph, the conclusion Q-*at*-t_2 is undefeated. But this is unreasonable. Q-*at*-t_2 is inferred from $(P \vee Q)$-*at*-t_2. $(P \vee Q)$ is expected to be true at t_2 only because it was true at t_0, and it was only true at t_0 because P was true at t_0. This makes it reasonable to believe $(P \vee Q)$-*at*-t_2 only insofar as it is reasonable to believe P-*at*-t_2, but the latter is defeated. This example illustrates clearly that temporal-projection does not work equally well for all propositions. In particular, the set of propositions for which temporal-projection works is not closed under disjunction. Let us label those propositions for which it does work *temporally-projectible*. A principle of temporal-projection must be restricted to temporally-projectible propositions:

TEMPORAL-PROJECTION

If P is temporally-projectible and $t_0 < t_1$, believing P-*at*-t_0 is a defeasible reason for the agent to believe P-*at*-t_1, the strength of the reason being a monotonic decreasing function of $t_1 - t_0$.

---

[22] Pollock (1997) and (1998a).

[23] For a discussion of reason-strength, see my (1995).

[24] For arguments to the effect that such reasoning is pervasive, see my (1997).

[25] An example with the same structure is presented by Myers and Smith (1988).

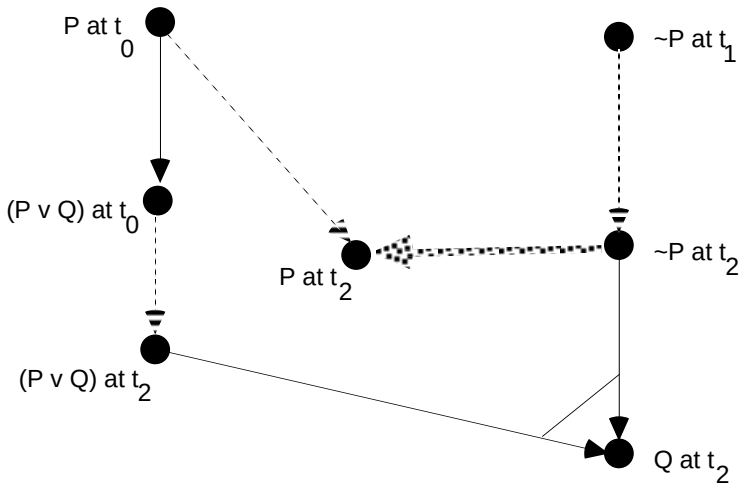

Figure 1. The need for a temporal projectibility constraint.

What are called "projectibility problems" arise in a number of places in philosophical epistemology. Goodman (1955) first showed that inductive reasoning does not work equally well for all properties—that principles of induction require a projectibility constraint. In my (1972) I showed that many projectibility problems result from attempting to employ induction with respect to disjunctions. In my (1990) I showed that similar projectibility problems arise in other contexts—the statistical syllogism, direct inference, and statistical induction. In all of these contexts, disjunctions create major difficulties. Apparently, the same conclusion must be drawn for temporal-projection.

The need for a projectibility constraint is clear, but the exact content of the constraint is not. Disjunctions create projectibility problems, but they are not the only culprits. It is easy to see that conjunctions of temporally-projectible propositions are temporally-projectible. If we have an undefeated reason for believing P-*at*-t_1 and an undefeated reason for believing Q-*at*-t_1, then we can infer (P & Q)-*at*-t_1 deductively, so the latter inference cannot be problematic. On the other hand, the negation of a conjunction is equivalent to a disjunction, so the negations of temporally-projectible propositions are not automatically temporally-projectible. This is clear for the negations of logically complex temporally-projectible propositions, but it also seems to be true for atomic propositions (however exactly this is to be understood). The ascriptions of "simple" properties to objects will generally be projectible, but the negations of such ascriptions need not be. For instance, "*x* is red" would seem to be temporally-projectible. But "*x* is not red" is equivalent to a disjunction "*x* is blue or green or yellow or orange or ...", and as such it would seem to be temporally-unprojectible. We can make many such observations about temporal-projectibility, but I do not have a general criterion of temporal-projectibility to propose. The literature contains no good theories of projectibility in any of its guises.[26] Constructing such a theory is at this time an unsolved philosophical problem.

The Frame Problem Resurrected

TEMPORAL-PROJECTION was originally proposed as a solution to the Frame Problem. However, TEMPORAL-PROJECTION turns out to be only part of the solution, as was first shown by Hanks and McDermott (1986). To illustrate (with a different example than theirs), suppose there is a causal law to the effect that if a match is dry and it is struck then it will burn. Suppose

[26] See Stalker (1994) for a compendium of work on projectibility.

we have a match that is initially known to be dry, at time t_0. Shortly thereafter, at time t_1, it is struck. We want to be able to conclude that it will light at some time t_2 (> t_1). It may seem that TEMPORAL-PROJECTION allows us to make this inference. The match was known to be dry at t_0, so TEMPORAL-PROJECTION gives us a reason for expecting it to still be dry at t_1. Then on the basis of the law we can infer that the match will burn at some time $t_2 > t_1$. However, as Hanks and McDermott observed, we also know that the match is not burning at time t_0, and so (assuming temporal-projectibility) TEMPORAL-PROJECTION gives us a reason for thinking it will not be burning at time t_2. This conflicts with the conclusion that it will burn at time t_2. Thus TEMPORAL-PROJECTION does not favor either the conclusion that the match will burn or the conclusion that the match will not burn. This is diagrammed in figure two. But, intuitively, we want to conclude defeasibly that the match will still be dry at t_1 and hence will burn at t_2. Thus TEMPORAL-PROJECTION does not solve the Frame Problem.[27]

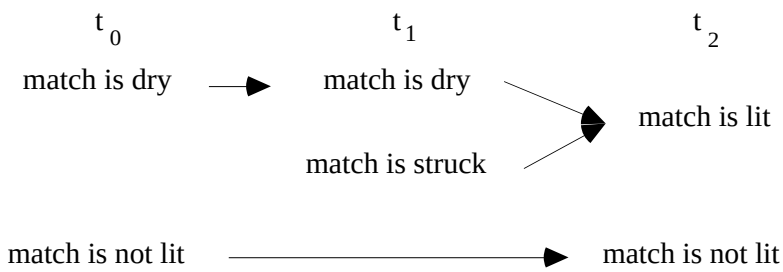

Figure 2. The Yale Match Lighting Problem

There is a kind of consensus among researchers dealing with the frame problem that the solution to this problem lies in performing the temporal-projections in temporal order.[28] We first use TEMPORAL-PROJECTION to infer that the match is still dry at time t_1. At that point, nothing has yet happened to block the application of TEMPORAL-PROJECTION, so we make this inference. From this we can infer that the match will burn at t_2. At time t_2, we can also try to use TEMPORAL-PROJECTION to infer that the match will not burn, but this time something has already happened (the match was struck while dry) to block the projection, and so we do not infer that the match will not burn. This general idea was first suggested by Shoham (1986), and subsequently endorsed by Hanks and McDermott (1987), Lifschitz (1987), and others. I will follow the literature in calling this *chronological minimalization* (changes are minimized in chronological order).

Attempts to formalize chronological minimalization have met with mixed success, largely, I think, because they were based upon inadequate theories of defeasible reasoning. In addition, Kautz (1986) proposed a troublesome counterexample which seems to show that there is something wrong with the fundamental idea underlying chronological minimalization. Modifying his example slightly, suppose I leave my car in a parking lot at time t_0. I return at time t_3 to find it missing. Suppose I know somehow that it was stolen either at time t_1 or time t_2, where $t_0 < t_1 < t_2 < t_3$. Intuitively, there should be no reason to favor one of these times over the other as the time

[27] This is formulated more precisely in my (1998a).

[28] See Hanks and McDermott (1987). A number of more recent papers explore this same idea. In planning theory there is a different kind of consensus, namely, that we should avoid trying to solve the Frame Problem in its fully generality and just run a program that gets around it. That was a large part of the motivation for both the STRIPS and ADL representation of actions. But one of the points of this paper is that it is worth taking the Frame Problem seriously in order to avoid some of the limitations of those representations.

the car was stolen. However, chronological minimalization would have us use temporal-projection first at t_1 to conclude that the car was still in the lot, and then because the car was stolen at either t_1 or t_2, we can conclude that the car was stolen at t_2. This seems completely unreasonable.

The difference between the cases in which chronological minimalization gives the intuitively correct answer and the cases in which it does not seems to be that in the former there is a set of temporal-projections that are rendered inconsistent by a causal connection between the propositions being projected. In the latter case, there is also a set of temporal-projections not all of which can be correct, but the inconsistency does not result from a causal connection. So, for example, the match case is causal, but the stolen car case is not.

In my (1996), (1997) and (1998a) I suggested a way of making this precise and implementing the reasoning. For present purposes, most of the details of that account are irrelevant. It is useful, however, to consider one aspect of the account given there. Thus far, I have talked about "planning-conditionals", without trying to say just what kind of conditionals these are. It is clear, however, that to make goal-regression planning work, they must express the kind of causal-connections that are involved in the solution to the Frame Problem. My proposal is that the requisite causal connection is that expressed by a *law of nature*. These are exceptionless generalizations that reflect the ultimate causal structure of the world. It is customary in philosophy to contrast these with "accidental generalizations", which happen to be true but might have been false. For example, suppose in all the history of the world there has been just one green-eyed mathematician named "Bartholemew", and there will never be another one. Suppose Bartholemew disliked coffee. Then it is true that every green-eyed mathematician named "Bartholemew" dislikes coffee. But this is only accidentally true, and is not a law of nature. Accidental generalizations do not support causal connections. For example, renaming a green-eyed mathematician "Bartholemew" will not cause him to dislike coffee.

Laws of nature entail ordinary universal generalizations, but not every universal generalization counts as a law of nature. There is an extensive philosophical literature on the topic of laws of nature, but this is not the place to review it. I will simply assume the account given in my (1990). Laws of nature can be formulated using *nomic generalizations*. These relate predicates and relations, and can be expressed in the form "Any *A* would be a *B*". Where φ and θ are open formulas, we can write the nomic generalization "Any φ would be a θ" as "$\varphi \Rightarrow \theta$". '$\Rightarrow$' is a variable-binding operator, binding all free occurrences of variables in φ and θ. Let us define the modal operators $\Diamond_p$ and $\Box_p$ of "physical possibility" and "physical necessity" by taking $\Diamond_p P$ to mean that P is logically consistent with the set of all true nomic generalizations, and $\Box_p P$ to mean $\sim\Diamond_p\sim P$. '$\Box_p$' turns out to be an S5 modal operator. Let $x_1,...,x_n$ be the variables having free occurrences in φ and θ. It can be shown that if $\Diamond_p\exists x_1,...,x_n\varphi$ then $\varphi \Rightarrow \theta$ iff $\Box_p\forall x_1,...,x_n(\varphi \supset \theta)$.[29] If $\Diamond_p\exists x_1,...,x_n\varphi$ holds, the generalization $\varphi \Rightarrow \theta$ is said to be "non-counterlegal". In planning, we are only interested in non-counterlegal generalizations, because we are only interested in generalizations concerning goals and subgoals that could actually be achieved.

What appears to be crucial to the nomic generalizations that can be used for the kind of causal reasoning that occurs in the Frame Problem is that there are built-in temporal references in φ and θ and φ is about earlier times than θ.[30] This is required for us to be able to do the causal reasoning by performing temporal-projections in chronological order. In my (1998a), I explored the logical structure of causal reasoning employing nomic generalizations of the form

---

[29] This is all shown in my (1990).

[30] See my (1998a) and (1997) for more detail about the exact form of these temporal references.

(CS) $\{(A\text{-at-}t \ \& \ C\text{-}at\text{-}t) \Rightarrow (\exists\delta)G\text{-throughout-}(t+\varepsilon\ ,\ t+\varepsilon+\delta]\}$.[31]

This says that performing A when C is true is causally sufficient for making G true after an interval ε. (CS) takes account of the fact that causation can take time, a fact that has been heretofore ignored in this paper. (In effect, I have been assuming that $\varepsilon = 0$.) The interval $(t+\varepsilon\ ,\ t+\varepsilon+\delta]$ is open on the left and closed on the right.[32] Consequently, (CS) does not allow us to infer deductively that G is true at any particular time—only that it is true at *some* time succeeding $t+\varepsilon$. However, if G is temporally-projectible, TEMPORAL-PROJECTION allows us to go on to infer defeasibly that G is true at *any* time succeeding $t+\varepsilon$. I will take planning-conditionals to have the form of (CS), abbreviating (CS) as $(A/\ SG) \blacktriangleright_{\varepsilon} G$. When $\varepsilon = 0$, I will omit it, writing just '$\blacktriangleright$'.

To make it clear how planning depends upon temporal projection, I will make the temporal reference explicit in a plan-step. Rather than just writing the plan-step as "A", I will write it as "A-*at*-t", where t is either a real number (designating a time) or a variable. Then I will take the ordering constraints in a plan to be ordering constraints on the time designators rather than the plan-steps themselves. So, for instance, when we use the planning-conditionals $(A/\ C) \blacktriangleright_{\varepsilon} SG$ and $(A^*/\ SG) \blacktriangleright_{\varepsilon^*} G$ to construct a plan with a causal-link $A\text{-}at\text{-}t_1 \rightarrow SG\text{-}at\text{-}t_2 \rightarrow A^*\text{-}at\text{-}t_2 \rightarrow G\text{-}at\text{-}t_3$, the ordering-constraints that must be added to the plan are $\{t_1+\varepsilon < t_2,\ t_2+\varepsilon^* < t_3\}$.

The solution to a planning problem is a set of actions and constraints on the times those actions are performed such that causal reasoning of the sort involved in the Yale Match Lighting Problem will enable us to infer defeasibly that the goal will be true at the desired time. In effect, a plan corresponds to a defeasible causal argument, and the planning problem is to find premises for such an argument from which we can infer defeasibly that the desired goal will be true. Note that the argument corresponding to the plan nowhere mentions the plan itself. The argument is just about actions and their consequences. The plan is an artifact of the way in which we go about finding appropriate premises for the argument. Finding the premises is a difficult problem, and standard planning procedures solve that problem by introducing plans as structures and reasoning about them.[33]

Thus far I have been assuming that an action A is performed *at* an instant t, and that a planning conditional will require C to be true at that same instant. However, realistically, most actions must be performed over an interval rather than at an instant. Performing the action is a process that takes time, and C may be required at some intermediate point in that process rather than at the beginning of the process. For instance, think of serving the ball in tennis. One must first throw the ball into the air so that it will arrive at a point x at a time t. Then one swings the tennis racket so that it will hit the ball at point x at time t. But one must begin the swing earlier than time t. So the ball's being at point x is required midway through the performance of the swing-action rather at the beginning. This can be accommodated by employing more complex planning-conditionals having the logical form

$$\{(A^*\text{-}at\text{-}t \ \& \ SG\text{-}at\text{-}t+\alpha) \Rightarrow (\exists\delta)G\text{-}throughout\text{-}(t+\varepsilon\ ,\ t+\varepsilon+\delta]\}.$$

---

[31] $G\text{-}throughout\text{-}[t,t^*]$ is defined to mean $(\forall x)[t < x \leq t^* \supset G\text{-}at\text{-}t]$.

[32] That is, it is the set of real numbers x such that $t+\varepsilon < x \leq t+\varepsilon+\delta$.

[33] SATPLAN (Kautz and Selman 1996) is an exception to this. See the discussion of SATPLAN in section twelve.

Employing this in place of $(A^*/\ SG) \blacktriangleright_{\varepsilon^*} G$ in the above planning example will produce a plan with the causal-link $A\text{-}at\text{-}t_1 \rightarrow SG\text{-}at\text{-}t_2 \rightarrow A^*\text{-}at\text{-}t_3 \rightarrow G\text{-}at\text{-}t_4$ and the ordering-constraints $\{t_1+\varepsilon < t_2,\ t_2 = t_3+\alpha,\ t_3+\varepsilon^* < t_4\}$. To avoid unnecessary complexity, for the bulk of this paper I will not consider planning with these more complex planning-conditionals, confining myself to the use of planning-conditionals of the form (CS) with $\varepsilon = 0$. However, the theory can be readily extended to handle these more complex conditionals, and I will return to this topic briefly in section 11 and discuss how to do it.

6. Planning and Temporal-Projectibility

The conventional theory of goal-regression planning as developed in section two was based upon the Soundness Assumption, which was formulated as follows:

Soundness Assumption

A linear plan will achieve a goal *G* relative to a state *start-state* iff *G* is a result of the sequence of actions prescribed by its plan-steps relative to *start-state* and the set of all true planning-conditionals.

This employs the concept of a result of a sequence of actions, which was defined in (R1). To make the temporal reference explicit, let us rewrite (R1) as follows:

(R1) Where *start-state* is a state of affairs, *conditionals* is a set of planning-conditionals, and $t_0 < \ldots < t_{n+1}$ is a sequence of times, $P\text{-}at\text{-}t_{n+1}$ is a **result** of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$ relative to *start-state* and *conditionals* iff either:

(i) $n = 0$ and $P\text{-}at\text{-}t_0$ is true in *start-state*; or

(ii) $n > 0$ and *conditionals* contains a conditional $(A_n/C) \blacktriangleright P$ such that $C\text{-}at\text{-}t_n$ is a result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$; or

(iii) $n > 0$, $P\text{-}at\text{-}t_n$ is a result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, and *conditionals* does not contain a conditional of the form $(A_n/C) \blacktriangleright \sim Q$ such that Q is either P or a conjunct of P and $C\text{-}at\text{-}t_n$ is a result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$; or

(iv) $n > 0$ and $P\text{-}at\text{-}t_{n+1}$ is a conjunction whose conjuncts are results of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$.

The observations of section five require modifications to both the Soundness Assumption itself and to the definition of "result", and these in turn require modifications to the rules of goal-regression planning that are based upon the Soundness Assumption. First, the inferences underlying the definition of "result" are only defeasible inferences, based upon TEMPORAL-PROJECTION. As such, the term "result" is a misnomer. We are not characterizing what *will definitely* happen if the action-sequence is performed. We are just characterizing what can be reasonably (and defeasibly) expected to happen. More precisely, we are characterizing the set of *warranted* expectations, in the sense of section four. So it would be better to use the term "expectable-result". Second, because the inferences are based upon TEMPORAL-PROJECTION, they are subject to temporal-projectibility constraints. We must make some changes to the definition to accommodate these constraints.

Clause (i) tells us to expect $P\text{-}at\text{-}t_1$ to be true if $P\text{-}at\text{-}t_0$ is true. This is just an instance of TEMPORAL-PROJECTION. Accordingly, it requires the addition of a temporal-projectibility constraint on *P*:

(i) $n = 0$, P is temporally-projectibile, and P-*at*-t_0 is true in *start-state*

It follows that PROPOSE-NULL-PLAN also requires a temporal-projectibility-constraint:

PROPOSE-NULL-PLAN
Given an interest in finding a plan for achieving *goal-at-t*, if *goal-at-t* is temporally-projectible and *goal-at-*t_0 is true where $t_0 < t$, infer defeasibly that a null-plan will achieve *goal-at-t*.

Note that PROPOSE-NULL-PLAN becomes a defeasible inference-rule, because it builds in an application of TEMPORAL-PROJECTION.

Clause (iv) requires no modification to accommodate temporal-projectibility, because it only concerns relations between expectable-results that have already been projected forward to the appropriate times, and hence does not presuppose any new application of TEMPORAL-PROJECTION. However, clauses (ii) and (iii) require temporal-projectibility constraints.

Goal-Regression

Clause (ii) describes causal inferences of the sort to which the Frame Problem is relevant. If we can expect C-*at*-t_n to be achieved by executing the sequence of actions $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n\text{-}1}\text{-}at\text{-}t_{n\text{-}1}\rangle$, then given the conditional $(A_n / C) \blacktriangleright P$ we can infer that for some time t^* between t_n and t_{n+1}, and P-*at*-t^* will be made true by performing A_n-*at*-t_n. If P is temporally-projectible, we can then infer by temporal-projection that P will still be true at t_{n+1}. For this reasoning to work, a projectibility constraint must be added to clause (ii):

(ii) $n > 0$, P is temporally-projectibile, and *conditionals* contains a conditional $(A_n / C) \blacktriangleright P$ such that C-*at*-t_n is an expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n\text{-}1}\text{-}at\text{-}t_{n\text{-}1}\rangle$ and $t_{n+1} > t_n$.

GOAL-REGRESSION is based directly on clause (ii), so it must contain a corresponding constraint, and it becomes defeasible:

GOAL-REGRESSION
Given an interest in finding a plan for achieving *G-at-t*, if *G is temporally-projectible* , adopt interest in finding planning-conditionals $(A / C) \blacktriangleright G$ having G as their consequent. Given such a conditional, adopt an interest in finding a plan for achieving C-*at*-t^*. If it is concluded that a plan *subplan* will achieve C-*at*-t^*, construct a plan by (1) adding a new step to the end of *subplan* where the new step prescribes the action A-*at*-t^*, (2) adding the constraint ($t^* < t$) to the ordering-constraints of *subplan*, and (3) adjusting the causal-links appropriately. Infer defeasibly that the new plan will achieve *G-at-t*.

Similar constraints are required in EMBEDDED-GOAL-REGRESSION and EMBEDDED-NULL-PLAN.

Undermining Causal-Links

Clause (iii) is, in effect, a statement of TEMPORAL-PROJECTION applied to the expectable-results of an action-sequence, together with the statement of a defeater for the application of TEMPORAL-PROJECTION. For TEMPORAL-PROJECTION to be applicable, we must require that P be temporally-projectible. Given that constraint, if it is defeasibly reasonable to expect P to be true after executing $A_1\text{-}at\text{-}t_1,\ldots,A_{n\text{-}1}\text{-}at\text{-}t_{n\text{-}1}$, then it is defeasibly reasonable to expect P to remain true after executing A_n-*at*-t_n as well. A defeater for this defeasible expectation consists of having a reason for thinking that P-*at*-t_n will not be true. Given that it is defeasibly reasonable to expect

C to be true after executing A_1-*at*-t_1,...,A_{n-1}-*at*-t_{n-1}, it follows in accordance with the preceding discussion that, given the conditional $(A_n/C) \blacktriangleright \sim Q$, it is defeasibly reasonable to expect Q to become false at some time t^* after executing A_n-*at*-t_n, where $t^* > t_n$. If Q's being false requires P to be false, then as in the Yale Match-Lighting Problem, this defeats the temporal-projection to the conclusion that P will still be true at t_{n+1}. Note that this reasoning does not require that $\sim Q$ be temporally-projectible. So (iii) should be reformulated as follows:

(iii) $n > 0$, P-*at*-t_n is a temporally-projectible expectable-result of $\langle A_1$-*at*-t_1,...,A_{n-1}-*at*-$t_{n-1}\rangle$, and *conditionals* does not contain a conditional of the form $(A_n/C) \blacktriangleright \sim Q$ such that Q is either P or a conjunct of P, C-*at*-t_n is an expectable-result of $\langle A_1$-*at*-t_1,...,A_{n-1}-*at*-$t_{n-1}\rangle$, and $t_{n+1} > t_n$.

Clause (iii) underlies the defeat of SPLIT-CONJUNCTIVE-GOAL by finding that the merged plan undermines one of its own causal-links. It follows that in the search for embellishments, the same projectibility-constraints must be observed as in the original search for plans.

Clause (iii) also underlies CONFRONTATION, but CONFRONTATION is correct as previously formulated. The observation that in clause (iii) $\sim Q$ need not be temporally-projectible is an important one. In CONFRONTATION, we search for a plan for ~*subgoal* (or for the negation of a conjunct of *subgoal*). We know that *subgoal* will be temporally-projectible, but there is no reason to expect its negation to be.

Combining these observations, we are led to the following definition of "expectable-result":

(R2) Where *start-state* is a state of affairs, *conditionals* is a set of planning-conditionals, and $t_0 < \ldots < t_{n+1}$ is a sequence of times, P-*at*-t_{n+1} is an **expectable-result** of $\langle A_1$-*at*-t_1,...,A_n-*at*-$t_n\rangle$ relative to *start-state* and *conditionals* iff either:

(i) $n = 0$, P is temporally-projectibile, and P-*at*-t_0 is true in *start-state*; or

(ii) $n > 0$, P is temporally-projectibile, and *conditionals* contains a conditional $(A_n/C) \blacktriangleright P$ such that C-*at*-t_n is an expectable-result of $\langle A_1$-*at*-t_1,...,A_{n-1}-*at*-$t_{n-1}\rangle$ and $t_{n+1} > t_n$; or

(iii) $n > 0$, P-*at*-t_n is a temporally-projectible expectable-result of $\langle A_1$-*at*-t_1,...,A_{n-1}-*at*-$t_{n-1}\rangle$, and *conditionals* does not contain a conditional of the form $(A_n/C) \blacktriangleright \sim Q$ such that Q is either P or a conjunct of P, C-*at*-t_n is an expectable-result of $\langle A_1$-*at*-t_1,...,A_{n-1}-*at*-$t_{n-1}\rangle$, and $t_{n+1} > t_n$; or

(iv) $n > 0$ and P-*at*-t_{n+1} is a conjunction whose conjuncts are expectable-results of $\langle A_1$-*at*-t_1,...,A_n-*at*-$t_n\rangle$.

The Soundness Assumption must now be reinterpreted as giving us merely a defeasible expectation that a plan will achieve its goal:

Soundness Assumption

Executing a linear plan can be defeasibly expected to achieve a goal G relative to a state *start-state* iff G is an expectable-result of the sequence of actions prescribed by the plan-steps of the plan relative to *start-state* and the set of all warranted planning-conditionals.

The Importance of the Temporal-Projectibility Constraints

It is of some interest to illustrate the importance of the temporal-projectibility constraints. They have significant impacts on goal-regression planning. Suppose C_1, C_2, C_3, and D_1 are

temporally-projectible, we know that C_1-*at*-t_0 and D_1-*at*-t_0 are true, and we are given the planning-conditionals $(A_1/C_1) \blacktriangleright C_2$, $(A_2/C_2) \blacktriangleright G_1$, $(A_2/C_3) \blacktriangleright G_1$, $(A_3/D_1) \blacktriangleright G_2$ and $(A_3/D_1) \blacktriangleright \sim C_2$. Suppose our goal is $(G_1$ & $G_2)$-*at*-*t*. We can construct plan #1 (figure 3) for G_1-*at*-*t*, and plan #2 (figure 4) for G_2-*at*-*t*: To construct a plan for the conjunctive goal $(G_1$ & $G_2)$-*at*-*t*, we merge plans #1 and #2 to produce plan #3 (figure 5).

Plan #1
PLAN-STEPS:
(1) A_1-*at*-t_1
causal-links:
start → C_1-*at*-t_1 → (1) → C_2-*at*-t_2
(2) A_2-*at*-t_2
causal-links:
(1) → C_2-*at*-t_2 → (2) → G_1-*at*-*t*
ordering-constraints:
$t_1 < t_2$
GOAL: G_1-*at*-*t*
established by:
(2) → G_1-*at*-*t*

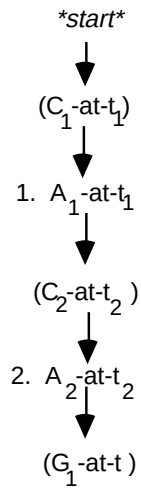

Figure 3. Plan #1

Plan #2
PLAN-STEPS:
(3) A_3-*at*-t_3
causal-links:
start → D_1-*at*-t_3 → (3) → G_2-*at*-*t*
GOAL: G_2-*at*-*t*
established by:
(3) → G_2-*at*-*t*

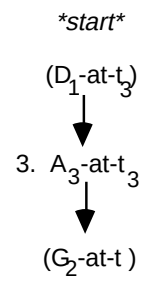

Figure 4. Plan #2

Plan #3
PLAN-STEPS:
(1) A_1-*at*-t_1
causal-links:
start ➞ C_1-*at*-t_1 ➞ (1) ➞ C_2-*at*-t_2
(2) A_2-*at*-t_2
causal-links:
(1) ➞ C_2-*at*-t_2 ➞ (2) ➞ G_1-*at*-*t*
ordering-constraints:
$t_1 < t_2$
(3) A_3-*at*-t_3
causal-links:
start ➞ D_1-*at*-t_3 ➞ (3) ➞ G_2-*at*-*t*
GOAL: (G_1 & G_2)-*at*-*t*
established by:
(2) ➞ G_1-*at*-*t*
(3) ➞ G_2-*at*-*t*

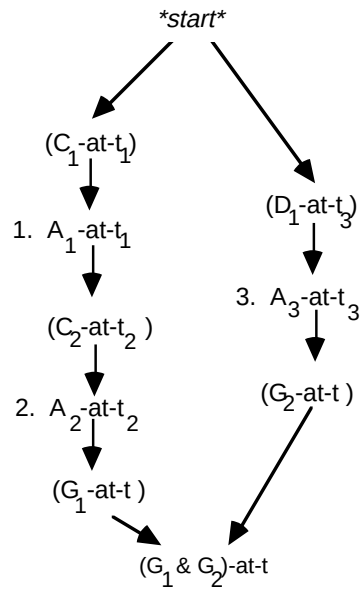

Figure 5. Plan #3

We must then investigate whether plan #3 undermines one of its own causal-links. Because we also have the planning-conditional (A_3/D_1) ➤ ~C_2, we find that plan-step (3), which is not ordered with respect to plan-steps (1) and (2), undermines the causal-link (1) ➞ C_2-*at*-t_2 ➞ (2) ➞ G_1-*at*-*t*. The plan can be repaired by adding an ordering-constraint requiring plan-step (3) to be executed either before plan-step (1) or after plan-step (2), producing plan #4 (figure 6).

Plan #4
PLAN-STEPS:
(1) A_1-*at*-t_1
causal-links:
start ➞ C_1-*at*-t_1 ➞ (1) ➞ C_2-*at*-t_2
(2) A_2-*at*-t_2
causal-links:
(1) ➞ C_2-*at*-t_2 ➞ (2) ➞ G_1-*at*-*t*
ordering-constraints:
$t_1 < t_2$
(3) A_3-*at*-t_3
causal-links:
start ➞ D_1-*at*-t_3 ➞ (3) ➞ G_2-*at*-*t*
ordering-constraints:
$t_3 < t_1$ or $t_2 < t_3$
GOAL: (G_1 & G_2)
established by:
GOAL: (G_1 & G_2)-*at*-*t*
established by:
(2) ➞ G_1-*at*-*t*
(3) ➞ G_2-*at*-*t*

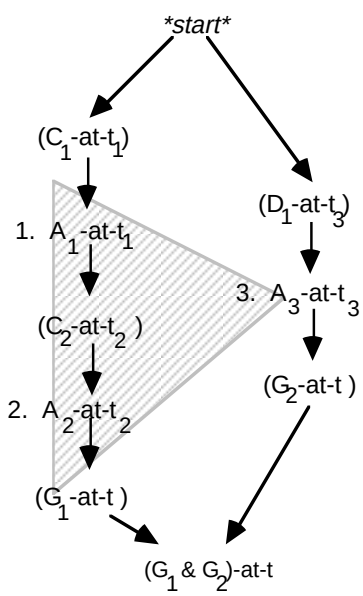

Figure 6. Plan #4

Now, suppose we ignore the temporal-projectibility constraints. From $(A_1/C_1) \blacktriangleright C_2$ we can deduce $(A_1/C_1) \blacktriangleright (C_2 \vee C_3)$, and from $(A_2/C_2) \blacktriangleright G_1$ and $(A_2/C_3) \blacktriangleright G_1$ we can deduce the conditional $(A_2/(C_2 \vee C_3)) \blacktriangleright G_1$.[34] Without the temporal-projectibility constraints, we could construct plan #5 (figure 7) for G_1-*at-t*: To construct a plan for the conjunctive goal $(G_1 \,\&\, G_2)$-*at-t*, we merge plans #5 and #2 to produce plan #6 (figure 8).

Plan #5
- PLAN-STEPS:
  - (1) A_1
    - causal-links:
      - *start* → C_1-*at*-t_1 → (1) → $(C_2 \vee C_3)$-*at*-t_2
  - (2) A_2
    - causal-links:
      - (1) → $(C_2 \vee C_3)$-*at*-t_2 → (2) → G_1-*at-t*
    - ordering-constraints:
      - $t_1 < t_2$
- GOAL: G_1
  - established by:
    - (2) → G_1-*at-t*

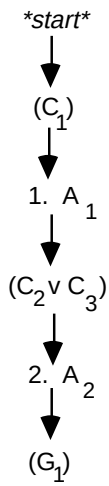

Figure 7. Plan #5

Plan #6
- PLAN-STEPS:
  - (1) A_1
    - causal-links:
      - *start* → C_1-*at*-t_1 → (1) → $(C_2 \vee C_3)$-*at*-t_2
  - (2) A_2
    - causal-links:
      - (1) → $(C_2 \vee C_3)$-*at*-t_2 → (2) → G_1-*at-t*
    - ordering-constraints:
      - $t_1 < t_2$
  - (3) A_3
    - causal-links:
      - *start* → D_1-*at*-t_3 → (3) → G_2-*at-t*
- GOAL: $(G_1 \,\&\, G_2)$-*at-t*
  - established by:
    - (2) → G_1-*at-t*
    - (3) → G_2-*at-t*

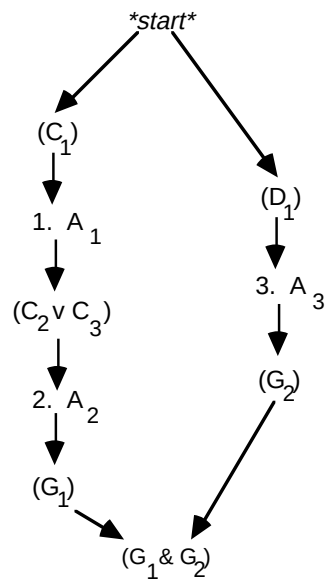

Figure 8. Plan #6

Note that plans #5 and #6 differ from plans #1 and #3 only in their causal-links. They

[34] This planning-conditional violates the syntactical constraint that the precondition be a conjunction of literals, but such syntactical constraints can always be circumvented by simply introducing a new atomic formula equivalent to $(C_2 \vee C_3)$. The relaxation of the syntactical constraints will be explored in the next section.

prescribe the same actions in the same order. Now when we investigate whether plan #6 undermines one of its own causal-links, we find that, unlike plan #3, it does not. This is because the conditional $(A_3/D_1) \blacktriangleright \sim C_2$ does not entail $(A_3/D_1) \blacktriangleright \sim(C_2 \vee C_3)$. Thus without the projectibility constraints, we would be led to adopt plan #6, whose execution is the same as plan #3. But as we have seen, plan #3 cannot be expected to achieve its goal. If step (3) is executed between steps (1) and (2), the plan will fail. The execution of plan #6 is precisely the same as the execution of plan #3, so it cannot be expected to achieve its goal either, but this is not revealed by looking for undermined causal-links.

7. Relaxing the Syntactical Constraints

Once the account of section two is modified as indicated in section six, it constitutes a provably correct theory of goal-regression planning provided all goals and planning-conditionals satisfy the syntactical constraint that goals and subgoals must be literals or conjunctions of literals. But now it is time to re-examine both that constraint and the definition of "expectable-result".

There is something artificial about the definition (R2) of "expectable-result". It ought to be the case that logical consequences of expectable-results are expectable-results. This suggests revising the previous definition (R2) as follows:

(R3) Where *start-state* is a state of affairs, *conditionals* is a set of planning-conditionals, and $t_0 < \ldots < t_{n+1}$ is a sequence of times, $P\text{-}at\text{-}t_{n+1}$ is an **expectable-result** of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$ relative to *start-state* and *conditionals* iff either:

(i) $n = 0$, P is temporally-projectibile, and $P\text{-}at\text{-}t_0$ is true in *start-state*; or

(ii) $n > 0$, P is temporally-projectibile, and *conditionals* contains a conditional $(A_n/C) \blacktriangleright P$ such that $C\text{-}at\text{-}t_n$ is an expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$ and $t_{n+1} > t_n$; or

(iii) $n > 0$, $P\text{-}at\text{-}t_n$ is a temporally-projectible expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, and *conditionals* does not contain a conditional of the form $(A_n/C) \blacktriangleright \sim Q$ such that Q is either P or a conjunct of P, $C\text{-}at\text{-}t_n$ is an expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, and $t_{n+1} > t_n$; or

(iv) $n > 0$ and $P\text{-}at\text{-}t_{n+1}$ is a logical consequence of expectable-results of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$.

If we assume that the compactness theorem[35] holds for the logical consequence relation (as it does, e.g., in first-order logic, but not in second-order logic), this definition can be written equivalently as follows:

(R4) Where *start-state* is a state of affairs, *conditionals* is a set of planning-conditionals, and $t_0 < \ldots < t_{n+1}$ is a sequence of times, $P\text{-}at\text{-}t_{n+1}$ is an **expectable-result** of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$ relative to *start-state* and *conditionals* iff either:

(i) $n = 0$, P is temporally-projectibile, and $P\text{-}at\text{-}t_0$ is true in *start-state*; or

(ii) $n > 0$, P is temporally-projectibile, and *conditionals* contains a conditional $(A_n/C) \blacktriangleright P$ such that $C\text{-}at\text{-}t_n$ is an expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$ and $t_{n+1} > t_n$; or

[35] The compactness theorem says that if a formula is a logical consequence of a set X of formulas, it is also a logical consequence of some finite subset of X.

(iii) $n > 0$, $P\text{-}at\text{-}t_n$ is a temporally-projectible expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, and *conditionals* does not contain a conditional of the form $(A_n/C) \blacktriangleright \sim Q$ such that Q is either P or a conjunct of P, $C\text{-}at\text{-}t_n$ is an expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, and $t_{n+1} > t_n$; or

(iv) $n > 0$ and $P\text{-}at\text{-}t_{n+1}$ is a conjunction whose conjuncts are expectable-results of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$; or

(v) $n > 0$ and $P\text{-}at\text{-}t_{n+1}$ is a logical consequence of some expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$.

It is worth noting that even if the logical consequence relation is not compact, these two definitions will give rise to the same planning because, presumably, we can only plan for finitely many subgoals in the course of any planning problem. For this reason, I will focus on the latter definition of "expectable-result".

Thus far we have required goals and subgoals to be conjunctions of literals. If we identify the logical consequence relation with first-order consequence, and we restrict our attention to conjunctions of literals, then the definitions (R2), (R3), and (R4) are all equivalent. This is a justification for the using (R1) in the "conventional" theory of goal-regression planning, and it carries over to using (R2) in the present theory. However, we can reasonably object both to these syntactical constraints and to the identification of logical-consequence with first-order consequence.

Problems with Literals

First, consider the syntactical constraints. Initially, it may seem that they constitute a way of incorporating the temporal-projectibility constraint. Unfortunately, this simple approach to the problem does not work, for several reasons:

- The fact that something is symbolized by an atomic formula does not imply anything about whether it is projectible. *Anything* can be symbolized as an atomic formula, including wildly unprojectible logical compounds. We might try to get around this by requiring that atomic formulas express "logically simple" propositions. Unfortunately, not everyone agrees that this notion makes sense. It is often alleged that logical form belongs only to sentences, not propositions.
- Even if the notion of a logically simple proposition does make sense, it is doubtful that all logically simple propositions will be temporally-projectible. For example, "The time is now 3 PM" might plausibly be regarded as logically simple, but its being true now is no reason to think it will still be true an hour from now.
- We might build into the semantics of our language the restriction that atomic formulas express temporally-projectible logically simple propositions. That, however, does not guarantee that the negation of an atomic formula will be temporally-projectible. We have seen that negations of temporally-projectible propositions are not automatically temporally-projectible, and that is true even for what are presumably logically simple propositions. The example given above was "x is red". Thus there will still be no guarantee that literals will be temporally-projectible. Furthermore, this would seem to preclude our being able to express temporally-unprojectible logically simple propositions like "The time is now 3 PM".
- Not all logically complex propositions are temporally-unprojectible. Requiring that goals and subgoals be conjunctions of literals will rule out logically complex propositions that, intuitively, are perfectly fine objects of goal-regression planning. This will be illustrated in the next paragraph.

It must be concluded that the restriction to conjunctions of literals is artificial and should be eliminated. Temporal-projectibility is a semantical notion and cannot be captured by syntax

alone.

Problems with First-Order Consequence

Logical consequence cannot be identified with first-order consequence. Let us consider a particularly simple example of this, which simultaneously illustrates the importance of temporally-projectible formulas that are logically complex. Nomic generalizations cannot change truth value as time passes. They are fixed features of the world. Thus they are temporally-projectible. Furthermore, there cannot be true planning-conditionals whose consequents are the negations of true nomic generalizations. Thus it follows from (R4) that any true nomic generalization is an expectable-result of any action-sequence. Now consider a non-counterlegal nomic generalization ($Fx \Rightarrow Gx$). This entails (but is not entailed by) $(\forall x)(Fx \supset Gx)$. Consequently, $[Fa \ \& \ (Fx \Rightarrow Gx)]$ entails Ga, but this is not a first-order implication. It can, however, be an important implication for planning. If our goal is Ga, one way to achieve that goal is to achieve the subgoal Fa.

More generally, let us define:

> *P nomically implies Q* iff there is a set Γ of true nomic generalizations such that $\Gamma \cup \{P\}$ entails Q.

If the logical-consequence relation in (R4) is construed sufficiently broadly to include the entailments employed in the definition of nomic implication, it follows as above that if P is an expectable-result of an action-sequence and the planning agent is able to reason to the warranted conclusion that P nomically implies Q, then Q is also an expectable-result. It will be syntactically convenient to abbreviate "the planning agent is able to reason to the warranted conclusion that P nomically implies Q" as "P is known to nomically imply Q", but it must be acknowledge that this is not a literal use of "known". Then we can express (R4) equivalently as follows:

(R5) Where *start-state* is a state of affairs, *conditionals* is a set of planning-conditionals, and $t_0 < \ldots < t_{n+1}$ is a sequence of times, $P\text{-}at\text{-}t_{n+1}$ is an **expectable-result** of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$ relative to *start-state* and *conditionals* iff either:

- (i) $n = 0$, P is temporally-projectibile, and $P\text{-}at\text{-}t_0$ is true in *start-state*; or
- (ii) $n > 0$, P is temporally-projectibile, and *conditionals* contains a conditional $(A_n/C) \blacktriangleright P$ such that $C\text{-}at\text{-}t_n$ is an expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$ and $t_{n+1} > t_n$; or
- (iii) $n > 0$, $P\text{-}at\text{-}t_n$ is a temporally-projectible expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, and *conditionals* does not contain a conditional of the form $(A_n/C) \blacktriangleright \sim Q$ such that Q is either P or a conjunct of P, $C\text{-}at\text{-}t_n$ is an expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, and $t_{n+1} > t_n$; or
- (iv) $n > 0$ and $P\text{-}at\text{-}t_{n+1}$ is a conjunction whose conjuncts are expectable-results of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$; or
- (v) $n > 0$ and $P\text{-}at\text{-}t_{n+1}$ is a logical consequence of some expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$; or
- (vi) $n > 0$ and $P\text{-}at\text{-}t_{n+1}$ is known to be nomically implied by some expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$.

Logical consequence is the limiting case of nomic implication, so clause (v) is redundant in (R5), but it will be useful to keep it in the definition.[36]

---

[36] I assume that the set of warranted conclusions is closed under logical consequence, so clause (v) need

Collective Undermining

Should (R5) be adopted as our final definition of "expectable-result"? No, because relaxing the syntactical constraints has the effect of making clause (iii) unreasonable. Clause (iii) plays a double role. It tells us that we can defeasibly project an expectable-result P forwards through the plan, and it tells us when the structure of the plan blocks that projection. According to clause (iii), the projection becomes unreasonable only when later steps of the plan can be expected to make P false (i.e., undermine it). However, having now allowed planning-conditionals to have consequents of arbitrary logical form, the possibility arises that the consequent of a planning-conditional might be the negation of a conjunction, $\sim(P_1 \,\&\, ... \,\&\, P_n)$, of previously expectable-results. In accordance with the reasoning resolving the Frame Problem, we can infer that the consequent is true, and hence one of these previously expectable-results will not obtain. But we have no way of determining which it is that will not obtain, so we should be agnostic and refrain from concluding of any of them that it will obtain.[37] In other words, they all cease to be expectable-results. This is a kind of collective undermining.

This point can be made logically more rigorous as follows. Suppose $P_1,...,P_n$ are expectable-results of $\langle A_1\text{-}at\text{-}t_1,...,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, but we have a planning conditional $(A_n/C) \blacktriangleright \sim(P_1 \,\&\, ... \,\&\, P_n)$ such that $C\text{-}at\text{-}t_n$ is an expectable-result of $\langle A_1\text{-}at\text{-}t_1,...,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$. By clause (ii), $\sim(P_1 \,\&\, ... \,\&\, P_n)$ is an expectable-result of $\langle A_1\text{-}at\text{-}t_1,...,A_n\text{-}at\text{-}t_n\rangle$. We can infer that each P_i is an expectable-result of $\langle A_1\text{-}at\text{-}t_1,...,A_n\text{-}at\text{-}t_n\rangle$, and then by clause (iv) we get that $(P_1 \,\&\, ... \,\&\, P_n)$ is an expectable-result of $\langle A_1\text{-}at\text{-}t_1,...,A_n\text{-}at\text{-}t_n\rangle$. So we have a contradiction being an expectable-result, and then by clause (v), everything becomes an expectable-result.

(R5) must be revised. We do not want $(P_1 \,\&\, ... \,\&\, P_n)$ to be an expectable-result of $\langle A_1\text{-}at\text{-}t_1,...,A_n\text{-}at\text{-}t_n\rangle$, and the only way to deny that is to deny that the P_i's are expectable-results of $\langle A_1\text{-}at\text{-}t_1,...,A_n\text{-}at\text{-}t_n\rangle$. The projection of the P_i's must be blocked by our having the planning conditional $(A_n/C) \blacktriangleright \sim(P_1 \,\&\, ... \,\&\, P_n)$ where $C\text{-}at\text{-}t_n$ is an expectable-result of $\langle A_1\text{-}at\text{-}t_1,...,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$. In other words, we must have collective undermining.

This suggests revising clause (iii) as follows (where ΠX is the conjunction of a set X of propositions):

> $n > 0$, $P\text{-}at\text{-}t_n$ is a temporally-projectible expectable-result of $\langle A_1\text{-}at\text{-}t_1,...,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, and *conditionals* does not contain a conditional of the form $(A_n/C) \blacktriangleright \sim\Pi X$ such that $C\text{-}at\text{-}t_n$ is a temporally-projectible expectable-result of $\langle A_1\text{-}at\text{-}t_1,...,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, $t_{n+1} > t_n$, and X is a set of temporally-projectible expectable-results of $\langle A_1\text{-}at\text{-}t_1,...,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, one of which is P.

This will not quite do, however. The difficulty is that it follows from the logic of nomic generalizations that if $(A_n/C) \blacktriangleright \sim\Pi X$ then for any P, $(A_n/C) \blacktriangleright \sim\Pi(X \cup \{P\})$. That is, you can always weaken the consequent of a nomic generalization by adding disjuncts. Thus (iii), as revised above, would result in every previously expectable-result being defeated if any is. To avoid this we must require that X be a *minimal* conjunction of expectable-results:

(iii) $n > 0$, $P\text{-}at\text{-}t_n$ is a temporally-projectible expectable-result of $\langle A_1\text{-}at\text{-}t_1,...,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$,

---

not be restricted to known logical consequences.

[37] Technically, in the logic of defeasible reasoning this is a case of collective defeat.

and if *conditionals* contains a conditional of the form $(A_n/C) \blacktriangleright \sim\Pi X$ such that $C\text{-}at\text{-}t_n$ is a temporally-projectible expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, $t_{n+1} > t_n$, X is a set of temporally-projectible expectable-results of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, and $P \in X$, then there is a conditional $(A_n/C^*) \blacktriangleright \sim\Pi X_0$ in *conditionals* such that $C^*\text{-}at\text{-}t_n$ is a temporally-projectible expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, $X_0 \subseteq X$, and $P \notin X_0$.

With this change, we can produce what seems to be an adequate definition of "expectable-result":

(R6) Where *start-state* is a state of affairs, *conditionals* is a set of planning-conditionals, and $t_0 < \ldots < t_{n+1}$ is a sequence of times, $P\text{-}at\text{-}t_{n+1}$ is an **expectable-result** of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$ relative to *start-state* and *conditionals* iff either:

(i) $n = 0$, P is temporally-projectibile, and $P\text{-}at\text{-}t_0$ is true in *start-state*; or

(ii) $n > 0$, P is temporally-projectibile, and *conditionals* contains a conditional $(A_n/C) \blacktriangleright P$ such that $C\text{-}at\text{-}t_n$ is an expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$ and $t_{n+1} > t_n$; or

(iii) $n > 0$, $P\text{-}at\text{-}t_n$ is a temporally-projectible expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, and if *conditionals* contains a conditional of the form $(A_n/C) \blacktriangleright \sim\Pi X$ such that $C\text{-}at\text{-}t_n$ is a temporally-projectible expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, $t_{n+1} > t_n$, X is a set of temporally-projectible expectable-results of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, and $P \in X$, then there is a conditional $(A_n/C^*) \blacktriangleright \sim\Pi X_0$ in *conditionals* such that $C^*\text{-}at\text{-}t_n$ is a temporally-projectible expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, $X_0 \subseteq X$, and $P \notin X_0$; or

(iv) $n > 0$ and $P\text{-}at\text{-}t_{n+1}$ is a conjunction whose conjuncts are expectable-results of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$; or

(v) $n > 0$ and $P\text{-}at\text{-}t_{n+1}$ is a logical consequence of some expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$; or

(vi) $n > 0$ and $P\text{-}at\text{-}t_{n+1}$ is known to be nomically implied by some expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$.

8. Revised Rules for Goal-Regression Planning

If we add to (R6) the following definition of "expectable-result" for partial-order plans:

> P is an expectable-result of a partial-order plan iff it is an expectable-result of every linearization of the plan.

this provides a semantics for goal-regression planning. We can then look for a set of procedures for constructing plans and attempt to prove the soundness and completeness of the set of procedures. The procedures described in sections two and five provide the starting point for the construction of a sound and complete set of procedures for goal-regression planning. They must be modified somewhat to accommodate the differences between (R6) and (R2).

Planning for implacanda

There are two differences between (R2) and (R6) — (R6) introduces collective undermining, and it includes expectable-results that are logically or nomically implied by other expectable-results. Let us focus on the latter difference first. (R2) recognized just one case of expectable-results

being derived logically from other expectable-results. That was the case in which a conjunction is achieved by achieving its conjuncts. Such logical derivations were recorded in plans by including multiple goals and subgoals in a causal-link. We can extend the use of causal-links by allowing causal-links of the form $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$ where the relationship between $subgoal_i$ and $subgoal_{i+1}$ or $goal_i$ and $goal_{i+1}$ is one of logical entailment or nomic implication. To accommodate this, we can begin with the three basic rules of goal-regression planning, PROPOSE-NULL-PLAN, GOAL-REGRESSION, and SPLIT-CONJUNCTIVE-GOAL. To these we add one new rule:

PLAN-FOR-IMPLICANDA[38]

Given an interest in finding a plan for achieving *goal*, and given a nomic implication *subgoal* ⇒ *goal*, adopt an interest in finding a plan for achieving *subgoal*. If a plan *subplan* is proposed for achieving *subgoal*, construct a new plan by replacing each causal-link $n \rightarrow subgoal_1 \rightarrow ... subgoal_n \rightarrow$ **finish** $\rightarrow goal_1 ... \rightarrow goal_m \rightarrow subgoal$ of *subplan* by the causal-link $n \rightarrow subgoal_1 \rightarrow ... subgoal_n \rightarrow$ **finish** $\rightarrow goal_1 ... \rightarrow goal_m \rightarrow subgoal \rightarrow goal$. Infer defeasibly that the new plan will achieve *goal*.

A special case of this rule occurs when the nomic implication is a logical entailment.

As formulated, PLAN-FOR-IMPLICANDA is a computational nightmare. It makes planning *much* more difficult. However, it turns out that all but one special case of this rule can be eliminated. This can be seen by reflecting on the causal-link structure of plans that result from its use. Consider the following cases:

We never need a structure of the form

$$\begin{matrix} P \searrow \\ Q \nearrow \end{matrix} \; (P \,\&\, Q) \rightarrow R \rightarrow S \rightarrow n_2$$

because, due to the transitivity of nomic implication, it can always be replaced by the simpler structure

$$\begin{matrix} P \searrow \\ Q \nearrow \end{matrix} \; (P \,\&\, Q) \rightarrow S \rightarrow n_2$$

Similarly, where A is the action prescribed by n_2, we never need a structure of the form

$$\begin{matrix} P \searrow \\ Q \nearrow \end{matrix} \; (P \,\&\, Q) \rightarrow S \rightarrow n_2 \rightarrow R$$

because if $(P \,\&\, Q) \Rightarrow S$ and $(A \& S) \Rightarrow R$, we also have $(A \,\&\, (P \,\&\, Q)) \Rightarrow R$, and so can build a plan with the simpler structure

$$\begin{matrix} P \searrow \\ Q \nearrow \end{matrix} \; (P \,\&\, Q) \rightarrow n_2 \rightarrow R$$

We never need a structure of the form

$$\begin{matrix} P \searrow \\ Q \nearrow \end{matrix} \; (P \,\&\, Q) \rightarrow R \begin{matrix} \searrow \\ S \nearrow \end{matrix} \; (R \,\&\, S) \rightarrow n_2 \rightarrow G$$

because if $(P \,\&\, Q) \Rightarrow R$ and $(A \,\&\, (R \,\&\, S)) \Rightarrow G$, we also have $(A \,\&\, ((P \,\&\, Q) \,\&\, S)) \Rightarrow G$, and so can

[38] In traditional logic, if P implies Q, P is the *implicans* and Q the *implicandum*.

build a plan with the simpler structure

$$\begin{array}{l} P \searrow \\ Q \nearrow \end{array} (P \,\&\, Q) \begin{array}{l} \searrow \\ S \nearrow \end{array} ((P \,\&\, Q) \,\&\, S) \rightarrow n_2 \rightarrow G$$

We never need a structure of the form

$$P \rightarrow n_2 \rightarrow Q \rightarrow R$$

because if $(A \,\&\, P) \Rightarrow Q$ and $Q \Rightarrow R$ then $(A \,\&\, P) \Rightarrow R$, and so we can build a plan with the simpler structure

$$P \rightarrow n_2 \rightarrow R$$

We never need a structure of the form

$$^{*}start^{*} \rightarrow Q \rightarrow R$$

because if Q is true and $Q \Rightarrow R$ then R is true, and hence we can build a plan with the simpler structure

$$^{*}start^{*} \rightarrow R$$

Only one case remains. Where *goal* is the ultimate goal of the planning exercise (not just a subgoal generated in the course of the planning), a structure of the form

$$\begin{array}{l} n_1 \rightarrow P \searrow \\ n_2 \rightarrow Q \nearrow \end{array} (P \,\&\, Q) \rightarrow goal$$

cannot be eliminated. So a restriction can be imposed on PLAN-FOR-IMPLICANDA allowing its use only as the first step of the plan search.

It is clear that further restrictions must be imposed on PLAN-FOR-IMPLICANDA to make its application efficient. For example, given an interest in finding a plan for achieving G, we do not want to automatically adopt interest in finding plans for achieving every goal of the form $(G \,\&\, P)$ for arbitrary P. However, I will not pursue this point here.

Collective undermining

As before, an application of SPLIT-CONJUNCTIVE-GOAL produces a plan which may undermine some of its own causal-links. We must adopt rules to search for underminings and repair them if possible. Now, however, undermining becomes more complicated because relaxing the syntactical constraints gives rise to collective undermining. To work out the logical details of this, let us revise the definition of presumptive-soundness as follows:

> A plan is **presumptively-sound relative to** a set *conditionals* of planning-conditionals and a state *start-state* iff
>
> (1) where *goal* is the goal of the plan, the plan contains a causal-link $n_1 \rightarrow subgoal_1 \rightarrow \ldots \rightarrow subgoal_n \rightarrow goal \rightarrow {}^{*}finish^{*} \rightarrow goal$, and
>
> (2) for every causal-link $n_1 \rightarrow subgoal_1 \rightarrow \ldots \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \ldots \rightarrow goal_m$ of the plan:
>
> (a) if $n \neq 1$, then for each i such that $1 \leq i < n$, either:
>
> (i) $subgoal_{i+1}$ is a conjunction, $subgoal_i$ is one of its conjuncts, and the plan also contains a causal-link $n_1{}^{*} \rightarrow subgoal_1{}^{*} \rightarrow \ldots \rightarrow subgoal_j{}^{*} \rightarrow subgoal_{i+1} \rightarrow \ldots \rightarrow subgoal_n \rightarrow n_2{}^{*} \rightarrow goal_1 \ldots \rightarrow goal_m$ where $subgoal_j{}^{*}$ is the other conjunct of $subgoal_{i+1}$; or
>
> (ii) $subgoal_{i+1}$ is nomically implied by $subgoal_i$;
>
> (b) if $m \neq 1$, then for each i such that $1 \leq i < m$, either:
>
> (i) $goal_{i+1}$ is a conjunction, $goal_i$ is one of its conjuncts, and the plan also contains

a causal-link $n_1^* \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2^* \rightarrow goal_1^* ... \rightarrow ... goal_j^* \rightarrow goal_{i+1} \rightarrow ... \rightarrow goal_m$ where $goal_j^*$ is the other conjunct of $goal_{i+1}$; or

(ii) $goal_{i+1}$ is nomically implied by $goal_i$;

(c) if $n_1 = {}^*start^*$ then $subgoal_1$ is true in *start-state*;

(d) if $n_2 \neq {}^*finish^*$ then if A is the action of n_2 , "$(A/subgoal_n) \blacktriangleright goal_1$" is a member of *conditionals*;

(e) if $n_1 \neq {}^*start^*$ then the plan contains a causal-link $n_3 \rightarrow subgoal_1^* \rightarrow ... \rightarrow subgoal_{n^*}^* \rightarrow n_4 \rightarrow goal_1^* ... \rightarrow goal_{m^*}^* \rightarrow subgoal_1$;

(f) n_1 is ordered before n_2 by the ordering-constraints of the plan; and

(g) $subgoal_n$ is temporally-projectible.

To accommodate collective undermining, let us define:

A plan-step n of a plan **collectively undermines a causal-link** $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$ iff there is a linearization of the plan such that:

(i) n occurs between n_1 and n_2,

(ii) there is a set X of temporally-projectible expectable-results of the sequence of actions prescribed by the plan-steps preceding n in the linearization such that:

(a) $subgoal_1 \in X$;

(b) $\sim\Pi X$ is an expectable-result of the sequence of actions prescribed by the plan-steps $^*start^*,...,n$ in the linearization;

(c) there is no X_0 such that $X_0 \subseteq X$, $subgoal_1 \notin X_0$, and $\sim\Pi X_0$ is an expectable-result of the sequence of actions prescribed by the plan-steps $^*start^*,...,n$ in the linearization.

Given a causal-link $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$, let n_1 be its *root*, n_2 its *target*, and $subgoal_1$ its *initial subgoal*. The most manageable case of collective undermining occurs when the set X is a set of initial subgoals of causal-links. Let us define:

A plan-step s of a plan **collectively undermines a set $\mathcal{L}$ of causal-links** iff there is a linearization of the plan such that:

(i) s occurs between the root and the target of each member of $\mathcal{L}$,

(ii) the initial-subgoal of each member of $\mathcal{L}$ is temporally-projectible;

(iii) the negation of the conjunction of the initial-subgoals of members of $\mathcal{L}$ is an expectable-result of the sequence of actions prescribed by the plan-steps $^*start^*,...,s$ in the linearization;

(iv) there is no set $\mathcal{L}^*$ of causal-links of *plan* such that $\mathcal{L}^* \subset \mathcal{L}$ and the negation of the conjunction of the initial-subgoals of members of $\mathcal{L}^*$ is an expectable-result of the sequence of actions prescribed by the plan-steps $^*start^*,...,s$ in the linearization.

To illustrate with a contrived example, consider a game in which you throw a ball at a target mounted on a wall, and then the ball falls into one of two buckets. I am a good enough pitcher to be able to hit the target reliably, but not good enough to be able to determine into which bucket the ball will fall. Suppose the buckets are initially empty, and the goal is to hit the target and have the buckets both empty. Consider the plan to accomplish this by throwing the ball and hitting the target. The causal-links of this plan have the form ***start**** → *have-ball* → ***throw-ball*** → *hit-target*, ***start**** → *bucket-#1-empty* → *both-buckets-empty* → ****finish**** → *both-buckets-empty*, and ***start**** → *bucket-#2-empty* → *both-buckets-empty* → ****finish**** → *both-buckets-empty*. This is obviously not a good plan, because although the buckets are initially empty and throwing the ball will

result in the target being hit, the ball will then fall into one of the buckets and they will not both be empty after all. Given that I cannot predict into which bucket the ball will fall, no causal-link of the plan is undermined in the sense of section two, but the *set* of causal-links consisting of ***start**** → *bucket-#1-empty* → *both-buckets-empty* → ****finish**** → *both-buckets-empty*, and ***start**** → *bucket-#2-empty* → *both-buckets-empty* → ****finish**** → *both-buckets-empty* is collectively-undermined in the above sense, and hence each member of that set is collectively-undermined.

Unfortunately, not all cases of collective undermining result from collectively undermining a pre-existing set of causal-links. Suppose that in the preceding example my goal is simply to hit the target and have bucket #1 empty. Again, consider the plan to do that by throwing the ball and hitting the target. In this case the plan has just two causal-links, ****start**** → *have-ball* → ***throw-ball*** → *hit-target* and ***start**** → *bucket-#1-empty* → ****finish**** → *bucket-#1-empty*. The plan is still not a good plan, because the ball might fall into bucket #1. But now the collective undermining of the causal-link ***start**** → *bucket-#1-empty* → ****finish**** → *bucket-#1-empty* does not result from collectively undermining a set of causal-links. Instead, it results from collectively undermining the set of two expectable-results *bucket-#1-empty* and *bucket-#2-empty*, only one of which is an initial-subgoal of a causal-link in the plan.

Let us revise the definition of causal-soundness to appeal to collective undermining:

> A plan is **causally-sound** iff it is presumptively-sound relative to a set of true planning-conditionals and the plan does not collectively-undermine any of its own causal-links.

We can can then prove theorems 1 and 2 much as before. Let us also define:

> An embellishment $plan_0$ of *plan* **presumptively undermines** a set of causal-links $\mathcal{L}$ of *plan* iff (1) $plan_0$ is presumptively-sound, (2) the goal of $plan_0$ is the negation of the conjunction of the initial-subgoals of members of $\mathcal{L}$, (3) $plan_0$ has a single penultimate plan-step *s*, and (4) there is a linearization of $plan_0$ in which *s* occurs between the root and the target of every member of $\mathcal{L}$.

As illustrated above, the collective undermining of a causal-link need not result from the undermining of a set of causal-links, because some of the expectable-results involved in the collective undermining may not be involved in the plan. However, it is always possible to extend the plan by adding causal-links to those expectable-results (making them both the goal and subgoal of the causal-link), and adding attendant ordering-constraints, so that the collective undermining does consist of a undermining of a set of causal-links of the extended plan. That produces an embellishment of the plan. We can then prove:

Theorem 8:

> A plan *plan* collectively-undermines its causal-link *L* iff there is an embellishment *plan+* of *plan* and a set $\mathcal{L}$ of causal-links of *plan+* such that *plan+* presumptively-undermines $\mathcal{L}$, and (1) $L \in \mathcal{L}$, (2) *plan+* is causally-sound, and (3) there is no embellishment $plan_0$ of *plan+* and set $\mathcal{L}^*$ of causal-links of *plan+* such that (a) $plan_0$ presumptively-undermines $\mathcal{L}^*$, (b) $\mathcal{L}^* \subseteq \mathcal{L}$, (c) $L \notin \mathcal{L}^*$, and (d) $plan_0$ is causally-sound.

An immediate corollary of theorems 1 and 8 is:

Theorem 9:

> A plan *plan* collectively-undermines its causal-link *L* iff there is a presumptively-sound

embellishment *plan+* of *plan* and a set $\mathcal{L}$ of causal-links of *plan+* such that *plan+* presumptively-undermines $\mathcal{L}$, and (1) $L \in \mathcal{L}$, (2) there is no presumptively-sound embellishment $plan_0$ of *plan+* and set $\mathcal{L}^*$ of causal-links of *plan+* such that (a) $plan_0$ presumptively-undermines $\mathcal{L}^*$, (b) $\mathcal{L}^* \subseteq \mathcal{L}$, (c) $L \notin \mathcal{L}^*$, and (d) $plan_0$ does not collectively-undermine any of its causal-links.

Translating theorem 9 into a set of rules for defeating applications of SPLIT-CONJUNCTIVE-GOAL is a bit complicated. The search for defeaters will consist of finding an appropriate planning-conditional $(A/C) \Rightarrow \sim\Pi X$, such that (1) A is the action prescribed by some plan-step s, (2) the initial-subgoal of some causal-link L is in X, (3) the ordering of the plan-steps can be extended so that s occurs between the root and target of L, (4) an embellishment achieving C can be constructed in such a way that all the plan-steps precede s, and (5) embellishments can be constructed that achieve the other members of X and merged with the original embellishment in such a way that s occurs between the root and target of every link that establishes one of these members of X. The defeater produced in this way is itself defeasible, being defeated by finding a proper subset X_0 of X that satisfies (1), (4) and (5) but not (2).

We can implement this by replacing UNDERMINE-CAUSAL-LINKS and UNDERMINE-CAUSAL-LINK by the following triple of rules:

COLLECTIVELY-UNDERMINE-CAUSAL-LINKS

Given an inference in accordance with SPLIT-CONJUNCTIVE-GOAL, ADD-ORDERING-CONSTRAINT, or CONFRONTATION to the conclusion that *plan&* will achieve $(G_1 \,\&\, G_2)$-*at*-*t*, adopt interest in establishing that *plan&* collectively-undermines one of its own causal-links. If it is determined that it does collectively-undermine one of its own causal-links, take the inference to the conclusion that *plan&* will achieve $(G_1 \,\&\, G_2)$-*at*-*t* to be defeated.

UNDERMINE-CAUSAL-LINK

Given an interest in establishing that *plan&* collectively-undermines one of its own causal-links, for each causal-link $n_1 \rightarrow subgoal_1 \rightarrow \ldots \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow \ldots \rightarrow goal_m$ of *plan&*, adopt interest in finding an embellishment $plan_0$ of *plan&* that achieves $\sim subgoal_1$ between n_1 and n_2 consistent with the ordering-constraints of *plan&*,. Given $plan_0$, infer nondefeasibly that *plan&* collectively-undermines one of its own causal-links.

COLLECTIVELY-UNDERMINE-CAUSAL-LINK

Given an interest in establishing that *plan&* collectively-undermines one of its own causal-links, for each causal-link $n_1 \rightarrow subgoal_1 \rightarrow \ldots \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow \ldots \rightarrow goal_m$ of *plan&*, adopt interest in finding a g and an embellishment $plan_0$ of *plan&* such that (1) $plan_0$ achieves $\sim(subgoal_1 \,\&\, g)$ between n_1 and n_2 consistent with the ordering-constraints of *plan&*, and (2) there is an embellishment $plan^*$ of $plan_0$ that achieves g before some penultimate node of $plan_0$. Given $plan_0$ and $plan^*$, infer nondefeasibly that *plan&* collectively-undermines one of its own causal-links.

Here g is $\Pi(X - \{subgoal_1\})$. The defeasibility of the defeat results from the defeasibility of the inference to the conclusion that $plan^*$ will achieve g.

UNDERMINE-EMBEDDED-CAUSAL-LINKS and UNDERMINE-EMBEDDED-CAUSAL-LINK are replaced by an analogous triple of rules.

There are two ways of repairing plans that collectively-undermine some of their own causal-

links:

ADD-ORDERING-CONSTRAINT

Given an interest in finding a plan for achieving a conjunctive goal $(g_1 \& g_2)$-*at-t*, and plans $plan_1$ for g_1-*at-t* and $plan_2$ for g_2-*at-t*,, if *plan&* is a putative plan for $(g_1 \& g_2)$-*at-t*, constructed by merging plans $plan_1$ and $plan_2$ (and possibly other plans), but a plan-step *n* of *plan&* collectively-undermines one of its own causal-links $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$ by collectively-undermining a set of causal-links $\mathcal{L}$, construct a plan *plan+* by adding an ordering-constraint to the effect that that *n* not occur between the root and target of some member of $\mathcal{L}$ (if this can be done consistently) and infer defeasibly that *plan+* will achieve $(g_1 \& g_2)$-*at-t*.

CONFRONTATION

Given an interest in finding a plan for achieving a conjunctive goal $(g_1 \& g_2)$-*at-t*, and plans $plan_1$ for g_1-*at-t*, and $plan_2$ for g_2-*at-t*, if *plan&* is a putative plan for $(g_1 \& g_2)$-*at-t* constructed by merging plans $plan_1$ and $plan_2$ (and possibly other plans), but a plan-step *n* of *plan&* collectively-undermines one of its own causal-links $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$ by collectively-undermining a set of causal-links $\mathcal{L}$ by virtue of there being an embellishment $plan_0$ that achieves *P* where *P* is the negation of the initial subgoal of some member of $\mathcal{L}$, then for each causal-link $n_0 \rightarrow G_1 \rightarrow ... \rightarrow G_m \rightarrow n \rightarrow P$ of $plan_0$, adopt interest in finding a plan for achieving $\sim G_1$. If a plan *repair-plan* is proposed for achieving $\sim G_1$, construct a new plan *plan+* by adding to *plan&* the plan-steps, ordering-constraints, and causal-links of *repair-plan*, with the following exception. Replace each causal-link of the form $n^* \rightarrow SG_1 \rightarrow ... \rightarrow SG_n \rightarrow$ **finish** $\rightarrow SG_n$ in *repair-plan* by the causal-link $n^* \rightarrow SG_1 \rightarrow ... \rightarrow SG_n \rightarrow n \rightarrow SG_n$ and order n^* between n_0 and *n*. If this ordering is consistent, infer defeasibly that *plan+* will achieve $(g_1 \& g_2)$-*at-t*.

CONFRONTATION works by undermining the embellishment that collectively-undermines the causal-link of the original plan. Note, however, that CONFRONTATION requires outright undermining, not collective undermining. Collective undermining would leave open the possibility that the embellishment will achieve its goal (and undermine the causal-link of the original plan) in some cases even if does not do so in all cases.

Similar modifications are required to ADD-EMBEDDED-ORDERING-CONSTRAINT and EMBEDDED-CONFRONTATION.

This set of rules for goal-regression planning is sound and complete, for the same reason the set of rules of section four was sound and complete relative to (R1). Thus we have successfully removed the syntactical constraints of section two.

Implementation

With the exception of PLAN-FOR-IMPLICANS, the inference-schemes described above have been implemented in the OSCAR defeasible reasoner to produce an implemented defeasible planner. The implementation is straightforward using OSCAR's macro language for the construction of inference-schemes. For example, GOAL-REGRESSION is implemented by defining:

```
(def-backwards-reason GOAL-REGRESSION
   :conclusions "(plan-for plan goal)"
   :condition (interest-variable plan)
   :backwards-premises
      "((precondition & action) => goal)"
      (:condition (temporally-projectible precondition))
      "(plan-for subplan precondition new-goals nodes nodes-used links)"
```

```
    "(define plan (extend-plan action goal subplan))"
    (:condition (not (null plan)))
  :variables precondition action goal plan subplan)
```

where EXTEND-PLAN is defined by a piece of LISP code. The details of this implementation are described in a technical report (Pollock 1998), and both the technical report and the implemented planner can be downloaded from my website (http://www.u.arizona.edu/~pollock).

Simple Extensions

Most planners are based on special-purpose inference engines dedicated exclusively to planning. The OSCAR planner is based instead on a general-purpose defeasible reasoner. The reasoner is applied to planning by providing it with reasoning-schemas that concern plans, but the structure of the reasoner itself remains unchanged. A simple illustration of this is that the planning rules need not mention unification or variable binding, because that is all handled automatically by the defeasible reasoner. One of the consequences of this approach is that we can represent causal information about actions very simply in terms of conditionals, rather than adopting a special action representation language like STRIPS or ADL. This makes it easy for the OSCAR planner to combine reasoning about other matters with its planning, and it makes it easy to extend the planning rules to accommodate more complex planning considerations that are beyond the scope of most planners. The next three sections illustrate this by showing how the OSCAR planner can reason aboout concurrent actions, quantified preconditions and effects, domains in which objects are created and destroyed, and causal connections involving complex temporal relationships.

9. Concurrent Actions

It is often observed that either the STRIPS or ADL representation of actions precludes the possibility of concurrent actions, because concurrent actions can conspire to produce "cooperative results" that are not results of the individual actions. Pednault (1987) gives the example of lifting both sides of a table simultaneously, and Chapman (1987) gives the example of pressing down on both sides of a Lego die. It is perhaps of interest that representing the results of actions instead in terms of planning-conditionals makes it quite easy to both describe the results of concurrent actions and construct plans that use those results. To accomplish this, we simply allow A in the planning conditional $(A/C) \blacktriangleright G$ to be either an action or a conjunction of actions performed simultaneously. Then we can describe the cooperative results of concurrent actions by employing planning conditionals like ((*lift right side of table at time t* & *lift left side of table at time t*) / *table weighs less than 100 pounds*) ➤ (*table suspended above floor at time t**)). Where A and A^* are conjunctions of actions, I will write $A \subseteq A^*$ iff every conjunct of A is a conjunct of A^*.[39] Because planning-conditionals are nomic generalizations, if $A \subseteq A^*$ and $(A/C) \blacktriangleright G$, it follows that $(A^*/C) \blacktriangleright G$, so adding conjuncts can only give rise to new results—not conflicting results.

To accommodate concurrent actions in our definition of *expectable-result*, we allow each A_i in an action-sequence to be either a single action or a conjunction of actions, and modify clauses (ii) and (iii) accordingly:

(R7) Where *start-state* is a state of affairs, *conditionals* is a set of planning-conditionals, and $t_0 < \ldots < t_{n+1}$ is a sequence of times, $P\text{-}at\text{-}t_{n+1}$ is an **expectable-result** of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$

[39] If A is not a conjunction, I take it to be its own only conjunct.

relative to *start-state* and *conditionals* iff either:

(i) $n = 0$, P is temporally-projectibile, and P-*at*-t_0 is true in *start-state*; or

(ii) $n > 0$, P is temporally-projectibile, and for some $A_n^* \subseteq A_n$, *conditionals* contains a conditional $(A_n^*/C) \blacktriangleright P$ such that C-*at*-t_n is an expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots, A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$ and $t_{n+1} > t_n$; or

(iii) $n > 0$, P-*at*-t_n is a temporally-projectible expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, and if for some $A_n^* \subseteq A_n$, *conditionals* contains a conditional of the form $(A_n/C) \blacktriangleright \sim\Pi X$ such that C-*at*-t_n is a temporally-projectible expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots, A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, $t_{n+1} > t_n$, X is a set of temporally-projectible expectable-results of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, and $P \in X$, then for some $A_n^* \subseteq A_n$, *conditionals* contains a conditional $(A_n^*/C^*) \blacktriangleright \sim\Pi X_0$ such that C^*-*at*-t_n is a temporally-projectible expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_{n-1}\text{-}at\text{-}t_{n-1}\rangle$, $X_0 \subseteq X$, and $P \notin X_0$; or

(iv) $n > 0$ and P-*at*-t_{n+1} is a conjunction whose conjuncts are expectable-results of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$; or

(v) $n > 0$ and P-*at*-t_{n+1} is a logical consequence of some expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots, A_n\text{-}at\text{-}t_n\rangle$; or

(vi) $n > 0$ and P-*at*-t_{n+1} is known to be nomically implied by some expectable-result of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$.

To construct plans with this semantics, we record each conjunct of a conjunctive action in a separate plan-node (but all with the same time reference). Then we modify GOAL-REGRESSION as follows:

GOAL-REGRESSION

Given an interest in finding a plan for achieving G-*at*-t, if G is temporally-projectible , adopt interest in finding planning-conditionals $(A/C) \blacktriangleright G$ having G as their consequent. Given such a conditional, adopt an interest in finding a plan for achieving C-*at*-t^*. If it is concluded that a plan *subplan* will achieve C-*at*-t^*, construct a plan by (1) adding new steps to the end of *subplan* where each new step prescribes an action A_i-*at*-t^* where A_i is a conjunct of A, (2) adding the constraint $(t^* < t)$ to the ordering-constraints of *subplan*, and (3) adjusting the causal-links appropriately. Infer defeasibly that the new plan will achieve G-*at*-t.

EMBEDDED-GOAL-REGRESSION must be modified similarly. Only one other change required. In ADD-ORDERING-CONSTRAINT, the requirement that the time t of the undermining step not occur between the times t_1 and t_2 of the root and target of the undermined causal-link was previously taken to mean that either $t < t_1$ or $t_2 < t$. Given the possibility of concurrent actions, it must mean instead that either $t \leq t_1$ or $t_2 \leq t$.

10. Quantifiers, Creation and Destruction

The preconditions and consequents of planning-conditionals often involve quantification. *A* standard example (Weld 1994) is that a precondition for putting block *A* on block *B* is that there isn't already a block on block *B*. Thus in order to put block *A* on block *B*, if block *C* is already on block *B*, GOAL-REGRESSION will lead us to try to find a plan for removing it. There

are two radically different ways of doing this. We might put block C elsewhere, or we might zap it with a block-destroying ray gun.

ADL makes the assumption that nothing is created or destroyed in the course of executing a plan, thus precluding the second way of solving the above planning problem (Pednault 1987, pg. 69). Accordingly, the standard way of handling quantification in planning is to expand universally or existentially quantified formulas into finite conjunctions or disjunctions ranging over all the actual objects in the fixed planning domain (Weld 1994).

Given the semantics for planning provided by (R7), there is no need to restrict planning problems to either fixed or finite domains. Let $E!(x,t)$ mean "x exists at time t". Then $[(\forall x)\varphi]$-at-t is taken to be equivalent to $(\forall x)[E!(x,t) \supset \varphi\text{-at-}t]$. A planning-conditional governing putting one block on another is:

(10.1) (put A on B at t / $\sim(\exists x)(x \text{ on } B)\text{-at-}t$) ➤ ($A$ on B at t^*).

The goal of having A on B at time τ will then produce the subgoal of having

(10.2) $\sim(\exists x)(x \text{ on } B)\text{-at-}\tau_1$

for some $\tau_1 < \tau$. Suppose that C is the only block on B at the start-time t_0:

(10.3) $(\forall x)[(x \text{ on } B) \equiv x = C]\text{-at-}t_0$.

(10.3) logically entails

(10.4) $(\forall x)[(x \text{ on } B) \supset x = C]\text{-at-}t_0$.

Assuming temporal-projectibility, we can infer defeasibly:

(10.5) $(\forall x)[(x \text{ on } B) \supset x = C]\text{-at-}\tau_1$.

(10.5) conjoined with either

(10.6) $\sim(C \text{ on } B)\text{-at-}\tau_1$

or

(10.6*) $\sim E!(C,\tau_1)$

logicallly entails (10.2). Thus PLAN-FOR-IMPLICANDA gives rise to (10.6) and (10.6*) as subgoals (conjoined with (10.5), which also becomes a subgoal but is achieved by PROPOSE-NULL-PLAN). We can achieve (10.6) by moving C to another location, or we can achieve (10.6*) by zapping C.

This example illustrates that the expansion of quantified formulas into conjunctions and disjunctions can foil a planning problem by making solutions involving the creation or destruction of objects unavailable. But sometimes the expansion seems to be intuitively correct. If my goal is to have all the lights in the room turned on, it would be perverse to try to achieve this by either removing all the lights from the room or destroying them. The problem is that the English formulation of the goal as "make it the case that all of the lights in the room are turned on" is ambiguous between "make it that case that $(\forall x)(x$ is a light in the room at $t \supset x$ is turned on at

t)" and "make it the case that $(\forall x)$(x is a light in the room *at* $t_0 \supset x$ is turned on at t)". The former goal can be achieved by removing or destroying lights, but the latter goal can only be achieved by turning on all the lights currently in the room. However, none of this is a problem for the OSCAR planning system as long as the goal is formulated precisely to distinguish between these two readings.

It is worth noting that the expansion of quantified formulas assumes a finite domain. But there is no finiteness requirement for planning with quantified formulas in the OSCAR system of planning. The above example did assume that there are only finitely many things on B, but that is just a feature of that particular example. If instead there were infinitely many blocks on B, but we had a block-zapper that would simultaneously destroy all the blocks atop a given block, then we could solve the planning problem even more simply and without using (10.4).

11. Accommodating Complex Temporal Relationships

Most current planners employ a STRIPS or ADL representation of actions. As Pednault (1987) shows, ADL is expressively equivalent to the situation calculus (McCarthy and Hayes 1969). In particular, time is represented by discrete time points. Throughout this paper I have been implicitly assuming metric time, times being represented by real numbers, but none of the planning rules formulated so far in this paper turn upon that, because all '➤' requires is a simple ordering of times. However, I remarked at the end of section five that the causal connections between actions, preconditions, and effects, can involve more complex temporal relationships than those symbolized using '➤'. The example of serving a tennis ball illustrated the use of a nomic conditional of the form

$$\{(A^*\text{-at-}t \ \&\ SG\text{-}at\text{-}t{+}\alpha) \Rightarrow (\exists\delta)G\text{-throughout-}(t{+}\varepsilon\,,\ t{+}\varepsilon{+}\delta]\},$$

where the time of the action is the time at which it begins. More generally, the causal reasoning involved in planning can employ conditionals of the form

$$\{(A^*\text{-at-}t \ \&\ (SG_1\text{-}throughout\text{-}[t{+}\alpha_1,t{+}\beta_1] \ \&\ \ldots \ \&\ SG_m\text{-}throughout\text{-}[t{+}\alpha_m,t{+}\beta_m]))$$
$$\Rightarrow (\exists\delta)G\text{-throughout-}(t{+}\varepsilon\,,\ t{+}\varepsilon{+}\delta]\}.$$

I will abbreviate this as $(A\ /\ SG_1, \ldots, SG_m) \blacktriangleright_{\varepsilon,\langle\alpha 1,\ldots,\alpha m\rangle,\langle\beta 1,\ldots,\beta m\rangle} G$, or more simply as $(A/SG) \blacktriangleright_{\varepsilon,\alpha,\beta} G$ where SG, α, and β are finite sequences. If β is a sequence of zeros, I will omit it. If α is a sequence of zeros as well, I will omit both. If $\varepsilon = 0$ too, I will omit all three subscripts. It is of interest to see how these more complex planning-conditionals can be used in planning, and to show that they present no serious obstacle to the OSCAR planner.

We must generalize our semantics to accommodate the use of these conditionals. That semantics proceeds in terms of the concept of an *expectable-result*, but that can no longer be defined using the kind of definition exemplified by (R2)-(R7) which recurses on the sequence of actions. The problem is that time delays in causation may result in action $A_{n\text{-}1}$-*at*-$t_{n\text{-}1}$ making P true, but only *after* t_n. We can instead define the *expectable-interval* over which P can be defeasibly expected to be true as a result of performing the actions in the sequence. Even this is complicated, because there can be multiple planning-conditionals appealing to a single action, with the result that the single action can first make P true and later make it false. E.g., eating dinner early may make me not hungry an hour from now but hungry again 3 hours from now. To handle all of this we consider the sequence of all times at which the expectable-intervals may change as a result of something happening, and define the expectable-intervals recursively on the basis of

that sequence. This sequence of times is characterized as follows:

> Given a set of planning-conditionals *conditionals*, an action-sequence $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$, and a starting-time t_0, the sequence of *associated-times* consists of t_0 together with all times $t_i+\varepsilon$ such that $1 \le i \le n$ and *conditionals* contains some conditional $(A_i/C) \blacktriangleright_{\varepsilon,\alpha,\beta} P$. The sequence of associated-times is ordered by '≤'.

(R8) Where *conditionals* is a set of planning-conditionals, $t_0 < \ldots < t_{n+1}$ is a sequence of times, and $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$ is an action-sequence, let $\langle\tau_0,\ldots,\tau_k\rangle$ be the sequence of associated times. For each temporally-projectible P, the *expectable-interval* $\mathcal{I}(P,\langle\tau_0,\ldots,\tau_i\rangle)$ is defined recursively as follows:

$$\mathcal{I}(P,\langle\tau_0\rangle) = \begin{cases} [\tau_0,\infty) \text{ if } P\text{-}at\text{-}\tau_0 \text{ is true} \\ \varnothing \text{ otherwise} \end{cases}$$

if $i > 0$

$$\mathcal{I}_0(P,\langle\tau_o,\ldots,\tau_i\rangle) = \begin{cases} \mathcal{I}(P,\langle\tau_o,\ldots,\tau_{i-1}\rangle) \cup (\tau_i,\infty) \text{ if for some } A_n^* \subseteq A_n, \textit{ conditionals} \text{ contains a} \\ \text{conditional } (A_n^*/C) \blacktriangleright_{\varepsilon,\alpha,\beta} P \text{ such that for each } C_j \text{ in } C, \\ [\tau_i+\alpha_i-\varepsilon,\tau_i+\beta_i-\varepsilon] \subseteq \mathcal{I}(C_j,\langle\tau_o,\ldots,\tau_{i-1}\rangle); \\ \\ \mathcal{I}(P,\langle\tau_o,\ldots,\tau_{i-1}\rangle) -(\tau_i,\infty) \text{ if for some } A_n^* \subseteq A_n, \textit{ conditionals} \text{ contains a} \\ \text{conditional } (A_n^*/C) \blacktriangleright_{\varepsilon,\alpha,\beta} \sim\Pi X \text{ such that for each } C_j \text{ in } C, \\ [\tau_i+\alpha_i-\varepsilon,\tau_i+\beta_i-\varepsilon] \subseteq \mathcal{I}(C_j,\langle\tau_o,\ldots,\tau_{i-1}\rangle), P\in X, \text{ and there is no} \\ A_n^* \subseteq A_n \text{ such that } \textit{conditionals} \text{ contains a conditional} \\ (A_n^*/C^*) \blacktriangleright_{\delta,\gamma,\lambda} \sim\Pi X_0 \text{ such that for each } C^*_j \text{ in } C^*, \\ [\tau_i+\gamma_i-\delta,\tau_i+\lambda_i-\delta] \subseteq \mathcal{I}(C^*_j,\langle\tau_o,\ldots,\tau_{i-1}\rangle), X_0 \subseteq X, \text{ and } P\notin X_0; \\ \\ \mathcal{I}(P,\langle\tau_o,\ldots,\tau_{i-1}\rangle) \text{ otherwise.} \end{cases}$$

$$\mathcal{I}(P,\langle\tau_o,\ldots,\tau_i\rangle) = \cup\{\cap\{\mathcal{I}_0(Q,\langle\tau_o,\ldots,\tau_i\rangle) \mid Q\in X\} \mid \Pi X \text{ is known to nomically imply } P\}$$

P-at-t is an *expectable-result* of $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$ iff $t\in \mathcal{I}(P,\langle\tau_o,\ldots,\tau_i\rangle)$.
P-throughout-$[t,t^*]$ is an *expectable-result of* $\langle A_1\text{-}at\text{-}t_1,\ldots,A_n\text{-}at\text{-}t_n\rangle$ iff $[t,t^*] \subseteq \mathcal{I}(P,\langle\tau_o,\ldots,\tau_i\rangle)$.

To explain, there are three ways the expectable-interval for P can be altered at τ_i. First, there may be a conditional $(A_n^*/C) \blacktriangleright_{\varepsilon,\alpha,\beta} P$ that makes P expectably true from τ_i forwards. This results in our adding (τ_i,∞) to the previous expectable-interval $\mathcal{I}(P,\langle\tau_o,\ldots,\tau_{i-1}\rangle)$. Second, there may be a conditional $(A_n^*/C) \blacktriangleright_{\varepsilon,\alpha,\beta} \sim\Pi X$ making P no longer expectably true from τ_i forwards. This results in our deleting the times in (τ_i,∞) from the previous expectable-interval $\mathcal{I}(P,\langle\tau_o,\ldots,\tau_{i-1}\rangle)$. I assume that these first two cases are mutually exclusive. There cannot be true causal connections resulting in P and $\sim P$ both being caused at the same time. $\mathcal{I}_0(P,\langle\tau_o,\ldots,\tau_i\rangle)$ is the expectable-interval that results directly from planning-conditionals governing P. However, P can also be made expectably true by being nomically implied by by other expectably true propositions. Given any set X of propositions that are known to jointly nomically imply P, P is expectably true whenever all members of X are expectably true. So $\mathcal{I}(P,\langle\tau_o,\ldots,\tau_i\rangle)$ must include $\cap\{\mathcal{I}_0(Q,\langle\tau_o,\ldots,\tau_i\rangle) \mid Q\in X\}$. This holds for every set X of propositions that are known to jointly nomically imply P, so we take $\mathcal{I}(P,\langle\tau_o,\ldots,\tau_i\rangle)$ to be the union of all such intersections. Note that because $\Pi\{P\}$ nomically implies P, $\mathcal{I}_0(P,\langle\tau_o,\ldots,\tau_i\rangle) \subseteq \mathcal{I}(P,\langle\tau_o,\ldots,\tau_i\rangle)$. Furthermore, if there are no planning-conditionals directly affecting P, $\mathcal{I}_0(P,\langle\tau_o,\ldots,\tau_i\rangle) = \mathcal{I}(P,\langle\tau_o,\ldots,\tau_{i-1}\rangle)$, so $\mathcal{I}(P,\langle\tau_o,\ldots,\tau_{i-1}\rangle) \subseteq \mathcal{I}(P,\langle\tau_o,\ldots,\tau_i\rangle)$.

To illustrate this complex definition with a simple example, suppose C_1 and C_4 are true at the starting time t_0, and suppose we have the planning-conditionals $(A_1/C_1) \blacktriangleright_{\varepsilon1,\alpha1} C_2$, $(A_2/C_2) \blacktriangleright_{\varepsilon2,\alpha2} C_3$, $(A_2/C_4) \blacktriangleright_{\varepsilon3,\alpha3} \sim C_3$. Suppose actions A_1 and A_2 are performed at times t_1 and t_2, where $t_0 < t_1 < t_1+\alpha_1 < t_1+\varepsilon_1 < t_2 < t_2+\alpha_2 < t_2+\varepsilon_2 < t_2+\alpha_3 < t_2+\varepsilon_3$. Then the associated times are $\tau_0 = t_0$, $\tau_1 = t_1+\varepsilon_1$, $\tau_2 = t_2+\varepsilon_2$, and $\tau_3 = t_2+\varepsilon_3$. The expectable-intervals then evolve as follows:

<u>Expectable-intervals relative to $\langle\tau_0\rangle$:</u>

t_0
τ_0

C_1 $[\tau_0$ .....$\infty)$

C_2

C_3

C_4 $[\tau_0$ .....$\infty)$

<u>Expectable-intervals relative to $\langle\tau_0,\tau_1\rangle$:</u>

		A_1		
	t_0	t_1	$t_1+\alpha_1$	$t_1+\varepsilon_1$
	τ_0			τ_1
C_1	$[\tau_0$ .....$\infty)$			
C_2				$(\tau_1$ .....$\infty)$
C_3				
C_4	$[\tau_0$ .....$\infty)$			

<u>Expectable-intervals relative to $\langle\tau_0,\tau_1,\tau_2\rangle$:</u>

		A_1			A_2		
	t_0	t_1	$t_1+\alpha_1$	$t_1+\varepsilon_1$	t_2	$t_2+\alpha_2$	$t_2+\varepsilon_2$
	τ_0			τ_1			τ_2
C_1	$[\tau_0$ .....$\infty)$						
C_2				$(\tau_1$ .....$\infty)$			
C_3						$(\tau_2$ .....$\infty)$	
C_4	$[\tau_0$ .....$\infty)$						

<u>Expectable-intervals relative to $\langle\tau_0,\tau_1,\tau_2,\tau_3\rangle$:</u>

		A_1			A_2				
	t_0	t_1	$t_1+\alpha_1$	$t_1+\varepsilon_1$	t_2	$t_2+\alpha_2$	$t_2+\varepsilon_2$	$t_2+\alpha_3$	$t_2+\varepsilon_3$
	τ_0			τ_1			τ_2		τ_3
C_1	$[\tau_0$ .....$\infty)$								
C_2				$(\tau_1$ .....$\infty)$					

C_3 $(\tau_2$ .....$\tau_3]$

C_4 $[\tau_0$ .....$\infty)$

Now consider how this altered semantics affects the rules for planning. The changes are minor. First, the ordering-constraints imposed by GOAL-REGRESSION must be adjusted:

GOAL-REGRESSION

Given an interest in finding a plan for achieving *G-at-t*, if *G* is temporally-projectible, adopt interest in finding planning-conditionals $(A/C) \blacktriangleright_{\varepsilon,\alpha,\beta} G$ having *G* as their consequent. Given such a conditional, adopt an interest in finding a plan for achieving the conjunction of the C_i-*at*-t_i's for C_i in *C* (where each t_i is a new variable). If it is concluded that a plan *subplan* will achieve the conjunction of the C_i-*at*-t_i's, construct a plan by (1) adding new steps to the end of *subplan* where each new step prescribes an action A_j-*at*-t^* where A_j is a conjunct of *A*, (2) adding the constraint $(t^*+\varepsilon < t)$ and the constraints $(t_i < t^*+\alpha_i)$ to the ordering-constraints of *subplan*, and (3) adjusting the causal-links appropriately. Infer defeasibly that the new plan will achieve *G-at-t*.

EMBEDDED-GOAL-REGRESSION must be modified similarly.

The other changes required concern undermining causal-links. Consider a causal-link $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$ established using the planning-conditionals $(A_1/subgoal_1) \blacktriangleright_{\varepsilon,\alpha,\beta} goal_1$ and $(A_2/goal_1) \blacktriangleright_{\delta,\gamma,\lambda} G$.. There will be a time-lag ε between the time t_1 that n_1 is executed (A_1 is performed) and the achievement of $subgoal_1$. Furthermore, $subgoal_1$ will be required to remain true not just until the time t_2 that n_2 is executed (A_2 is performed), but until $t_2+\lambda$. ε is the *root-offset* and λ the *target-offset* of the causal-link. Let us record these offsets by rewriting the causal-link in the form $n_1 \rightarrow_\varepsilon subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow_\lambda goal_1 \rightarrow ... \rightarrow goal_m$. Then we revise UNDERMINE-CAUSAL-LINK and COLLECTIVELY-UNDERMINE-CAUSAL-LINK as follows:

UNDERMINE-CAUSAL-LINK

Given an interest in establishing that *plan&* collectively-undermines one of its own causal-links, for each causal-link $n_1 \rightarrow_\varepsilon subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow_\lambda goal_1 \rightarrow ... \rightarrow goal_m$ of *plan&*, where t_1 is the time of n_1 and t_2 is the time of n_2, adopt interest in finding an embellishment $plan_0$ of *plan&* that achieves $\sim subgoal_1$ between $t_1+\varepsilon$ and $t_2+\lambda$ consistent with the ordering-constraints of *plan&*,. Given $plan_0$, infer nondefeasibly that *plan&* collectively-undermines one of its own causal-links.

COLLECTIVELY-UNDERMINE-CAUSAL-LINK

Given an interest in establishing that *plan&* collectively-undermines one of its own causal-links, for each causal-link $n_1 \rightarrow_\varepsilon subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow_\lambda goal_1 \rightarrow ... \rightarrow goal_m$ of *plan&*, where t_1 is the time of n_1 and t_2 is the time of n_2, adopt interest in finding a *g* and an embellishment $plan_0$ of *plan&* such that (1) $plan_0$ achieves $\sim(subgoal_1 \ \& \ g)$ between $t_1+\varepsilon$ and $t_2+\lambda$ consistent with the ordering-constraints of *plan&*, and (2) there is an embellishment *plan** of $plan_0$ that achieves *g* before some penultimate node of $plan_0$. Given $plan_0$ and *plan**, infer nondefeasibly that *plan&* collectively-undermines one of its own causal-links.

UNDERMINE-EMBEDDED-CAUSAL-LINK and COLLECTIVELY-UNDERMINE-EMBEDDED-CAUSAL-LINK are modified analogously. The rules for finding embellishments must be modified in the obvious way to accommodate these slightly more complex ordering-constraints. The ordering-constraints added by ADD-ORDERING-CONSTRAINT, CONFRONTATION, ADD-EMBEDDED-ORDERING-CONSTRAINT, and EMBEDDED-CONFRONTATION must also be modified in the obvious way to take account of the causal-link offsets.

12. Conclusions

Goal-regression planning is aimed at the construction of plans. In this paper I have adopted a certain conception of a plan, according to which a plan is identified with a quadruple consisting of a set of plan-steps, a set of ordering-constraints, a set of causal-links, and a goal. The definition (R6) of "expectable-result" constitutes a semantics, relative to which we can prove the soundness and completeness of the set of rules for goal-regression planning described in section eight. This provides a secure logical foundation for goal-regression planning applicable to planning by autonomous rational agents in complex environments.

The theory of goal-regression planning developed in this paper draws heavily on conventional AI planning theory, but there are also important differences. One important difference is that the planner proposed here is not an r.e. planner. It has been argued that a planning agent embedded in a complex environment must interleave planning with epistemic cognition aimed at providing information needed for planning, and this makes r.e. planning logically impossible. Instead, planning must be done defeasibly. To accomplish this I have proposed taking goal-regression planning to be a species of epistemic cognition whose purpose is to generate defeasibly reasonable conclusions of the form "Plan *p* would achieve its goal if the prescribed plan-steps were executed in any order consistent with its ordering-constraints". A system that plans in this way has been implemented using the OSCAR system of defeasible reasoning.

Viewing planning as an epistemic endeavor generates a different kind of semantical foundation for planning. The conventional approach adopts the definition (R1) of "result", and then attempts to prove the soundness and completeness of a planning algorithm. This concept of "result" is intended to be an objective concept describing the way the world will be as a result of executing the plan. But once it is recognized that planning is based upon defeasible expectations rather than objectively determinate results of actions, it becomes apparent that no such definition of "result" is possible. The semantics of planning must instead be based upon the epistemic concept of an "expectable-result".

It has been argued here that goal-regression planning presupposes a certain kind of solution to the Frame Problem, and this generates the second important difference from conventional AI planning theory. The solution to the Frame Problem uses TEMPORAL-PROJECTION at crucial points in the process of inferring that a plan can be expected to achieve its goal, and that requires the imposition of temporal-projectibility constraints in the definition of "expectable-result" and on the subgoals generated by GOAL-REGRESSION and SPLIT-CONJUNCTIVE-GOAL.

Conventional AI planning theory imposes syntactical constraints on goals and planning-conditionals that, upon reflection, seem unreasonable. Once these constraints are relaxed, it becomes apparent that the definition (R2) of "expectable-result" is inadequate in various ways. The most obvious failure of (R2) is that the logical and nomic consequences of expectable-results are not automatically expectable-results. It also fails to accommodate collective undermining. Correcting these shortcomings leads to the final definition (R6).

The system of defeasible goal-regression planning described in section eight and based upon (R6) has been implemented in the OSCAR planner, which is based upon the OSCAR system of defeasible reasoning. The resulting planner should, however, be viewed more as a

proof of concept than as a serious attempt at building a practical planner. There has been no attempt to make it particularly efficient—just to make it work. Much of the research that has been directed at making classical planners more efficient is equally applicable to the OSCAR planner, and can in principle be applied to the prioritization of the reasoning underlying the plan search. However, that is a matter for further research.

One feature of the OSCAR planner requires particular mention in this connection. "Standard" partial order planners like SNLP and UCPOP produce partial plans as they regress backwards from the goal, and extend the partial plans by either adding new steps to achieve unachieved subgoals or adding steps or ordering-constraints to resolve threats. By contrast, the OSCAR planner produces only complete plans, starting with short complete plans for subgoals and then extending them to build a complete plan for the final goal. Partial plans are, in effect, represented by the structure of epistemic interests produced in the course of the plan reasoning, but they are not actually constructed as entities to be reasoned about. A side effect of this is that the search for threats (or underminings) must be postponed until the plans are produced. There is much recent work on the issue of the order in which plan construction (i.e., adding plan-steps) and threat resolution should occur to make a planner efficient. Of particular importance is Pollack, Joslin, and Paolucci (1997) and Gerevini and Schubert (1996).[40] Both papers produce substantial empirical evidence suggesting that it many cases it is best to postpone threat resolution until plan construction is otherwise complete. In explaining this, Gerevini and Schubert observe (and Pollack, Joslin, and Paolucci concur) that many threats will disappear during the course of planning as new ordering constraints and variable bindings are adodpted, and so planning effort is saved by ignoring them until we know that they will not disappear. Note that this is also an argument for postponing threat resolution until we know that a threat is real, i.e., an undermining, which is just what the OSCAR planner does. So it is unclear whether this feature of the OSCAR planner is an advantage or a deficit. This is also matter for further research.

A planning system that constructs plans by reasoning about them defeasibly will of necessity run more slowly than an r.e. planner because of the greater overhead of the defeasible reasoner. In light of these "built-in inefficiencies", I expect there to be considerable resistance to the proposal that we should base planners on the semantics provided by (R6) rather than the more traditional semantics provided by (R1). Traditional AI planning researchers may insist that classical planning algorithms provide an efficient solution to the planning problem, and my proposed changes undermine that inefficiency. The proper answer to this objection is that classical planning algorithms provide an efficient solution to a different problem. In an applied planning situation in which (1) all of the relevant factual knowledge can be compiled-in from the start, (2) temporal-projectibility problems can be avoided by carefully tailoring the formalism to avoid expressing temporally-unprojectible propositions, and (3) the syntactical constraints can be satisfied by careful tailoring of the formalism, then classical planning algorithms are a better solution than a defeasible planner based upon (R6). (R6) and the defeasible planner based upon it are instead solutions to two other problems. The first is the theoretical problem of understanding the logic of goal-reduction planning in its full generality. Classical planning algorithms based upon (R2) cannot provide such an understanding. The second is the problem of building a truly autonomous planning agent that can deal with a complex and unpredictable environment. For such an agent, assumptions (1) – (3) are indefensible. If we want such an agent to behave rationally in an uncooperative world, it must be based upon (R6) and engage in defeasible planning. Classical planning algorithms are inapplicable.

There has been much recent excitement about two new "non-traditional"

---

[40] See also Peot and Smith (1993, Joslin and Pollack (1994), Srinivasan and Howe (1995), and Williamson and Hanks (1996).

planners—Graphplan (Blum and Furst 1995, 1996) and SATPLAN (Kautz and Selman 1996). On many problems, these planners drammatically outperform traditional planners like UCPOP and PRODIGY, and they appear to do so without using classical planning ideas like goal-regression and threat-resolution. Does this threaten to make the classical ideas irrelevant? I don't think so, for at least two reasons. First, it is not clear why Graphplan and SATPLAN perform so well. For example, Brafman, Hoos, and Boutilier (1998) provide evidence that the success of SATPLAN is attributable more to its use of stochastic search than to the propositional encoding that makes it part company from classical planning ideas. They construct a hybrid planner LSPS that combines classical planning ideas with stochastic search and find that in many cases it outperforms SATPLAN. Second, Graphplan and SATPLAN make essential use of the assumption that we have complete knowledge of the planning domain, including knowledge of all existent objects and all effects of actions. The planners literally cannot function without that assumption, but that assumption is unreasonable for an autonomous agent operating in a complex environment. So these planners may prove very useful for applied planning problems in well controlled domains, but unable to provide the planning resources required by an autonomous rational agent.

The definition (R6) makes it possible to provide a firm logical foundation for goal-regression planning in autonomous agents. However, this work all presupposes the concept of a plan described above. This concept does not capture all of our planning endeavors. First, and most obviously, planning as discussed in this paper is restricted to "causal planning", wherein the connections between actions and the goals and subgoals they aim at achieving is a causal one. Such causal connections are expressed by nomic generalizations. This is a very strict notion of a causal connection. It can reasonably be objected that human planning agents are rarely in a position of knowing the nomic generalizations that underlie the causal connections they employ in planning. For example, suppose I plan to start a fire by striking a match and holding the lit match under some kindling. One causal connection I am relying upon in this plan is that between striking a match and its lighting. I know of certain conditions that are required for the match to light—the match must be dry, there must be sufficient oxygen, it must not be too windy, etc. But I cannot enumerate a complete list of conditions sufficient to nomically imply that the match will light when struck. This may seem to constitute a serious logical obstacle to applying the planning rules formulated above, but fortunately, it is not. Although I cannot fill out the list of conditions, I am confident that *there is* a condition *C* (possibly a long conjunction) which (1) when added to the conditions I know will generate a true planning-conditional, and (2) *C* describes conditions that actually hold, and are very general "background conditions" of a sort that can be expected to hold in virtually any case in which we try to light a match by striking it. Thus, even without knowing what *C* is, we can take it to be established by a null-plan, and we can assume that it is not undermined by subsequent plan-steps. This means, in effect, that *C* can be safely ignored in planning.

In causal planning, the plan is not guaranteed to achieve its goal, because TEMPORAL-PROJECTION is defeasible. But it is guaranteed to achieve its goal if the temporal-projections do not fail. In this sense, the causal structure is guaranteed to be correct, although presupposed persistences may fail. In this respect, causal planning contrasts with probabilistic planning, in which the connection between the actions and their effects are probabilistic rather causal. Perhaps most cases of probabilistic planning can be reduced to cases in which there are nomic generalizations as above, but the condition *C* is not known to be true. All we know is that *C* has a certain probability of being true, and that probability may change in light of the execution of some of the plan-steps. Probabilistic planning is a complicated matter,[41] but sometimes it is unavoidable.

---

[41] See, for example, BURIDAN, as described in Kushmerick, Hanks, and Weld (1995).

We sometimes find ourselves in situations in which we must plan, but lack the kind of causal knowledge required for causal planning. However, the investigation of such probabilistic goal-regression planning is beyond the scope of this paper.

The theory of goal-regression planning developed here has some more mundane shortcomings as well. This is most easily seen by reconsidering the role of disjunctions in planning. Once the syntactical constraints of section two were relaxed, the possibility of planning-conditionals with disjunctive consequents made everything more complicated by introducing the possibility of collective undermining. Section eight showed how to construct rules for goal-regression planning that accommodate collective undermining. However, the rules of section eight still fall short of constituting a complete theory of goal-regression planning. Those rules are complete for plans as thus far defined, but planning-conditionals with disjunctive consequents reveal an inadequacy in our concept of a plan.

It is useful to think briefly about the source of disjunctive consequents. There seem to be two ways in which planning-conditionals with disjunctive consequents can arise. The most straightforward source of disjunctive consequents is disjunctive antecedents. It follows from the logic of nomic generalizations that if $(A/P) \blacktriangleright R$ and $(A/Q) \blacktriangleright S$ then $[A/(P \vee Q)] \blacktriangleright (R \vee S)$. Such planning-conditionals will normally have temporally-unprojectible preconditions, and hence will be useless in planning according to the rules of section eight.

Sometimes the indeterminacy indicated by a disjunctive consequent results instead from indeterminacy built into the action. Real agents (e.g., human beings) are limited in the precision with which they can perform actions. This limited precision can produce different consequences when the same action is repeated in essentially the same circumstances. For instance, think of throwing a tennis ball "straight up" over the net of a tennis court. We never really succeed in throwing it *straight* up, so sometimes the ball will land in one court and sometimes in the other court, and we cannot predict which. This can produce a planning-conditional with a disjunctive consequent even though there is no indeterminacy in the precondition. Such planning-conditionals can play a role in undermining proposed plans, but will be of limited use in planning in accordance with the rules of section eight because their temporally-unprojectible consequents preclude their being strung together by GOAL-REGRESSION to produce a multi-step plan.

However, in real goal-regression planning, planning-conditionals with disjunctive antecedents and consequents are not useless. Disjunctions are often accurate representations of our knowledge of the circumstances in which an action may be performed. We may not know which disjunct is or will be true, and we must be able to plan in the face of such ignorance. How can we make use of disjunctive information in planning? Suppose P, Q, and R are temporally-projectible, we know P to be true, our goal is G, and we have the planning conditionals $(A_1/P) \blacktriangleright (Q \vee R)$, $(A_2/Q) \blacktriangleright G$, and $(A_3/R) \blacktriangleright G$. We know that if we perform A_1, either Q or R will result, but we cannot be sure which. We cannot plan by projecting $(Q \vee R)$ forwards to the time we perform the next step of a plan, but what we can do is consider the two possible results of the first step separately. If A_1 produces Q, we can project that forwards in time, perform A_2, and thereby achieve G. If instead A_1 produces R, we can project that forwards in time, perform A_3, and thereby achieve G. We cannot be sure which disjunct will result from performing A_1, but we can plan for each contingency separately and then decide which subplan to execute when we see what actually happens as a result of performing A_1. This is an example of *contingency planning*. Disjunctions are handled in contingency planning by planning separately for the possibility of each disjunct and then letting the results of executing the initial steps of the plan determine what subsequent steps are executed.[42]

---

[42] For work on contingency planning, see Warren (1976), Etzioni et al. (1992), Peot and Smith (1992), Goldman and Boddy (1994a, 1994b), and Pryor and Collins (1996).

What distinguishes contingency planning is that plan-steps can have conditionality built into them, and they may "call" different subplans depending upon what conditions are satisfied. Such plans are logically more complicated than the simple plans produced by the planning system described in this paper. This illustrates that the theory of goal-regression planning developed here is only complete for the rather simple conception of plans it assumes. Specifically, such plans consist of a set of plan-steps, a set of ordering-constraints, a set of causal-links (expressing nomic connections), a goal, and nothing more. Enriching the concept of a plan to add additional logical structure will require further modifications to the theory of goal-regression planning.

Appendix: Proofs of Theorems

Lemma 1a: If a plan is causally-sound, so are its linearizations.

Lemma 1b: If a linear plan is causally-sound, then it is sound.

Proof by induction on the length of a plan (i.e., the number of steps in the plan).

<u>Base case</u>: If a null-plan is causally-sound then it is presumptively-sound, and then its goal is true in the start-state. Hence by clause (i) of the definition of "result", its goal is a result of the null-sequence of actions. So the plan is sound.

<u>Induction case</u>: Suppose the lemma holds for plans of length $< n$. Suppose *plan* is a plan of length n. Let $s_1,...,s_n$ be the plan-steps of *plan*, and *goal* its goal. Let $\mathcal{L}$ be the set of all causal-links of *plan* having *goal* as their final-goal and having **finish** as their target. For each L in $\mathcal{L}$, if L is the causal-link $s_k \rightarrow g_L \rightarrow ... \rightarrow goal \rightarrow$ **finish** $\rightarrow goal$, construct a linear-plan $plan_L$ whose goal is g_L, whose plan-steps are $s_1,...,s_k$, whose ordering-constraints are those of *plan* restricted to $plan_L$, and whose causal-links are those of *plan* restricted to **start**,$s_1,...,s_k$, together with the causal-link $s_k \rightarrow g_L \rightarrow$ **finish** $\rightarrow g_L$.

$plan_L$ is presumptively-sound because *plan* is presumptively-sound and we have not removed any causal-links pertaining to the plan-steps of $plan_L$. The only linearization of $plan_L$ is $plan_L$ itself, so if $plan_L$ contained a causal-link $s_i \rightarrow subgoal \rightarrow ... \rightarrow s_j \rightarrow ...$ undermined by one of its plan-steps s_l, s_l would have to occur between s_i and s_j, and for some Q that is either ~*subgoal* or the negation of a conjunct of *subgoal*, Q would be a result of the sequence of actions prescribed by $s_1,...,s_l$. But then s_l would also undermine this same causal-link in *plan*. By hypothesis, *plan* is causally-sound and hence does not undermine any of its own causal-links, so there can be no such undermining in L either. Therefore $plan_L$ is causally-sound.

If $k < n$, it follows by the induction hypothesis that $plan_L$ is sound and hence g_L is a result of the sequence of actions prescribed by $s_1,...,s_k$. By hypothesis, the causal-link $s_k \rightarrow g_L \rightarrow ... \rightarrow goal \rightarrow$ **finish** $\rightarrow goal$ is not undermined by *plan*, so by clause (ii) of the definition of "result", g_L is a result of the sequence of actions prescribed by $s_1,...,s_n$.

For at least one L in $\mathcal{L}$, $k = n$. In that case, as *plan* is presumptively-sound, it has a causal-link of the form $s \rightarrow sg_1 \rightarrow ... \rightarrow sg_l \rightarrow s_n \rightarrow g_L \rightarrow ... \rightarrow goal$. As above, we can construct a plan of length $< n$ for sg_l, and as its causal-links are not undermined by *plan*, sg_l is a result of the sequence of actions prescribed by $s_1,...,s_n$. Then by clause (ii) of the definition of "result", g_L is a result of the sequence of actions prescribed by $s_1,...,s_n$.

Thus all of the g_L (for L in $\mathcal{L}$) are results of the sequence of actions prescribed by $s_1,...,s_n$. *goal*

is either the only g_L (if there is just one) or the conjunction of the g_L's. In the former case *goal* is a result of the sequence of actions prescribed by $s_1,...,s_n$, and in the latter case, by clause (iv) of the definition of "result", *goal* is a result of the sequence of actions prescribed by $s_1,...,s_n$. Hence *plan* is sound.

Theorem 1: If a plan is causally-sound then it is sound.
Proof: Suppose *plan* is causally-sound. By lemma 1a, all of its linearizations are causally-sound. By lemma 1b, all of its linearizations are sound. So by the definition of soundness for partial-order plans, *plan* is sound.

Theorem 2:

> If a goal is a result of an action-sequence $\langle A_1,...,A_n\rangle$, then there is a causally-sound plan for that goal some linearization of which prescribes the actions $A_1,...,A_n$ in that order.

Proof by induction on the length of the action sequence. Just let the plan be linear, and put in causal-links each time a planning-conditional is used to get a result.

Theorem 3:

> A plan-step n of a *plan* undermines a causal-link $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$ of *plan* iff there is a presumptively-sound embellishment $plan_0$ of *plan* whose goal is either $\sim subgoal_1$ or the negation of a conjunct of $subgoal_1$, and (1) n is the single penultimate plan-step of $plan_0$, (2) there is a linearization of *plan* consistent with the ordering imposed by $plan_0$ in which n occurs between n_1 and n_2, and (3) $plan_0$ is sound.

Proof: From right-to-left is immediate by clause (iii) of (R1). Conversely, suppose n undermines $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$. By definition, there is a linearization of plan in which n occurs between n_1 and n_2 and $\sim g$ is a result of the sequence of actions prescribed by the plan-steps $*start*,...,n$ in the linearization, where g is either $subgoal_1$ or a conjunct of $subgoal_1$. By theorem 2, there is a causally-sound plan for $\sim g$ having $start*,...,n$ as a linearization. Let $plan_0$ be the result of adding the rest of the steps of *plan* to the end of that linearization, in the order prescribed by *plan*. Those additional steps succeed n and so are not relevant to the soundness. Hence by theorem 1, $plan_0$ is sound, and by the definition of causal-soundness it is presumptively-sound.

Theorem 4:

> A plan-step n of a *plan* undermines a causal-link $n_1 \rightarrow subgoal_1 \rightarrow ... \rightarrow subgoal_n \rightarrow n_2 \rightarrow goal_1 \rightarrow ... \rightarrow goal_m$ of *plan* iff there is a presumptively-sound embellishment $plan_0$ of *plan* whose goal is either $\sim subgoal_1$ or the negation of a conjunct of $subgoal_1$, and (1) n is the single penultimate plan-step of $plan_0$, (2) there is a linearization of *plan* consistent with the ordering imposed by $plan_0$ in which n occurs between n_1 and n_2, and (3) $plan_0$ does not undermine any of its own causal-links.

Proof: From left-to-right, if $plan_0$ is sound, by theorem 2 it has a causally-sound linearization, so replace $plan_0$ by that linearization. From right-to-left, if $plan_0$ does not undermine any of its causal-links and it is presumptively-sound then it is causally-sound, so by theorem 1 it is sound.

Theorem 5:

> If the set of planning-conditionals is r.e. but not recursive, then the set of sound solution-pairs ⟨problem,solution⟩ is not r.e.

Proof: Suppose the set of planning-conditionals is r.e. and the set of ⟨problem,solution⟩ pairs is r.e. It will be shown that that it is then decidable whether a planning conditional $(C/A) \blacktriangleright G$ is in the set of planning-conditionals, and hence that set is recursive. To decide this, construct a planning problem by letting the start state consist of $\{C,\sim G\}$ and taking the goal to be $(\sim G \ \& \ G^*)$, and add the conditional $(C/A) \blacktriangleright G^*$ to the set of planning-conditionals. The resulting set of planning-conditionals is still r.e. but not recursive unless the original set was recursive. Consider the test plan diagrammed in figure nine. This plan is a solution to the planning problem iff plan-step 1 does not undermine the causal-link *start* → $\sim G$ → $(\sim G \ \& \ G^*)$ → **finish**, which in turn holds iff $(C/A) \blacktriangleright G$ is not in the set of planning-conditionals. Thus if the set of ⟨problem,solution⟩ pairs is r.e., we can discover that $(C/A) \blacktriangleright G$ is not in the set of planning-conditionals by running an algorithm generating the ⟨problem,solution⟩ pairs and waiting and seeing that the above plan is produced for this planning problem. Thus the complement of the set of planning-conditionals is r.e. By assumption, the set of planning-conditionals is r.e., so it follows that the set of planning-conditionals is recursive.

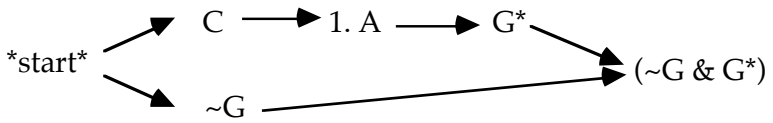

Figure 9. Test plan

Let us say that a presumptively-sound plan is *minimal* iff the removal of any causal-links or ordering-constraints will render it no longer presumptively-sound. Minimal causally-sound plans are defined analogously. The following three lemmas are obvious:

Lemma 6a:

OSCAR can find any minimal presumptively-sound plan by repeated uses of PROPOSE-NULL-PLAN, GOAL-REDUCTION, and SPLIT-CONJUNCTIVE-GOAL.

Lemma 6b:

OSCAR can find any minimal presumptively-sound embellishment by repeated-uses of EMBEDDED-NULL-PLAN, EMBEDDED-GOAL-REGRESSION, and SPLIT-EMBEDDED-CONJUNCTIVE-GOAL.

Lemma 6c:

OSCAR returns only presumptively-sound plans.

Theorem 6:

If OSCAR searches the space of potential inferences systematically then the OSCAR planner is sound.

Proof: Let the *proof-repair-depth* of a plan returned by the OSCAR planner be the number of undermining embellishments that OSCAR finds and repairs in the course of constructing the plan. We prove by induction on the proof-repair-depth of a plan that if OSCAR returns it then it is sound.

Suppose OSCAR returns a plan of proof-repair-depth 0. By lemma 6c it is presumptively-sound. By lemma 6b, there are no undermining embellishments, and so by theorem 1 the plan is sound.

Suppose the theorem holds for all plans of proof-repair-depth ≤ n, and OSCAR returns a plan of proof-repair-depth n+1. By lemma 6c, it is presumptively-sound. If it is not sound, by theorem 1, there must be a minimal undermining-embellishment. The embellishment is presumptively-sound, so by lemma 6b OSCAR will find it and try to repair it. The only way the plan can fail to be sound is if the repair is unsuccessful, i.e., OSCAR constructs an unsound repair plan. But the proof-repair-depth of the repair-plan will be ≤ n, so by the induction hypothesis, it cannot be unsound. Thus the plan must be sound.

Theorem 7:

If OSCAR searches the space of potential inferences systematically then the OSCAR planner is complete.

Proof: A minimal causally-sound plan contains a minimal presumptively-sound plan for the same goal. The causally-sound plan can then be generated from the minimal presumptively-sound plan by finding causal-underminings and repairing them. The repairs in question are either repairs to the presumptively-sound plan itself or repairs to repair-plans used to repair the presumptively-sound plan. Consider the set of all plans generated in this way, starting with the minimal presumptively-sound plan and ending with the minimal causally-sound plan, including the undermining embellishments, the repair-plans and their repair-plans, etc. For any plan in this set, define the *semantical-repair-depth* of the plan to be the minimal number of causal-underminings that must be repaired in constructing it out of the presumptively-sound plan. Then we prove by induction on the semantical-repair-depth that unless OSCAR's search terminates earlier by finding a different causally-sound plan for that goal, for every number n less than or equal to the semantical-repair-depth of the causally-sound plan, by repeated application of ADD-ORDERING-CONSTRAINT, CONFRONTATION, ADD-EMBEDDED-ORDERING-CONSTRAINT, and EMBEDDED-CONFRONTATION, OSCAR will find every plan in the set with semantical-repair-depth n, along with any undermining embellishments.

The only plan in the set with semantical-repair-depth 0 is the minimal presumptively-sound plan, and by lemma 5a, OSCAR can find it.

Suppose the theorem holds for all plans in the set of semantical-repair-depth ≤ n, and consider a plan of semantical-repair-depth n+1. This plan is constructed by repairing a subplan of semantical-repair-depth n, either by adding an ordering-constraint so that the undermining embellishment of the subplan is not an embellishment of the plan or by adding a repair-plan that undermines the undermining embellishment. By the induction hypothesis, OSCAR can find the undermined subplan, the undermining embellishment, and the repair-plan, so OSCAR can find the plan of semantical-repair-depth n+1.

It follows that OSCAR can find any minimal causally-sound plan regardless of its semantical-repair-depth. If there is a sound plan for the goal, by theorem 2 there is a causally-sound plan, and then there is a minimal causally-sound plan. So OSCAR is complete.

Theorem 8:

A plan *plan* collectively-undermines its causal-link L iff there is an embellishment *plan*+ of *plan* and a set $\mathcal{L}$ of causal-links of *plan*+ such that *plan*+ presumptively-undermines $\mathcal{L}$, and (1) $L \in \mathcal{L}$, (2) *plan*+ is causally-sound, and (3) there is no embellishment $plan_0$ of *plan*+ and set $\mathcal{L}^*$ of causal-links of *plan*+ such that (a) $plan_0$ presumptively-undermines $\mathcal{L}^*$, (b) $\mathcal{L}^* \subseteq \mathcal{L}$, (c) $L \notin \mathcal{L}^*$, and (d) $plan_0$ is causally-sound.

Proof: Analogous to theorem 2.

Theorem 9:

A plan *plan* collectively-undermines its causal-link L iff there is a presumptively-sound embellishment *plan+* of *plan* and a set $\mathcal{L}$ of causal-links of *plan+* such that *plan+* presumptively-undermines $\mathcal{L}$, and (1) $L \in \mathcal{L}$, (2) there is no presumptively-sound embellishment $plan_0$ of *plan+* and set $\mathcal{L}^*$ of causal-links of *plan+* such that (a) $plan_0$ presumptively-undermines $\mathcal{L}^*$, (b) $\mathcal{L}^* \subseteq \mathcal{L}$, (c) $L \notin \mathcal{L}^*$, and (d) $plan_0$ does not collectively-undermine any of its causal-links.

Proof: Analogous to theorem 3.

References

Allen, James
1987 "Formal models of planning". In *Readings in Planning*, ed. Allen, J., Hendler, J., and Tate, A. Morgan Kaufmann, Los Altos, Calif.

Barrett, A., and Weld, D.
1994 "Partial order planning: evaluating possible efficiency gains", *Artificial Intelligence* **67**, 71-112.

Blum, A., and M. L. Furst
1995 "Fast planning through planning graph analysis". In *Proceedings of the Fourteenth International Joint Conference on AI,*, 1636-42.
1997 "Fast planning through planning graph analysis", *Artificial Intelligence* **90**, 281-300.

Brafman, R. I., H. Hoos, and C. Boutilier
1998 "LPSP: a linear plan-level stochastic planner", unpublished, available from http://www.cs.ubc.ca/spider/cebly/craig.html.

Chapman, D.
1987 "Planning for conjunctive goals", *Artificial Intelligence* **32**, 333-377.

Erol, K., D. S. Nau, and V. S. Subrahmanian
1975 "Complexity, decidability and undecidability results for domain-independent planning", *Artificial Intelligence* **76**, 75-88.

Etzioni, O., and S. Hanks, D. Weld, D. Draper, N. Lesh, and M. Williamson
1992 "An approach to planning with incomplete information", *Proceedings of the Third International Conference on Knowledge Representation and Reasoning*, 115-125.

Ferguson, George, and James Allen
1994 "Arguing about plans: plan representation and reasoning for mixed-initiative planning", *Proceedings of the Second International Conference on AI Planning Systems.*

Fikes, R. E., and Nilsson, N. J.
1971 "STRIPS: a new approach to the application of theorem proving to problem solving", *Artificial Intelligence* **2** 189-208.

Gerevini, A., and L. Schubert
1996 "Accelerating partial-order planners: some techniques for effective search control and pruning", *Journal of Artificial Intelligence Research* **5**, 95-137.

Ginsberg, Matt
1995 "Approximate planning", *Artificial Intelligence* **76**, 89-124.

Goldman, R. P., and Boddy, M. S.
1994 "Conditional linear planning", *Proceedings of the Second International Conference on Artificial Intelligence Planning Systems*, 80-85.

Goodman, Nelson
1955 *Fact, Fiction, and Forecast*. Cambridge: Harvard University Press.
Hanks, Steve, and McDermott, Drew
1986 "Default reasoning, nonmonotonic logics, and the Frame Problem", *Proceedings of AAAI-86*.
1987 "Nonmonotonic logic and temporal projection", *Artificial Intelligence* **33**, 379-412.
Joslin, David, and Martha Pollack
1994 "Least-cost flaw repair: a plan-refinement strategy for partial-order planning", *Proceedings AAAI-94*, 1004-1009.
1995 "Passive and active decision postponement in plan generation", *Proceedings of the Third European Workshop and Planning*.
Kautz, H. A.
1986 "The logic of persistence", *Proceedings of AAAI-86*, 401-405.
Kautz, H. A., and B. Selman
1996 "Pushing the envelope: planning, propositional logic, and stochastic search", *Proceedings of AAAI-96*, 1194-1201.
Kushmerick, N., Hanks, S., and Weld, D.
1995 "An algorithm for probabilistic planning". *Artificial Intelligence* **76**, 239-286.
Lifschitz, Vladimir
1987 "Formal theories of action", in *The Frame Problem in Artificial Intelligence: Proceedings of the 1987 Workshop*, ed. F. Brown. Morgan Kaufmann, Los Altos, Calif.
1987a "On the semantics of STRIPS", in *Reasoning about Actions and Plans*, ed. Georgeff, M., and Lansky, A. Morgan Kaufmann, Los Altos, Calif.
McAllester, David, and Rosenblitt, David
1991 "Systematic nonlinear planning", *Proceedings of AAAI-91*, 634-639.
McCarthy, John, and Hayes, Patrick
1969 "Some philosophical problems from the standpoint of artificial intelligence". In B. Metzer & D. Michie (eds.), *Machine Intelligence 4*. Edinburgh: Edinburgh University Press.
Myers, Karen, and David Smith
1988 "The persistence of derived information", *Proceedings of AAAI-88*, 496-500.
Pednault, Edwin P.
1987 Toward a Mathematical Theory of Plan Synthesis, PhD dissertation, Stanford University.
Penberthy, J. Scott, and Weld, Daniel
1992 "UCPOP: a sound, complete, partial order planner for ADL". *Proceedings 3rd International Conference on Principles of Knowledge Representation and Reasoning*, 103-114.
Peot, M. A., and D. Smith
1992 "Conditional nonlinear planning", *Proceedings of the First International Conference on Artificial Intelligence Planning Systems*, 189-197.
1993 "Threat removal strategies for partial-order planning", *Proceedings AAAI-93*, 492-499.
Pollack, Martha, David Joslin, and Massimo Paolucci
1997 "Flaw selection strategies for partial-ordering planning", *Journal of Artificial Intelligence Research* **6**, 223-262.
Pollock, John
1972 "The logic of projectibility", *Philosophy of Science* **39**, 302-314.
1990 *Nomic Probability and the Foundations of Induction*, Oxford University Press.
1995 *Cognitive Carpentry*, MIT Press.
1996 "Reason in a changing world", in *Practical Reasoning*, ed. Dov M. Gabbay and Hans Jürgen Ohlbach, Springer, 495-509. This can be downloaded from http://www.u.arizona.edu/~pollock/.
1997 "Reasoning about change and persistence: a solution to the frame problem", *Nous* **31**, 143-169.

1998 "Reasoning defeasibly about plans". Technical report of the OSCAR Project. This can be downloaded from http://www.u.arizona.edu/~pollock/.
1998a "Perceiving and reasoning about a changing world", *Computational Intelligence*. **14** (1998), 498-562.
Pryor, Louise, and Gregg Collins
1996 "Planning for contingencies: a decision-based approach", *Journal of Artificial Intelligence Research* **4**, 287-339.
Shoham, Yoav
1986 *Time and Causation from the standpoint of artificial intelligence*, Computer Science Research Report No. 507, Yale University, New Haven, CT.
1987 *Reasoning about Change*, MIT Press.
Srinivasan, R., and A. E. Howe
1995 "Comparison of methods for improving search efficiency in a partial-order planner", *Proceedings IJCA-95*, 162--1626.
Stalker, Douglas
1994 *Grue! The New Riddle of Induction*, ed, Open Court.
Warren, D.
1976 "Generating conditional plans and programs", in *Proceedings of the Summer Conference on Artificial Intelligence and the Simulation of Behavior*, 344-354.
Weld, Daniel
1994 "An introduction to least commitment planning", *AI Magazine*, **15** 27-62.
Williamson, M., and S. Hanks
1996 "Flaw selection strategies for value-directed planning", *Proceedings of the Third International Conference on Artificial Intelligence Planning Systems*, 237-244.