

Part Six: Reasoning Defeasibly About the World

- Our interest in building a defeasible reasoner was to have an inference-engine for a rational agent capable of getting around in the real world.
- This requires it to engage in defeasible reasoning regarding at least:
 - the results of perception, which are not always accurate
 - inductive generalizations, which can be wrong because they are based on a restricted sample
 - conclusions drawn on the basis of high probabilities
 - sophisticated cognizers must reason defeasibly about time, projecting conclusions drawn at one time forward to future times.
 - sophisticated cognizers must be able to reason about the causal consequences of both their own actions and other events in the world. This reasoning turns out to be defeasible as well.

"Perceiving and reasoning about a changing world", *Comp. Intelligence*, Nov., 1998.

Perception

- Perception provides the source of new information about the world.
- The agent's perceptual apparatus provide *percepts*, which I take to have dates and propositional contents.

PERCEPTION

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

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.
- 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.

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 an undercutting defeater for PERCEPTION as a reason of strength $> r$.

The Projectibility Constraint

- 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 Projectibility Constraint

- The difficulty is that if I am in circumstances of type C_1 , then I am also in the disjunctive circumstances ($C_1 \vee C_2$).
- Furthermore, the probability of being in circumstances of type C_2 given that one is in circumstances of type ($C_1 \vee C_2$) 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 ($C_1 \vee C_2$).
- Consequently, if ($C_1 \vee C_2$) 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.

Grue

- Projectibility constraints were first noted by Nelson Goodman (1955).
- The Nicod Principle:
 - For any predicates A and B , observing a sample of A 's all of which are B 's is a defeasible reason for believing that all A 's are B 's.
- Goodman's counterexample:
 - "x is grue" means "either x is green and first observed before the year 2000, or x is blue and not first observed before the year 2000"
 - All the emeralds we have observed have been green, and therefore grue.
 - By the Nicod principle, that gives us reasons for thinking that all emeralds are green, and also all emeralds are grue.
 - But that entails that no new emeralds will be observed beginning with the year 2000, which is absurd.
 - The conclusion is that "grue" is not appropriate for use in induction — it is not projectible.

The Projectibility Constraint

- 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).
- 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.
- 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.
- 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.

Discounted Perception

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. 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.

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.

"R-at-t & prob(P/R & (I have a percept of P)) $\leq s$ "

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 defeater for DISCOUNTED-PERCEPTION.

Reason-schemas

- **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.
- **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.

Reason-schemas

- **Backwards-premises** are data-structures encoding the following information:
 - bp-formula
 - bp-kind
 - bp-condition — an optional constraint that must be satisfied by an inference-node for it to instantiate this premise.
- **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.

Reason-Defining Macros

```
(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 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))

(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))
```

Implementing PERCEPTION

PERCEPTION

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

(def-forwards-reason PERCEPTION

```
:forwards-premises "(p at time)" (:kind :percept)
:conclusion "(p at time)"
:variables p time
:defeasible? t
:strength .98)
```

The strength of .98 has been chosen arbitrarily.

Implementing PERCEPTUAL-RELIABILITY

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 an undercutting defeater for PERCEPTION as a reason of strength $> r$.

(def-backwards-undercutter PERCEPTUAL-RELIABILITY

:defeatee perception

:forwards-premises

"((the probability of p given ((I have a percept with content p) & R)) $\leq s$)"

(:condition (and ($s < 0.99$) (projectible R)))

:backwards-premises

"(R at time)"

:variables p time R s

:defeasible? t

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.
 - 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.
 - 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.
 - 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.

Temporal Projection

- To say that a property is stable is to say that there is a high probability ρ that if an object has the property at time t then it still has it at $t+1$.
- More generally, the probability that an object has the property at $t+\Delta t$ given that it has the property at t is $\rho^{\Delta t}$.
- An agent must assume defeasibly that that world tends to be stable to degree ρ where ρ is a constant (the temporal decay factor).

Temporal Projection

A probability of $\rho^{\Delta t}$ corresponds to a reason-strength of $2 \cdot (\rho^{\Delta t} - .5)$.

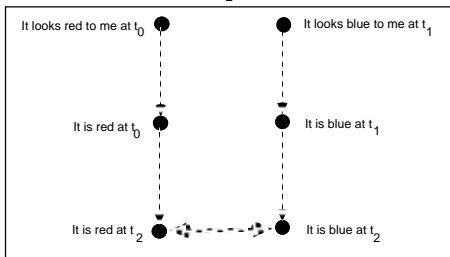
$\rho^{\Delta t} > .5$ iff $\Delta t > \log(.5)/\log(\rho)$.

So we need a principle something like the following:

When $\Delta t > \log(.5)/\log(\rho)$, believing P-at-t is a defeasible reason of strength $2 \cdot (\rho^{\Delta t} - .5)$ for the agent to believe P-at-($t+\Delta t$).

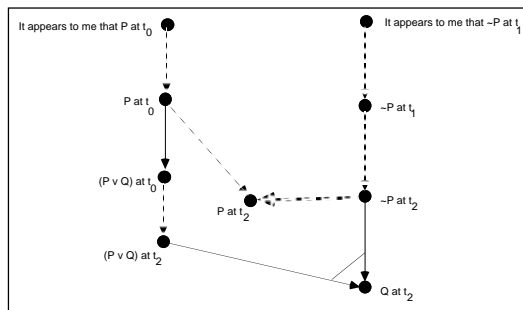
Perceptual Updating

- Suppose an object looks red at time t_0 and blue at a later time t_1 , and we know nothing else about it. We should conclude defeasibly that it has changed color, and so at a still later time t_2 it will still be blue.



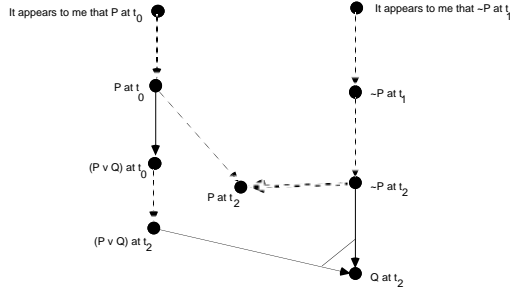
Temporal Projectibility

But this reasoning is intuitively wrong.



Temporal Projectibility

The disjunction is problematic. We rule it out by requiring "temporal-projectibility".



Temporal Projection

TEMPORAL-PROJECTION

If P is temporally-projectible and $\Delta t > \log(.5)/\log(\rho)$, believing P -at- t is a defeasible reason of strength $2 \cdot (\rho^{\Delta t} - .5)$ for the agent to believe P -at- $(t+\Delta t)$.

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.

Implementing Temporal Projection

TEMPORAL-PROJECTION

If P is temporally-projectible and $\Delta t > \log(.5)/\log(\rho)$, believing P -at- t is a defeasible reason of strength $2 \cdot (\rho^{\Delta t} - .5)$ for the agent to believe P -at- $(t+\Delta t)$.

- 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 $\Delta t > 0$.

Implementing Temporal Projection

TEMPORAL-PROJECTION

If P is temporally-projectible and $\Delta t > \log(.5)/\log(\rho)$, believing P -at- t is a defeasible reason of strength $2 \cdot (\rho^{\Delta t} - .5)$ for the agent to believe P -at- $(t+\Delta t)$.

```
(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)"
  "((time* - time0) < log(.5)/log('temporal-decay'))"
:variables p time0 time
:defeasible? T
:strength (- (* 2 (expt 'temporal-decay' (- time time0))) 1))
```

This requires the reasoner to engage in explicit arithmetical reasoning.

Implementing Temporal Projection

We can instead let LISP do the arithmetical computation in the background.

```
(def-backwards-reason TEMPORAL-PROJECTION
:conclusion "(p at time)"
:condition (and (temporally-projectible p) (numberp time))
:forwards-premises
  "(p at time0)"
  (:condition (and (time0 < time*)
                  ((time* - time0) < log(.5)/log('temporal-decay'))))

:backwards-premises
  "(time0 < time)"
  "((time* - time0) < log(.5)/log('temporal-decay'))"

:variables p time0 time
:defeasible? T
:strength (- (* 2 (expt 'temporal-decay' (- time time0))) 1))
```

PROBABILISTIC-DEFEAT-FOR-TEMPORAL-PROJECTION

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.

```
(def-backwards-undercutter
PROBABILISTIC-DEFEAT-FOR-TEMPORAL-PROJECTION
:defeater 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)
```

Illustration of OSCAR'S Defeasible Reasoning

This is the Perceptual-Updating Problem.

First, Fred looks red to me.

Later still, Fred looks blue to me.

What should I conclude about the color of Fred?

see OSCAR do it

Time = 0

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

What color is Fred?

given

Time = 1

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

(It appears to me that the color of Fred is red) at 1

What color is Fred?

Percept acquired

Time = 2

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

(It appears to me that the color of Fred is red) at 1

The color of Fred is red

What color is Fred?

by PERCEPTION

Time = 3

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

(It appears to me that the color of Fred is red) at 1

The color of Fred is red

-The color of Fred is red

What color is Fred?

Interest in rebutter

Time = 4

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

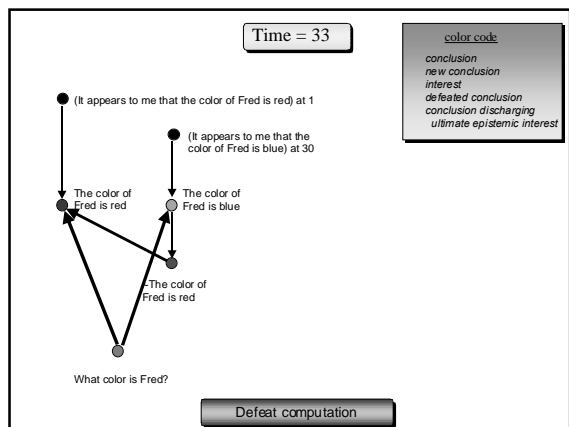
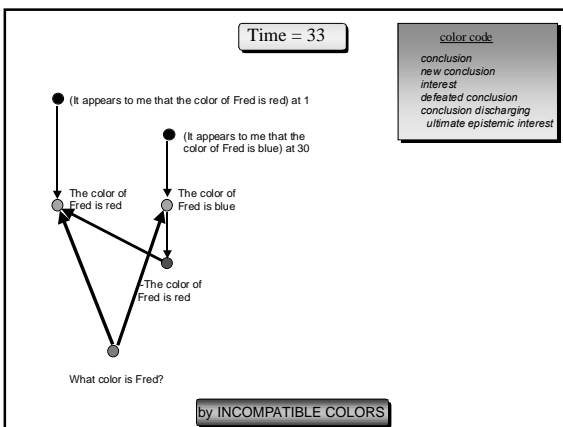
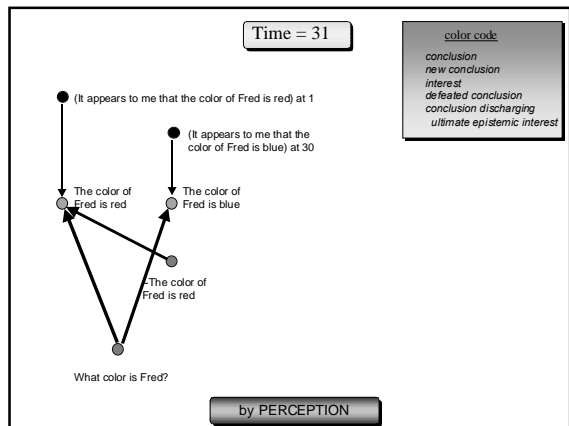
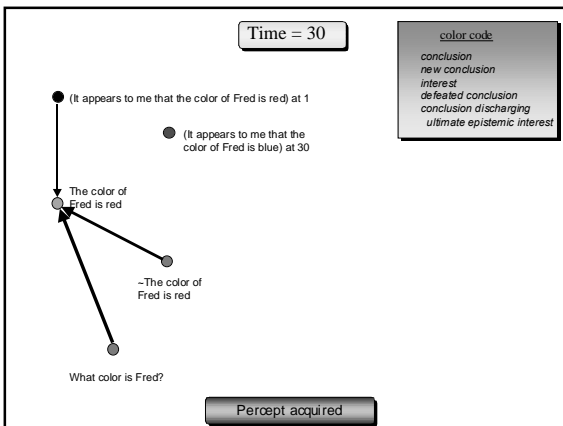
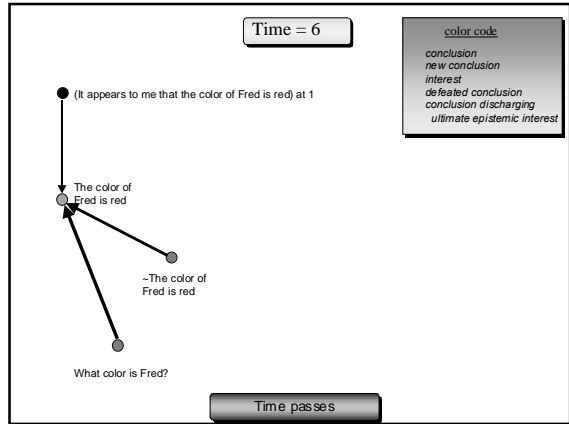
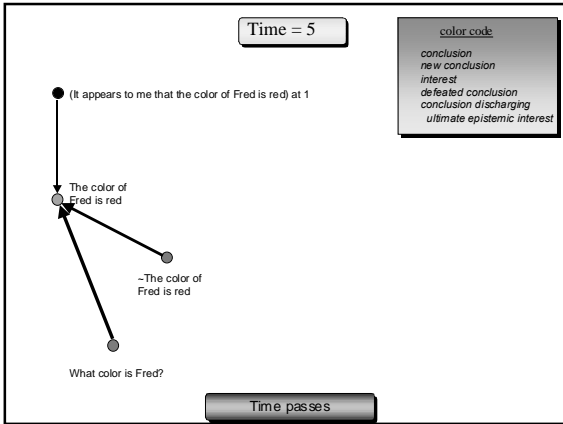
(It appears to me that the color of Fred is red) at 1

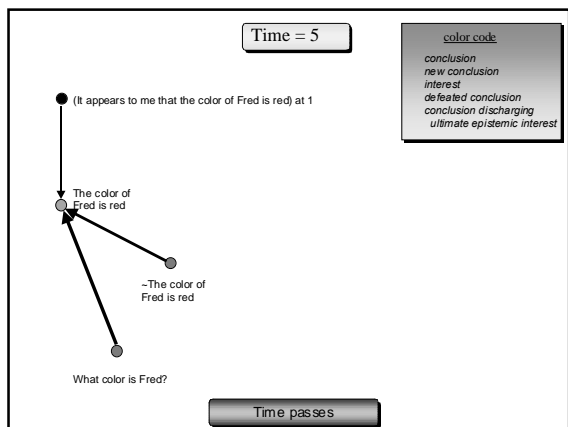
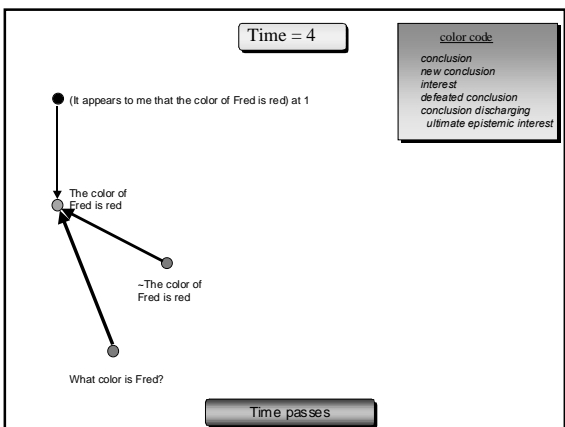
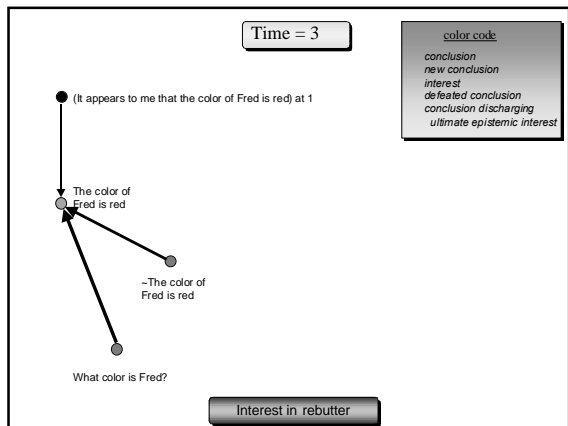
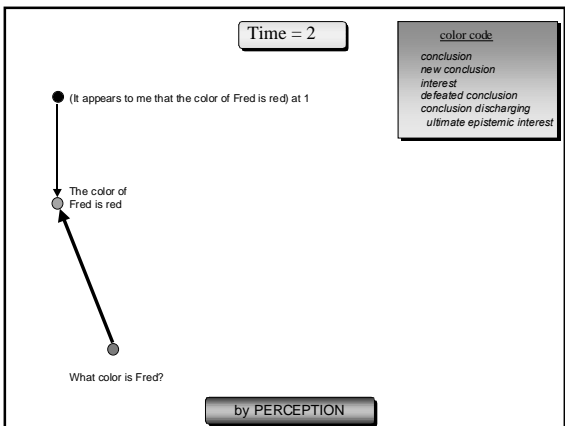
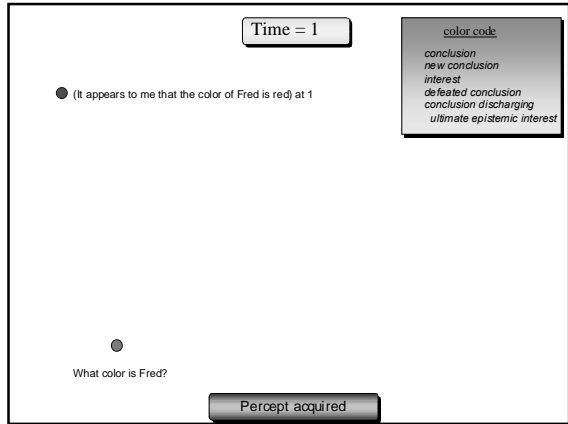
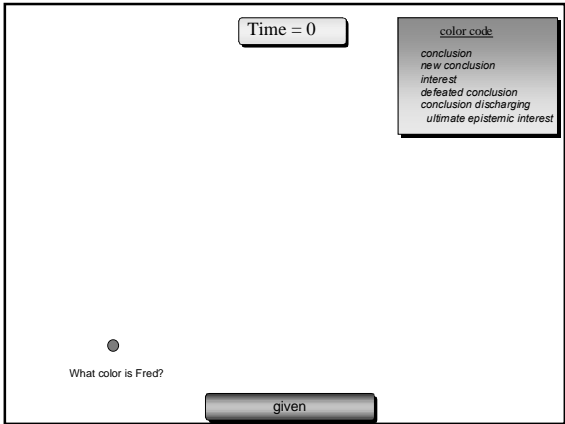
The color of Fred is red

-The color of Fred is red

What color is Fred?

Time passes





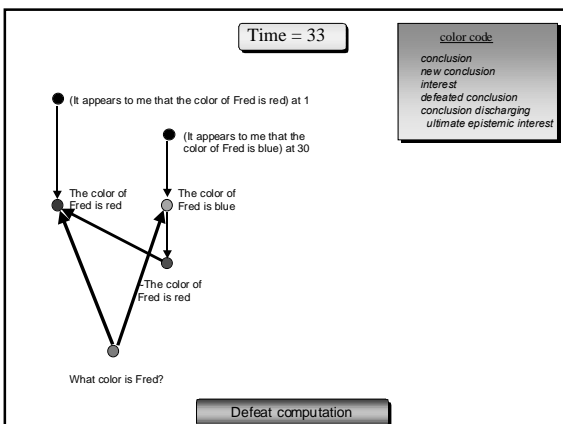
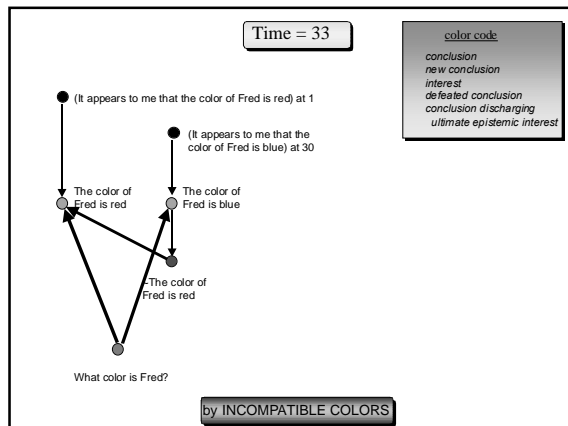
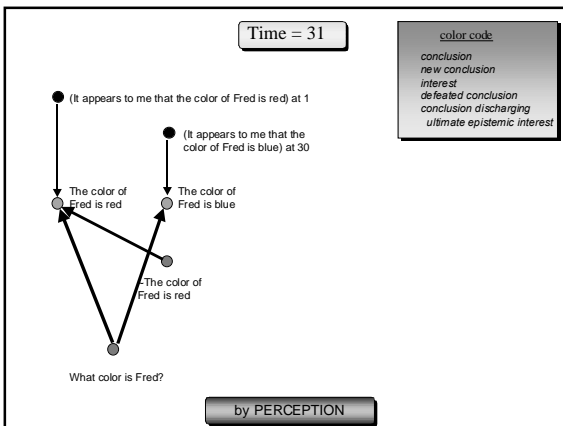
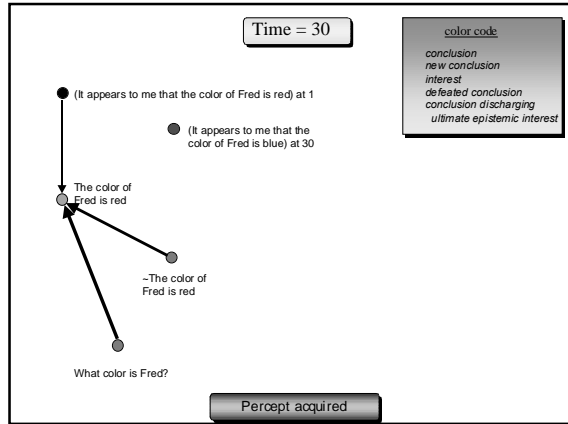
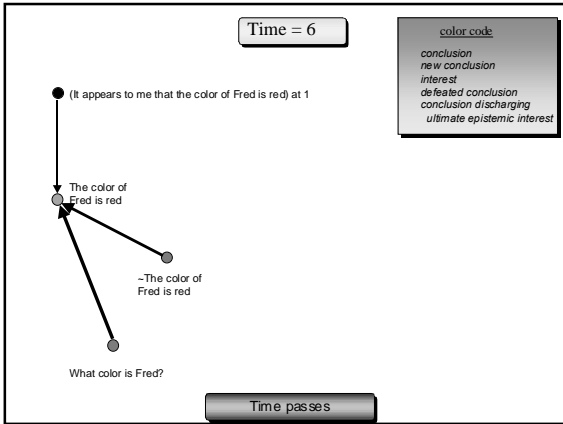


Illustration of OSCAR'S Defeasible Reasoning

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 Fred's appearing blue is not a reliable indicator of Fred's being blue when I am wearing blue-tinted glasses.

What should I conclude about the color of Fred?

see OSCAR do it

