VII
PERCEIVING AND REASONING ABOUT A CHANGING WORLD

Previous chapters have detailed the construction of the OSCAR architecture. This chapter begins the application of that architecture to concrete problems in agent building. These are problems that must be faced in order to build a rational agent that is able to get around in a realistically complex world, but can be solved largely by using the OSCAR architecture as a tool rather than by fiddling with the architecture itself. By and large, such problems will be addressed by constructing reason-schemas that will enable OSCAR to reason about various aspects of the world.

A traditional problem of philosophical epistemology is that of explaining how it is possible for human beings to acquire knowledge of the external world. Essentially the same problem arises for artificial rational agents. The designers of such agents must provide them with procedures for accessing the world and forming reliable beliefs about it. Some knowledge may be built in, but in a complex changing environment, an agent cannot be equipped from its inception with all the information it needs. It must be capable of gathering new information by sensing its surroundings. This is perception, in a generic sense. All of an agent's knowledge of the world must be inferred from perception and background knowledge. The problems that an agent designer faces are essentially similar to those faced by the philosopher addressing the problem of our knowledge of the external world. These problems are at least threefold. First, perception need not be veridical—the world can be other than it appears. Second, perception is really a form of sampling. An agent cannot perceptually monitor the entire state of the world at all time. The best perception can do is provide the agent with images of small parts of the world at discrete times or over short time intervals. Perception provides momentary snapshots of scattered nooks and crannies at disparate times, and it is up to the agent's cognitive faculties to make inferences from these to a coherent picture of the world. Third, the world changes. The agent must be able to make inferences that enable it to keep track of an evolving world. For this it must be able to reason about both persistence and change, using knowledge of causal processes. Building an artificial agent that is able to perform these cognitive feats is no less difficult than solving the philosophical problem of our knowledge of the external world. In fact, the best way to solve the engineering problem is most likely to figure out how humans perform these tasks and then build AI systems that work similarly. This paper makes a start at providing this kind of analysis. The analysis is based upon decades of work in philosophical epistemology. The procedures that will be proposed are reason-schemas for defeasible reasoning. They have been implemented using the system of defeasible

reasoning that is incorporated into the OSCAR architecture for rational agents.[1] Along the way, solutions will be proposed for the Frame Problem, the Qualification Problem, the Ramification Problem, and the Yale Shooting Problem.

1. Reasoning from Percepts

An agent must be capable of gathering new information by sensing its surroundings. This is perception, in a generic sense. Perception is a process that begins with the stimulation of sensors, and ends with beliefs about the agent's immediate surroundings. In artificial agents, this should be understood sufficiently broadly to include the input of information by a human operator. It is useful to draw a line between the last non-doxastic (non-belief) states in this process and the first beliefs. The production, in human beings, of the non-doxastic states is the subject of psychology and neuroscience. In AI it is the subject of research in machine vision. The reasoning from the beliefs is studied partly by epistemology and partly by psychology.[2] What is at issue here is the theory of the interface between the non-doxastic states and the beliefs.

I will refer to the final non-doxastic states from which beliefs are obtained in perception as percepts. Two mechanisms are possible for moving from percepts to beliefs. On the one hand, an agent could implement a purely automatic process whereby percepts give rise to beliefs automatically, and reasoning begins with the beliefs thus generated. This has the consequence that the beliefs thus produced are not inferred from anything, and hence are not rationally correctable. This has been a favorite view of philosophers regarding the nature of human cognition. But perception need not be veridical, and humans can discover that particular percepts are not accurate representations of the world, so the beliefs that are the automatic progeny of percepts cannot be beliefs about the world as such—they must be beliefs about the perceiver's sensory input. On this view, beliefs about physical objects (tables, chairs, people, plants, buildings, etc.) are inferred from beliefs about sensory input. I have attacked this view of human perception elsewhere.[3] The basic observation to be made about human beings is that when we perceive our surroundings, the resulting beliefs are usually beliefs about physical objects (tables, chairs, people, plants, buildings, etc.), and not beliefs about our own inner perceptual experiences. We can focus our attention on our perceptual experiences, forming beliefs about them by introspection, but that requires an explicit change of attention. On the other hand, because we can perform such a change of

---

[1] This architecture is detailed in Pollock (1995) and (1995a).

[2] Just as in linguistics, there is a competence/performance distinction to be made in the study of human reasoning. Epistemology constructs normative theories of competence, and psychology studies human performance. For more on this, see Pollock (1995).

[3] Pollock (1987) and (1995).

attention, we can evaluate the quality of the inference from those experiences to the beliefs about physical objects to which they give rise. This suggests that we should regard our reasoning as beginning from the percepts themselves, and not from beliefs about our percepts. Some philosophers have objected to this view on the grounds that reasoning is, by definition, a process of making transitions from beliefs to beliefs, and hence percepts cannot enter into reasoning. But this is just a verbal quibble. I have urged that the structure and evaluation of the transitions from percepts to beliefs about physical objects is sufficiently inference-like to warrant calling them inferences.[4] I propose to further substantiate that claim here by giving a precise description of the inferences and implementing them within OSCAR for incorporation into an artificial agent.

The preceding observations are just about human beings, but there are lessons to be drawn from them about rational agents in general. I see no reason why we couldn't build rational agents by having percepts automatically give rise to beliefs about percepts, and having all reasoning begin from those beliefs. But that seems like needless duplication. If, as I will argue below, we can construct a system of perceptual reasoning that begins directly from the percepts, then beliefs about the percepts will in most cases be otiose and their production will be a needless burden on cognitive resources.

Accordingly, I will take the basic inference in perceptual reasoning to be from a percept to a conclusion about the world. This enables us to assign propositional contents to percepts. The content of a percept will be taken to be the same as the content of the belief for which the percept provides a reason. But of course, the percept is not the same thing as the belief. Having the percept consists of having a perceptual experience that is as if the belief were true. Given this understanding of the content of percepts, we can, as a first approximation, formulate the reasoning from percepts to beliefs as follows:

(1) Having a percept with the content P is a defeasible reason for the agent to believe P.[5]

In this principle, the variable 'P' ranges over the possible contents of percepts. That range will depend upon the perceptual apparatus of the agent in question.

---

[4] See Pollock (1987).

[5] This is based upon proposals in my (1967), (1971), (1974), (1987), and (1995).

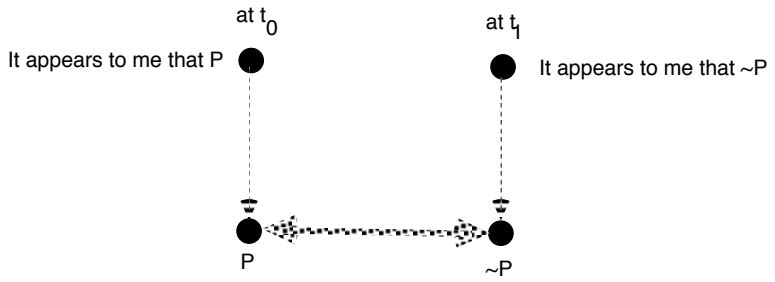

Figure 1. Perceptual updating.

This formulation of the reasoning captures the obvious but important point that perceptual reasoning must be defeasible—appearances can be deceptive. However, that this formulation is not entirely adequate becomes apparent when we consider perceptual updating. The world changes, and accordingly percepts produced at different times can support inferences to conflicting conclusions. We can diagram this roughly as in figure 1, where t_1 is a later time than t_0, '⇢' symbolizes defeasible inference, and '⇢' symbolizes defeat relations. The reasoning seems to produce a case of collective defeat—the inferences to P and ~P defeat each other. But this should not be a case of collective defeat. The initial percept supports the belief that P holds at t_0, and the second percept supports the belief that P does not hold at t_1. These conclusions do not conflict. We can hold both beliefs simply by acknowledging that the world has changed.

We can accommodate this by building temporal reference into the belief produced by perception, and giving the percept a date. This allows us to reformulate the above defeasible reason as follows:

PERCEPTION
Having a percept at time t with the content P is a defeasible reason for the agent to believe P-at-t.

We can then redraw the diagram of the reasoning as in figure 2, with the result that the apparent conflict has gone away.

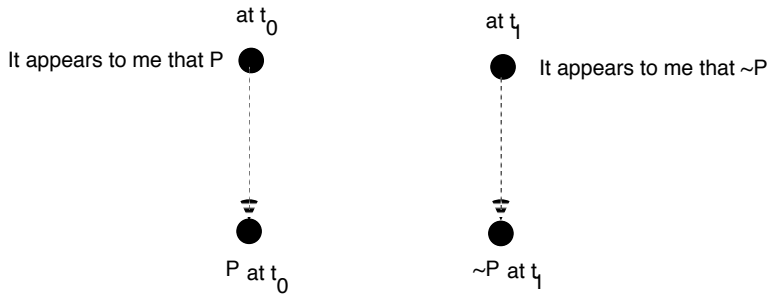

Figure 2. Perceptual updating revised.

There is a large literature on how, exactly, to build temporal reference into beliefs. This includes the literature on the situation calculus, temporal logic, possible worlds, etc. However, for present purposes, nothing very complicated is required. I will simply attach a term designating the time to the formula. No assumptions will be made about time other than that it is linearly ordered. More complex kinds of temporal reasoning may require a more sophisticated treatment, but I presume that nothing in this paper will be incompatible with such a treatment.

2. Perceptual Reliability

When giving an account of a species of defeasible reasoning, it is as important to characterize the defeaters for the defeasible reasons as it is to state the reasons themselves. This paper assumes the theory of defeasible reasons and reasoning implemented in OSCAR and described in detail in Pollock (1995) and (1995a). One of the central doctrines of that theory is that there are just two kinds of defeaters—rebutting defeaters and undercutting defeaters. Any reason for denying P-at-t is a rebutting defeater for PERCEPTION. An undercutting defeater for an inference from a belief in P to a belief in Q attacks the connection between P and Q rather than merely denying the conclusion. An undercutting defeater is a reason for the formula (P ⊗ Q) (read "It is false that P would not be true unless Q were true", or abbreviated as "P does not guarantee Q"). For an inference from a percept, an undercutting defeater will have the analogous form "(It appears to me that P-at-t) ⊗ Q)". The only obvious undercutting defeater for PERCEPTION is a reliability defeater, which is of a general sort applicable to all defeasible reasons. Reliability defeaters result from observing that the inference from P to Q is not, under the present circumstances, reliable. To make this precise it is necessary to understand how reason-strengths work in OSCAR. Some reasons are better than others. In OSCAR, reason-strengths range from 0 to 1. Reason-strengths are calibrated by comparing them with the statistical syllogism. According to the statistical syllogism, when $r > 0.5$, "Bc & prob(A/B) = r" is a defeasible reason for "Ac", the strength of the reason being a function of r.[6] A reason of strength r is taken to have the same strength as an instance of the statistical syllogism from a probability of $2(r - 0.5)$, thus mapping the interval [.5,1] onto the interval [0,1]. The inference rule PERCEPTION will have some strength r, although this may vary from agent to agent. The value of r should correspond roughly to the reliability of an agent's system of perceptual input in the circumstances in which it normally functions. A reason-strength r corresponds to a

[6] This is a slight oversimplification. See my (1990) for a detailed discussion of the statistical syllogism.

probability 0.5(r + 1). PERCEPTUAL-RELIABILITY constitutes a defeater by informing us that under the present circumstances, perception is not as reliable as it is normally assumed to be:

PERCEPTUAL-RELIABILITY
Where R is projectible, r is the strength of PERCEPTION , and $s < 0.5(r + 1)$, "R-at-t, and the probability is less than or equal to s of P's being true given R and that I have a percept with content P" is a conclusive undercutting defeater for PERCEPTION as a reason of strength $\geq$ r.

The projectibility constraint in this principle is a perplexing one. To illustrate its need, suppose I have a percept of a red object, and am in improbable but irrelevant circumstances of some type C_1. For instance, C_1 might consist of my having been born in the first second of the first minute of the first hour of the first year of the twentieth century. Let C_2 be circumstances consisting of wearing rose-colored glasses. When I am wearing rose-colored glasses, the probability is not particularly high that an object is red just because it looks red, so if I were in circumstances of type C_2, that would quite properly be a reliability defeater for a judgment that there is a red object before me. However, if I am in circumstances of type C_1 but not of C_2, there should be no reliability defeater. The difficulty is that if I am in circumstances of type C_1, then I am also in the disjunctive circumstances (C_1 / C_2). Furthermore, the probability of being in circumstances of type C_2 given that one is in circumstances of type (C_1 / C_2) is very high, so the probability is not high that an object is red given that it looks red to me but I am in circumstances (C_1 / C_2). Consequently, if (C_1 / C_2) were allowed as an instantiation of R in PERCEPTUAL-RELIABILITY, being in circumstances of type C_1 would suffice to indirectly defeat the perceptual judgment.

The preceding examples show that the set of circumstance-types appropriate for use in PERCEPTUAL-RELIABILITY is not closed under disjunction. This is a general characteristic of projectibility constraints. The need for a projectibility constraint in induction is familiar to most philosophers (although unrecognized in many other fields).[7] I showed in Pollock (1990) that the same constraint occurs throughout probabilistic reasoning, and the constraint on induction can be regarded as derivative from a constraint on the statistical syllogism.[8] However, similar constraints occur in other contexts and do not appear to be derivative from the constraints on the statistical syllogism. The constraint on reliability defeaters is one example of this, and another example will be given below. Unfortunately, at this time there is no generally acceptable theory of projectibility. The term "projectible" serves more as the label for a problem than as an indication of the solution to the problem.

PERCEPTUAL-RELIABILITY constitutes a defeater by informing us that under the

---

[7] The need for the projectibility constraint on induction was first noted by Goodman (1955).

[8] The material on projectibility in my (1990) has been collected into a paper and reprinted in my (1994).

present circumstances, perception is not as reliable as it is normally assumed to be. Notice, however, that this should not prevent our drawing conclusions with a weaker level of justification. The probability recorded in PERCEPTUAL-RELIABILITY should function merely to weaken the strength of the perceptual inference rather than completely blocking it. This can be accomplished by supplementing PERCEPTION with the following rule:

DISCOUNTED-PERCEPTION
Where R is projectible, r is the strength of PERCEPTION , and $0.5 < s < 0.5(r + 1)$, having a percept at time t with the content P and the belief "R-at-t, and the probability is less than s of P's being true given R and that I have a percept with content P" is a defeasible reason of strength $2(s – 0.5)$ for the agent to believe P-at-t.

DISCOUNTED-PERCEPTION must be defeasible in the same way PERCEPTION is:

PERCEPTUAL-UNRELIABILITY
Where A is projectible and $s^* < s$, "A-at-t, and the probability is less than or equal to s^* of P's being true given A and that I have a percept with content P" is a conclusive defeater for DISCOUNTED-PERCEPTION.

In a particular situation, the agent may know that a number of facts hold each of which is sufficient to lower the reliability of perception. The preceding principles have the consequence that the only undefeated inference from the percept will be that made in accordance with the weakest instance of DISCOUNTED-PERCEPTION.[9]

3. Implementation

This paper is about certain kinds of problems that arise in reasoning about a changing world. The problems are specifically about reasoning, and thus are to be solved by formulating a theory of such reasoning. The test of such a theory is that it produces the intuitively correct reasoning when applied to concrete examples. The theory proposed here for reasoning about change and persistrence is formulated within the framework provided by the OSCAR theory of defeasible reasoning. It is important to distinguish between the theory of defeasible reasoning that underlies OSCAR and the implementation. The theory of defeasible reasoning is a theory about the general structure of defeasibility, how it relies upon defeasible reasons and defeaters, and how defeat statuses are computed on the basis of a set of interacting arguments some of

[9] In such a case, we might also know that the reliability of perception on the combination of facts is higher than it is on the individual facts (interfering considerations might cancel out). In that case, we should be able to make an inference from the percept to its content in accordance with that higher probability. However, that inference can be made straightforwardly using the statistical syllogism, and does not require any further principles specifically about perception.

which defeat others. This general theory has been described in detail elsewhere (see particulary my (1995)), and I will not repeat it here. The abstract reasoning schemas formulated in this paper are formulated within the OSCAR formalism, but could probably be reformulated in many other formalisms as well (e.g., default logic). The implementation that I will describe next can be viewed as making the abstract proposals more precise, and making them more objectively testable by making it an objective matter of fact what consequences they have for specific examples.

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. Conclusions are stored as nodes in the inference-graph (inference-nodes).

Reasoning proceeds in terms of reasons. Backwards-reasons are used in reasoning backwards, and forwards-reasons are used in reasoning forwards. Forwards-reasons are data-structures with the following fields:

- reason-name.
- forwards-premises — a list of forwards-premises.
- backwards-premises — a list of backwards-premises.
- reason-conclusion — a formula.
- defeasible-rule — t if the reason is a defeasible reason, nil otherwise.
- reason-variables — variables used in pattern-matching to find instances of the reason-premises.
- reason-strength — a real number between 0 and 1, or an expression containing some of the reason-variables and evaluating to a number.
- reason-description — an optional string describing the reason.

Forwards-premises are data-structures encoding the following information:

- fp-formula — a formula.
- fp-kind — :inference, :percept, or :desire (the default is :inference)
- fp-condition — an optional constraint that must be satisfied by an inference-node for it to instantiate this premise.
- clue? — explained below.

Similarly, backwards-premises are data-structures encoding the following information:

- bp-formula
- bp-kind

The use of the premise-kind is to check whether the formula from which a forwards inference proceeds represents a desire, percept, or the result of an inference. The contents of percepts, desires, and inferences are all encoded as formulas, but the inferences that can be made from them depend upon which kind of item they are. For

example, we reason quite differently from the desire that x be red, the percept of x's being red, and the conclusion that x is red.

Backwards-reasons will be data-structures encoding the following information:

- reason-name.
- forwards-premises.
- backwards-premises.
- reason-conclusion — a formula.
- reason-variables — variables used in pattern-matching to find instances of the reason-premises.
- strength — a real number between 0 and 1, or an expression containing some of the reason-variables and evaluating to a number.
- defeasible-rule — t if the reason is a defeasible reason, nil otherwise.
- reason-condition — a condition that must be satisfied by an interest before the reason is deployed.

Simple forwards-reasons have no backwards-premises, and simple backwards-reasons have no forwards-premises. Given inference-nodes that instantiate the premises of a simple forwards-reason, the reasoner infers the corresponding instance of the conclusion. Similarly, given an interest that instantiates the conclusion of a simple backwards-reason, the reasoner adopts interest in the corresponding instances of the backwards-premises. Given inference-nodes that discharge those interests, an inference is made to the conclusion from those inference-nodes.

In deductive reasoning, with the exception of a rule of reductio ad absurdum, we are unlikely to encounter any but simple forwards- and backwards-reasons.[10] However, the use of backwards-premises in forwards-reasons and the use of forwards-premises in backwards-reasons provides an invaluable form of control over the way reasoning progresses. This will be illustrated at length below. Mixed forwards- and backwards-reasons are those having both forwards- and backwards-premises. Given inference-nodes that instantiate the forwards-premises of a mixed forwards-reason, the reasoner does not immediately infer the conclusion. Instead the reasoner adopts interest in the corresponding instances of the backwards-premises, and an inference is made only when those interests are discharged. Similarly, given an interest instantiating the conclusion of a mixed backwards-reason, interests are not immediately adopted in the backwards-premises. Interests in the backwards-premises are adopted only when inference-nodes are constructed that instantiate the forwards-premises.

There can also be degenerate backwards-reasons that have only forwards-premises. In a degenerate backwards-reason, given an interest instantiating the conclusion, the reasoner then becomes "sensitive to" inference-nodes instantiating the forwards-premises, but does not adopt interest in them (and thereby actively search for arguments to establish them). If appropriate inference-nodes are produced by other reasoning, then an inference is made to the conclusion. Degenerate backwards-reasons are thus much

---

[10] This is discussed at greater length in Chapter Two of my (1995a).

like simple forwards-reasons, except that the conclusion is only drawn if there is an interest in it.

Reasons are most easily defined in OSCAR using the macros def-forwards-reason and def-backwards-reason:

```
(def-forwards-reason symbol
   :forwards-premises list of formulas optionally interspersed with expressions of the form (:kind ...) or
   (:condition ...)
   :backwards-premises  list of formulas optionally interspersed with expressions of the form (:kind ...) or
   (:condition ...)
  :conclusion  formula
   :strength number or a an expression containing some of the reason-variables and evaluating to a number.
   :variables list of symbols
   :defeasible? T or NIL (NIL is the default)
   :description an optional string (quoted) describing the reason)

(def-backwards-reason symbol
  :conclusion  list of formulas
   :forwards-premises list of formulas optionally interspersed with expressions of the form (:kind ...) or
   (:condition ...)
   :backwards-premises  list of formulas optionally interspersed with expressions of the form (:kind ...) or
   (:condition ...)
   :condition  this is a predicate applied to the binding produced by the target sequent
   :strength number or an expression containing some of the reason-variables and evaluating to a number.
   :variables list of symbols
   :defeasible? T or NIL (NIL is the default)
   :description an optional string (quoted) describing the reason)
```

Epistemic reasoning begins from contingent information input into the system in the form of percepts. Percepts are encoded as structures with the following fields:

- percept-content—a formula, without temporal reference built in.
- percept-clarity—a number between 0 and 1, indicating how strong a reason the percept provides for the conclusion of a perceptual inference.
- percept-date—a number.

When a new percept is presented to OSCAR, an inference-node of kind :percept is constructed, having a node-formula that is the percept-content of the percept (this includes the percept-date). This inference-node is then inserted into the inference-queue for processing.

Using the tools described above, we can implement PERCEPTION as a simple forwards-reason:

```
(def-forwards-reason perception
   :forwards-premises "(p at time)" (:kind :percept)
   :conclusion "(p at time)"
```

```
  :variables p time
  :defeasible? t
  :strength .98
  :description "When information is input, it is defeasibly reasonable to believe it.")
```

The strength of .98 has been chosen arbitrarily.

PERCEPTUAL-RELIABILITY was formulated as follows:

> PERCEPTUAL-RELIABILITY
> Where R is projectible, r is the strength of PERCEPTION , and $s < 0.5(r + 1)$, "R-at-t, and the probability is less than or equal to s of P's being true given R and that I have a percept with content P" is a conclusive defeater for PERCEPTION.

It seems clear that this should be treated as a backwards-reason. That is, given an interest in the undercutting defeater for PERCEPTION, this reason schema should be activated, but if the reasoner is not interested in the undercutting defeater, this reason schema should have no effect on the reasoner. However, treating this as a simple backwards-reason is impossible, because there are no constraints (other than projectibility) on R. We do not want interest in the undercutting defeater to lead to interest in every projectible R. Nor do we want the reasoner to spend its time trying to determine the reliability of perception given everything it happens to know about the situation. This can be avoided by making this a degenerate backwards-reason (no backwards-premises), taking R-at-t (where t is the time of the percept) and the probability premise to be forwards-premises. This suggests to the following definition:

```
(def-backwards-undercutter PERCEPTUAL-RELIABILITY
  :defeatee perception
  :forwards-premises
    "((the probability of p given ((I have a percept with content p) & R)) <= s)"
     (:condition  (and (s < 0.99) (projectible R)))
     "(R at time)"
  :variables p time R s
  :description "When perception is unreliable, it is not reasonable to accept its representations.")
```

(def-backwards-undercutter is a variant of def-backwards-reason that computes the reason-conclusion for us.) For now, I will take the projectible formulas to be any conjunctions of literals, although it must be recognized that this is simplistic and must ultimately be refined.

A problem remains for this implementation. PERCEPTUAL-RELIABILITY requires us to know that R is true at the time of the percept. We will typically know this only by inferring it from the fact that R was true earlier. The nature of this inference is the topic of the next section. Without this inference, it is not possible to give interesting illustrations of the implementation just described, so that will be postponed until section six.

4. Temporal Projection

The reason-schema PERCEPTION enables an agent to draw conclusions about its current surroundings on the basis of its current percepts. However, that is of little use unless the agent can also draw conclusions about its current surroundings on the basis of earlier (at least fairly recent) percepts. For instance, imagine a robot whose task is to visually check the readings of two meters and then press one of two buttons depending upon which reading is higher. This should not be a hard task, but if we assume that the robot can only look at one meter at a time, it will not be able to acquire the requisite information about the meters using only the reason-schema PERCEPTION. The robot can look at one meter and draw a conclusion about its value, but when the robot turns to read the other meter, it no longer has a percept of the first and so is no longer in a position to hold a justified belief about what that meter reads now. This is a reflection of the observation made at the beginning of the paper that perception samples bits and pieces of the world at disparate times, and an agent must be supplied with cognitive faculties enabling it to build a coherent picture of the world out of those bits and pieces. In the case of our robot, what is needed is some basis for believing that the first meter still reads what it read a moment ago. In other words, the robot must have some basis for regarding the meter reading as a stable property—one that tends not to change quickly over time.

It is natural to suppose that a rational agent, endowed with the ability to learn by induction, can discover inductively that some properties (like the meter readings) are stable to varying degrees, and can bring that knowledge to bear on tasks like the meter-reading task. However, I argued long ago (Pollock (1974)) that such inductive learning is not epistemically possible—it presupposes the very stability that is the object of the learning. The argument for this somewhat surprising conclusion is as follows. To say that a property is stable is to say that objects possessing it tend to retain it. To confirm this inductively, an agent would have to re-examine the same object at different times and determine whether the property has changed. The difficulty is that in order to do this, the agent must be able to reidentify the object as the same object at different times. Although this is a complex matter, it seems clear that the agent makes essential use of the perceptible properties of objects in reidentifying them. If all perceptible properties fluctuated wildly, we would be unable to reidentify anything. If objects tended to exchange their perceptible properties abruptly and unpredictably, we would be unable to tell which object was which.[11] The upshot of this is that it is epistemically impossible to investigate the stability of perceptible properties inductively without presupposing that most of them tend to be stable. If we make that general supposition, then we can use induction to refine it by discovering that some perceptible properties are more stable than others, that particular properties tend to be unstable under specifiable circumstances, etc. But our conceptual framework must include a general presumption

[11] A more detailed presentation of this argument can be found in chapter six of Pollock (1974).

of stability for perceptible properties before any of this refinement can take place. In other words, the built-in epistemic arsenal of a rational agent must include reason-schemas of the following sort for at least some choices of P:

(2) If $t_0 < t_1$, believing P-at-t_0 is a defeasible reason for the agent to believe P-at-t_1.

Principle (2) amounts to a presumption that P's being true is a stable property of a time (i.e., a stable fluent, to use the jargon of the situation calculus). 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. Let ρ be the probability that P will hold at time t+1 given that it holds at time t. It is shown in appendix 2 that, assuming independence, it follows that the probability that P will hold at time $(t+\Delta t)$ given that it holds at time t is $\frac{1}{2}(2\rho - 1)^{\Delta t} + \frac{1}{2}$. In other words, the strength of the presumption that a stable property will continue to hold over time decays as the time interval increases. This is important for understanding the logic of reasoning about stable properties. To illustrate, consider what I will call the perceptual updating problem. Suppose an agent has a percept of P at time t_0, and a percept of ~P at a later time t_1. What an agent should conclude (defeasibly) under these circumstances is that the world has changed between t_0 and t_1, and although P was true at t_0, it is no longer true at t_1 and hence no longer true at a later time t_2. If we attempt to reconstruct this reasoning using principle (2), we do not seem to get the right answer. Principle (2) produces the inference-graph of figure 3, and it is a straightforward case of collective defeat. This is intuitively incorrect.

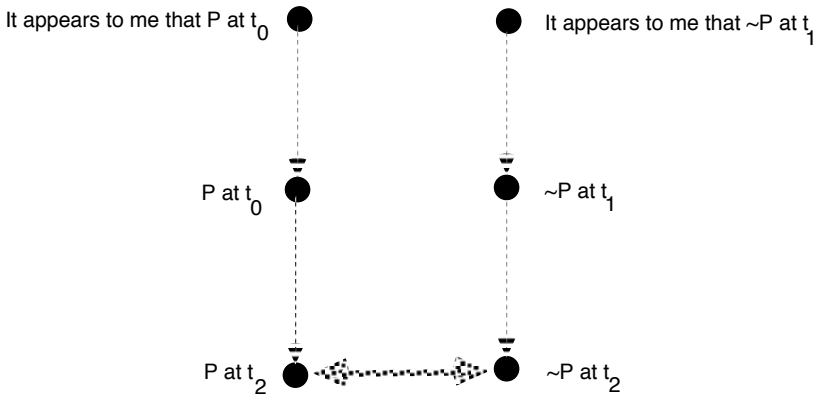

Figure 3. The perceptual updating problem.

The solution to getting the reasoning to come out right is to embrace the observation that the strength of the presumption that a stable property will continue to hold over time decays as the time interval increases, and build this into principle (2). The strength

of the reason provided by principle (2) must decrease as $(t_1 - t_0)$ increases. This will have the result that the support for ~P-at-t_2 is greater than the support for P-at-t_2, and hence the latter is defeated but the former is not. I propose then that we take the reason-strength of temporal projection to have the form $\rho^{\Delta t}$ where ρ is a constant I will call the temporal-decay factor. I have arbitrarily set ρ to .999 (OSCAR uses an arbitrary time-scale anyway—the only constraint is that reasoning should be completed in a reasonable amount of time). This produces the decay curve of figure 4.

A probability of $\frac{1}{2}(2\rho - 1)^{\Delta t} + \frac{1}{2}$ corresponds to a reason-strength of $(2\rho - 1)^{\Delta t}$. So the proposal is that we reformulate (1) as follows:

(3) Believing P-at-t is a defeasible reason of strength $(2\rho - 1)^{\Delta t}$ for the agent to believe P-at-$(t+\Delta t)$.

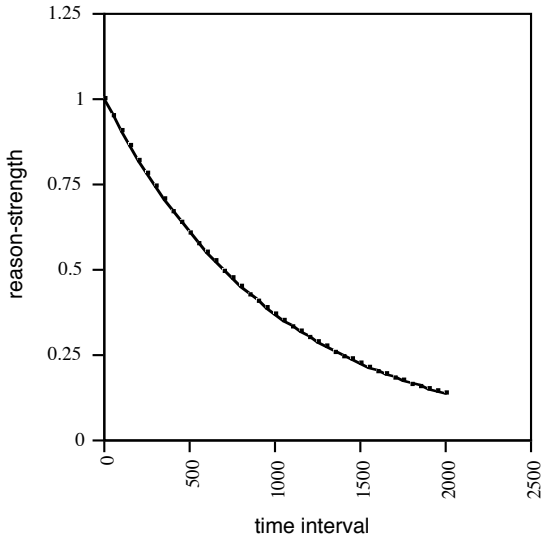

Figure 4. Decaying reason-strengths.

5. Temporal Projectibility

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. However, from P-at-t_0 we can deductively infer (PvQ)-at-t_0. Without any restrictions on the proposition-variable in TEMPORAL-PROJECTION, (P/Q)-at-t_0 gives us a defeasible reason for expecting (PvQ)-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 5, 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 (PvQ)-at-t_2. (PvQ) 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 (PvQ)-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, it does not work for disjunctions.

This appears to be a projectibility problem, analogous to that discussed above in connection with reliability defeaters. In temporal projection, the use of arbitrary disjunctions, and other non-projectible constructions, must be precluded. It is unclear precisely what the connection is between the projectibility constraint involved in temporal projection and that involved in induction, so I will refer to it neutrally as "temporal-projectibility". Notice that in temporal-unprojectibility, disjunctions are not the only culprits. The ascriptions of 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. On the other hand, there are "bivalent" properties, like "dead" and "alive" for which the negation of an ascription is projectible because it is equivalent to ascribing the other (temporally-projectible) property.

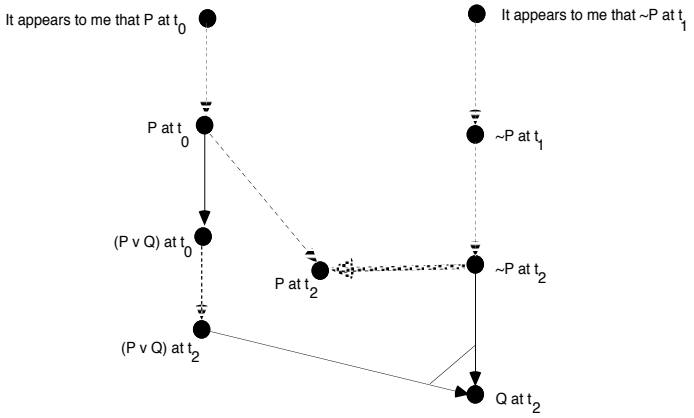

Figure 5. The need for a temporal projectibility constraint.

Using the concept of temporal-projectibility, temporal projection should be reformulated as follows:

TEMPORAL-PROJECTION
Believing P-at-t is a defeasible reason of strength $(2\rho - 1)^{\Delta t}$ for the agent to believe P-at-$(t+\Delta t)$.

It is unfortunate that this principle has to be formulated in terms of an unanalyzed concept of temporal-projectibility. This is a problem for the proposal. But notice that it must equally be a problem for any theory of temporal projection. Other authors have not been bothered by it simply because they overlooked it. One of the contributions of this paper is to call attention to this general problem.

TEMPORAL-PROJECTION is based on an a-priori presumption of stability for temporally-projectible properties. However, it must be possible to override or modify the presumption by discovering that the probability of P's being true at time t+1 given that P is true at time t is something other than the constant ρ. This requires the following defeater:

PROBABILISTIC-DEFEAT-FOR-TEMPORAL-PROJECTION
"The probability of P-at-(t+1) given P-at-t $\neq \rho$" is a conclusive undercutting defeater for TEMPORAL-PROJECTION.

If we know that the probability of P-at-(t+1) given P-at-t is σ where $\sigma \neq \rho$, we may still be able to project P forwards in time, but now the inference will be based upon the statistical syllogism and the known high probability rather than an a-priori principle of temporal projection.

6. Implementing Temporal Projection

In order to implement TEMPORAL-PROJECTION, we must have a test for the temporal-projectibility of formulas. This is a problem, because as indicated above, I do not have a theory of temporal-projectibility to propose. For present purposes, I will finesse this by assuming that atomic formulas, negations of atomic formulas whose predicates are on a list *bivalent-predicates*, and conjunctions of the above, are temporally-projectible. This will almost certainly be inadequate in the long run, but it will suffice for testing the proposed reason-schemas.

It seems clear that TEMPORAL-PROJECTION must be treated as a backwards-reason. That is, given some fact P-at-t, we do not want the reasoner to automatically infer P-at-$(t+\Delta t)$ for every one of the infinitely many times $\Delta t > 0$. An agent should only make such an inference when the conclusion is of interest. For the same reason, the premise P-at-t should be a forwards-premise rather than a backwards-premise—we do not want the reasoner adopting interest in P-at-$(t-\Delta t)$ for every $\Delta t > 0$. I propose to handle this by implementing it as a backwards reason. This will have the effect that when the reasoner

adopts interest in P-at-t, it will check to see whether it already has a conclusion of the form P-at-t_0 for $t_0 < t$, and if so it will infer P-at-t. This can be done in either of two ways. We could use a mixed-backwards-reason:

```
(def-backwards-reason TEMPORAL-PROJECTION
   :conclusion  "(p at time)"
   :condition  (and (temporally-projectible p) (numberp time))
   :forwards-premises
      "(p at time0)"
   :backwards-premises
       "(time0 < time)"
   :variables p time0 time
   :defeasible? T
   :strength  (expt (1- (* 2 *temporal-decay*)) (- time time0))
   :description
   "It is defeasibly reasonable to expect temporally projectible truths to remain unchanged.")
```

This requires the reasoner to engage in explicit arithmetical reasoning about whether (time0 < time). It is more efficient to make this a condition on the forwards-premise rather than an independent premise. This produces a degenerate backwards-reason:

```
(def-backwards-reason TEMPORAL-PROJECTION
   :conclusion  "(p at time)"
   :condition  (and (temporally-projectible p) (numberp time))
   :forwards-premises
      "(p at time0)"
      (:condition (time0 < time*)
   :variables p time0 time
   :defeasible? T
   :strength  (expt (1- (* 2 *temporal-decay*)) (- time time0))
   :description
   "It is defeasibly reasonable to expect temporally projectible truths to remain unchanged.")
```

PROBABILISTIC-DEFEAT-FOR-TEMPORAL-PROJECTION is implemented as a conclusive degenerate backwards-reason:

```
(def-backwards-undercutter PROBABILISTIC-DEFEAT-FOR-TEMPORAL-PROJECTION
  :defeatee  TEMPORAL-PROJECTION
   :forwards-premises
     "((the probability of (p at (t + 1)) given (p at t)) = s)"
       (:condition (not (s = *temporal-decay*)))
   :variables  p s time0 time)
```

To illustrate, consider the perceptual updating problem. OSCAR runs the problem as follows:

==

Problem number 6:[12] This is the perceptual updating problem. First, Fred looks red to me. Later, Fred looks blue to me. What should I conclude about the color of Fred?

Forwards-substantive-reasons:

perception

Backwards-substantive-reasons:

TEMPORAL-PROJECTION

incompatible-colors

Inputs:

(the color of Fred is red) : at cycle 1 with justification 1.0

(the color of Fred is blue) : at cycle 30 with justification 1.0

Ultimate epistemic interests:

(? x)((the color of Fred is x) at 50)   degree of interest = 0.5

==

THE FOLLOWING IS THE REASONING INVOLVED IN THE SOLUTION

Nodes marked DEFEATED have that status at the end of the reasoning.

```
                    # 1
                    interest: ((the color of Fred is y0) at 50)
                    This is of ultimate interest
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
It appears to me that ((the color of Fred is red) at 1)
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 # 1
 It appears to me that ((the color of Fred is red) at 1)
 # 2
 ((the color of Fred is red) at 1)
 Inferred by:
        support-link #1 from { 1 } by perception
 undefeated-degree-of-support = 0.98
 # 3
 ((the color of Fred is red) at 50)          DEFEATED
 undefeated-degree-of-support = 0.904
 Inferred by:
        support-link #2 from { 2 } by TEMPORAL-PROJECTION  defeaters: { 7 }   DEFEATED
 This discharges interest 1
                    # 5
                    interest: ~((the color of Fred is red) at 50)
```

[12] The problem number refers to the problem set that accompanies the LISP code for OSCAR and the reasoning described in this paper, all of which can be downloaded from http://www.u.arizona.edu/~pollock/.

```
                    Of interest as a defeater for support-link 2 for node 3
          ==========================================
          Justified belief in ((the color of Fred is red) at 50)
          with undefeated-degree-of-support 0.904
          answers #<Query 1: (? x)((the color of Fred is x) at 50)>
          ==========================================
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
It appears to me that ((the color of Fred is blue) at 30)
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 # 4
 It appears to me that ((the color of Fred is blue) at 30)
 # 5
 ((the color of Fred is blue) at 30)
 Inferred by:
           support-link #3 from { 4 } by perception
 undefeated-degree-of-support = 0.98
 # 6
 ((the color of Fred is blue) at 50)
 Inferred by:
           support-link #4 from { 5 } by TEMPORAL-PROJECTION  defeaters: { 8 }
 undefeated-degree-of-support = 0.960
 This discharges interest 1
                    # 9
                    interest: ~((the color of Fred is blue) at 50)
                    Of interest as a defeater for support-link 4 for node 6
          ==========================================
          Justified belief in ((the color of Fred is blue) at 50)
          with undefeated-degree-of-support 0.960
          answers #<Query 1: (? x)((the color of Fred is x) at 50)>
          ==========================================
 # 7
 ~((the color of Fred is red) at 50)
 Inferred by:
           support-link #5 from { 6 } by incompatible-colors
 undefeated-degree-of-support = 0.960
 defeatees: { link 2 for node 3 }
          vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
          #<Node 3> has become defeated.
          vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
          ==========================================
          Lowering the undefeated-degree-of-support of ((the color of Fred is red) at 50)
          retracts the previous answer to #<Query 1: (? x)((the color of Fred is x) at 50)>
          ==========================================
========== ULTIMATE EPISTEMIC INTERESTS ========
 Interest in (? x)((the color of Fred is x) at 50)
 is answered by node 6:  ((the color of Fred is blue) at 50)
---------------------------------------------------
```

Now let us return to the problem noted above for PERCEPTUAL-RELIABILITY. This is that we will typically know R-at-t only by inferring it from R-at-t_0 for some $t_0 < t$ (by TEMPORAL-PROJECTION). TEMPORAL-PROJECTION is a backwards-reason. That is, given some fact P-at-t, the reasoner only infers P-at-t* (for t* > t) when that conclusion is of interest. Unfortunately, in PERCEPTUAL-RELIABILITY, R-at-t is not an interest, and so it will not be inferred from R-at-t_0 by TEMPORAL-PROJECTION. This difficulty can be circumvented by formulating PERCEPTUAL-RELIABILITY with an extra forwards-premise R-at-t_0 which is marked as a clue, and a backwards-premise R-at-t:

```
(def-backwards-undercutter PERCEPTUAL-RELIABILITY
   :defeatee *perception*
   :forwards-premises
      "((the probability of p given ((I have a percept with content p) & R)) <= s)"
      (:condition (and (projectible R) (s < 0.99)))
      "(R at time0)"
      (:condition (time0 < time))
      (:clue? t)
   :backwards-premises "(R at time)"
   :variables p time R time0 s
   :description "When perception is unreliable, it is not reasonable to accept its representations.")
```

The difference between ordinary forwards-premises and clues is that when a clue is instantiated by a node, that node is not inserted into the basis for the inference. The function of clues is to guide the reasoning. Thus in an application of PERCEPTUAL-RELIABILITY, if R-at-t_0 is concluded, this suggests that R-at-t is true and leads to an interest in it, which can then be inferred from R-at-t_0 by TEMPORAL-PROJECTION. An example of such reasoning follows:

```
=====================================================Problem number 1:  Fred looks red to me.  However, I also know that my surroundings are illuminated by red light.  All along, I know that the probability is not high of Fred being red given that Fred looks red to me, but my surroundings are illuminated by red light.  What should I conclude about the color of Fred?

Forwards-substantive-reasons:
      perception

Backwards-substantive-reasons:
      PERCEPTUAL-RELIABILITY
      TEMPORAL-PROJECTION

Inputs:
      (The color of Fred is red) : at cycle 1 with justification 1.0

Given:
      (My surroundings are illuminated by red light (at 0))     justification = 1.0
```

```
        ((The probability of (The color of Fred is red) given
          ((I have a percept with content (The color of Fred is red)) & My surroundings are illuminated by red
light))
         <= 0.8)     justification = 1.0

Ultimate epistemic interests:
    (? x)((the color of Fred is x) at 1)    degree of interest = 0.75
===========================================================THE FOLLOWING IS THE
REASONING INVOLVED IN THE SOLUTION
Nodes marked DEFEATED have that status at the end of the reasoning.

 # 1
 (my surroundings are illuminated by red light at 0)
 given
 # 2
 ((the probability of (the color of Fred is red) given ((I have a percept with content (the color of Fred is red))
& my surroundings are illuminated by red light)) <= 0.8)
 given
                                # 1
                                interest: ((the color of Fred is y0) at 1)
                                This is of ultimate interest
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
It appears to me that ((the color of Fred is red) at 1)
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 # 3
 It appears to me that ((the color of Fred is red) at 1)
 # 4
 ((the color of Fred is red) at 1)                DEFEATED
 Inferred by:
            support-link #3 from { 3 } by perception  defeaters: { 6 }   DEFEATED
 This discharges interest 1
                                # 2
                                interest: (((it appears to me that (the color of Fred is red)) at 1) ⊗ ((the color of
                                  Fred is red) at 1))
                                Of interest as a defeater for support-link 3 for node 4
            ==========================================
            Justified belief in ((the color of Fred is red) at 1)
            answers #<Query 1: (? x)((the color of Fred is x) at 1)>
            ==========================================
                                # 4
                                interest: (my surroundings are illuminated by red light at 1)
                                For interest 2 by PERCEPTUAL-RELIABILITY
                                This interest is discharged by node 5
 # 5
 (my surroundings are illuminated by red light at 1)
```

```
  Inferred by:
          support-link #4 from { 1 } by TEMPORAL-PROJECTION
  This discharges interest 4
  # 6
  (((it appears to me that (the color of Fred is red)) at 1) ⊗ ((the color of Fred is red) at 1))
  Inferred by:
          support-link #5 from { 2 , 5 } by PERCEPTUAL-RELIABILITY using {1}
defeatees: { link 3 for node 4 }
  This node is inferred by discharging interest #2
          vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
          #<Node 4> has become defeated.
          vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
          ========================================
          Lowering the undefeated-degree-of-support of ((the color of Fred is red) at 1)
          retracts the previous answer to #<Query 1: (? x)((the color of Fred is x) at 1)>
          ========================================
========== ULTIMATE EPISTEMIC INTERESTS ========
  Interest in (? x)((the color of Fred is x) at 1)
  is unsatisfied.
--------------------------------------------------
```

Note that node 1 is not listed as a premise of the inference to node 6.

DISCOUNTED-PERCEPTION and PERCEPTUAL-UNRELIABILITY can be implemented similarly:

```
(def-forwards-reason DISCOUNTED-PERCEPTION
   :forwards-premises
   "((the probability of p given ((I have a percept with content p) & R)) <= s)"
      (:condition (and (projectible R) (0.5 < s) (s < 0.99)))
   "(p at time)"
     (:kind :percept)
   "(R at time0)"
      (:condition (time0 < time))
      (:clue? t)
   :backwards-premises "(R at time)"
   :conclusion "(p at time)"
   :variables p time R time0 s
   :strength  (2 * (s - 0.5))
   :defeasible? t
```

```
   :description "When information is input, it is defeasibly reasonable to believe it.")

(def-backwards-undercutter PERCEPTUAL-UNRELIABILITY
   :defeatee DISCOUNTED-PERCEPTION
   :forwards-premises
   "((the probability of p given ((I have a percept with content p) & A)) <= s*)"
     (:condition (and (projectible A) (s* < s)))
   "(A at time1)"
     (:condition (time1 <= time))
     (:clue? t)
   :backwards-premises "(A at time)"
   :variables p time R A time0 time1 s s*
   :defeasible? t
   :description "When perception is unreliable, it is not reasonable to accept its representations.")
```

These rules are illustrated by the following example:

```
=====================================================
Problem number 9:  This illustrates the use of discounted-perception and perceptual-unreliability.

Forwards-substantive-reasons:
        perception
        DISCOUNTED-PERCEPTION

Backwards-substantive-reasons:
        PERCEPTUAL-RELIABILITY
        PERCEPTUAL-UNRELIABILITY
        TEMPORAL-PROJECTION
        neg-at-intro

Inputs:
        (the color of Fred is red) : at cycle 10 with justification 1.0

Given:
        ((the probability of (the color of Fred is red) given ((I have a percept with content (the color of Fred is
red)) &
            my surroundings are illuminated by red light)) <= 0.7)  :  with justification = 1.0
        ((the probability of (the color of Fred is red) given ((I have a percept with content (the color of Fred is
red)) &
```

I am wearing red tinted glasses)) <= 0.8) : with justification = 1.0

(I am wearing red tinted glasses at 1) : at cycle 15 with justification = 1.0

(my surroundings are illuminated by red light at 1) : at cycle 30 with justification = 1.0

(~my surroundings are illuminated by red light at 8) : at cycle 50 with justification = 1.0

Ultimate epistemic interests:

((the color of Fred is red) at 10)   degree of interest = 0.5

===

THE FOLLOWING IS THE REASONING INVOLVED IN THE SOLUTION

Nodes marked DEFEATED have that status at the end of the reasoning.

1

((the probability of (the color of Fred is red) given ((I have a percept with content (the color of Fred is red)) & my surroundings are illuminated by red light)) <= 0.7)

given

2

((the probability of (the color of Fred is red) given ((I have a percept with content (the color of Fred is red)) & I am wearing red tinted glasses)) <= 0.8)

given

1

interest: ((the color of Fred is red) at 10)

This is of ultimate interest

||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||

It appears to me that ((the color of Fred is red) at 10)

||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||

3

It appears to me that ((the color of Fred is red) at 10)

4

((the color of Fred is red) at 10)

Inferred by:

support-link #3 from { 3 } by perception defeaters: { 7 }   DEFEATED

This node is inferred by discharging interest 1

2

interest: (((it appears to me that (the color of Fred is red)) at 10) ⊗ ((the color of Fred is red) at 10))

Of interest as a defeater for support-link 3 for node 4

==

Justified belief in ((the color of Fred is red) at 10)

with undefeated-degree-of-support 0.98

answers #<Query 1: ((the color of Fred is red) at 10)>

==========

5

(I am wearing red tinted glasses at 1)

given

5

interest: (I am wearing red tinted glasses at 10)

For interest 1 by DISCOUNTED-PERCEPTION

For interest 2 by PERCEPTUAL-RELIABILITY

This interest is discharged by node 6

6

(I am wearing red tinted glasses at 10)

Inferred by:

support-link #5 from { 5 } by TEMPORAL-PROJECTION

This discharges interest 5

4

((the color of Fred is red) at 10)

Inferred by:

support-link #6 from { 2 , 3 , 6 } by DISCOUNTED-PERCEPTION using {5} defeaters: { 10 }

support-link #3 from { 3 } by perception defeaters: { 7 } DEFEATED

This node is inferred by discharging interests (1 1)

8

interest: ((((the probability of (the color of Fred is red) given ((I have a percept with content (the color of Fred is red)) & I am wearing red tinted glasses)) <= 0.8) & (((it appears to me that (the color of Fred is red)) at 10) & (I am wearing red tinted glasses at 10))) ⊗ ((the color of Fred is red) at 10))

Of interest as a defeater for support-link 6 for node 4

7

(((it appears to me that (the color of Fred is red)) at 10) ⊗ ((the color of Fred is red) at 10))

Inferred by:

support-link #7 from { 2 , 6 } by PERCEPTUAL-RELIABILITY

defeatees: { link 3 for node 4 }

This node is inferred by discharging interests (2 2)

vvvvvvvvvvvvvvvvvvvvvvvvvvvvv

The undefeated-degree-of-support of #<Node 4> has decreased to 0.6

vvvvvvvvvvvvvvvvvvvvvvvvvvvvv

8

(my surroundings are illuminated by red light at 1)

given

12
interest: (my surroundings are illuminated by red light at 10)
For interest 1 by DISCOUNTED-PERCEPTION
For interest 2 by PERCEPTUAL-RELIABILITY
For interest 8 by PERCEPTUAL-UNRELIABILITY
This interest is discharged by node 9

9
(my surroundings are illuminated by red light at 10) DEFEATED
Inferred by:
support-link #9 from { 8 } by TEMPORAL-PROJECTION defeaters: { 13 } DEFEATED
This discharges interest 12

14
interest: ~(my surroundings are illuminated by red light at 10)
Of interest as a defeater for support-link 9 for node 9

4
((the color of Fred is red) at 10)
Inferred by:
support-link #10 from { 1 , 3 , 9 } by DISCOUNTED-PERCEPTION using {8} DEFEATED
support-link #6 from { 2 , 3 , 6 } by DISCOUNTED-PERCEPTION using {5} defeaters: { 10 }
support-link #3 from { 3 } by perception defeaters: { 7 } DEFEATED
This node is inferred by discharging interests (1 1)

7
(((it appears to me that (the color of Fred is red)) at 10) ⊗ ((the color of Fred is red) at 10))
Inferred by:
support-link #11 from { 1 , 9 } by PERCEPTUAL-RELIABILITY using {8} DEFEATED
support-link #7 from { 2 , 6 } by PERCEPTUAL-RELIABILITY
defeatees: { link 3 for node 4 }
This node is inferred by discharging interests (2 2)

10
((((the probability of (the color of Fred is red) given ((I have a percept with content (the color of Fred is red)) & I am wearing red tinted glasses)) <= 0.8) & (((it appears to me that (the color of Fred is red)) at 10) & (I am wearing red tinted glasses at 10))) ⊗ ((the color of Fred is red) at 10)) DEFEATED
Inferred by:
support-link #12 from { 1 , 9 } by PERCEPTUAL-UNRELIABILITY DEFEATED
defeatees: { link 6 for node 4 }
This node is inferred by discharging interest #8
vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
The undefeated-degree-of-support of #<Node 4> has decreased to 0.4
vvvvvvvvvvvvvvvvvvvvvvvvvvvvv

```
        ========================================
        Lowering the undefeated-degree-of-support of ((the color of Fred is red) at 10)
        retracts the previous answer to #<Query 1: ((the color of Fred is red) at 10)>
        ========================================
                        # 19
                        interest: (~my surroundings are illuminated by red light at 10)
                        For interest 14 by neg-at-intro
                        This interest is discharged by node 12
 # 11
 (~my surroundings are illuminated by red light at 8)
 given
 # 12
 (~my surroundings are illuminated by red light at 10)
 Inferred by:
          support-link #14 from { 11 } by TEMPORAL-PROJECTION  defeaters: { 14 }
 This discharges interest 19
                        # 21
                        interest: ~(~my surroundings are illuminated by red light at 10)
                        Of interest as a defeater for support-link 14 for node 12
 # 13
 ~(my surroundings are illuminated by red light at 10)
 Inferred by:
          support-link #15 from { 12 } by neg-at-intro
 defeatees: { link 9 for node 9 }
 This node is inferred by discharging interest #14
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
        The undefeated-degree-of-support of #<Node 4> has increased to 0.6
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
        #<Node 9> has become defeated.
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
        #<Node 10> has become defeated.
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
        ========================================
        Justified belief in ((the color of Fred is red) at 10)
        with undefeated-degree-of-support 0.6
        answers #<Query 1: ((the color of Fred is red) at 10)>
        ========================================
========== ULTIMATE EPISTEMIC INTERESTS ========
 Interest in ((the color of Fred is red) at 10)
 is answered affirmatively by node 4
```

---

7. Extending Temporal Projection

Sometimes we want to reason about something being true throughout an interval rather than at an instant. For example, given that Fred is red at 10, it is reasonable to conclude that for each time t between 20 and 30, Fred is red at t, and hence that Fred is red throughout the interval [20,30]. This conclusion can be expressed using quantifiers over time as:

$$(\forall t)[(20 \leq t \leq 30) \rightarrow ((\text{Fred is red}) \text{ at } t)].$$

Furthermore, if we ignore considerations of reason-strength, OSCAR can perform this reasoning using the existing principle of TEMPORAL-PROJECTION:

```
==========================================================
Forwards-substantive-reasons:

Backwards-substantive-reasons:
      TEMPORAL-PROJECTION+
      arithmetical-inequality
      inequality-transitivity

Inputs:

Given:
      ((Fred is red) at 10)  :  with justification = 1.0

Ultimate epistemic interests:
   (∀time)(((20 ≤ time) & (time ≤ 30)) ∅ ((Fred is red) at time))    degree of interest = 0.75
==========================================================
 # 1
 ((Fred is red) at 10)
 given
                              # 1
                              interest: (∀time)(((20 ≤ time) & (time ≤ 30)) ∅ ((Fred is red) at time))
                              This is of ultimate interest
                              # 2
                              interest: (((20 ≤ x0) & (x0 ≤ 30)) ∅ ((Fred is red) at x0))
                              For interest 1 by UG
                              This interest is discharged by node 8
 # 2
 ((20 ≤ x0) & (x0 ≤ 30))    supposition: { ((20 ≤ x0) & (x0 ≤ 30)) }
```

supposition
generated by interest 2
3
interest: ((Fred is red) at x0) supposition: { ((20 ≤ x0) & (x0 ≤ 30)) }
For interest 2 by conditionalization
This interest is discharged by node 7
3
(20 ≤ x0) supposition: { ((20 ≤ x0) & (x0 ≤ 30)) }
Inferred by:
support-link #2 from { 2 } by simp
4
interest: (10 ≤ x0) supposition: { ((20 ≤ x0) & (x0 ≤ 30)) }
For interest 3 by TEMPORAL-PROJECTION+
This interest is discharged by node 6
5
interest: (10 ≤ 20) supposition: { ((20 ≤ x0) & (x0 ≤ 30)) }
For interest 4 by inequality-transitivity
This interest is discharged by node 5
5
(10 ≤ 20)
Inferred by:
support-link #4 from { } by arithmetical-inequality
This discharges interest 5
6
(10 ≤ x0) supposition: { ((20 ≤ x0) & (x0 ≤ 30)) }
Inferred by:
support-link #5 from { 3 , 5 } by inequality-transitivity
This node is inferred by discharging interest #4
7
((Fred is red) at x0) supposition: { ((20 ≤ x0) & (x0 ≤ 30)) }
Inferred by:
support-link #6 from { 1 , 6 } by TEMPORAL-PROJECTION+
This node is inferred by discharging interest #3
8
(((20 ≤ x0) & (x0 ≤ 30)) ∅ ((Fred is red) at x0))
Inferred by:
support-link #7 from { 7 } by conditionalization
This node is inferred by discharging interest #2
9
(∀time)(((20 ≤ time) & (time ≤ 30)) ∅ ((Fred is red) at time))
Inferred by:
support-link #8 from { 8 } by UG
This node is inferred by discharging interest #1
==========
Justified belief in (∀time)(((20 ≤ time) & (time ≤ 30)) ∅ ((Fred is red) at time))
answers #<Query 1: (∀time)(((20 ≤ time) & (time ≤ 30)) ∅ ((Fred is red) at time))>

```
          ========================================
==================================================
```

However, in getting OSCAR to perform this reasoning, I have replaced TEMPORAL-PROJECTION by TEMPORAL-PROJECTION+, which is just like TEMPORAL-PROJECTION except that the reason-strength is left unspecified (and hence defaults to 1.0). OSCAR cannot perform this reasoning using TEMPORAL-PROJECTION unmodified, because the applicability of that principle requires that the times be numbers, whereas in this example OSCAR must reason about variable times. There is no way to modify TEMPORAL-PROJECTION to allow reasoning about variable times, because if the times are not specified as numbers then there is no way to compute an appropriate reason-strength.

In this example, it is clear what the strength of support should be for the conclusion. It should be the weakest support for any conclusion of the form ((Fred is red) at t) where $20 \leq t \leq 30$, and that in turn occurs when t = 30. Consequently, we can capture the appropriate form of this reasoning by adopting an analogue of TEMPORAL-PROJECTION for intervals:

```
(def-backwards-reason INTERVAL-PROJECTION
   :conclusion  "(p throughout (time* time))"
   :condition  (and (temporally-projectible p) (numberp time*) (numberp time) (<= time* time))
   :forwards-premises
   "(p at time0)"
      (:condition (time0 < time)
   :variables p time0 time* time
   :defeasible? T
   :strength  (expt (1- (* 2 *temporal-decay*)) (- time time0))
   :description
   "It is defeasibly reasonable to expect temporally projectible truths to remain unchanged.")
```

OSCAR can then reason trivially as follows:

```
==================================================
 # 1
 ((Fred is red) at 10)
 given
                          # 1
                          interest: ((Fred is red) throughout (20 30))
                          This is of ultimate interest
 # 2
 ((Fred is red) throughout (20 30))
 Inferred by:
          support-link #2 from { 1 } by INTERVAL-PROJECTION
 This discharges interest 1
==================================================
```

We can simplify things still further by noting that TEMPORAL-PROJECTION can be regarded as a special case of INTERVAL-PROJECTION. Using an idea of Shoham (1987), we can take (P at t) to mean (P throughtout [t , t]). Then we can dispense with INTERVAL-PROJECTION and redefine TEMPORAL-PROJECTION as we defined INTERVAL-PROJECTION above.

A further generalization is desirable. INTERVAL-PROJECTION projects a conclusion throughout a closed interval. There will be cases in which we want to project a conclusion throughout an open interval (an interval of the form (x , y)) or a clopen interval (of the form (x , y]) rather than a closed interval. As TEMPORAL-PROJECTION is an backwards-reason, we can have it license inferences to any of these conclusions. To accomplish this, let us symbolize open intervals as (open x y), closed intervals as (closed x y), and clopen intervals as (clopen x y). These will be printed as "(x , y)", "[x , y]", and "(x , y]", respectively. When I want to refer to an interval without specifying whether it is open, clopen, or closed, I will write it in the form "⟨x , y⟩". Then we can redefine TEMPORAL-PROJECTION as follows:

TEMPORAL-PROJECTION
If P is temporally-projectible and $t < t^* \leq t^{**}$ then believing P-at-t is a defeasible reason of strength $(2\rho - 1)^{(t^{**} - t)}$ for the agent to believe P-throughout-⟨t* , t**⟩.

and impliment it as follows:

```
(def-backwards-reason TEMPORAL-PROJECTION
   :conclusion  "(p throughout (op time* time))"
   :condition  (and (temporally-projectible p) (numberp time*) (numberp time) (<= time* time)
                    (or (eq op 'open) (eq op 'closed) (eq op 'clopen)))
   :forwards-premises
      "(p at time0)"
      (:condition (time0 < time)
   :variables p time0 time* time op
   :defeasible? T
   :strength  (expt (1- (* 2 *temporal-decay*)) (- time time0))
   :description
   "It is defeasibly reasonable to expect temporally projectible truths to remain unchanged.")
```

8. Temporal Indexicals

An agent that did all of its temporal reasoning using the reason-schemas described above would be led into crippling computational complexities. Every time the agent wanted to reuse a belief about its surroundings, it would have to reinfer it for the present time. Inference takes time, so by the time it had reinferred the belief, other beliefs with which the agent might want to combine this belief in further inference would themselves no longer be current. To get around this difficulty, the agent would have to make

inferences about some time in the near future rather than the present, inferring a number of properties of that time, and then combine those properties to make a further inference about that time, and finally project that new property into the future for use in further inference. This would not be computationally impossible, but it would make life difficult for an agent that had to reason in this way.

Human beings achieve the same result in a more efficient way by employing the temporal indexical "now". Rather than repeatedly reinferring a property as time advances, they infer it once as holding now, and that single belief is retained until it becomes defeated. The mental representation (i.e., formula) believed remains unchanged as time advances, but the content of the belief changes continuously in the sense that at each instant it is a belief about that instant. This has the effect of continuously updating the agent's beliefs in accordance with TEMPORAL-PROJECTION, but no actual reasoning need occur.

The use of "now" can be implicit or explicit. That is, we can either write "x is red" or "x is red now". The latter is used primarily for emphasis. The representation is simpler if we drop the temporal reference rather than putting in the "now", so that is the course that will be followed below.

Percepts are always percepts of the agent's present situation, so we can regard them as providing defeasible reasons for inferences about the present:

INDEXICAL-PERCEPTION
Having a percept at time t with the content P is a defeasible reason for the agent to believe P.

INDEXICAL-PERCEPTION is defeated by considerations of reliability just as PERCEPTION is:

INDEXICAL-PERCEPTUAL-RELIABILITY
Where R is projectible, r is the strength of INDEXICAL-PERCEPTION, "R-at-t, and the probability is less than $0.5(r + 1)$ of P's being true given R and that I have a percept with content P at t" is a defeasible undercutting defeater for INDEXICAL-PERCEPTION.

The conclusion P is automatically projected into the future just by retaining it, so INDEXICAL-PERCEPTION can be viewed as combining PERCEPTION and TEMPORAL-PROJECTION into a single reason-scheme. The projectibility constraint on temporal projection has not been included here, on the assumption that only temporally-projectible propositions can be the contents of percepts.

Because INDEXICAL-PERCEPTION builds in an application of TEMPORAL-PROJECTION, it must be defeated by the same considerations that defeat TEMPORAL-PROJECTION:

PROBABILISTIC-DEFEAT-FOR-INDEXICAL-PERCEPTION
"The probability of P-at-(t+1) given P-at-t $\neq \rho$" is a conclusive undercutting defeater for INDEXICAL-PERCEPTION.

A complication arises for the reasoner in dealing with conclusions containing an

implicit or explicit "now" representing an implicit use of TEMPORAL-PROJECTION. The degree of support for a conclusion inferred by TEMPORAL-PROJECTION decreases as the time interval increases. Retaining a conclusion containing "now" is equivalent to making an inference by TEMPORAL-PROJECTION. Accordingly, the degree of support for that conclusion must decay over time, just as if it were being continually re-inferred by TEMPORAL-PROJECTION. To handle this, we must make a distinction between temporal and atemporal conclusions, where the former are those containing an explicit or implicit "now". Atemporal conclusions have fixed degrees-of-support. The strength of a temporal conclusion inferred from an atemporal conclusion is initially fixed by the strength α of the argument supporting it, but as time passes the value of $(2\rho - 1)^{\Delta t}$ decays, and when the latter value becomes less than α, the strength of the conclusion decays with it. On the other hand, if the temporal conclusion is inferred from an earlier temporal conclusion of strength $(2\rho - 1)^{x}$, the strength of the new conclusion will start out at $(2\rho - 1)^{x}$, and continue to decay as $(2\rho - 1)^{x+\Delta t}$. To handle this, inference-nodes are marked "temporal" or "atemporal", and their construction times are stored with them.

To implement this, reasons are given a new field temporal? that determines whether the application of the reason produces a temporal conclusion. This allows us to implement INDEXICAL-PERCEPTION as follows:

```
(def-forwards-reason INDEXICAL-PERCEPTION
   :forwards-premises "(p at time)"
        (:kind :percept)
   :conclusion "p"
   :variables p time
   :strength  (max .98 (expt (1- (* 2 *temporal-decay*)) (- now time)))
   :defeasible? t
   :temporal? t
   :description  "When information is input, it is defeasibly reasonable to believe it.")
```

In effect, INDEXICAL-PERCEPTION combines an application of PERCEPTION and an application of TEMPORAL-PROJECTION. INDEXICAL-PERCEPTUAL-RELIABILITY defeats INDEXICAL-PERCEPTION by defeating the imbedded application of PERCEPTION. However, unlike PERCEPTION, the strength of INDEXICAL-PERCEPTION decays as the time interval increases. The ability of INDEXICAL-PERCEPTUAL-RELIABILITY to defeat an application of INDEXICAL-PERCEPTION should not increase as the time interval increases, so the strength of INDEXICAL-PERCEPTUAL-RELIABILITY must also decay:

```
(def-backwards-undercutter INDEXICAL-PERCEPTUAL-RELIABILITY
   :defeatee *indexical-perception*
   :forwards-premises
   "((the probability of p given ((I have a percept with content p) & R)) <= s)"
     (:condition (and (projectible R) (s < 0.99)))
   "(R at time0)"
      (:condition (time0 < time)
```

```
    (:clue? t)
  :backwards-premises "(R at time)"
  :variables p time R time0 s
  :description "When perception is unreliable, it is not reasonable to accept its representations.")
```

Here is an example that combines PERCEPTION, INDEXICAL-PERCEPTION, and INDEXICAL-PERCEPTUAL-RELIABILITY (and also an undiscussed principle about reliable testimony):

```
==================================================
Problem number 8:  First, Fred looks red to me.  Later, I am informed by Merrill that I am then wearing blue-tinted glasses.  Later still, Fred looks blue to me.  All along, I know that the probability is not high of Fred being blue given that Fred looks blue to me but I am wearing blue-tinted glasses.  What should I conclude about the color of Fred?

Forwards-substantive-reasons:
        indexical-perception
        perception
        reliable-informant

Backwards-substantive-reasons:
        indexical-perceptual-reliability
        PERCEPTUAL-RELIABILITY
        TEMPORAL-PROJECTION
        indexical-incompatible-colors

Inputs:
        (the color of Fred is red) : at cycle 1 with justification 0.8
        (Merrill reports that I am wearing blue tinted glasses) : at cycle 20 with justification 1.0
        (the color of Fred is blue) : at cycle 30 with justification 0.8

Given:
        ((the probability of (the color of Fred is blue) given ((I have a percept with content (the color of Fred is blue)) & I am wearing blue tinted glasses)) <= 0.8)  :  with justification = 1.0
        (Merrill is a reliable informant)  :  with justification = 1.0

Ultimate epistemic interests:
    (? x)(the color of Fred is x)    degree of interest = 0.65
==================================================
THE FOLLOWING IS THE REASONING INVOLVED IN THE SOLUTION
Nodes marked DEFEATED have that status at the end of the reasoning.

  # 1
  ((the probability of (the color of Fred is blue) given ((I have a percept with content (the color of Fred is blue)) & I am wearing blue tinted glasses)) <= 0.8)
  given
  undefeated-degree-of-support = 1.0 at cycle 1.
```

```
 # 2
 (Merrill is a reliable informant)
 given
 undefeated-degree-of-support = 1.0 at cycle 1.
                                        # 1
                                        interest: (the color of Fred is y0)
                                        This is of ultimate interest
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
It appears to me that ((the color of Fred is red) at 1)
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 # 3
 It appears to me that ((the color of Fred is red) at 1)
 # 5
 (the color of Fred is red)
 Inferred by:
               support-link #4 from { 3 } by indexical-perception  defeaters: { 13 }
 undefeated-degree-of-support = 0.8 at cycle 2.
 This discharges interest 1
                                        # 5
                                        interest: ~(the color of Fred is red)
                                        Of interest as a defeater for support-link 4 for node 5
               =========================================
               Justified belief in (the color of Fred is red)
               with undefeated-degree-of-support 0.8
               answers #<Query 1: (? x)(the color of Fred is x)>
               =========================================
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
It appears to me that ((Merrill reports that I am wearing blue tinted glasses) at 20)
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 # 6
 It appears to me that ((Merrill reports that I am wearing blue tinted glasses) at 20)
 # 7
 ((Merrill reports that I am wearing blue tinted glasses) at 20)
 Inferred by:
               support-link #5 from { 6 } by perception
 undefeated-degree-of-support = 0.99 at cycle 20.
 # 9
 (I am wearing blue tinted glasses at 20)
 Inferred by:
               support-link #7 from { 2 , 7 } by reliable-informant
 undefeated-degree-of-support = 0.98 at cycle 22.
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
It appears to me that ((the color of Fred is blue) at 30)
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 # 10
 It appears to me that ((the color of Fred is blue) at 30)
                                        # 13
```

interest: (I am wearing blue tinted glasses at 30)
For interest 15 by indexical-perceptual-reliability
This interest is discharged by node 15

12
(the color of Fred is blue) DEFEATED
Inferred by:
support-link #9 from { 10 } by indexical-perception defeaters: { 16 , 14 } DEFEATED
undefeated-degree-of-support = 0.8 at cycle 30.
This discharges interest 1
15
interest: (((it appears to me that (the color of Fred is blue)) at 30) ⊗ (the color of Fred is blue))
Of interest as a defeater for support-link 9 for node 12
16
interest: ~(the color of Fred is blue)
Of interest as a defeater for support-link 9 for node 12

==========
Justified belief in (the color of Fred is blue)
with undefeated-degree-of-support 0.8
answers #<Query 1: (? x)(the color of Fred is x)>
==========

13
~(the color of Fred is red) DEFEATED
Inferred by:
support-link #10 from { 12 } by indexical-incompatible-colors DEFEATED
undefeated-degree-of-support = 0.7982 at cycle 31.
defeatees: { link 4 for node 5 }
vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
#<Node 5> has become defeated.
vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
==========
Lowering the undefeated-degree-of-support of (the color of Fred is red)
retracts the previous answer to #<Query 1: (? x)(the color of Fred is x)>
==========

14
~(the color of Fred is blue)
Inferred by:
support-link #11 from { 5 } by inversion from contradictory nodes 13 and 5
defeatees: { link 9 for node 12 }
undefeated-degree-of-support = 0.0 at cycle 31.
vvvvvvvvvvvvvvvvvvvvvvvvvvvvv

```
        #<Node 14> has become defeated.
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
 # 15
 (I am wearing blue tinted glasses at 30)
 Inferred by:
        support-link #12 from { 9 } by TEMPORAL-PROJECTION
 undefeated-degree-of-support = 0.98 at cycle 34.
 This discharges interest 13
 # 16
 (((it appears to me that (the color of Fred is blue)) at 30) ⊗ (the color of Fred is blue))
 Inferred by:
        support-link #13 from { 1 , 15 } by indexical-perceptual-reliability using {9}
 defeatees: { link 9 for node 12 }
 undefeated-degree-of-support = 0.98 at cycle 34.
 This node is inferred by discharging interest #15
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
        The undefeated-degree-of-support of #<Node 14> has increased to 0.773
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
        The undefeated-degree-of-support of #<Node 5> has increased to 0.773
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
        #<Node 12> has become defeated.
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
        #<Node 13> has become defeated.
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
        ========================================
        Justified belief in (the color of Fred is red)
        with undefeated-degree-of-support 0.773
        answers #<Query 1: (? x)(the color of Fred is x)>
        ========================================
        ========================================
        Lowering the undefeated-degree-of-support of (the color of Fred is blue)
        retracts the previous answer to #<Query 1: (? x)(the color of Fred is x)>
        ========================================
========== ULTIMATE EPISTEMIC INTERESTS ========
 Interest in (? x)(the color of Fred is x)
 is answered by node 5:  (the color of Fred is red)
--------------------------------------------------
```

It may be illuminating to draw the inference-graph for this latter problem, as in figure 6.

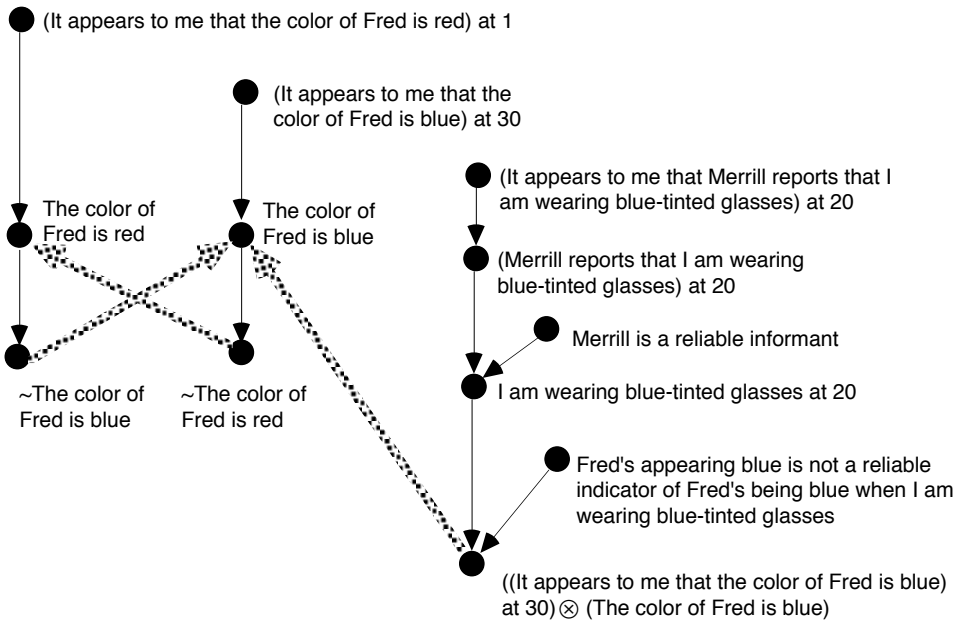

Figure 6. Inference graph

The use of temporal indexicals calls for indexical versions of TEMPORAL-PROJECTION, DISCOUNTED-PERCEPTION, and PERCEPTUAL-UNRELIABILITY:

INDEXICAL-TEMPORAL-PROJECTION
If P is temporally-projectible and t < now then believing P-at-t is a defeasible reason of strength $(2\rho - 1)^{(\text{now} - t)}$ for the agent to believe P.

DISCOUNTED-INDEXICAL-PERCEPTION
Where R is projectible, r is the strength of INDEXICAL-PERCEPTION , $0.5 < s < 0.5(r + 1)$, having a percept at time t with the content P and the belief "R-at-t, and the probability is less than s of P's being true given R and that I have a percept with content P" is a defeasible reason of strength $\max\{2(s - 0.5), (2\rho - 1)^{(\text{now} - t)}\}$ for the agent to believe P-at-t.

INDEXICAL-PERCEPTUAL-UNRELIABILITY
Where A is projectible and $s^* < s$, then "A-at-t, and the probability is less than or equal to s^* of P's being true given A and that I have a percept with content P" is a conclusive undercutting defeater for DISCOUNTED-INDEXICAL-PERCEPTION.

INDEXICAL-TEMPORAL-PROJECTION must be defeated by the same considerations that

defeat TEMPORAL-PROJECTION:

> PROBABILISTIC-DEFEAT-FOR-INDEXICAL-TEMPORAL-PROJECTION
> "The probability of P-at-(t+1) given P-at-t ≠ ρ" is a conclusive undercutting defeater for INDEXICAL-TEMPORAL-PROJECTION.

These are implemented as follows:

```
(def-backwards-reason INDEXICAL-TEMPORAL-PROJECTION
   :conclusion  "p"
   :forwards-premises
   "(p at time0)"
   (:condition  (time0 < now)
   :condition  (and (temporally-projectible p) (not (occur 'at p)))
   :variables p time0
   :defeasible? T
   :temporal? T
   :strength   (expt (1- (* 2 *temporal-decay*)) (- now time0))
   :description
   "It is defeasibly reasonable to expect temporally projectible truths to remain unchanged.")

(def-backwards-undercutter PROBABILISTIC-DEFEAT-FOR-INDEXICAL-TEMPORAL-
PROJECTION
  :defeatee  INDEXICAL-TEMPORAL-PROJECTION
   :forwards-premises
     "((the probability of (p at (t + 1)) given (p at t)) = s)"
       (:condition (not (s = *temporal-decay*)))
   :variables  p s time0 time)

(def-forwards-reason DISCOUNTED-INDEXICAL-PERCEPTION
   :forwards-premises
   "((the probability of p given ((I have a percept with content p) & R)) <= s)"
      (:condition (and (projectible R) (0.5 < s) (s < 0.995)))
   "(p at time)"
     (:kind :percept)
   "(R at time0)"
      (:condition  (time0 < now)
      (:clue? t)
   :backwards-premises "(R at time)"
   :conclusion "p"
   :variables p R time0 time s
   :strength (max (* 2 (s - 0.5) (expt *temporal-decay*)) (- *cycle* time)))
   :defeasible? t
   :temporal? t
   :description "When information is input, it is defeasibly reasonable to believe it.")

(def-backwards-undercutter INDEXICAL-PERCEPTUAL-UNRELIABILITY
```

```
:defeatee DISCOUNTED-INDEXICAL-PERCEPTION
:forwards-premises
"((the probability of p given ((I have a percept with content p) & A)) <= s*)"
   (:condition (and (projectible A) (s* < s)))
"(A at time1)"
   (:condition (time1 <= now)
   (:clue? t)
:backwards-premises "(A at time)"
:variables p time R A time0 time1 s s*
:description "When perception is unreliable, it is not reasonable to accept its representations.")
```

9. Reasoning about Change

Reasoning about what will change if an action is performed or some other change occurs often presupposes knowing what will not change. Early attempts to model such reasoning deductively proceeded by adopting a large number of "frame axioms", which were axioms to the effect that if something occurs then something else will not change. For instance, in a blocks world one of the frame axioms might be "If a block is moved, its color will not change". It soon became apparent that complicated situations required more frame axioms than axioms about change, and most of the system resources were being occupied by proofs that various properties did not change. In a realistically complicated situation, this became unmanageable. What became known as the Frame Problem is the problem of reorganizing reasoning about change so that reasoning about non-change can be done efficiently (McCarthy and Hayes (1969); Janlert, (1987)).

AI hackers, as Hayes (1987) calls them, avoided this problem by adopting the "sleeping dog strategy" (Haugeland (1987)). Starting with STRIPS, actual planning systems maintained databases of what was true in a situation, and with each possible action they stored lists of what changes those actions would produce. For planning systems intended to operate only in narrowly circumscribed situations, this approach is effective, although for general-purpose planning it quickly becomes unwieldy. In the attempt to provide a more general theory that justifies this approach as a special case, several authors (Sandewall (1972), McDermott (1982), McCarthy (1986)) proposed reasoning about change defeasibly and adopting some sort of defeasible inference scheme to the effect that it is reasonable to believe that something doesn't change unless you are forced to conclude otherwise. But to make the idea work, one needs both a precise framework for defeasible reasoning and a precise formulation of the requisite defeasible inference schemes. That proved to be a difficult problem.

The temporal projection principles defended in sections five and seven can be regarded as a precise formulation of the defeasible inference schemes sought. Unfortunately, these principles do not solve the Frame Problem. Steve Hanks and Drew McDermott (1986) were the first to observe that even with defeasible principles of non-change, a reasoner will often be unable to determine what changes and what does not. They illustrated this with what has become known as "the Yale shooting problem". The

general form of the problem is this. Suppose we have a causal law to the effect that if P is true at a time t and action A is performed at that time, then Q will be true shortly thereafter. (More generally, A could be anything that becomes true at a certain time. What is significant about actions is that they are changes.) Suppose we know that P is true now, and Q false. What should we conclude about the results of performing action A in the immediate future? Hanks and McDermott illustrate this by taking P to be "The gun is loaded, in working condition, and pointed at Jones", Q to be "Jones is dead", and A to be the action of pulling the trigger. We suppose (simplistically) that there is a causal law dictating that if the trigger is pulled on a loaded gun that is in working condition and pointed at someone, that person will shortly be dead. Under these circumstances, it seems clear that we should conclude that Jones will be dead shortly after the trigger is pulled.

The difficulty is that all we can infer from what we are given is that when A is performed either P will no longer be true or Q will be true shortly thereafter. Intuitively, we want to conclude (at least defeasibly) that P will remain true at the time A is performed and Q will therefore become true shortly thereafter. But none of our current machinery enables us to distinguish between P and Q. Because P is now true and Q is now false, we have a defeasible reason for believing that P will still be true when A is performed, and we have a defeasible reason for believing that Q will still be false shortly thereafter. This is diagrammed in figure 7. We know that one of these defeasible conclusions will be false, but we have no basis for choosing between them, so this becomes a case of collective defeat. That, however, is the intuitively wrong answer.

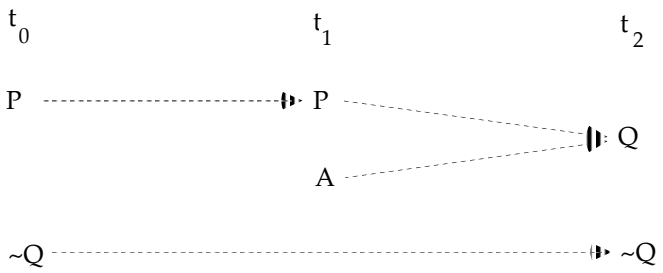

Figure 7. The Yale Shooting Problem

There is a kind of consensus that the solution to this problem lies in performing the TEMPORAL-PROJECTIONs in temporal order.[13] We first use TEMPORAL-PROJECTION to infer that the gun is still loaded 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

[13] See Hanks and McDermott (1987). A number of more recent papers explore this same idea.

we can infer that Jones will be dead at t_2. At time t_2, we can also try to use TEMPORAL-PROJECTION to infer that Jones will be alive, but this time something has already happened (the gun was fired) to block the projection, and so we do not infer that Jones will be alive. This general idea was first suggested by Shoham (1986,1987), and subsequently endorsed by Hanks and McDermott (1987), Lifschitz (1987), Gelfond and Lifschitz (1993), and others.[14] 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 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 a set of TEMPORAL-PROJECTIONs not all of which can be correct, but the inconsistency does not result simply from a causal connection. The shooting case is causal, but the stolen car case is not.

When we reason about causal mechanisms, we think of the world as "unfolding" temporally, and changes only occur when they are forced to occur by what has already happened. Thus when causal mechanisms force there to be a change, we conclude defeasibly that the change occurs in the later states rather than the earlier states. This seems to be part of what we mean by describing something as a causal mechanism. Causal mechanisms are systems that force changes, where "force" is to be understood in terms of temporal unfolding.[15]

When reasoning about such a causal system, part of the force of describing it as causal must be that the defeasible presumption against the effect occurring is somehow removed. Thus, although we normally expect Jones to remain alive, we do not expect this any longer when he is shot. To remove a defeasible presumption is to defeat it. This

---

[14] A related idea underlies my analysis of counterfactual conditionals in Pollock (1979) and (1984).

[15] This intuition is reminiscent of Shoham's (1987) "logic of chronological ignorance", although unlike Shoham, I propose to capture the intuition without modifying the structure of the system of defeasible reasoning.

suggests that there is some kind of general "causal" defeater for the temporal projection principles adumbrated above. The problem is to state this defeater precisely. As a first approximation we might try:

(4) For every $\varepsilon \geq 0$ and $\delta > 0$, "A&P-at-$(t+\varepsilon)$ & (A&P causes Q)" is an undercutting defeater for the defeasible inference from ~Q-at-t to ~Q-at-$(t+\varepsilon+\delta)$ by TEMPORAL-PROJECTION.

The temporal-unfolding view of causal reasoning requires causation to be temporally asymmetric. That is, "A&P causes Q" means, in part, that if A&P becomes true then Q will shortly become true. This precludes simultaneous causation, in which Q is caused to be true at t by A&P being true at t, because in such a case temporal ordering would provide no basis for preferring the temporal projection of P over that of ~Q. This may seem problematic, on the grounds that simultaneous causation occurs throughout the real world. For instance, colliding billiard balls in classical physics might seem to illustrate simultaneous causation. However, this is a mistake. If two billiard balls collide at time t with velocity vectors pointing towards each other, they do not also have velocity vectors pointing away from each other at the very same time. Instead, this illustrates what I have elsewhere (1984) called instantaneous causation. Instantaneous causation requires that if A&P becomes true at t, then for some $\delta > 0$, Q will be true throughout the clopen interval $(t\ ,\ t+\delta]$.[16] I believe that instantaneous causation is all that is required for describing the real world.

I have followed AI-convention here in talking about causal change in terms of causation. However, that introduces unnecessary complexities. For example, it is generally assumed in the philosophical literature on causation that if P causes Q then Q would not have been true if P were not true.[17] This has the consequence that when there are two independent factors each of which would be sufficient by itself to cause the same effect, if both occur then neither causes it. These are cases of causal overdetermination. A familiar example of causal overdetermination occurs when two assailants shoot a common victim at the same time. Either shot would be fatal. The result is that neither shot is such that if it had not occurred then the victim would not have died, and hence, it is generally maintained, neither shot caused the death of the victim. However, this kind of failure of causation ought to be irrelevant to the kind of causal reasoning under discussion in connection with change. Principle (4) ought to apply to cases of causal overdetermination as well as to genuine cases of causation. This indicates that the intricacies of the analysis of "cause" are irrelevant in the present context.

I take it (and have argued in my (1984)) that all varieties of causation (including causal overdetermination) arise from the instantiation of "causal laws". These are what I have

---

[16] I assume that time has the structure of the reals.

[17] See Lewis (1973). See my (1984) for more details about the relationship between causes and counterfactual conditionals.

dubbed nomic generalizations, and have discussed at length in my (1990). Nomic generalizations are symbolized as "P ⇒ Q", where P and Q are formulas and '⇒' is a variable-binding operator, binding all free occurrences of variables in P and Q. An informal gloss on "P ⇒ Q" is "Any physically-possible P would be a Q". For example, the law that electrons are negatively charged could be written "(x is an electron) ⇒ (x is negatively charged)". The free occurrences of 'x' are bound by '⇒'.

A rule of universal instantiation applies to nomic generalizations, allowing us to derive less general nomic generalizations:

> If 'x' is free in P and Q, and P(x/a) and Q(x/a) result from substituting the constant term 'a' for 'x', then (P ⇒ Q) entails (P(x/a) ⇒ Q(x/a)).

I propose that we replace "(A&P causes Q)" in (4) by "(A&P ⇒ (Q will shortly be true))", where the latter typically results from instantiating more general laws. More precisely, let us define "A when P is causally sufficient for Q after an interval ε" to mean

(∀t){(A-at-t & P-at-t) ⇒ (∃δ)Q-throughout-(t+ε , t+ε+δ]}.

Instantaneous causation is causal sufficiency with an interval 0.

My proposal is to replace "causes" by "causal sufficiency" in (4). Modifying it to take account of the interval over which the causation occurs:

> CAUSAL-UNDERCUTTER
> Where $t_0 \leq t_1$ and $(t_1+\varepsilon) < t$, "A-at-t_1 & Q-at-t_1 & (A when Q is causally sufficient for ~P after an interval ε)" is a defeasible undercutting defeater for the inference from P-at-t_0 to P-throughout- ⟨t* t⟩ by TEMPORAL-PROJECTION.

This can be implemented as follows:

```
(def-backwards-undercutter causal-undercutter
   :defeatee TEMPORAL-PROJECTION
   :forwards-premises
      "(A when Q is causally sufficient for ~P after an interval interval)"
      "(A at time1)"
        (:condition (and (time0 <= time1) ((time1 + interval) < time)))
   :backwards-premises
      "(Q at time1)"
   :variables  A Q P time0 time time* time1 interval op
   :defeasible?  T)
```

We can also construct an indexical version of this principle is as follows:

```
(def-backwards-undercutter INDEXICAL-CAUSAL-UNDERCUTTER
   :defeatee INDEXICAL-TEMPORAL-PROJECTION
   :forwards-premises
```

```
    "(A when Q is causally sufficient for ~P after an interval interval)"
    "(A at time1)"
     (:condition (time0 <= time1)
  :backwards-premises
    "(Q at time1)"
  :variables  A Q P time0 time1 interval
  :defeasible?  T
  :temporal? T)
```

For causal reasoning, we want to use the causal connection to support inferences about what will happen. This is more complicated than it might initially seem. The difficulty is that, for example,

> the gun is fired when the gun is loaded is causally sufficient for ~(Jones is alive) after an interval 20

does not imply that if the gun is fired at t and the gun is loaded at t then Jones is dead at t+20. Recall the discussion of instantaneous causation. All that is implied is that Jones is dead over some interval open on the left with t+20 as the lower bound. We can conclude that there is at time > t+20 at which Jones is dead, but it does not follow as a matter of logic that Jones is dead at any particular time because, at least as far as this causal law is concerned, Jones could become alive again after becoming dead. To infer that Jones is dead at a particular time after t+20, we must combine the causal sufficiency with temporal projection. This yields the following principle:

CAUSAL-IMPLICATION

- If Q is temporally-projectible, and $((t+\varepsilon) \leq t^* < t^{**})$, then "(A when P is causally sufficient for Q after an interval ε) & A-at-t & P-at-t" is a defeasible reason for "Q-throughout-$(t^* , t^{**}]$" and for "Q-throughout-(t^* , t^{**})".
- If Q is temporally-projectible, and $((t+\varepsilon) < t^* \leq t^{**})$, then "(A when P is causally sufficient for Q after an interval ε) & A-at-t & P-at-t" is a defeasible reason for "Q-throughout-$[t^* , t^{**}]$".

This is implemented as follows:

```
(def-backwards-reason CAUSAL-IMPLICATION
   :conclusion  "(Q throughout (op time* time**))"
   :condition (<= time* time**)
   :forwards-premises
   "(A when P is causally sufficient for Q after an interval interval)"
     (:condition (every #'temporally-projectible (conjuncts Q)))
   "(A at time)"
   (:condition
     (or (and (eq op 'clopen) ((time + interval) <= time*) (time* < time**))
         (and (eq op 'closed) ((time + interval) < time*) (time* <= time**))
```

```
      (and (eq op 'open) ((time + interval) <= time*) (time* < time**))))
  :backwards-premises
  "(P at time)"
  :variables  A P Q interval time time* time** op
  :strength   (expt (1- (* 2 *temporal-decay*)) (- time** time))
  :defeasible?  T)
```

Because CAUSAL-IMPLICATION, in effect, builds in an application of TEMPORAL-PROJECTION, it should be defeated by causal-undercutting in the same way TEMPORAL-PROJECTION is. Without this cases involving sequential causes would not work properly. For instance, suppose we add to the Yale Shooting Problem that "Jones is resuscitated when he is not alive" is causally sufficient for "Jones is alive" after an interval 5, and we add that Jones is resuscitated at 50. We then want to know whether Jones is alive at 60. The difficulty is that we can use CAUSAL-IMPLICATION to argue that Jones is not alive at 60, just as we used it to argue that Jones is not alive at 50. We can also use CAUSAL-IMPLICATION to argue that Jones is alive at 60 because he was not alive at 50 and he was resuscitated at 50. But this yields collective defeat, and no way to resolve it, just as in the original problem. This is diagramed in figure 8.

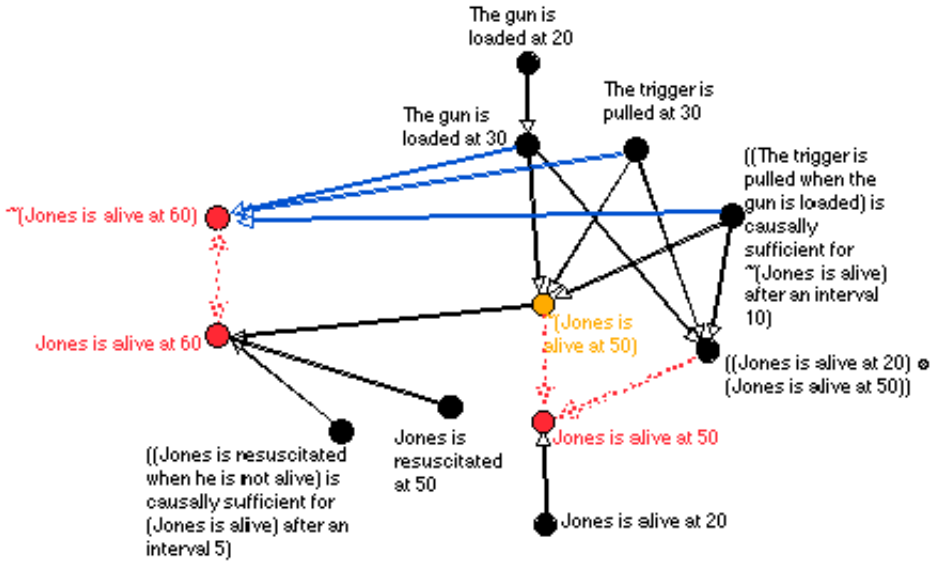

Figure 8. Shooting Problem with Resuscitation

This problem is resolved by adopting a version of causal-undercutting for CAUSAL-IMPLICATION:

CAUSAL-UNDERCUTTER-FOR-CAUSAL-IMPLICATION
Where $(t+\varepsilon) \leq t_1, (t_1+\varepsilon*) < t^{**}$, and $t_{00} \leq t_1$, "A*-at-t_1 & Q-at-t_{00} & (A* when R is causally

sufficient for ~Q after an interval ε*)" is a defeasible undercutting defeater for the inference from Q-at-t_0 to Q-throughout- ⟨t* t**⟩ by CAUSAL-IMPLICATION.

This is implemented as follows:

```
(def-backwards-undercutter CAUSAL-UNDERCUTTER-FOR-CAUSAL-IMPLICATION
   :defeatee *causal-implication*
   :forwards-premises
   "(define -q (neg q))"
   "(A* when R is causally sufficient for -q after an interval interval*)"
   "(A* at time1)"
     (:condition (and ((time + interval) <= time1) ((time1 + interval*) < time**)))
   :backwards-premises
   "(R at time00)"
   (:condition (time00 <= time1))
   :variables  A P Q interval time time* time** op A* R -q interval* time1 time00
   :defeasible?  T)
```

With the addition of this defeater, the problem is resolved as in figure 9.

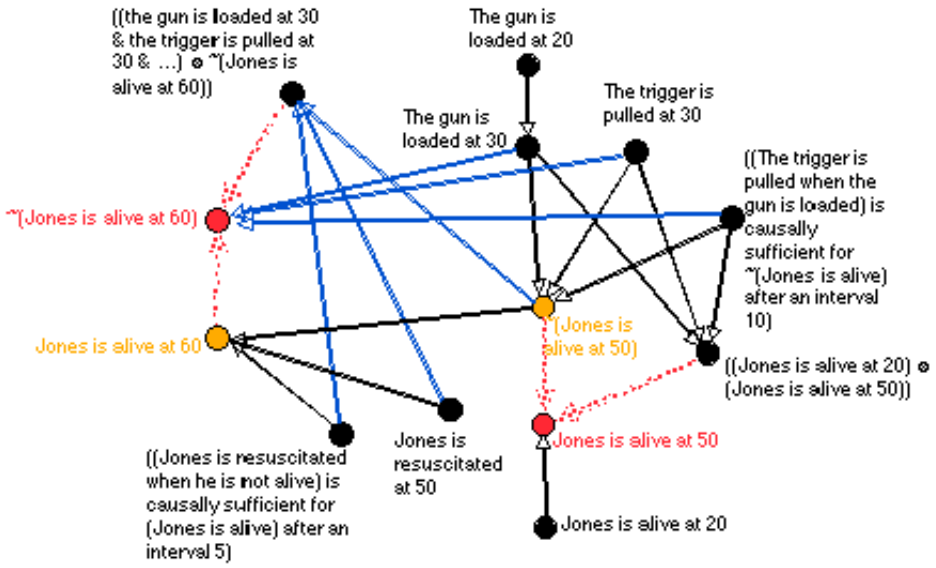

Figure 9. Resolution of the Shooting Problem with Resuscitation

We also need an indexical version of CAUSAL-IMPLICATION and CAUSAL-UNDERCUTTER-FOR-CAUSAL-IMPLICATION:

```
(def-backwards-reason indexical-CAUSAL-IMPLICATION
```

```
   :conclusion  "Q"
   :forwards-premises
   "(A when P is causally sufficient for Q after an interval interval)"
     (:condition (every #'temporally-projectible (conjuncts Q)))
   "(A at time)"
   (:condition ((time + interval) < now)
   :backwards-premises
   "(P at time)"
   :variables  A P Q interval time
   :defeasible?  T
   :strength (expt (1- (* 2 *temporal-decay*)) (- now time))
   :temporal? T)

(def-backwards-undercutter INDEXICAL-CAUSAL-UNDERCUTTER-FOR-CAUSAL-IMPLICATION
   :defeatee *indexical-temporal-projection*
   :forwards-premises
   "(define -q (neg q))"
   "(A* when R is causally sufficient for -q after an interval interval*)"
   "(A* at time1)"
     (:condition (and ((time + interval) <= time1) ((time1 + interval*) < now)))
   :backwards-premises
   "(Q at time00)"
   (:condition (time00 <= time1))
   :variables  A P Q interval time time* op A* R -q interval* time1 time00
   :defeasible?  T
   :temporal? T)
```

These principles can be illustrated by applying them to the Yale Shooting Problem with Resuscitation:

```
====================================================

Problem number 14:  This illustrates sequential causation. This requires causal undercutting for
causal implication.  I know that the gun being fired while loaded
will cause Jones to become dead.  I know that the gun is initially loaded, and Jones is initially
alive.  Later, the gun is fired. But I also know that he will be resuscitated later, and then he will
be alive.  Should I conclude that Jones becomes dead?

Forwards-substantive-reasons:
        neg-at-elimination

Backwards-substantive-reasons:
        *temporal-projection*
        *causal-undercutter*
        *causal-implication*
        *causal-undercutter-for-causal-implication*
        neg-at-intro
        neg-at-intro2

Inputs:

Given:
```

```
        (the_gun_is_loaded at 20)  :  with justification = 1.0
        ((Jones is alive) at 20)  :  with justification = 1.0
        (the_gun_is_fired at 30)  :  with justification = 1.0
        (Jones_is_resuscitated at 45)  :  with justification = 1.0
        (the_gun_is_fired when the_gun_is_loaded is causally sufficient for ~(Jones is alive) after an interval
10)  :  with justification = 1.0
        (Jones_is_resuscitated when ~(Jones is alive) is causally sufficient for (Jones is alive) after an
interval 5)  :  with justification = 1.0

Ultimate epistemic interests:
    (? ((Jones is alive) at 60))    degree of interest = 0.75
=====================================================
 # 1
 (the_gun_is_loaded at 20)
 given
 This discharges interest 5
 # 2
 ((Jones is alive) at 20)
 given
 # 3
 (the_gun_is_fired at 30)
 given
 # 4
 (Jones_is_resuscitated at 45)
 given
 # 5
 (the_gun_is_fired when the_gun_is_loaded is causally sufficient for ~(Jones is alive) after an interval 10)
 given
 # 6
 (Jones_is_resuscitated when ~(Jones is alive) is causally sufficient for (Jones is alive) after an interval 5)
 given
                              # 1
                              interest: ((Jones is alive) at 60)
                              This is of ultimate interest
                              # 2
                              interest: ~((Jones is alive) at 60)
                              This is of ultimate interest
                              Of interest as a defeater for support-links: { link 24 for node 7 , link 12 for node 7 ,
link 7 for node 7 }
                              # 3
                              interest: (~(Jones is alive) at 60)
                              For interest 2 by neg-at-intro
                              This interest is discharged by node 11
 # 7
 ((Jones is alive) at 60)
 Inferred by:
           support-link #7 from { 2 } by *temporal-projection*  defeaters: { 13 , 8 }   DEFEATED
 This node is inferred by discharging a link to interest #1
                              # 4
                              interest: (((Jones is alive) at 20) @ ((Jones is alive) at 60))
                              Of interest as a defeater for support-link 7 for node 7
          vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
          The undefeated-degree-of-support of #<Node 7> is 0.9920311211060142
          vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
          ========================================
          Justified belief in ((Jones is alive) at 60)
```

```
        with undefeated-degree-of-support 0.9920311211060142
        answers #<Query 1: (? ((Jones is alive) at 60))>
        ========================================
                        # 5
                        interest: (the_gun_is_loaded at ^@y0)
                        For interest 4 by *causal-undercutter* using nodes (3 5)
                        For interest 27 by *causal-undercutter* using nodes (3 5)
                        This interest is discharged by nodes (1 9)
# 8   NOT STRICTLY RELEVANT
(((Jones is alive) at 20) @ ((Jones is alive) at 60))
Inferred by:
         support-link #8 from { 5 , 3 , 1 } by *causal-undercutter*
defeatees: { link 7 for node 7 }
This node is inferred by discharging links to interests (4 4)
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
        #<Node 7> has become defeated.
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
        ========================================
        Lowering the undefeated-degree-of-support of ((Jones is alive) at 60)
        retracts the previous answer to #<Query 1: (? ((Jones is alive) at 60))>
        ========================================
                        # 8
                        interest: (~(Jones is alive) at 45)
                        For interest 1 by *causal-implication* using nodes (4 6)
                        For interest 1 by *causal-implication* using nodes (4 6)
                        This interest is discharged by node 10
                        # 9
                        interest: (the_gun_is_loaded at 30)
                        For interest 8 by *causal-implication* using nodes (3 5)
                        For interest 3 by *causal-implication* using nodes (3 5)
                        For interest 8 by *causal-implication* using nodes (3 5)
                        This interest is discharged by node 9
# 9
(the_gun_is_loaded at 30)
Inferred by:
        support-link #9 from { 1 } by *temporal-projection*
This discharges interests (5 9)
# 8   NOT STRICTLY RELEVANT
(((Jones is alive) at 20) @ ((Jones is alive) at 60))
Inferred by:
        support-link #10 from { 5 , 3 , 9 } by *causal-undercutter*
        support-link #8 from { 5 , 3 , 1 } by *causal-undercutter*
defeatees: { link 7 for node 7 }
This node is inferred by discharging links to interests (4 4)
# 10
(~(Jones is alive) at 45)
Inferred by:
        support-link #11 from { 5 , 3 , 9 } by *causal-implication*  defeaters: { 18 }
This node is inferred by discharging a link to interest #8
This discharges interests (17 22)
                        # 14
                        interest: ~(~(Jones is alive) at 45)
                        Of interest as a defeater for support-link 11 for node 10
# 7
((Jones is alive) at 60)
Inferred by:
```

```
        support-link #12 from { 6 , 4 , 10 } by *causal-implication*  defeaters: { 13 }
        support-link #7 from { 2 } by *temporal-projection*  defeaters: { 13 , 8 }   DEFEATED
This node is inferred by discharging a link to interest #1
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
        The undefeated-degree-of-support of #<Node 9> is 0.998001799040336
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
        The undefeated-degree-of-support of #<Node 10> is 0.9970041963621832
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
        The undefeated-degree-of-support of #<Node 7> has increased to 0.9970041963621832
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
        ========================================
        Justified belief in ((Jones is alive) at 60)
        with undefeated-degree-of-support 0.9970041963621832
        answers #<Query 1: (? ((Jones is alive) at 60))>
        ========================================
# 11
(~(Jones is alive) at 60)               DEFEATED
Inferred by:
        support-link #17 from { 5 , 3 , 9 } by *causal-implication*  defeaters: { 14 , 15 }   DEFEATED
        support-link #13 from { 10 } by *temporal-projection*  defeaters: { 14 , 12 }   DEFEATED
This node is inferred by discharging a link to interest #3
                        # 16
                        interest: ((~(Jones is alive) at 45) @ (~(Jones is alive) at 60))
                        Of interest as a defeater for support-link 13 for node 11
                        This interest is discharged by node 12
                        # 17
                        interest: (~(Jones is alive) at ^@y1)
                        For interest 16 by *causal-undercutter* using nodes (4 6)
                        This interest is discharged by node 10
# 12
((~(Jones is alive) at 45) @ (~(Jones is alive) at 60))
Inferred by:
        support-link #14 from { 6 , 4 , 10 } by *causal-undercutter*
defeatees: { link 13 for node 11 }
This node is inferred by discharging a link to interest #16
# 13
~((Jones is alive) at 60)               DEFEATED
Inferred by:
        support-link #15 from { 11 } by neg-at-intro   DEFEATED
defeatees: { link 24 for node 7 , link 7 for node 7 , link 12 for node 7 }
This node is inferred by discharging a link to interest #2
# 14
~(~(Jones is alive) at 60)
Inferred by:
        support-link #16 from { 7 } by inversion_from_contradictory_nodes_13_and_7
defeatees: { link 17 for node 11 , link 13 for node 11 }
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
        The undefeated-degree-of-support of #<Node 12> is 0.9970041963621832
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
        The undefeated-degree-of-support of #<Node 14> is 0.9970041963621832
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
        #<Node 11> has become defeated.
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
        #<Node 13> has become defeated.
        vvvvvvvvvvvvvvvvvvvvvvvvvvvvv
# 11
```

(~(Jones is alive) at 60) DEFEATED
Inferred by:
support-link #17 from { 5 , 3 , 9 } by *causal-implication* defeaters: { 14 , 15 } DEFEATED
support-link #13 from { 10 } by *temporal-projection* defeaters: { 14 , 12 } DEFEATED
This node is inferred by discharging a link to interest #3
21
interest: (((the_gun_is_fired when the_gun_is_loaded is causally sufficient for ~(Jones is alive) after an interval 10) & ((the_gun_is_fired at 30) & (the_gun_is_loaded at 30))) @ (~(Jones is alive) at 60))
Of interest as a defeater for support-link 17 for node 11
This interest is discharged by node 15
22
interest: (~(Jones is alive) at ^@y2)
For interest 21 by *causal-undercutter-for-causal-implication* using nodes (4 6)
This interest is discharged by node 10
15
(((the_gun_is_fired when the_gun_is_loaded is causally sufficient for ~(Jones is alive) after an interval 10) & ((the_gun_is_fired at 30) & (the_gun_is_loaded at 30))) @ (~(Jones is alive) at 60))
Inferred by:
support-link #18 from { 6 , 4 , 10 } by *causal-undercutter-for-causal-implication*
defeatees: { link 17 for node 11 }
This node is inferred by discharging a link to interest #21
vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
The undefeated-degree-of-support of #<Node 15> is 0.9970041963621832
vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
26
interest: ((Jones is alive) at 45)
For interest 14 by neg-at-intro2
This interest is discharged by node 16
16
((Jones is alive) at 45) DEFEATED
Inferred by:
support-link #19 from { 2 } by *temporal-projection* defeaters: { 19 , 17 } DEFEATED
This discharges interest 26
27
interest: (((Jones is alive) at 20) @ ((Jones is alive) at 45))
Of interest as a defeater for support-link 19 for node 16
This interest is discharged by node 17
17
(((Jones is alive) at 20) @ ((Jones is alive) at 45))
Inferred by:
support-link #21 from { 5 , 3 , 9 } by *causal-undercutter*
support-link #20 from { 5 , 3 , 1 } by *causal-undercutter*
defeatees: { link 19 for node 16 }
This node is inferred by discharging links to interests (27 27)
17
(((Jones is alive) at 20) @ ((Jones is alive) at 45))
Inferred by:
support-link #21 from { 5 , 3 , 9 } by *causal-undercutter*
support-link #20 from { 5 , 3 , 1 } by *causal-undercutter*
defeatees: { link 19 for node 16 }
This node is inferred by discharging links to interests (27 27)
18
~(~(Jones is alive) at 45) DEFEATED
Inferred by:
support-link #22 from { 16 } by neg-at-intro2 DEFEATED

```
 defeatees: { link 11 for node 10 }
 This node is inferred by discharging a link to interest #14
 # 19
 ~((Jones is alive) at 45)
 Inferred by:
          support-link #23 from { 10 } by inversion_from_contradictory_nodes_18_and_10
 defeatees: { link 19 for node 16 }
          vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
          The undefeated-degree-of-support of #<Node 19> is 0.9970041963621832
          vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
          #<Node 16> has become defeated.
          vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
          #<Node 18> has become defeated.
          vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
 # 7
 ((Jones is alive) at 60)
 Inferred by:
          support-link #24 from { 16 } by *temporal-projection*  defeaters: { 13 }   DEFEATED
          support-link #12 from { 6 , 4 , 10 } by *causal-implication*  defeaters: { 13 }
          support-link #7 from { 2 } by *temporal-projection*  defeaters: { 13 , 8 }   DEFEATED
 This node is inferred by discharging a link to interest #1
=====================================================


========== ULTIMATE EPISTEMIC INTERESTS ========
 Interest in (? ((Jones is alive) at 60))
 is answered by node 7:  ((Jones is alive) at 60)
-------------------------------------------------------
```

This, as far as I know, is the first implemented fully general solution to the Frame Problem.[18] It is noteworthy how simple it is to implement such principles in OSCAR, making OSCAR a potent tool for epistemological analysis.

10. The Qualification and Ramification Problems

Two problems that are often associated with the Frame Problem are the Qualification and Ramification Problems. Like the Frame Problem, these arose initially within the framework of attempts to reason about change deductively. The Frame Problem concerned the proliferation of frame axioms—axioms concerning what does not change.

---

[18] Partially implemented solutions have been proposed by Loui (1987) and Kartha and Lifschitz (1995). Loui's solution was based upon THEORIST (Poole, et al 1987, Poole 1988). However, THEORIST produces a recursively enumerable set of consequences in any defeasible reasoning problem, and so is not capable of dealing with full first-order logic in which consistency is not decidable and hence the set of defeasible consequences is only Δ_2 (see my 1995 for further discussionof this). In addition, Hanks and McDermott (1987) argue convincingly, to my mind, that Loui's solution does not really work. The Kartha and Lifschitz implementation is based upon circumscription, and hence only works in the special case in which the circumscribed theory can be reduced to a first-order theory. In any cases of realistic complexity, the latter would seem to be impossible.

The Qualification and Ramification Problems concerned the difficulty in correctly formulating axioms about what does change. The Qualification Problem is the problem of getting the antecedent right in axioms like "A match's being struck when it is dry, in the presence of oxygen, ... , is causally sufficient for it to light". The difficulty is that we are typically unable to fill in the ellipsis and give a complete list of the conditions required to cause a particular effect. McCarthy (1977) illustrated this with his famous "banana in the tailpipe" example. Most people are unaware that a car will not start if the tailpipe is blocked, e.g., by a banana. Thus if asked to state the conditions under which turning the ignition key will start the car, they will be unable to do so. An allied difficulty is that even if we could completely enumerate the conditions required, deductive reasoning about change would require us to then deductively verify that all of those conditions are satisfied—something that human beings clearly do not generally do.

Within the present framework, the solution to the Qualification Problem seems to be fairly simple. I defined "A when P is causally sufficient for Q after an interval ε" to mean

$$(\forall t)\{(\text{A-at-t} \ \& \ \text{P-at-t}) \Rightarrow (\exists \delta)\text{Q-throughout-}(t+\varepsilon \ , \ t+\varepsilon+\delta]\}.$$

So defined, the causal knowledge that we use in reasoning about change is not generally of this form. This is for two reasons. First, we rarely have more than a rough estimate of the value of ε. Second, we are rarely in a position to formulate P precisely. That latter is just the Qualification Problem. Our knowledge actually takes the form:

> $(\exists P^*)(\exists \varepsilon*)$[P* is true & $\varepsilon* \leq \varepsilon$ & (A when (P & P*) is causally sufficient for Q after an interval $\varepsilon*$)].

P formulates the known preconditions for the causal sufficiency, P* the unknown precondtions, and ε is the known upper bound on ε∗. Let us abbreviate this as "A when P is weakly causally sufficient for Q after an interval ε". We acquire knowledge of weak causal sufficiency inductively. For example, we learn inductively that striking a dry match is usually weakly causally sufficient for it to light after a negligible interval. If we then examine CAUSAL-UNDERCUTTER and CAUSAL-IMPLICATION, we find that both continue to hold if we reconstrue "causally suffficient" to mean "weakly causally sufficient". Thus we can reason about change in the same way even with incomplete causal knowledge. This resolves the Qualification Problem.

The Ramification Problem arises from the observation that in realistically complex environments, we cannot formulate axioms that completely specify the effects of actions or events. People sometimes refer to these as "actions with ramifications", as if these were peculiar actions. But in the real world, all actions have infinitely many ramifications stretching into the indefinite future. This is a problem for reasoning about change deductively, but does not seem to be a problem for reasoning about change defeasibly in the present framework. Consider how human beings deal with this difficulty. Our inability to enumerate all the effects of an action means that we cannot formulate true successor-state axioms (axioms that roll frame axioms and effect axioms into a single axiom in the situation calculus to completely describe the next situation). But we do not

have to. CAUSAL-UNDERCUTTER and CAUSAL-IMPLICATION allow us to reason defeasibly about change on the basis of our incomplete knowledge.

Another aspect of the Ramification Problem is that even if it were sufficient to formulate successor-state axioms using just the effects that we actually know to be produced by an action, listing all of the known effects would make the axiom so complex that a theorem prover would find it too unwieldy to use. For example, if we think about it for a while, we must enumerate among the effects of striking a match such things as displacing air around the match, marginally depleting the ozone layer, raising the temperature of the earth's atmosphere, marginally illuminating Alpha Centauri, making that match unavailable for future use, etc. These are effects that we typically do not care about, and so we do not reason about them. But this does not mean that we can omit them from a successor-state axiom with impunity, because occasionally we might care about one of them.

Within the present framework, this is not a problem. Reasoning in OSCAR is interest-driven, and CAUSAL-IMPLICATION is a backwards-reason. This means that we only reason about potential effects of actions and events when they are of interest to us. Whether they are of interest can vary from circumstance to circumstance, allowing our reasoning to vary similarly, without our having to revise our knowledge base or rules of inference to accommodate the change. Deductive reasoning in terms of successor-state axioms is too crude an implementation to reflect this feature of human reasoning, but the current framework handles it automatically. The conclusion is that the Ramification Problem simply does not arise in this framework.

11. The Extended Prediction Problem

The literature on the Frame Problem has generally focused on toy problems like the Yale Shooting Problem. Experience elsewhere in AI should make us wary of such an approach. Solutions to toy problems may not scale up to problems of realistic complexity. To see how the present proposals fare in this connection, consider "the extended prediction problem" introduced by Shoham (1987). He suggests that even if reasoners are able to reason about the immediate future, as in the Yale Shooting Problem, they will have difficulty using causal information to make predictions about the relatively distant future. He takes this to be illustrated by the traditional classical physics problem of predicting the future positions of colliding billiard balls. Consider two (dimensionless) billiard balls whose positions and velocities are known at initial times, and suppose they are rolling on a flat frictionless surface. Suppose further that they are on a collision course. The problem is to predict their positions at some particular time after the impending collision.

Although this seems like a simple problem, it is considerably more difficult than the toy problems addressed elsewhere in the literature. It is worth noting that even human physicists have one kind of difficulty solving it. Once it is recognized that a collision is going to occur, there is no difficulty solving the problem, but noticing the impending

collision is not a trivial task. If this is not obvious, consider trying to solve the same problem for 100 billiard balls rolling about on a billiard table. In trying to solve such a problem, a human being uses various heuristics to detect what may be collisions (e.g., seeing whether the directions of motion of two billiard balls cross), and then is able to determine by explicit reasoning whether these possible collisions are real collisions. What I will show is that if OSCAR's attention is drawn to the appropriate possible collisions, then OSCAR, equipped with the reasons formulated in this paper, can solve this problem as readily as human beings.

Solving this problem requires assuming that the velocity of a billiard ball (the vector quantity of the speed in a certain direction) will not change unless it collides with another billiard ball. This is intimately connected with Newton's laws of motion. It is of interest to think about Newtonian kinematics in connection with the preceding account of causal change. Newton's laws of motion tell us that the velocity of an object remains unchanged in the absense of external forces. If velocity is taken to be temporally-projectible, this is very much like applying TEMPORAL-PROJECTION to it. If it seems dubious that velocity should be temporally-projectible, notice that velocity and position must both be relativized to inertial frames of reference. This has the consequence that velocity remains unchanged relative to one inertial frame iff there is an inertial frame relative to which the velocity remains constantly zero and hence position remains unchanged. Thus velocity is a stable property iff position is. And it seems eminently plausible that position is temporally-projectible. Thus we can regard this part of Newtonian kinematics as a mathematicization of this aspect of commonsense reasoning. So construed, Newtonian force laws have the same status as the putative law about loaded guns that is employed in the Yale Shooting Problem. Correct reasoning about the consequences of billiard balls colliding encounters precisely the same difficulties as those encountered in predicting the consequences of pulling the triggers on loaded guns. Newtonian physics avoids this difficulty at a theoretical level by adopting a mathematical model of the entire world, and solving the "world equation". But of course, this is not something that a rational agent can really do in a complex environment (even if he is a Newtonian physicist). Using Newtonian physics to get around in the real world requires epistemic principles like TEMPORAL-PROJECTION and CAUSAL-UNDERCUTTER just as much as the use of more naive physical principles does.

Suppose, then, that we begin with two billiard balls whose positions and velocities are known at an initial time t_0. Suppose further that if those velocities remain unchanged until time t_1, the billiard balls will collide at that time. Assuming that it is a perfectly elastic collision, and the balls have the same mass, their new velocities are easily computed—the balls simply exchange trajectories. If we assume that the resulting velocities remain unchanged until time t_2, we can then compute the positions of the balls at t_2. This reasoning uses TEMPORAL-PROJECTION to conclude that the velocities are unchanged prior to the collision, and CAUSAL-UNDERCUTTER and CAUSAL-IMPLICATION (the latter builds in TEMPORAL-PROJECTION) to infer the velocities after the collision.

To make the problem concrete, suppose b_1 is at (0 3) at time 0 with velocity (.5 0), and b_2 is at (1 0) at time 0 with velocity (.4 .3). Then the balls should move as diagrammed in figure 10.

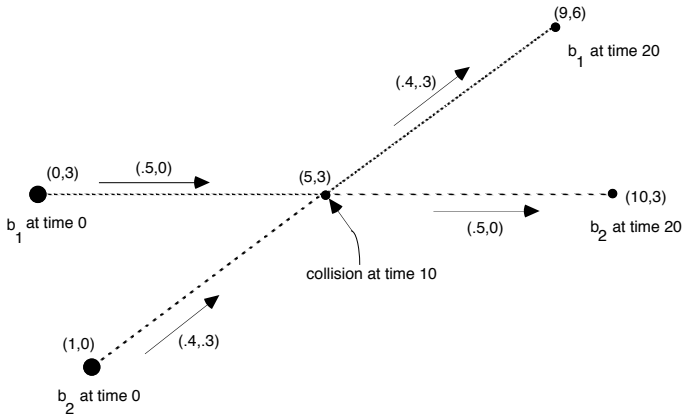

Figure 10. Colliding billiard balls.

Positions can be represented as pairs (x y) of x– and y–coordinates, and velocities as pairs (v_x v_y) of speeds along the x– and y–axes. To compute positions after an object has moved from an initial position with a constant velocity for a specified amount of time, we need the following kinematic principle:

NEW-POSITION
"(the position of s is (x_0 y_0))-at-t_0 & (the velocity of s is (v_x v_y) throughout (t_0 , t_1]" is a conclusive reason for "(the position of s is ((x_0 + (v_x(t_1 – t_0))) (y_0 + (v_y(t_1 – t_0)))))-at-t_1".

```
(def-backwards-reason new-position
   :conclusion  "((the position of b is (x y)) at time1)"
   :forwards-premises
      "((the position of b is (x0 y0)) at time0)"
       (:condition (time0 < time1))
   :backwards-premises
      "(∃vx)(∃vy)
       (& ((the velocity of b is (vx vy)) throughout (clopen time0 time1))
          (x = (x0 + (vx * (time1 - time0))))
          (y = (y0 + (vy * (time1 - time0)))))"
   :variables  b time1 x y x0 y0 time0)
```

We also need the following causal principle governing elastic collisions between billiard balls of the same mass. To keep the mathematics simple, I will just consider the case of dimensionless colliding billiard balls having the same mass. In that case, the colliding balls simply exchange velocities:

(5) $(\forall b_1)(\forall b_2)(\forall {}_1v_x)(\forall {}_1v_y)(\forall {}_2v_x)(\forall {}_2v_y)$
{[b_1 is a dimensionless billiard ball & b_2 is a dimensionless billiard ball & b_1 and b_2 have the same mass & $({}_1v_x^2 + {}_1v_y^2) = ({}_2v_x^2 + {}_2v_y^2)$] → [($b_1$ and b_2 collide) when (the velocity of b_1 is $({}_1v_x\ {}_1v_y)$) is causally sufficient for (the velocity of b_2 is $({}_1v_x\ {}_1v_y)$) after an interval 0]}.

Dimensionless billiard balls collide iff they have the same position:

COLLISION
"$(\exists x)$[(the position of b_1 is x)-at-t & (the position of b_2 is x)-at-t]" is a conclusive reason for "(b_1 and b_2 collide)-at-t".

```
(def-backwards-reason collision
  :conclusion  "((b1 and b2 collide) at time)"
  :backwards-premises
    "(∃x)(∃y)(((the position of b1 is (x y)) at time) & ((the position of b2 is (x y)) at time))"
  :variables   b1 b2 time)
```

Now suppose b_1 is at (0 3) at time 0 with velocity (.5 0), and b_2 is at (1 0) at time 0 with velocity (.4 .3). If the velocities remain unchanged, b_1 and b_2 should collide at time 10. If we pose this as a problem for OSCAR, by COLLISION, an interest in whether b_1 and b_2 collide at 10 generates an interest in their positions at 10. Because we know their positions at 0, NEW-POSITION generates an interest in their velocities between 0 and 10. We know the velocities at 0, and TEMPORAL-PROJECTION leads to an inference that those velocities remain unchanged between 0 and 10. From that we can compute the positions at 10, and infer that b1 and b2 collide at 10.

However, TEMPORAL-PROJECTION also leads to an inference that the positions at 10 are the same as those at 0. That inference must be defeated somehow, but the principles described so far will not accomplish that. This can be accomplished by adding another principle of causal sufficiency. It is a logically contingent but physically necessary fact (at least, according to Newtonian physics) that billiard balls, and other objects of nonzero mass, do not undergo infinite accelerations. As such, if a billiard ball has nonzero velocity at a time, that is causally sufficient for there being a subsequent time at which its position has changed. Thus we have:

(6) $(\forall b)(\forall x)(\forall y)(\forall v_x)(\forall v_y)$
{(the position of b is (x y)) when ((the velocity of b is $(v_x\ v_y)$) & ~$((v_x\ v_y) = (0.0\ 0.0))$) is causally sufficient for ~(the position of b is (x , y)) after an interval 0]

Because the velocities at 0 are nonzero, CAUSAL-UNDERCUTTER defeats the problematic inference that the billiard balls do not move from 0 to 10, leaving the single undefeated conclusion regarding the positions at 10 that both balls are at (5.0 3.0).

So far, we are given the positions and velocities of b_1 and b_2 at 0, and we have inferred their velocities between 0 and 10, their positions at 10, and have concluded that they collide at 10. Now suppose we want to know the position of b_1 at 20. Given a knowledge of the position of b_1 at 10, NEW-POSITION generates an interest in the velocity of b_1 between 10 and 20. By CAUSAL-IMPLICATION, we can infer that the velocity of b_1 between 10 and 20 is (.4 .3). From this NEW-POSITION enables us to infer (correctly) that the position of b_1 at 20 is (9.0 6.0).

However, there are conflicting inferences that can also be made, and they must be defeated. By TEMPORAL-PROJECTION, we can infer that the velocity of b_1 between 10 and 20 is the same as at 10. This is defeated as above by causal-undercutteR, using (5), because we know the velocities of b_1 and b_2 at 10 and know they collide. Similarly, we can use TEMPORAL-PROJECTION to infer that the velocity of b_1 between 0 and 20 is the same as at 0, which we know to be (.5 0). This is also defeated by causal-undercutteR, using (5).

By TEMPORAL-PROJECTION, we can infer that the position of b_1 at 20 is the same as at 0. But this is defeated by causal-undercutteR, using (6), because we know that the velocity of b_1 at 0 is nonzero. Finally, by TEMPORAL-PROJECTION, we can infer that the position of b_1 at 20 is the same as at 10. But this is also defeated by causal-undercutteR, using (6), because we know that the velocity of b_1 at 10 is nonzero. Thus all the undesirable inferences are defeated, leaving OSCAR with the desired conclusion that the position of b_1 at 20 is (9.0 6.0).

A display of OSCAR's reasoning on this problem appears in the appendix. Apparently the Extended Prediction Problem poses no new problems for OSCAR's ability to reason causally.

12. Conclusions and Comparisons

An agent operating in a complex changing environment must be able to acquire new information perceptually, project those observations forwards in time to draw conclusions about times when observations are not occurring, and reason about how the world will change as a result of changes either observed or wrought by the agent's own behavior. This paper has proposed defeasible reasons that license such reasoning, and described their implementation in OSCAR. perception, and the associated principles perceptual-reliability, discounted-perception, and perceptual-unreliability, enable an agent to acquire information perceptually while taking account of the fact that perception is less than totally reliable. TEMPORAL-PROJECTION and probabilistic-defeat-for-TEMPORAL-PROJECTION enable an agent to draw conclusions about times when observations are not occurring on the basis of observations at other times. CAUSAL-IMPLICATION and

causal-undercutter enable an agent to make use of causal information in forming expectations regarding how the world will evolve over time. The use of TEMPORAL-PROJECTION is pervasive, because observations occur at particular times and making one observation tends to preclude making another, but an agent's reasoning about the world will frequently depend upon the outcomes of a number of distinct observations. This is made more efficient by the introduction of temporal indexicals, which allow an agent to store and recall temporal conclusions rather than making new inferences.

Most work in AI that has been aimed at similar reasoning has proceeded within the framework of the situation calculus.[19] The situation calculus may be quite useful as a semantical tool, but it has always seemed to me to be needlessly complicated for use in actual reasoning. Automated reasoners have as much difficulty as human reasoners in dealing with axiomatizations of domains expressed in the situation calculus. It is noteworthy that human reasoners are able to reason about complex domains using much simpler formalizations. One aim of this paper has been to show that automated reasoners can do so as well.

The formal principles that have been fashioned in this paper were motivated by philosophical analyses, in the style of traditional philosophical epistemology, of human reasoning about perception, time, and causation. A second aim of this paper has been to illustrate the fecundity of such epistemological analysis in addressing problems in artificial intelligence.

The fundamental tool that makes this analysis possible is the OSCAR system of defeasible reasoning. This is the only implemented system of defeasible reasoning capable of reasoning successfully in a language like first-order logic where logical consistency is not decidable. Once again, this system of defeasible reasoning was motivated by an epistemological analysis of human defeasible reasoning. The problems that must be solved in the construction of such a reasoner are logical and computational, but solutions to these problems were suggested by considering how they are solved in human reasoning.[20] I take the success of the analyses of perceptual, temporal, and causal reasoning proposed in this paper to be a strong argument for the use of this system of defeasible reasoning in the design and construction of rational agents.

To the best of my knowledge, there is no work in AI with which to compare the analysis of perceptual reasoning proposed here.[21] This is a topic that has not previously received careful attention in AI. On the other hand, there has been a lot of work addressed at temporal projection, causal reasoning, the Frame Problem, and the Yale Shooting Problem. As remarked in section ten, the Frame Problem led Sandewall (1972)

---

[19] McCarthy and Hayes (1969).

[20] A sketch of these problems and how they are solved in OSCAR is presented in my (1996a). A more detailed account can be found in my (1995).

[21] There is a lot of relevant work in philosophical epistemology. The basic ideas of the present analysis originate in my (1967), (1971), and (1974). Competing philosophical theories are discussed at length in my (1987).

to propose reasoning about change defeasibly and adopting some sort of defeasible inference scheme to the effect that it is reasonable to believe that something doesn't change unless you are forced to conclude otherwise. In recent literature, this has come to be called "the commonsense law of inertia". This motivated much of the work in AI on nonmonotonic logic. I was led to a similar principle in my (1974) from a completely different direction—by thinking about the reidentification of objects (see section four above). The principles of temporal projection formulated in this paper are an attempt to make these common ideas precise within the framework of OSCAR.

The basic idea behind the treatment I have proposed for the combination of the Frame Problem and the Yale Shooting Problem is to think of the world as unfolding temporally, with changes occurring only when they are forced to occur by what has already happened. This is reminiscent of Shoham's logic of chronological ignorance (1987), and a series of papers stemming from Gelfond and Lifschitz (1993).[22] There is also a similarity to the notion of progressing a database discussed in Lin and Reiter (1994 and 1995). The latter is a monotonic solution to the frame problem, as is that of Schubert (1990, 1994). These monotonic solutions assume that we can formulate what Schubert calls "explanation closure axioms", to the effect that if some change occurs then it must have been caused in one of some small list of ways. But as Schubert readily acknowledges, "a monotonic theory of any realistically complex, dynamic world is bound to be an approximation", and for such worlds we need to go beyond first-order logic and employ nonmonotonic or probabilistic methods.[23] Shoham proposed to capture this idea by modifying the logic of defeasible reasoning. The resulting theory was complex, and has never been implemented. The work stemming from Gelfond and Lifschitz (1993) has proceeded by applying circumscription to axiomatizations within (extensions or modifications of) the situation calculus.[24] The resulting theories are unwieldy because of their reliance on the formalism of the situation calculus. But even more important, circumscription generates second-order theories, and so the proposed solutions are in principle not implementable in general. Shanahan (1996, 1997) describes an implemented solution to special cases of the Yale Shooting Problem that occur in robot navigation. However, being based upon circumscription, his proposed solution is not implementable in any but special cases. Kartha and Lifschitz (1995) also describe implemented solutions to special cases, based upon circumscription. It is noteworthy that their implementation works by attempting to reduce the circumscription to a set of first-order frame axioms, which is just what the appeal to nonmonotonic logic and defeasible reasoning was originally intended to avoid. It is also worth noting that none of these

---

[22] This work includes Kartha and Lifschitz (1995), Shanahan (1995), (1996), and (1997).

[23] Schubert (1994), pg. 698.

[24] Examples of such extensions and modifications are the event calculus of Kowalski and Sergot (1986) and the extension of that in Shanahan (1990).

approaches to temporal projection can solve the perceptual updating problem, because they have no way of decaying the strengths of conclusions. In addition, they fall prey to the projectibility problems discussed in the text. By way of contrast, the proposals made in this paper (1) are based upon a simple formalism, (2) are logically much simpler than the proposals based upon circumscription, and (3) are implemented in general. The current proposals result in reasoning sufficiently simple that we can expect an artificial agent to perform it in real time. For example, the reasoning underlying the solution to the Yale Shooting Problem was performed in .05 seconds on a Macintosh, and that underlying the Extended Prediction Problem (the most complex problem to which these proposals have been applied) required less than a second on a Macintosh.

The implementation of the reasoning described here should free automated planning from reliance on STRIPS-like representations of actions. STRIPS operators handle reasoning about change by precompiling the results of all possible actions under all possible circumstances. That works efficiently in narrowly circumscribed domains, but is impractical in most realistic domains. The difficulty has been that the only obvious alternative is to reason explicitly about change, and no one has known how to do that efficiently. The principles described here should take us at least part way along the path to a solution to this problem. That is the subject of current research in the OSCAR Project.

References

Gelfond, Michael, and Lifschitz, Vladimir

1993 "Representing action and change by logic programs", Journal of Logic Programming 17, 301-322.

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", AAAI-86.

1987 "Nonmonotonic logic and temporal projection", Artificial Intelligence 33, 379-412.

Haugland, John

1987 "An overview of the frame problem", in Z. Pylyshyn (ed.) The Robot's Dilemma,, MIT Press.

Hayes, Patrick

1987 "What the frame problem is and isn't", in Z. Pylyshyn (ed.) The Robot's Dilemma,, MIT Press.

Janlert, Lars-Erik

1987 "Modeling change—the frame problem", in Z. Pylyshyn (ed.) The Robot's Dilemma,, MIT Press.

Kartha, G. Neelakantan, and Lifschitz, Vladimir

1995 "A simple formalization of actions using circumscription". Proceedings of IJCAI 1995, 1970-1975.

Kautz, H. A.
1986 "The logic of persistence", Proceedings of AAAI-86, 401-405.
Kowalski, R. A., and Sergot, M. J.
1986 "A logic-based calculus of events", New Generation Computing 4, 67-95.
Lewis, David
1973 "Causation". Journal of Philosophy 90, 556-567.
Lifschitz, Vladimir
1987 "Formal theories of action", in M. L. Ginsberg (ed.), Readings in Non-monotonic Reasoning. Morgan-Kaufmann: Los Altos, CA.
Lin, Fangzhen, and Reiter, Raymond
1994 "How to progress a database (and why) I. Logical foundations." In Proceedings of the Fourth International Conference on Principles of Knowledge Representation (KR'94). 425-436.
1995 "How to progress a database II: The STRIPS connection." IJCAI-95. 2001-2007.
Loui, Ron
1987 "Response to Hanks and McDermott: temporal evolution of beliefs and beliefs about temporal evolution", Cognitive Science 11, 283-298.
McCarthy, John
1977 "Epistemological problems in artificial intelligence". IJCAI-77.
1986 "Applications of circumscription to formalizing common sense knowledge." Artificial Intelligence 26, 89-116.
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.
McDermott, Drew
1982 "A temporal logic for reasoning about processes and plans", Cognitive Science 6, 101-155.
Pollock, John
1967 "Criteria and our Knowledge of the Material World", The Philosophical Review, 76, 28-60.
1971 "Perceptual Knowledge", Philosophical Review, 80, 287-319.
1974 Knowledge and Justification, Princeton University Press.
1979 Subjunctive Reasoning, D. Reidel.
1984 The Foundations of Philosophical Semantics, Princeton University Press.
1987 Contemporary Theories of Knowledge, Rowman and Littlefield.
1990 Nomic Probability and the Foundations of Induction, Oxford University Press.
1994 "The projectibility constraint", in Grue! The New Riddle of Induction, ed. Douglas Stalker, Open Court, 135-152.
1995 Cognitive Carpentry, MIT Press.
1995a The OSCAR Manual, available online from http://www.u.arizona.edu/~pollock.
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/.

1996a "Implementing defeasible reasoning". Computational Dialectics Workshop at the International Conference on Formal and Applied Practical Reasoning, Bonn, Germany, 1996. This can be downloaded from http://www.u.arizona.edu/~pollock/.
1998 "Planning agents", to appear in Foundations of Rational Agency, ed. A. Rao and M. Wooldridge, Kluwer.
1998a "Reasoning defeasibly about plans", OSCAR project technical report. This can be downloaded from http://www.u.arizona.edu/~pollock/.
1998b "The logical foundations of goal-regression planning", OSCAR project technical report. This can be downloaded from http://www.u.arizona.edu/~pollock/.
Poole, David
1988 "A logical framework for default reasoning", Artificial Intelligence 36, 27-47.
Poole, David, Goebel, R. G., Aleliunas, R.
1987 "Theorist: a logical reasoning system for defaults and diagnosis", in N. Cercone and G. McCalla (eds.), The Knowledge Frontier: Essays in the Representation of Knowledge, Springer, New York.
Sandewall, Erik
1972 "An approach to the frame problem and its implementation". In B. Metzer & D. Michie (eds.), Machine Intelligence 7. Edinburgh: Edinburgh University Press.
Schubert, Len
1990 "Monotonic solution to the frame problem in the situation calculus", in H. Kyburg, R. Loui, and G Carlson (eds), Knowledge Representation and Defeasible Reasoning, Kluwer.
1994 "Explanation closure, action closure, and the Sandewall test suite for reasoning about change", Journal of Logic Computation 4, 679-700.
Shanahan, Murray
1990 "Representing continuous changes in the event calculus". ECAI-90, 598-603.
1995 "A circumscriptive calculus of events". Artificial Intelligence, 77, 249-284.
1996 "Robotics and the common sense informatic situation", Proceedings of the 12th European Conference on Artificial Intelligence, John Wiley & Sons.
1997 Solving the Frame Problem, MIT Press.
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.