THE OSCAR PROJECT

Rational Cognition in OSCAR

John L. Pollock
Department of Philosophy
University of Arizona
Tucson, Arizona 85721
pollock@arizona.edu
http://www.u.arizona.edu/~pollock

Part One: Evaluating Agent Architectures

- Stuart Russell: "rational agents are those that *do the right thing*".
  - The problem of designing a rational agent then becomes the problem of figuring out what the right thing is.
  - There are two approaches to the latter problem, depending upon the kind of agent we want to build.
- *Anthropomorphic Agents* — those that can help human beings rather directly in their intellectual endeavors.
  - These endeavors consist of decision making and data processing.
  - An agent that can help humans in these enterprises must make decisions and draw conclusions that are rational by human standards of rationality.
- *Goal-Oriented Agents* — those that can carry out certain narrowly-defined tasks in the world.
  - Here the objective is to get the job done, and it makes little difference how the agent achieves its design goal.

Evaluating Goal-Oriented Agents

- If the design goal of a goal-oriented agent is sufficiently simple, it may be possible to construct a metric the measures how well an agent achieves it.
  - Then the natural way of evaluating an agent architecture is in terms of the expected-value of that metric.
  - An ideally rational goal-oriented agent would be one whose design maximizes that expected-value.
  - The recent work on *bounded-optimality* (Russell and Subramanian; Horvitz; Zilberstein and Russell, etc.) derives from this approach to evaluating agent architectures.
- This approach will only be applicable in cases in which it is possible to construct a metric of success.
- If the design goal is sufficiently complex, that will be at least difficult, and perhaps impossible.

Evaluating Anthropomorphic Agents

- Here it is the individual decisions and conclusions of the agent that we want to be rational.
- In principle, we could regard an anthropomorphic agent as a special case of a goal-oriented agent, where now the goal is to make rational decisions and draw rational conclusions, but it is doubtful that we can produce a metric that measures the degree to which such an agent is successful in achieving these goals.
- Even if we could construct such a metric, it would not provide an analysis of rationality for such an agent, because the metric itself must presuppose prior standards of rationality governing the individual cognitive acts of the agent being evaluated.

Evaluating Anthropomorphic Agents

- In AI it is often supposed that the standards of rationality that apply to individual cognitive acts are straightforward and unproblematic:
  - Bayesian probability theory provides the standards of rationality for beliefs;
  - classical decision theory provides the standards of rationality for practical decisions.

- It may come as a surprise then that most philosophers reject Bayesian epistemology, and I believe there are compelling reasons for rejecting classical decision theory.

Bayesian Epistemology

- Bayesian epistemology asserts that the degree to which a rational agent is justified in believing something can be identified with a subjective probability. Belief updating is governed by conditionalization on new inputs.
  - There is an immense literature on this. Some of the objections to it are summarized in Pollock and Cruz, *Contemporary Theories of Knowledge*, 2nd edition (Rowman and Littlefield, 1999).
- Perhaps the simplest objection to Bayesian epistemology is that it implies that an agent is always rational in believing any truth of logic, because any such truth has probability 1.
  - However, this conflicts with common sense. Consider a complex tautology like $[P \leftrightarrow (Q \,\&\, \sim P)] \rightarrow \sim Q$. If one of my logic students picks this out of the air and believes it for no reason, we do not regard that as rational.
  - He should only believe it if he has good reason to believe it. In other words, rational belief requires reasons, and that conflicts with Bayesian epistemology.

Classical Decision Theory

- Classical decision theory has us choose acts one at a time on the basis of their expected values.

- It is *courses of action*, or *plans*, that must be evaluated decision-theoretically, and individual acts become rational by being prescribed by rationally adopted plans. (See Pollock, *Cognitive Carpentry*, chapter 5, MIT Press, 1995.)

- Furthermore, I will argue below that we cannot just take classical decision theory intact and apply it to plans. A plan is not automatically superior to a competitor just because it has a higher expected value.

Anthopomorphic Agents and Procedural Rationality

- The design of an anthropomorphic agent requires a general theory of rational cognition.

- The agent's cognition must be rational by human standards. Cognition is a *process*, so this generates an essentially procedural concept of rationality.

  – Many AI researchers have followed Herbert Simon in rejecting such a procedural account, endorsing instead a satisficing account based on goal-satisfaction, but that is not applicable to anthropomorphic agents.

Procedural Rationality

- We do not necessarily want an anthropomorphic agent to model human cognition exactly.
  - We want it to draw rational conclusions and make rational decisions, but it need not do so in exactly the same way humans do it. How can we make sense of this?

- Stuart Russell (following Herbert Simon) suggests that the appropriate concept of rationality should only apply to the *ultimate results* of cognition, and not the course of cognition.

Procedural Rationality

- A conclusion or decision is *warranted* (relative to a system of cognition) iff it is endorsed "in the limit".
  - i.e., there is some stage of cognition at which it becomes endorsed and beyond which the endorsement is never retracted.

- We might require an agent architecture to have the same theory of warrant as human rational cognition.
  - This is to evaluate its behavior in the limit.
  - An agent that drew conclusions and made decisions at random for the first ten million years, and then started over again reasoning just like human beings would have the same theory of warrant, but it would not be a good agent design.
  - This is a problem for any assessment of agents in terms of the results of cognition in the limit.

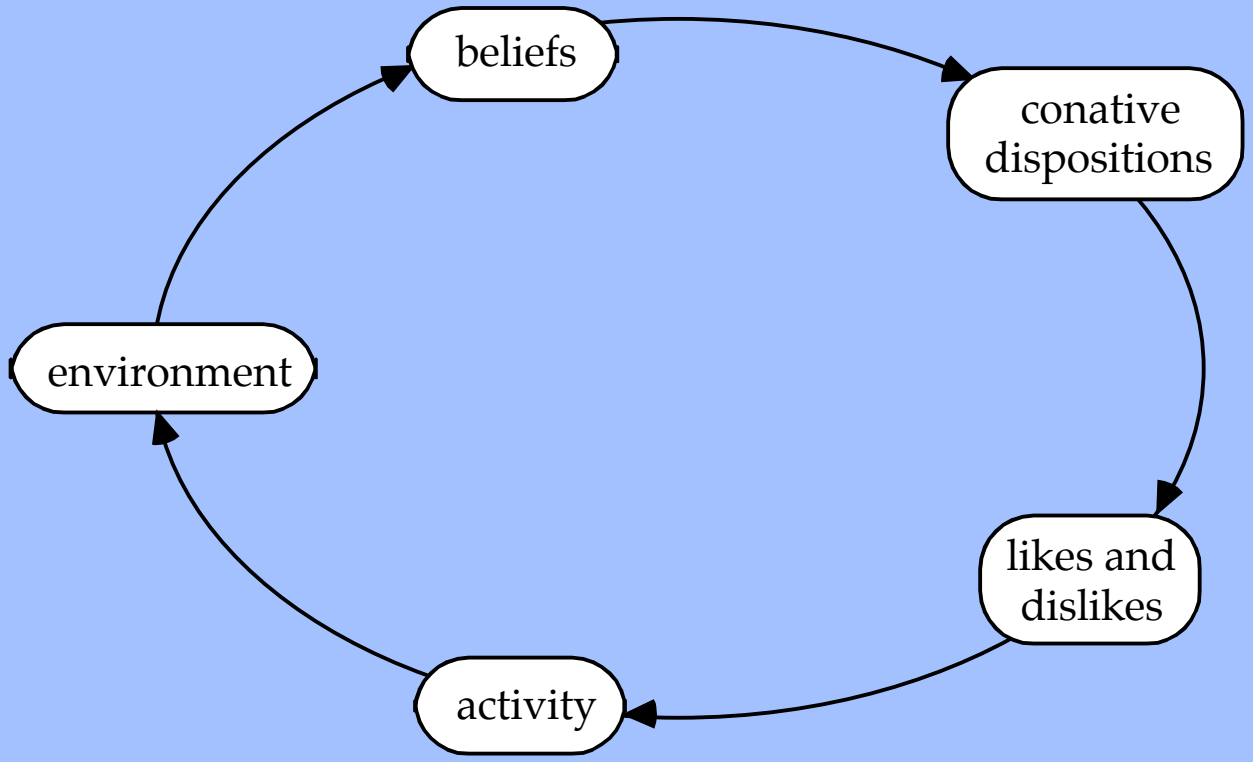
Procedural Rationality

- It looks like the best we can do is require that the agent's reasoning never strays very far from the course of human reasoning.
  - If humans will draw a conclusion within a certain number of steps, the agent will do so within a "comparable" number of steps, and if a human will retract the conclusion within a certain number of further steps, the agent will do so within a "comparable" number of further steps. This is admittedly vague.
  - We might require that the worst-case difference be polynomial in the number of steps, or something like that. However, this proposal does not handle the case of the agent that draws conclusions randomly for the first ten million years.
- I am not going to endorse a solution to this problem. I just want to call attention to it, and urge that whatever the solution is, it seems reasonable to think that the kind of architecture I am about to describe satisfies the requisite constraints.

Part Two:
The OSCAR Architecture

- **OSCAR is an architecture for rational agents based upon an evolving philosophical theory of rational cognition.**
  - **The general architecture is described in *Cognitive Carpentry* (MIT Press, 1995).**
  - **Related papers can be downloaded from http://www.u.arizona.edu/~pollock**

Schematic Rational Cognition



The Doxastic-Conative Loop

An Architecture for Rational Cognition

- **Epistemic cognition — about what to believe.**
- **Practical cognition — about what to do.**
  - **Epistemic cognition is skeptical; practical cognition is credulous.**

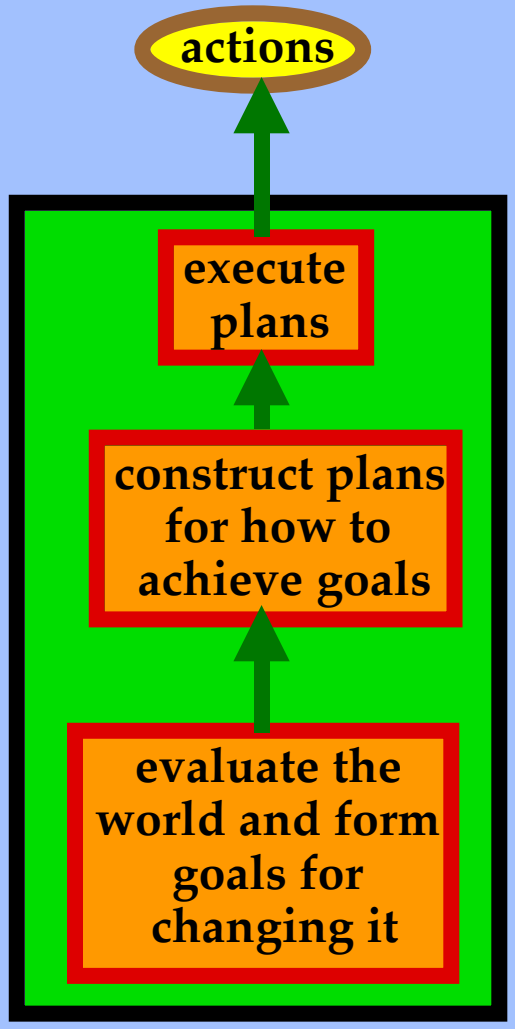
The Pre-eminence of Practical Cognition

- **Most work on rational agents in AI has focussed on practical cognition rather than epistemic cognition, and for good reason.**
- **The whole point of an agent is *to do something*, to interact with the world, and such interaction is driven by practical cognition.**
- **From this perspective, epistemic cognition is subservient to practical cognition.**

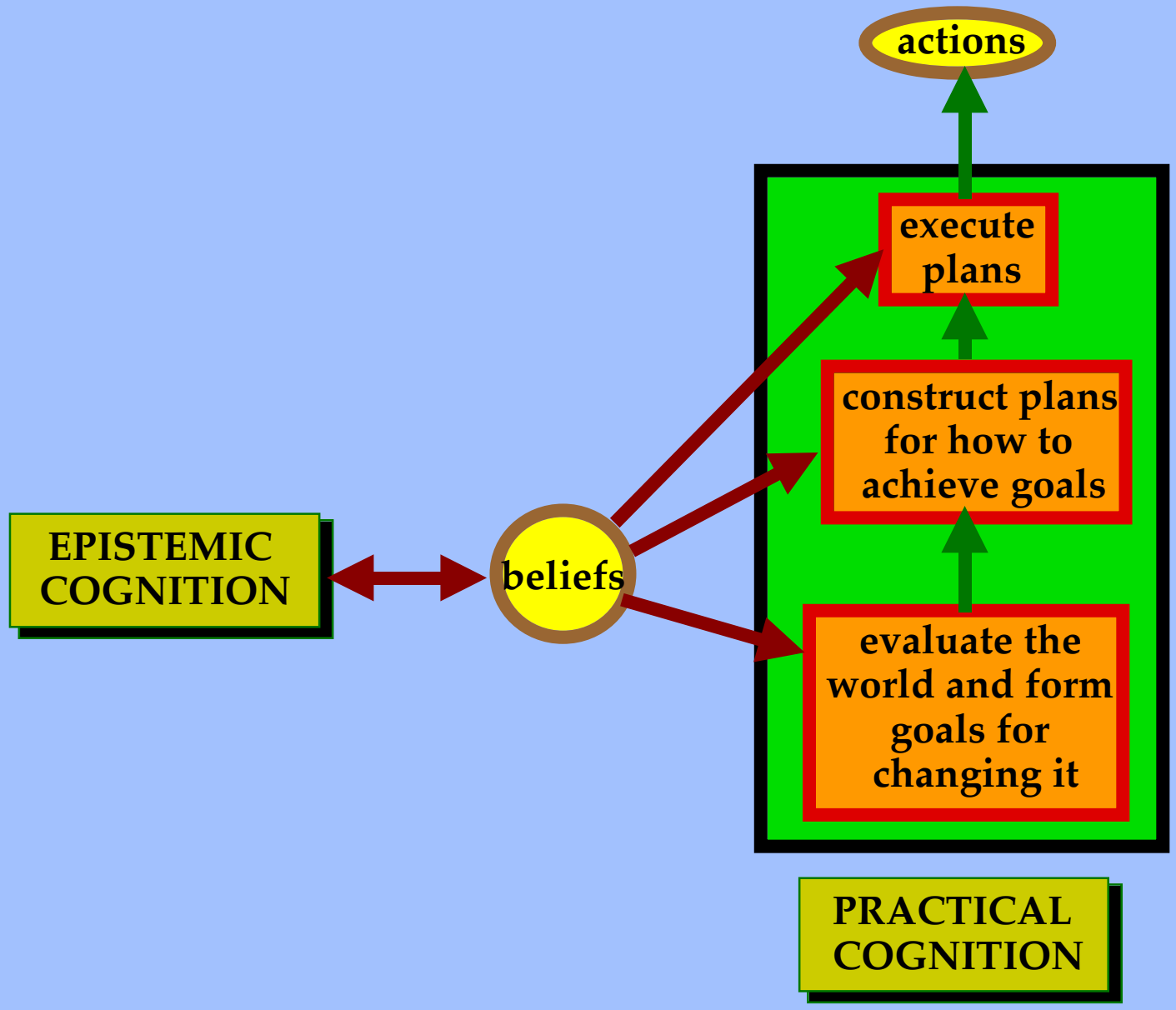
The Importance of Epistemic Cognition

- **The OSCAR architecture differs from most agent architectures in that, although it is still practical cognition that directs the agent's interaction with the world, most of the work in rational cognition is performed by epistemic cognition.**
  - **Practical cognition evaluates the world (as represented by the agent's beliefs), and then poses queries concerning how to make it better.**
  - **These queries are passed to epistemic cognition, which tries to answer them.**
  - **Competing plans are evaluated and selected on the basis of their expected utilities, but those expected utilities are again computed by epistemic cognition.**
  - **Finally, plan execution generally requires a certain amount of monitoring to verify that things are going as planned, and that monitoring is again carried out by epistemic cognition.**
  - **In general, choices are made by practical cognition, but the information on which the choices are based is the product of epistemic cognition, and the bulk of the work in rational cognition goes into providing that information.**

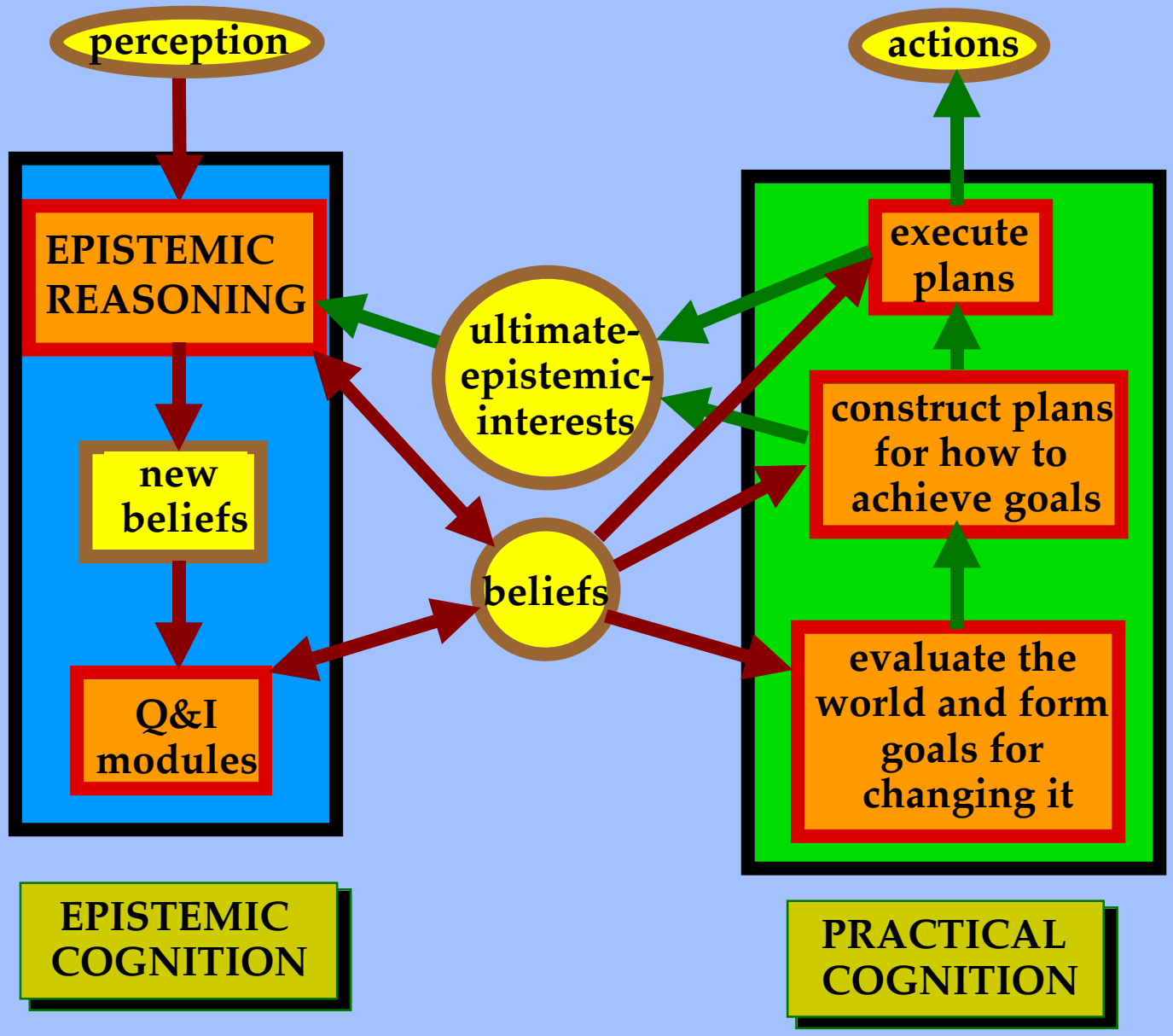
Practical Cognition



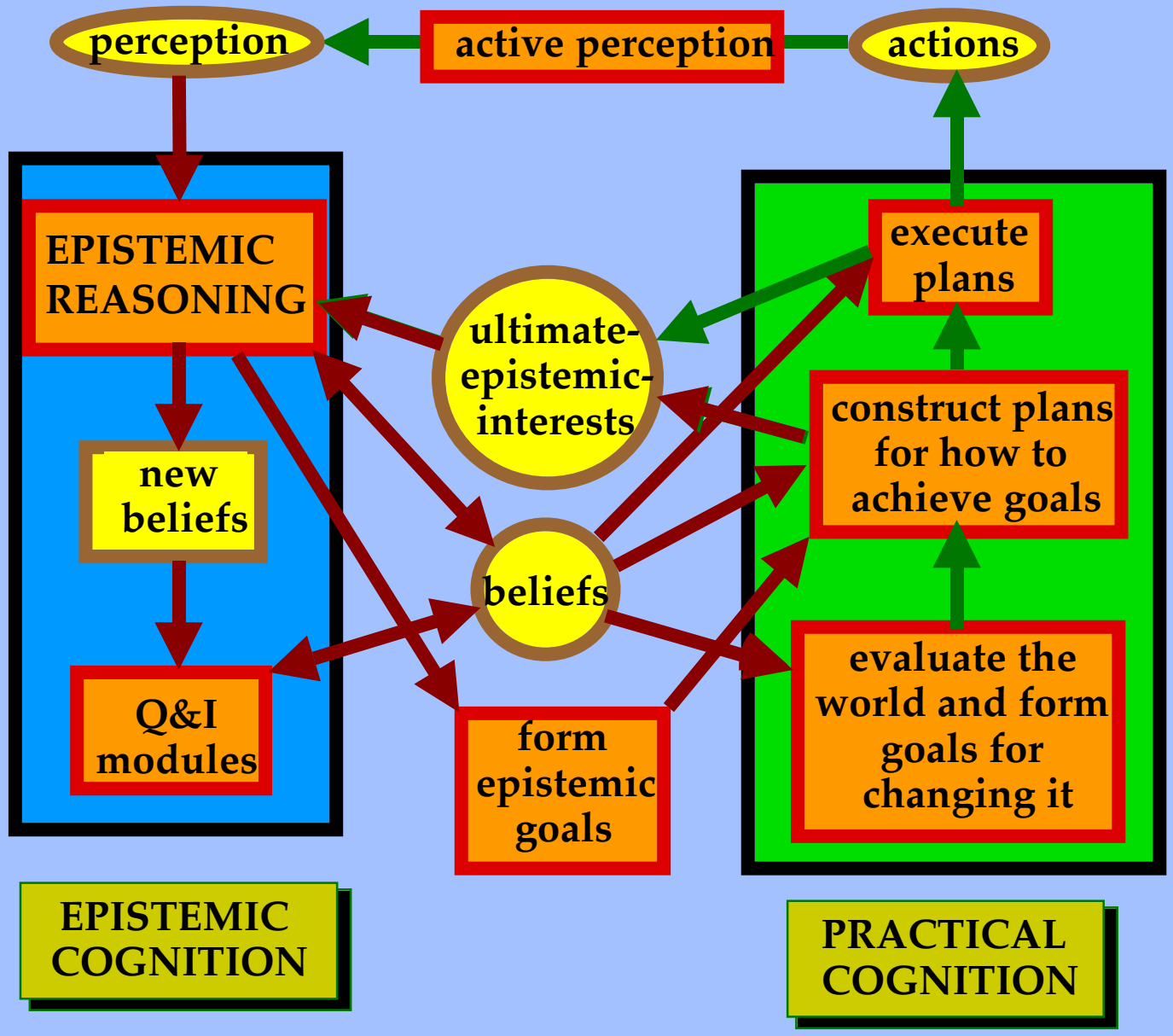
The Subservient Role of Epistemic Cognition



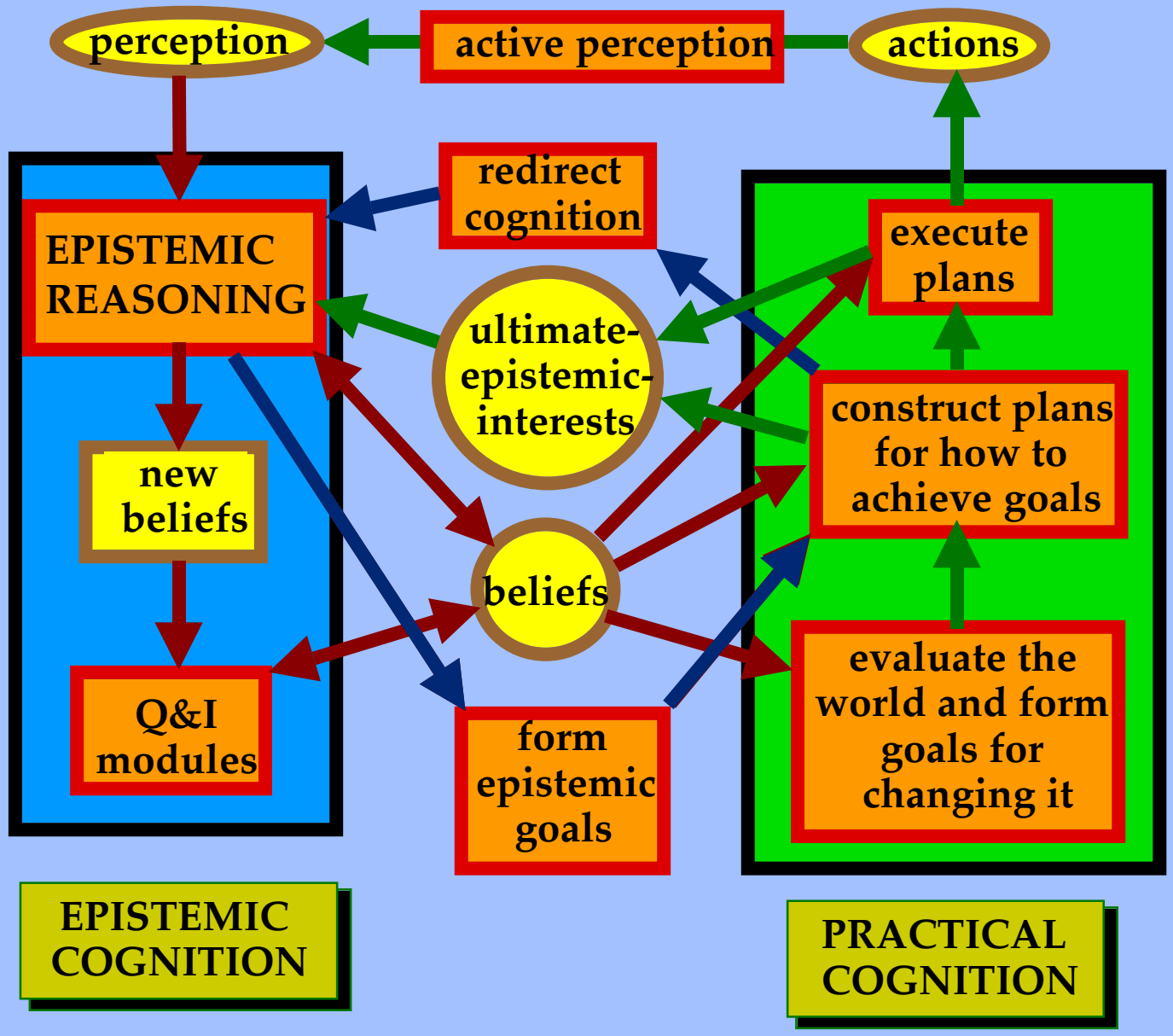
The Basic Interface



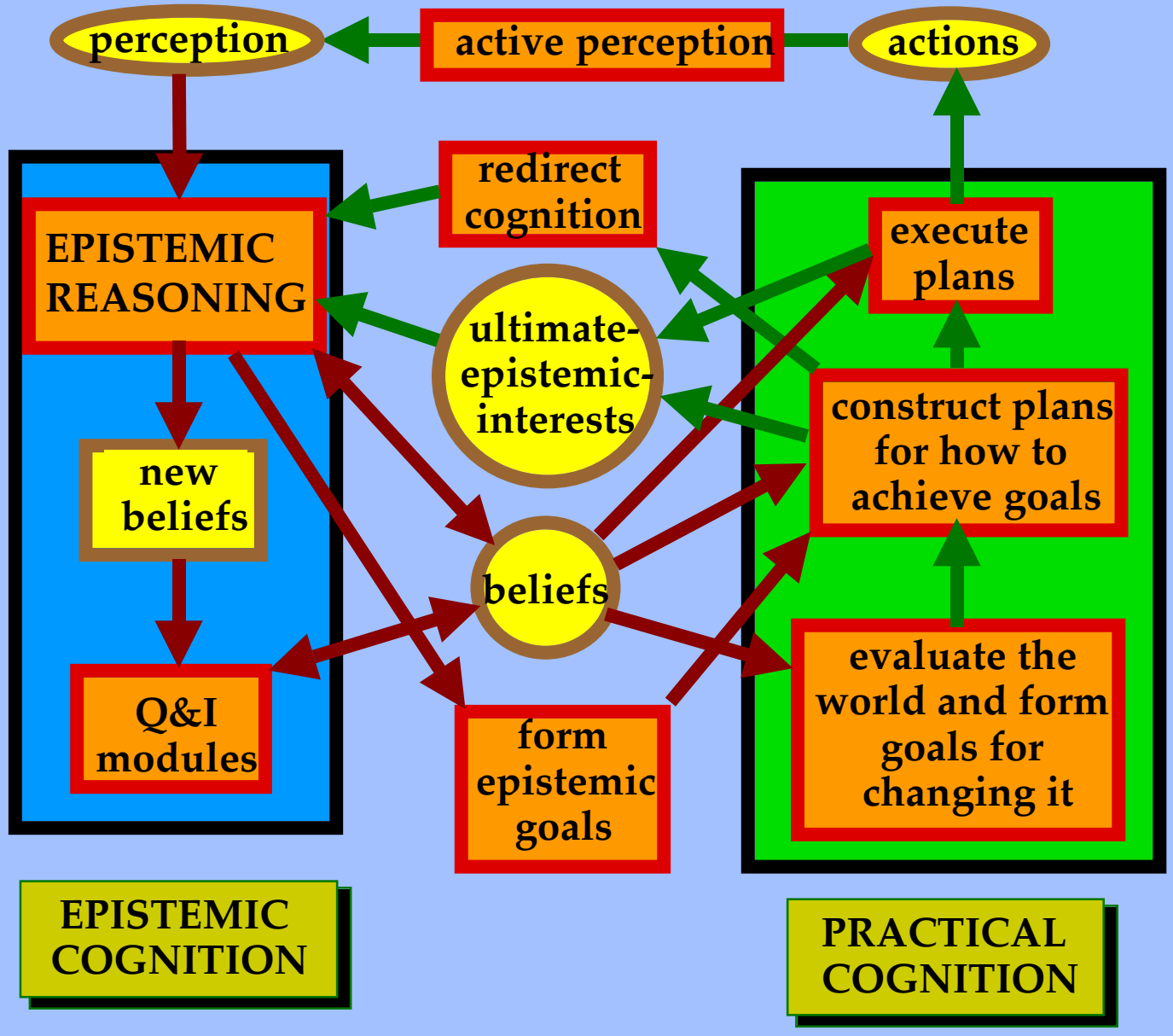
Empirical Investigation



Reflexive Cognition



An Architecture for Rational Cognition

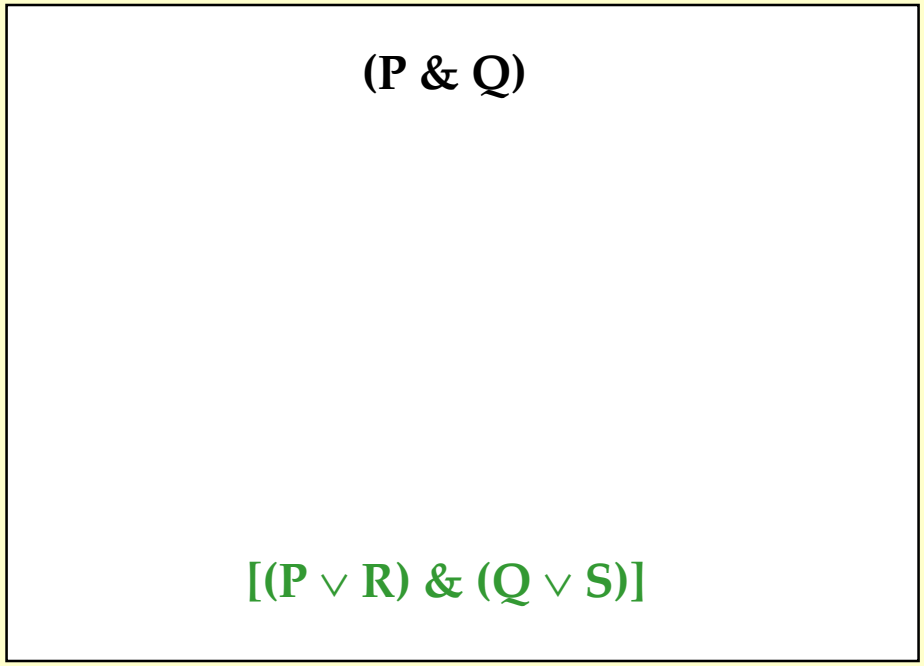


Part Three:
Epistemic Reasoning

- **Epistemic reasoning is driven by both input from perception and queries passed from practical cognition.**
- **The way in which epistemic interests effect the course of cognition is by initiating backward reasoning.**
- **Example of bidirectional reasoning**

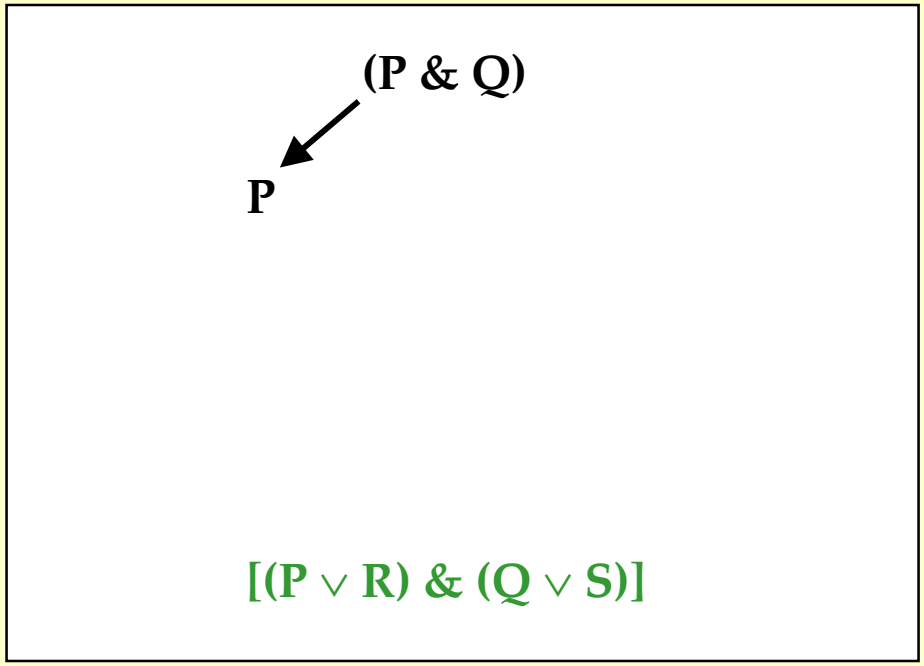
Epistemic Reasoning

- **Epistemic reasoning is driven by both input from perception and queries passed from practical cognition.**
- **The way in which epistemic interests effect the course of cognition is by initiating backward reasoning.**
- **Example of bidirectional reasoning**



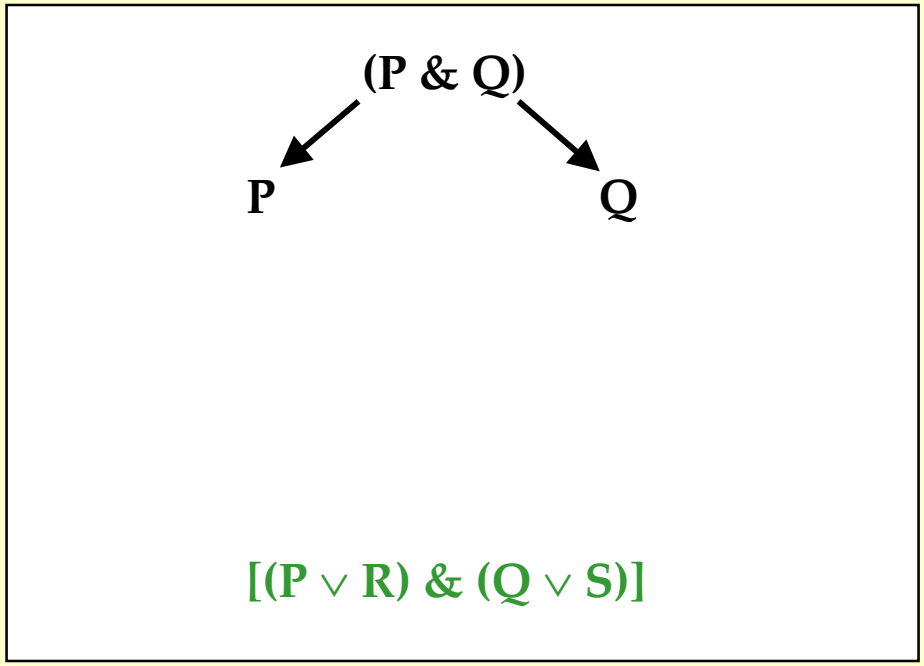
Epistemic Reasoning

- Epistemic reasoning is driven by both input from perception and queries passed from practical cognition.
- The way in which epistemic interests effect the course of cognition is by initiating backward reasoning.
- Example of bidirectional reasoning



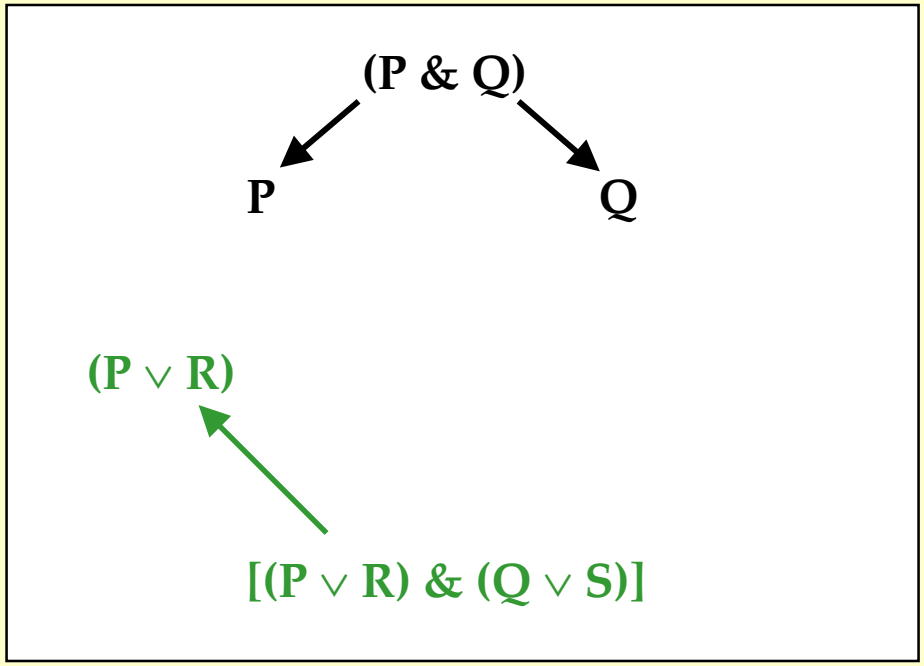
Epistemic Reasoning

- Epistemic reasoning is driven by both input from perception and queries passed from practical cognition.
- The way in which epistemic interests effect the course of cognition is by initiating backward reasoning.
- Example of bidirectional reasoning



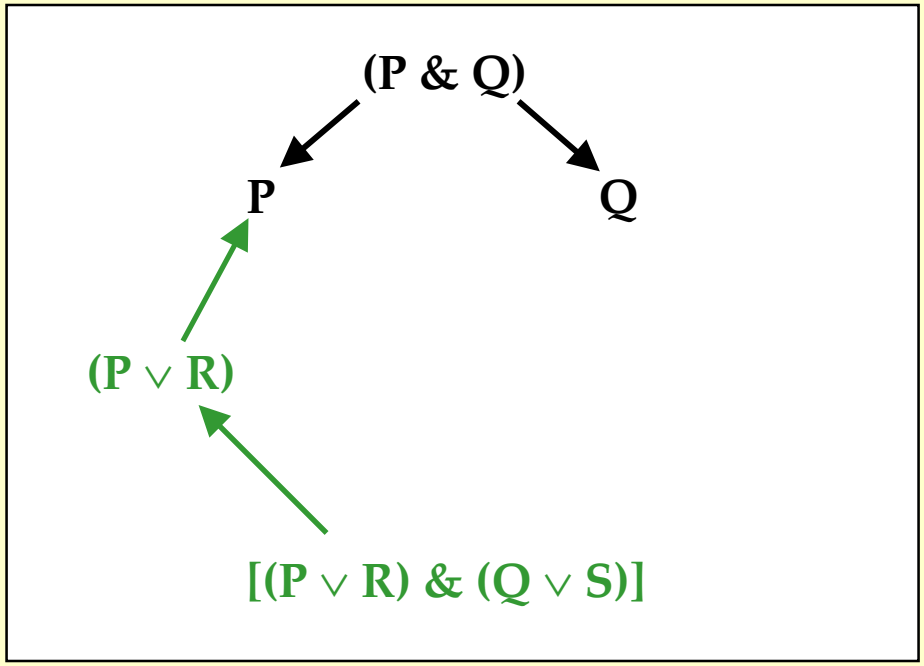
Epistemic Reasoning

- Epistemic reasoning is driven by both input from perception and queries passed from practical cognition.
- The way in which epistemic interests effect the course of cognition is by initiating backward reasoning.
- Example of bidirectional reasoning



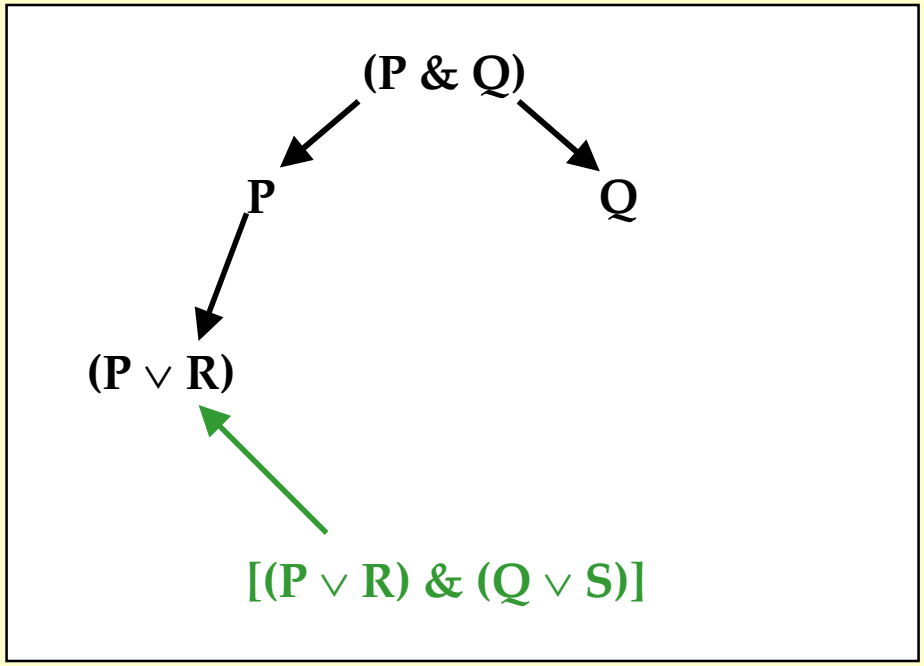
Epistemic Reasoning

- Epistemic reasoning is driven by both input from perception and queries passed from practical cognition.
- The way in which epistemic interests effect the course of cognition is by initiating backward reasoning.
- Example of bidirectional reasoning



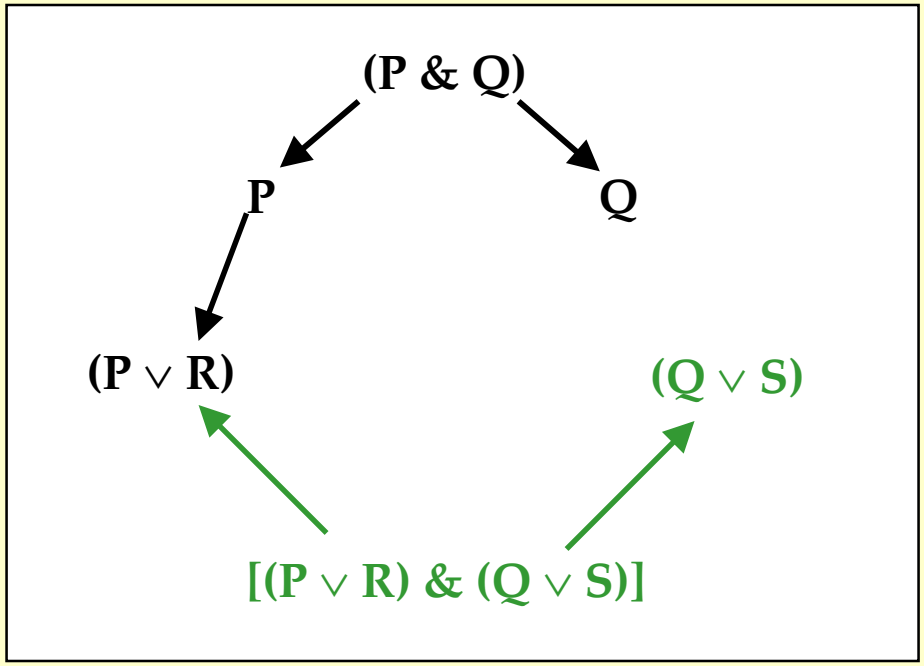
Epistemic Reasoning

- Epistemic reasoning is driven by both input from perception and queries passed from practical cognition.
- The way in which epistemic interests effect the course of cognition is by initiating backward reasoning.
- Example of bidirectional reasoning



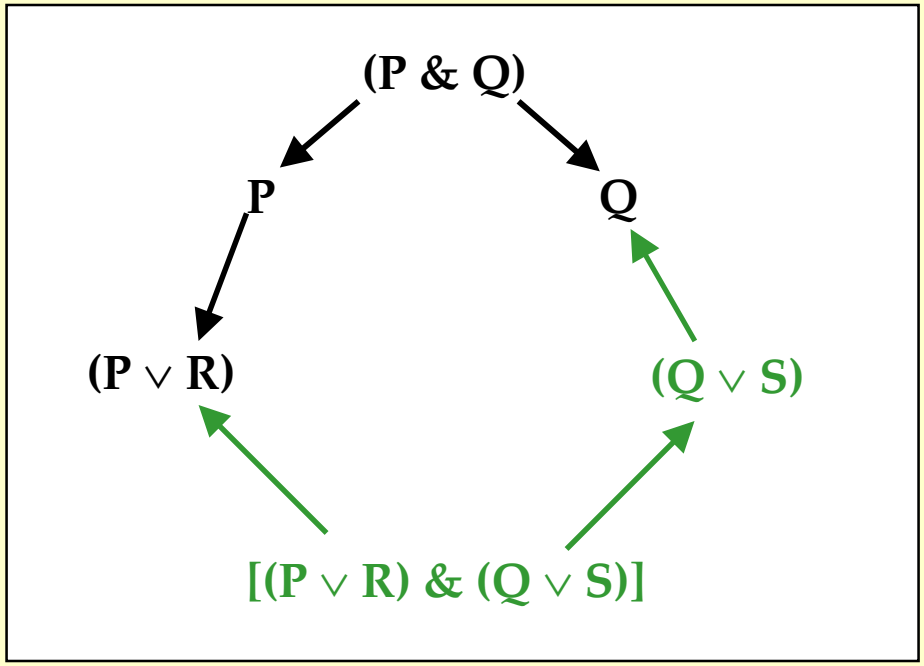
Epistemic Reasoning

- Epistemic reasoning is driven by both input from perception and queries passed from practical cognition.
- The way in which epistemic interests effect the course of cognition is by initiating backward reasoning.
- Example of bidirectional reasoning



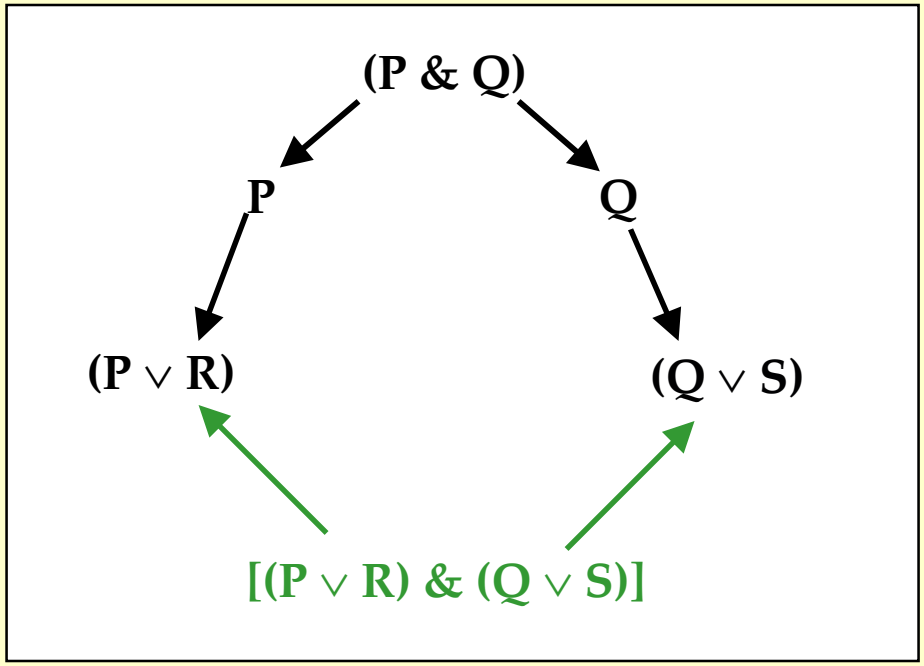
Epistemic Reasoning

- Epistemic reasoning is driven by both input from perception and queries passed from practical cognition.
- The way in which epistemic interests effect the course of cognition is by initiating backward reasoning.
- Example of bidirectional reasoning



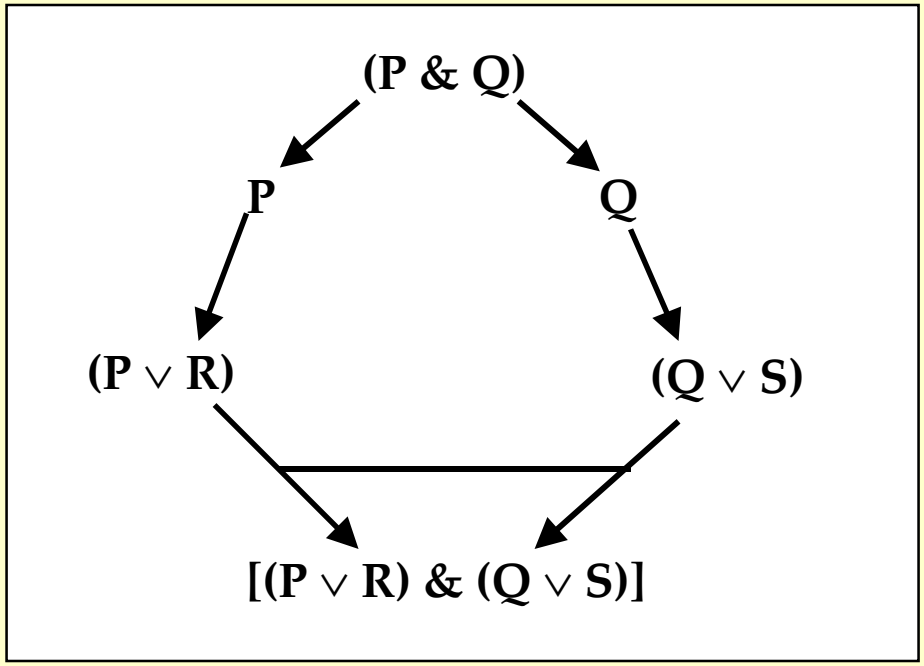
Epistemic Reasoning

- Epistemic reasoning is driven by both input from perception and queries passed from practical cognition.
- The way in which epistemic interests effect the course of cognition is by initiating backward reasoning.
- Example of bidirectional reasoning



Epistemic Reasoning

- **Epistemic reasoning is driven by both input from perception and queries passed from practical cognition.**
- **The way in which epistemic interests effect the course of cognition is by initiating backward reasoning.**
- **Example of bidirectional reasoning**



Epistemic Reasoning

- **OSCAR's reasoning is in the style of "natural deduction".**
- **Reason-schemas are segregated into backward and forward schemas.**
  - **forward schemas lead from conclusions to conclusions**

    From (P & Q), infer P.
  - **backward schemas lead from interests to interests**

    From P, Q infer (P & Q).
- **OSCAR is surprisingly efficient as a deductive reasoner.**
  - **In a recent comparison with the the highly respected OTTER resolution-refutation theorem prover on a set of 163 problems chosen by Geoff Sutcliffe from the TPTP theorem proving library:**
    - **» OTTER failed to get 16**
    - **» OSCAR failed to get 3**
    - **» On problems solved by both theorem provers, OSCAR (written in LISP) was on the average 40 times faster than OTTER (written in C)**

Defeasible Reasoning

- **Deductive reasoning guarantees the truth of the conclusion given the truth of the premises.**
- **Defeasible reasoning makes it reasonable to accept the conclusion, but does not provide an irrevocable guarantee of its truth.**
  - **conclusions supported defeasibly might have to be withdrawn later in the face of new information.**
- **All sophisticated epistemic cognizers must reason defeasibly:**
  - **perception is not always accurate**
  - **inductive reasoning must be defeasible**
  - **sophisticated cognizers must reason defeasibly about time, projecting conclusions drawn at one time forwards to future times.**
  - **it will be argued below that certain aspects of planning must be done defeasibly**

Defeasible Reasoning

- **Defeasible reasoning is performed using defeasible reason-schemas.**
- **What makes a reason-schema defeasible is that it can be defeated by having defeaters.**
- **Two kinds of defeaters**
  - **Rebutting defeaters attack the conclusion of the inference**
  - **Undercutting defeaters attack the connection between the premise and the conclusion.**
    - » **An undercutting defeater for an inference from P to Q is a reason for believing it false that P would not be true unless Q were true. This is symbolized (P ⊗ Q).**
    - » **More simply, (P ⊗ Q) can be read "P does not guarantee Q".**
  - **Example: something's looking red gives us a defeasible reason for thinking it is red.**
    - » **A reason for thinking it isn't red is a rebutting defeater.**
    - » **It's being illuminated by red lights provides an undercutting defeater.**

Defeasible Reasoning

- **Reasoning defeasibly has two parts**
  - **constructing arguments for conclusions**
  - **evaluating defeat statuses, and computing degrees of justification, given the set of arguments constructed**
    - » **OSCAR does this by using a defeat-status computation described in *Cognitive Carpentry*.**
    - » ***Justified beliefs* are those undefeated given the current stage of argument construction.**
    - » ***Warranted conclusions* are those that are undefeated relative to the set of all possible arguments that can be constructed given the current inputs.**

Defeasible Reasoning

- An *inference-graph* is a data structure recording a set of arguments.
- A *partial-status-assignment* for an inference-graph G is an assignment of "defeated" and "undefeated" to a subset of the arguments in G such that for each argument A in G:

  1. if a defeating argument for an inference in A is assigned "undefeated", A is assigned "defeated";

  2. if all defeating arguments for inferences in A are assigned "defeated", A is assigned "undefeated".
- A *status-assignment* for an inference-graph G is a maximal partial-status-assignment, i.e., a partial-status-assignment not properly contained in any other partial-status-assignment.
- An argument A is *undefeated* relative to an inference-graph G of which it is a member if and only if every status-assignment for G assigns "undefeated" to A.
- A belief is *justified* if and only if it is supported by an argument that is undefeated relative to the inference-graph that represents the agent's current epistemological state.

(For comparison with other approaches, see Henry Prakken and Gerard Vreeswijk, "Logics for Defeasible Argumentation", to appear in *Handbook of Philosophical Logic*, 2nd Edition, ed. D. Gabbay.)

Defeasible Reasoning

- **Raymond Reiter and David Israel both observed in 1981 that when reasoning defeasibly in a rich logical theory like first-order logic, the set of warranted conclusions will not generally be recursively enumerable.**
- **This has the consequence that it is impossible to build an automated defeasible reasoner that produces all and only warranted conclusions.**
  - **The most we can require is that the reasoner systematically modify its belief set so that it comes to approximate the set of warranted conclusions more and more closely.**
  - **The rules for reasoning should be such that:**

    (1) if a proposition *P* is warranted then the reasoner will eventually reach a stage where *P* is justified and stays justified;

    (2) if a proposition *P* is unwarranted then the reasoner will eventually reach a stage where *P* is unjustified and stays unjustified.
  - **This is possible if the reason-schemas are "well behaved".**

(See *Cognitive Carpentry*, chapter three, MIT Press, 1995.)

Some Defeasible Reason-Schemas

PERCEPTION

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

PERCEPTUAL-RELIABILITY

"R is true and having a percept with content P is not a reliable indicator of P's being true when R is true" is an undercutting defeater for PERCEPTION.

TEMPORAL-PROJECTION

"P-at-t" is a defeasible reason for "P-at-(t+Δt)", the strength of the reason being a monotonic decreasing function of Δt.

STATISTICAL-SYLLOGISM

"c is a B & prob(A/B) is high" is a defeasible reason for "c is an A".

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

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?

Time = 0

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

Merrill is a reliable informant

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

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

Merrill is a reliable informant

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

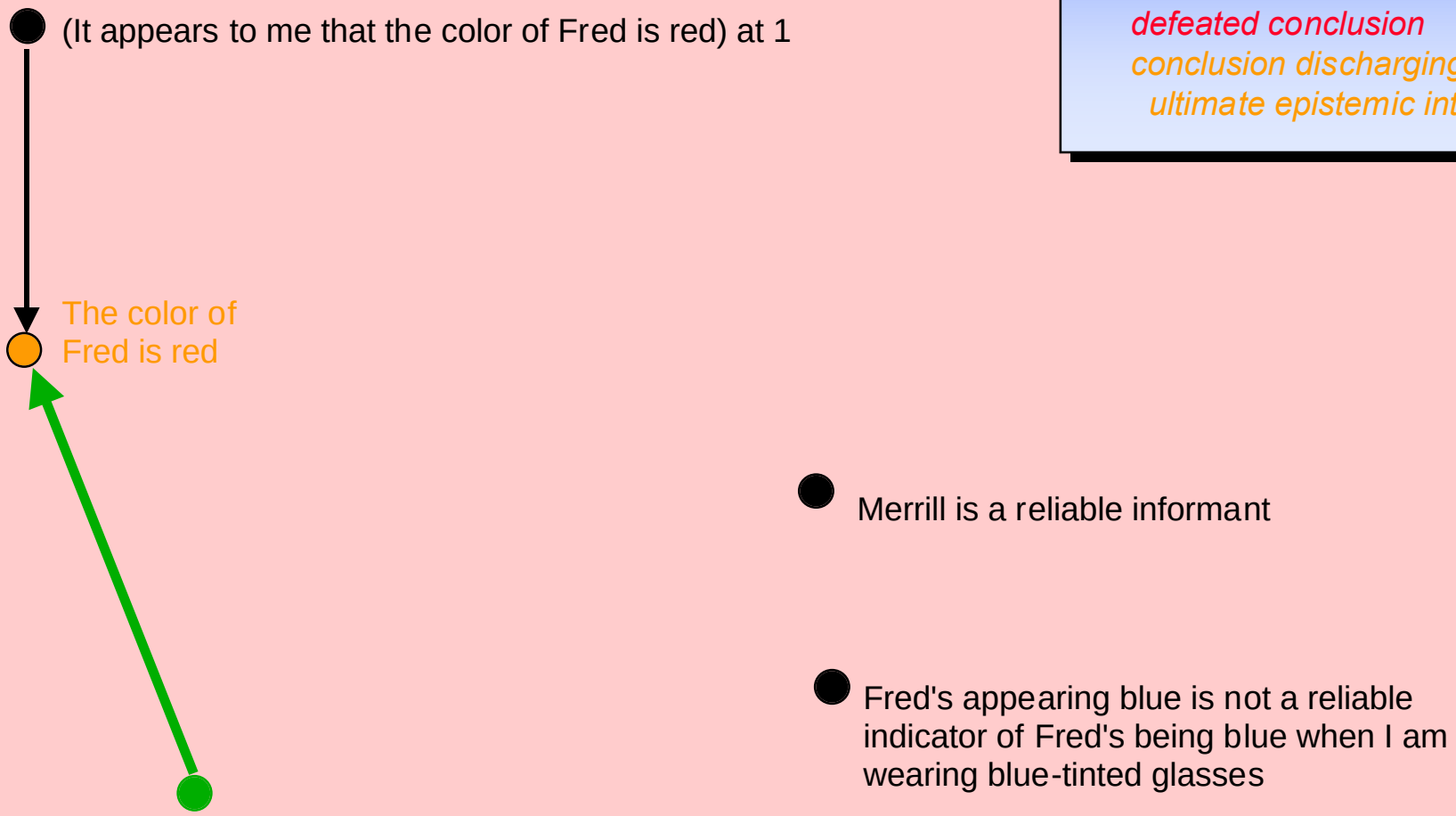
What color is Fred?

Percept acquired

Time = 2

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

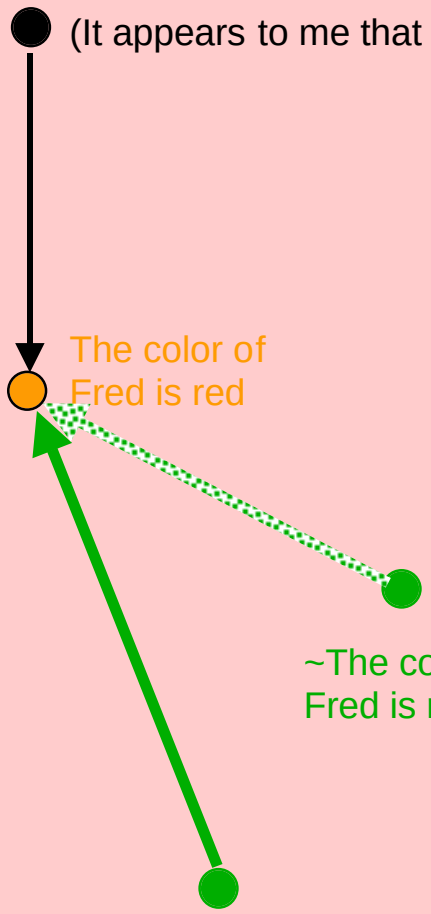


by PERCEPTION

Time = 3

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest



Merrill is a reliable informant

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

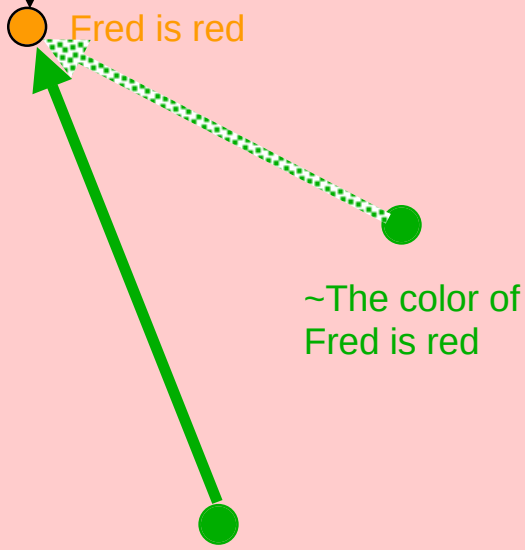
Interest in rebutter

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



Merrill is a reliable informant

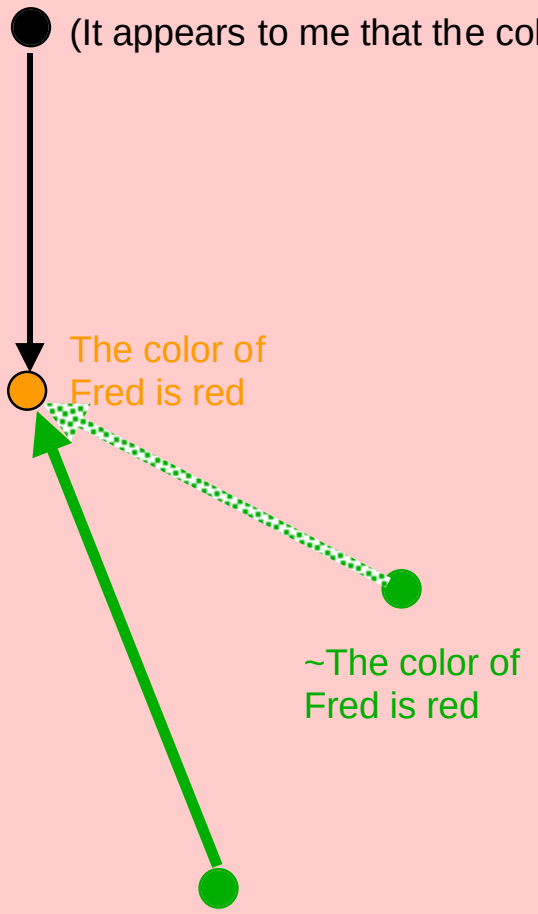
Fred's appearing blue is not a reliable indicator of Fred's being blue when I am wearing blue-tinted glasses

Time passes

Time = 5

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest



Merrill is a reliable informant

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

Time passes

Time = 6

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

Merrill is a reliable informant

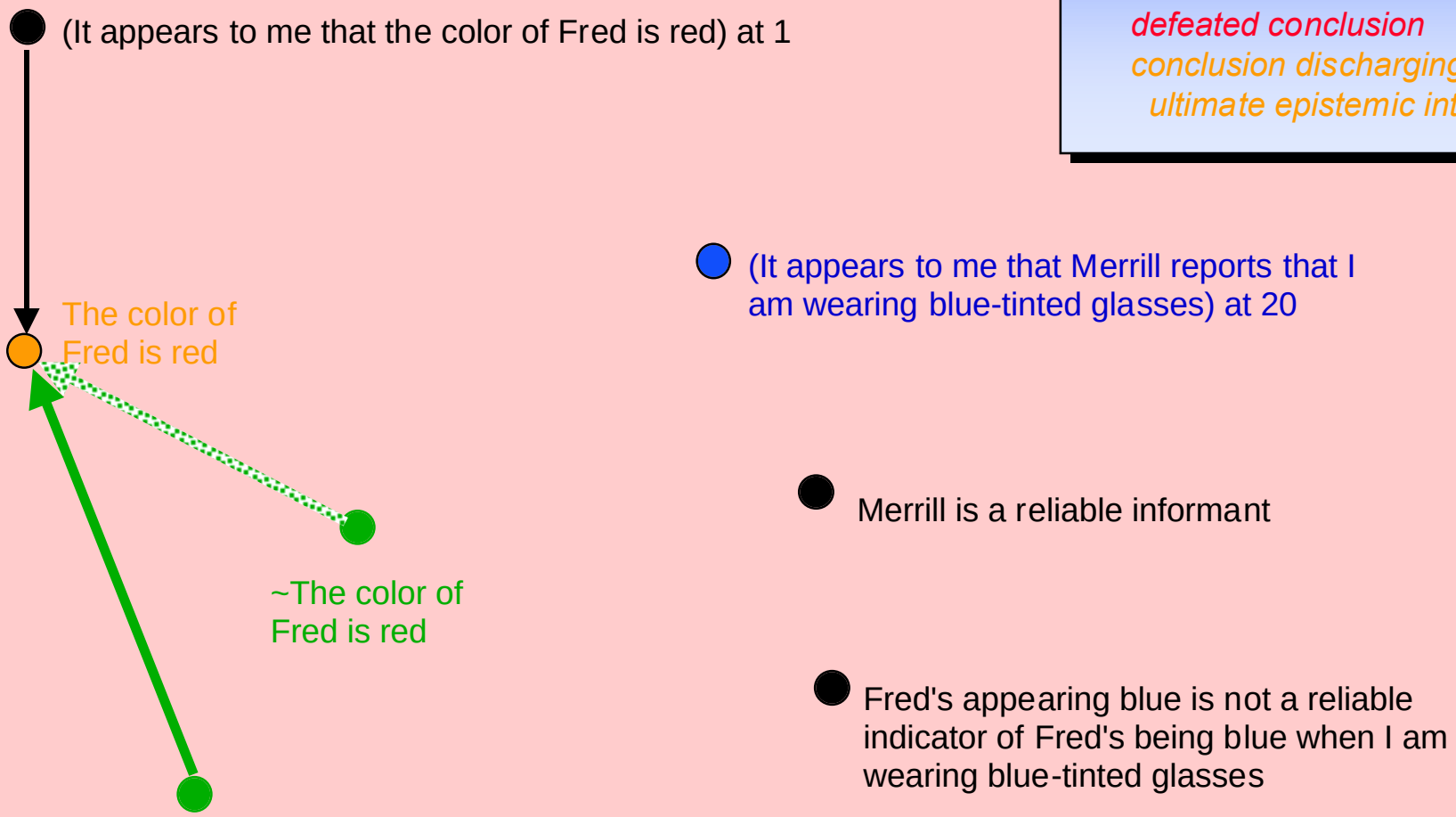
Fred's appearing blue is not a reliable indicator of Fred's being blue when I am wearing blue-tinted glasses

Time passes

Time = 20

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

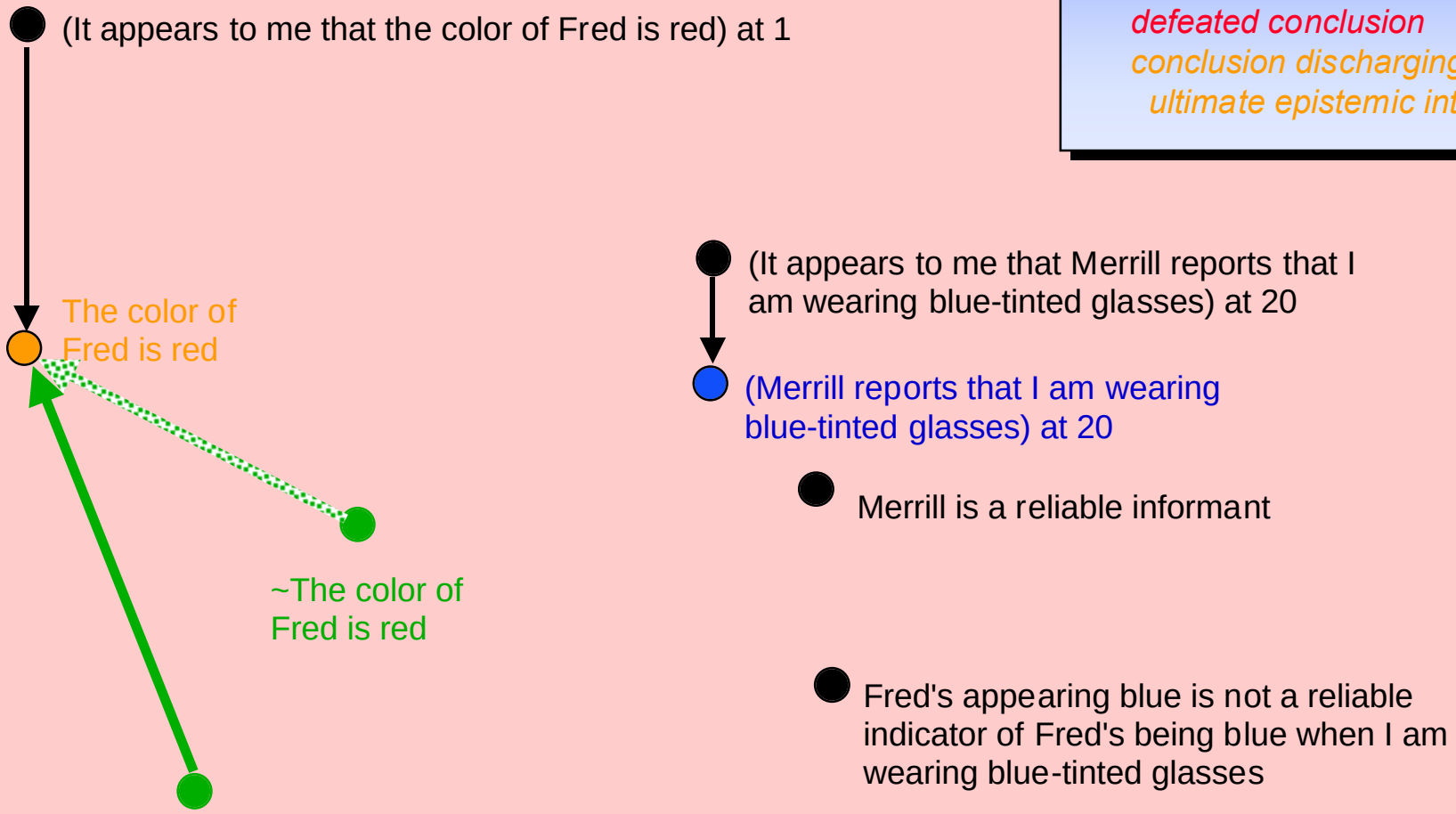


Percept acquired

Time = 21

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

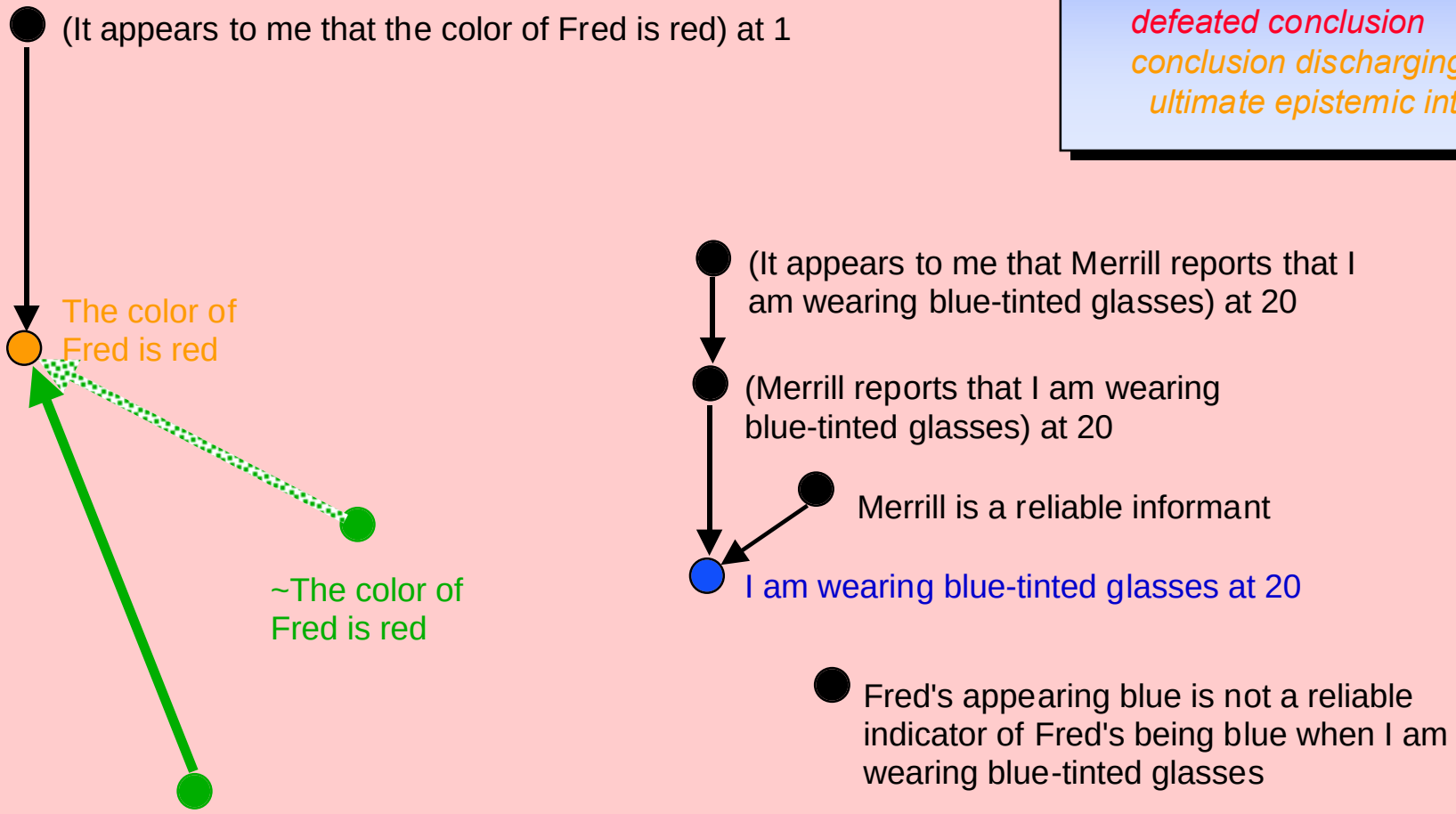


by PERCEPTION

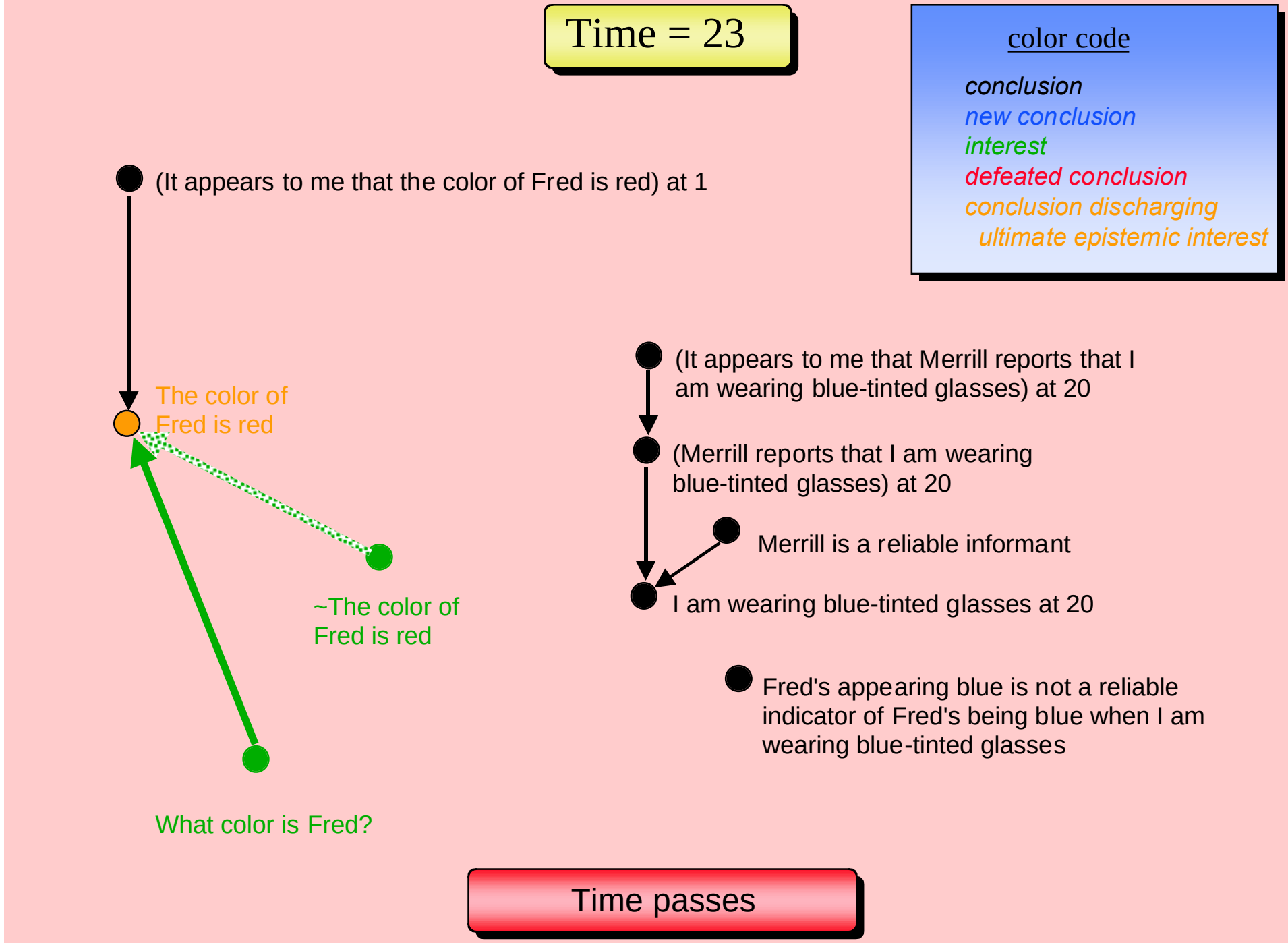
Time = 22

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest



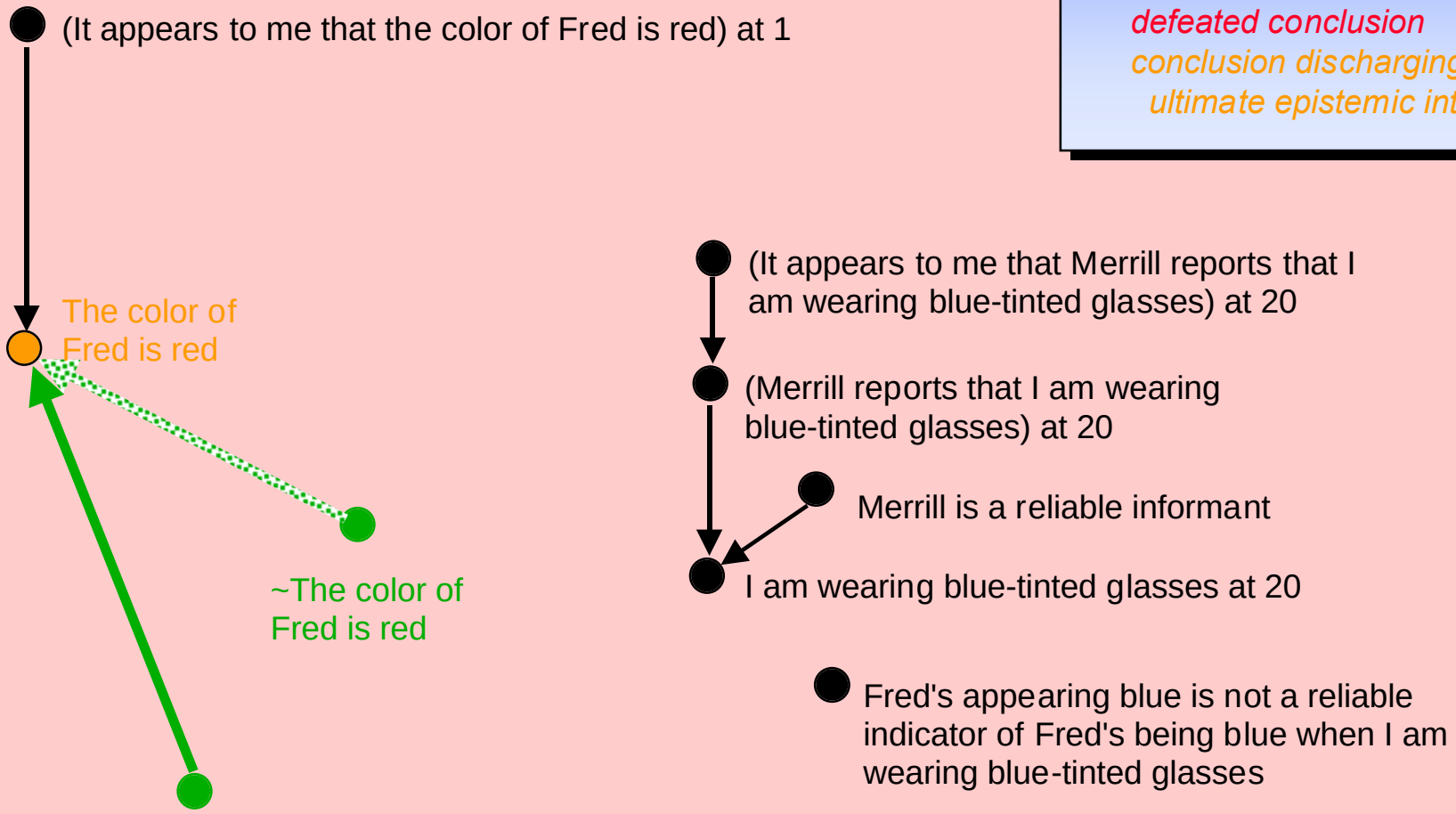
by STATISTICAL-SYLLOGISM


Time = 23
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?
(It appears to me that Merrill reports that I am wearing blue-tinted glasses) at 20
(Merrill reports that I am wearing blue-tinted glasses) at 20
Merrill is a reliable informant
I am wearing blue-tinted glasses at 20
Fred's appearing blue is not a reliable indicator of Fred's being blue when I am wearing blue-tinted glasses
Time passes

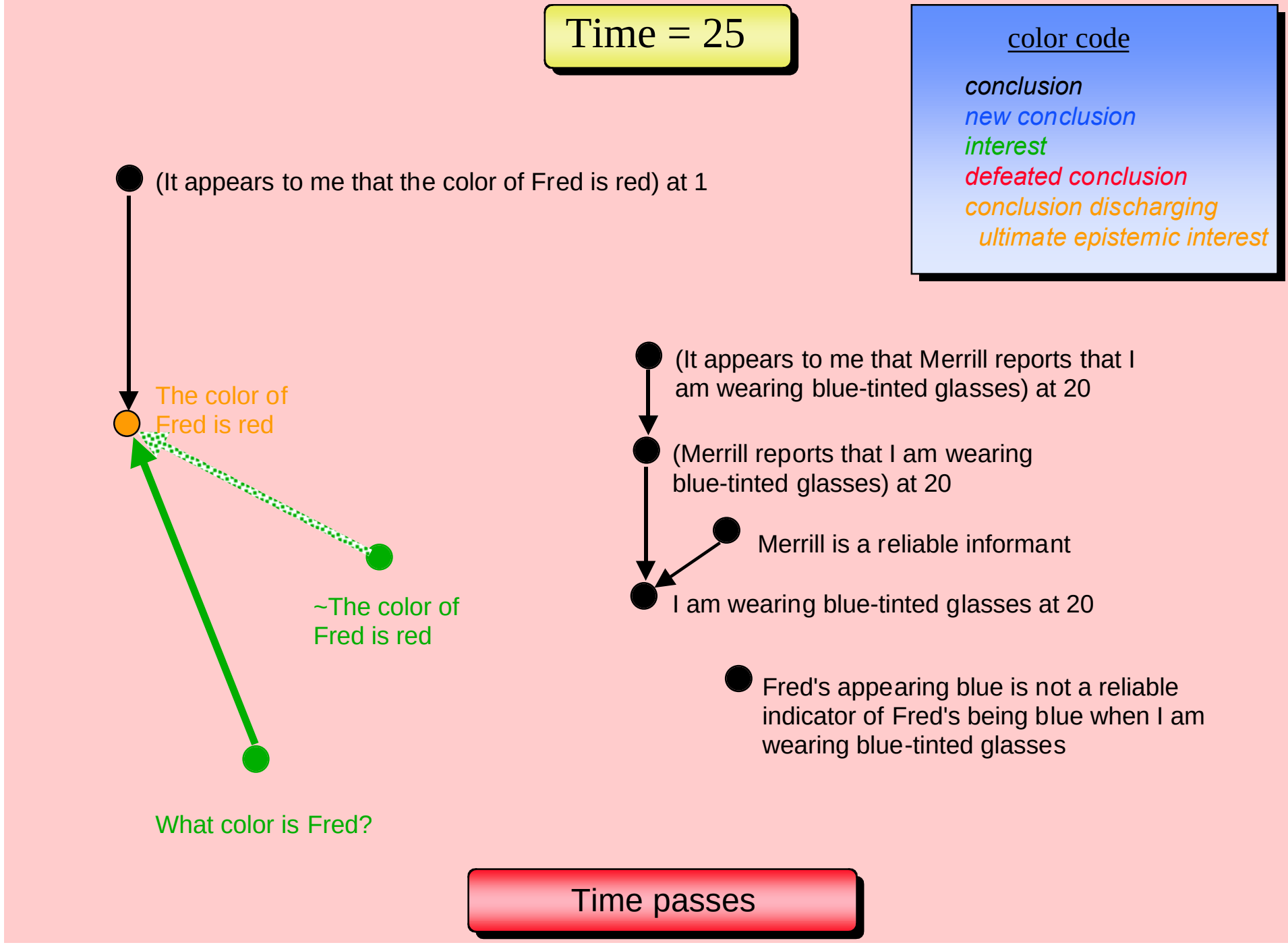
Time = 24

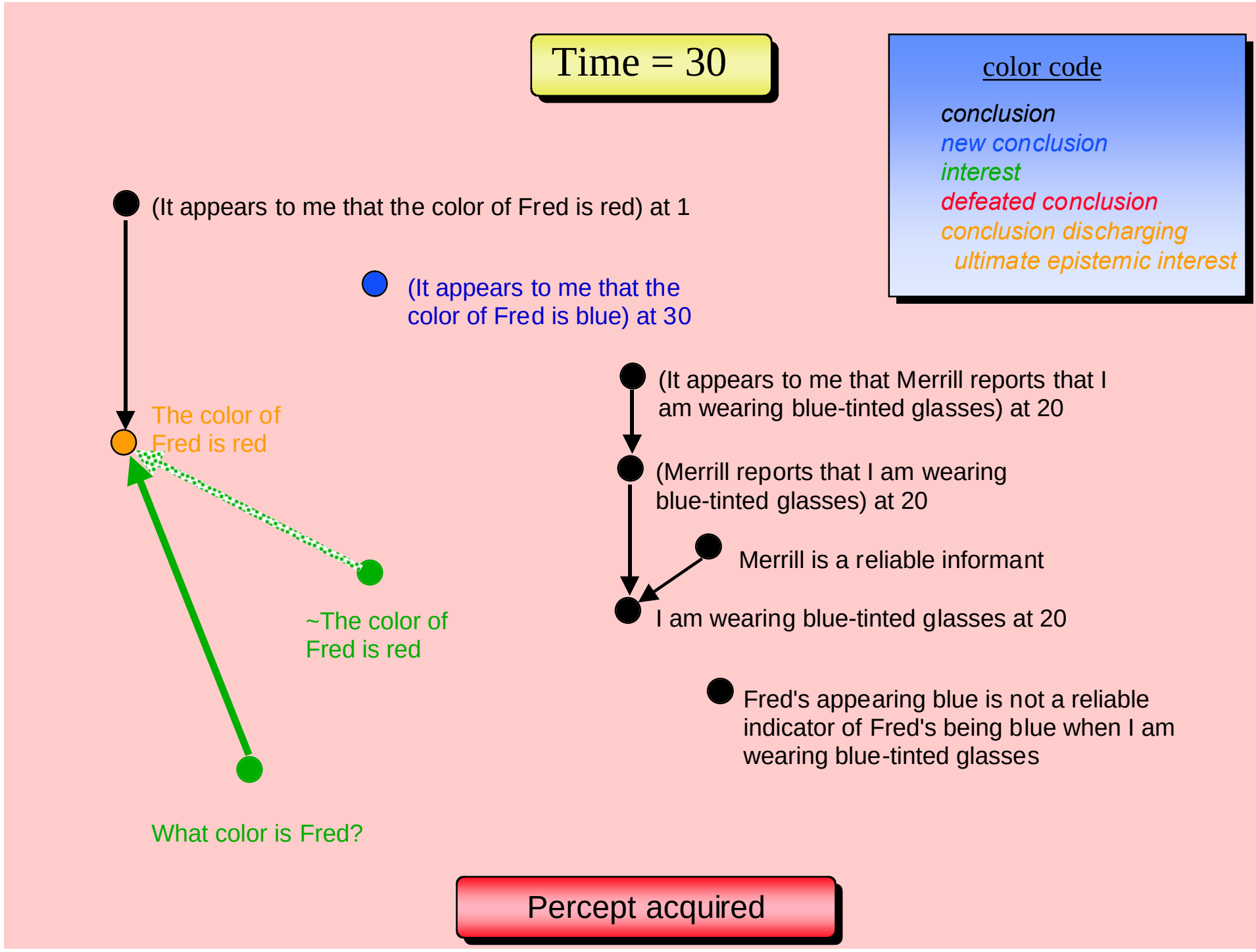
color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest



Time passes

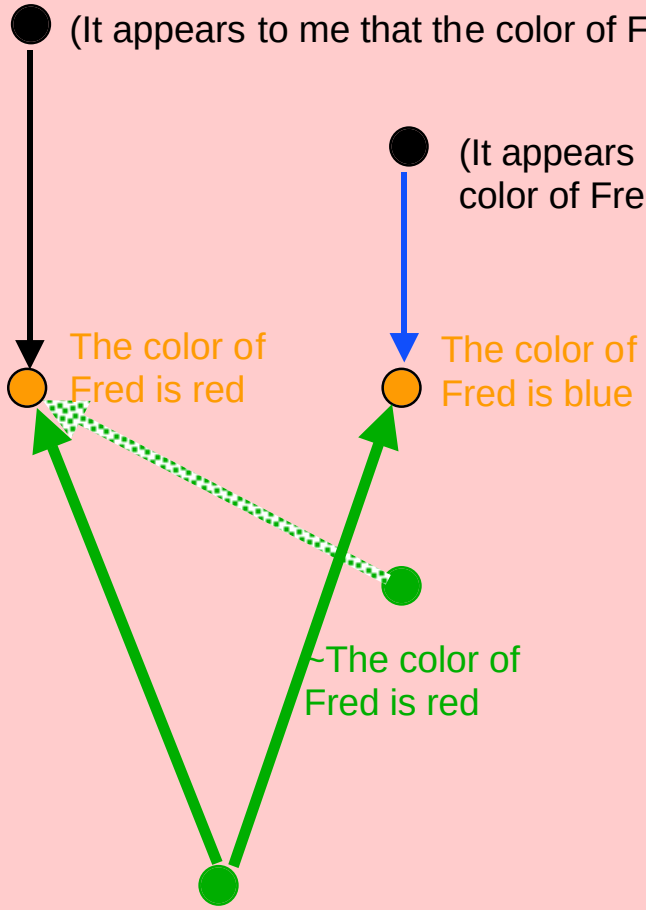

Time = 25
color code
conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest
(It appears to me that the color of Fred is red) at 1
The color of Fred is red
~The color of Fred is red
What color is Fred?
(It appears to me that Merrill reports that I am wearing blue-tinted glasses) at 20
(Merrill reports that I am wearing blue-tinted glasses) at 20
Merrill is a reliable informant
I am wearing blue-tinted glasses at 20
Fred's appearing blue is not a reliable indicator of Fred's being blue when I am wearing blue-tinted glasses
Time passes


Time = 30
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
(It appears to me that the color of Fred is blue) at 30
(It appears to me that Merrill reports that I am wearing blue-tinted glasses) at 20
The color of Fred is red
(Merrill reports that I am wearing blue-tinted glasses) at 20
Merrill is a reliable informant
~The color of Fred is red
I am wearing blue-tinted glasses at 20
Fred's appearing blue is not a reliable indicator of Fred's being blue when I am wearing blue-tinted glasses
What color is Fred?
Percept acquired

Time = 31

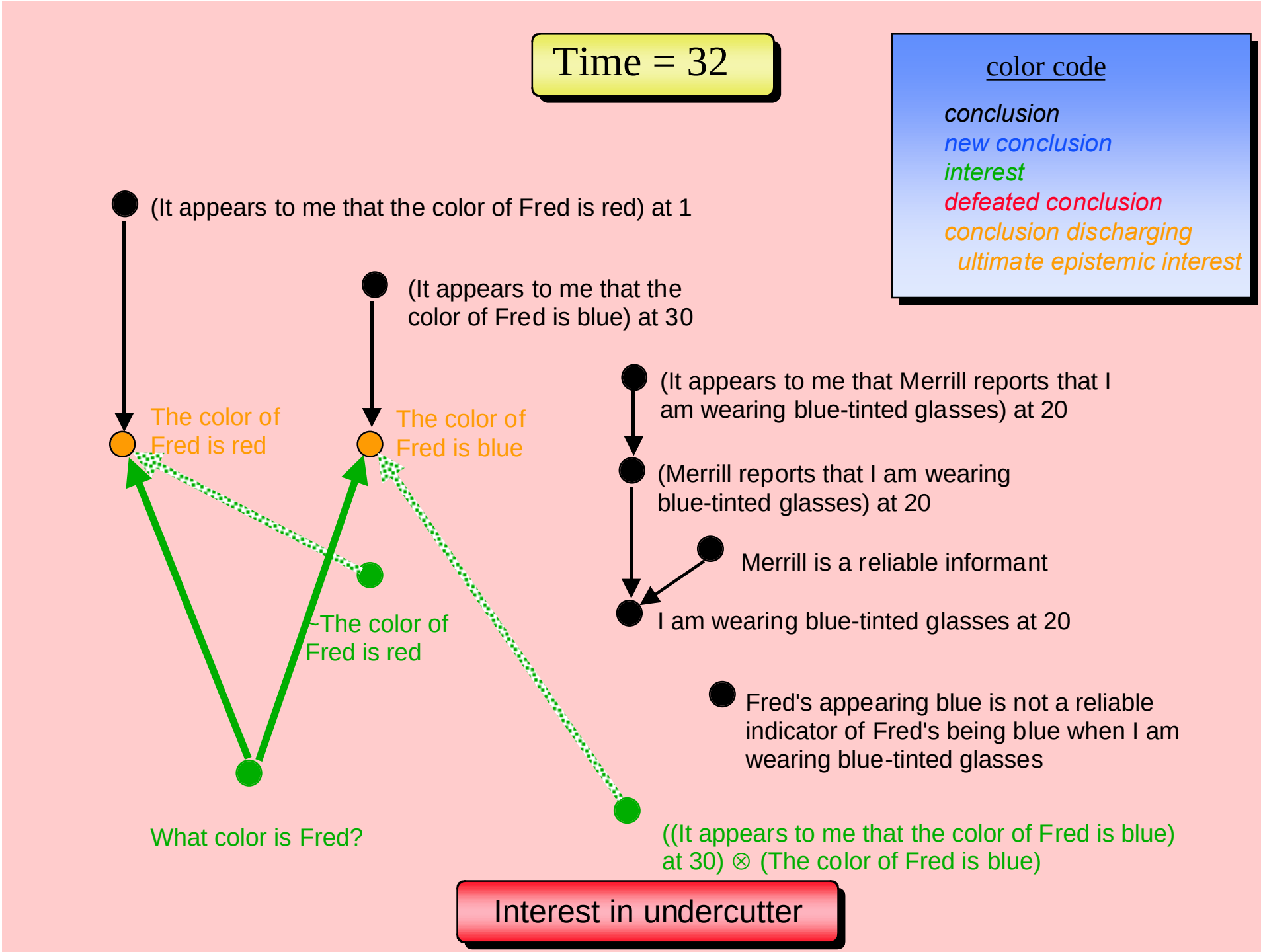
color code

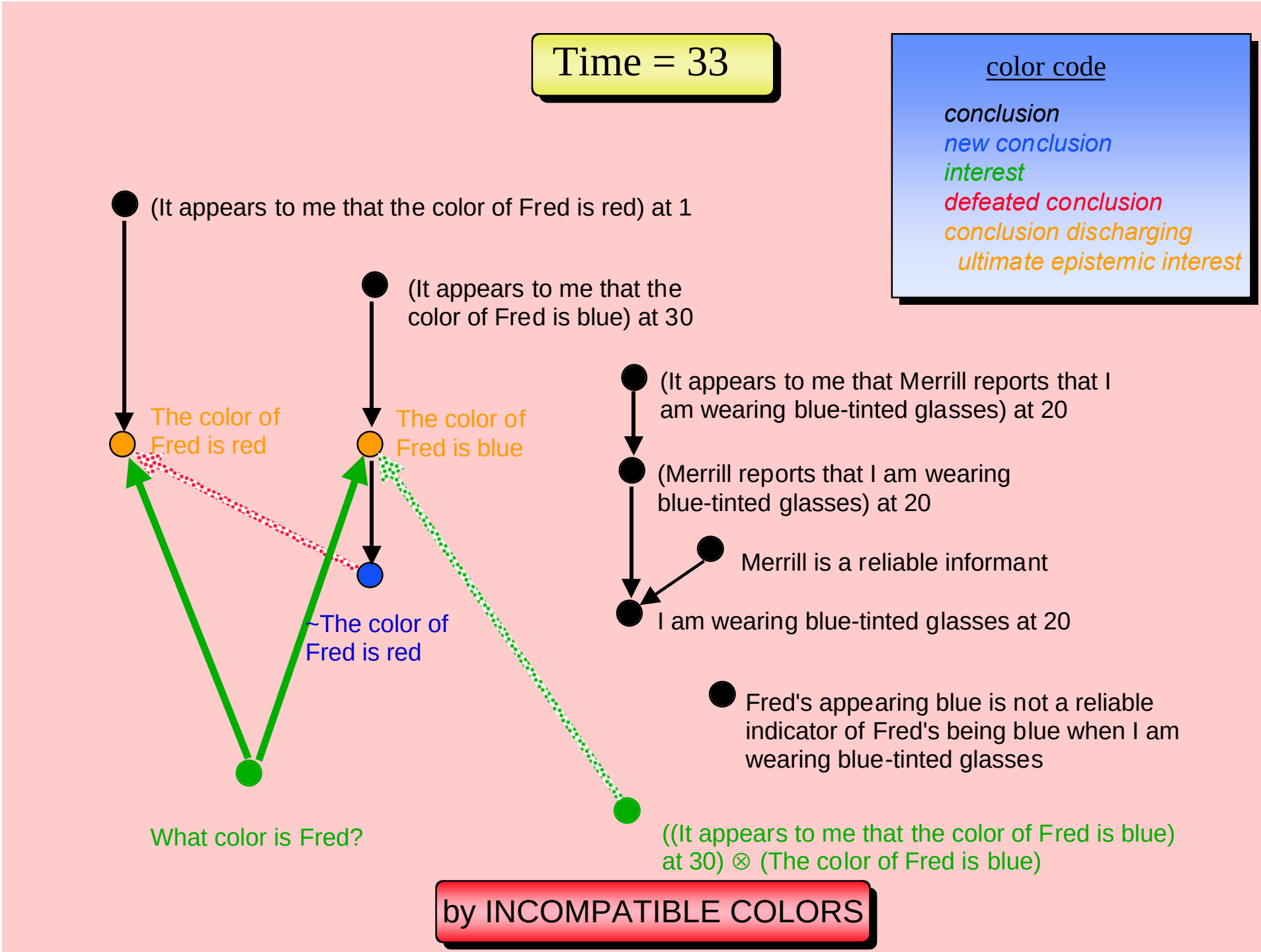
conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

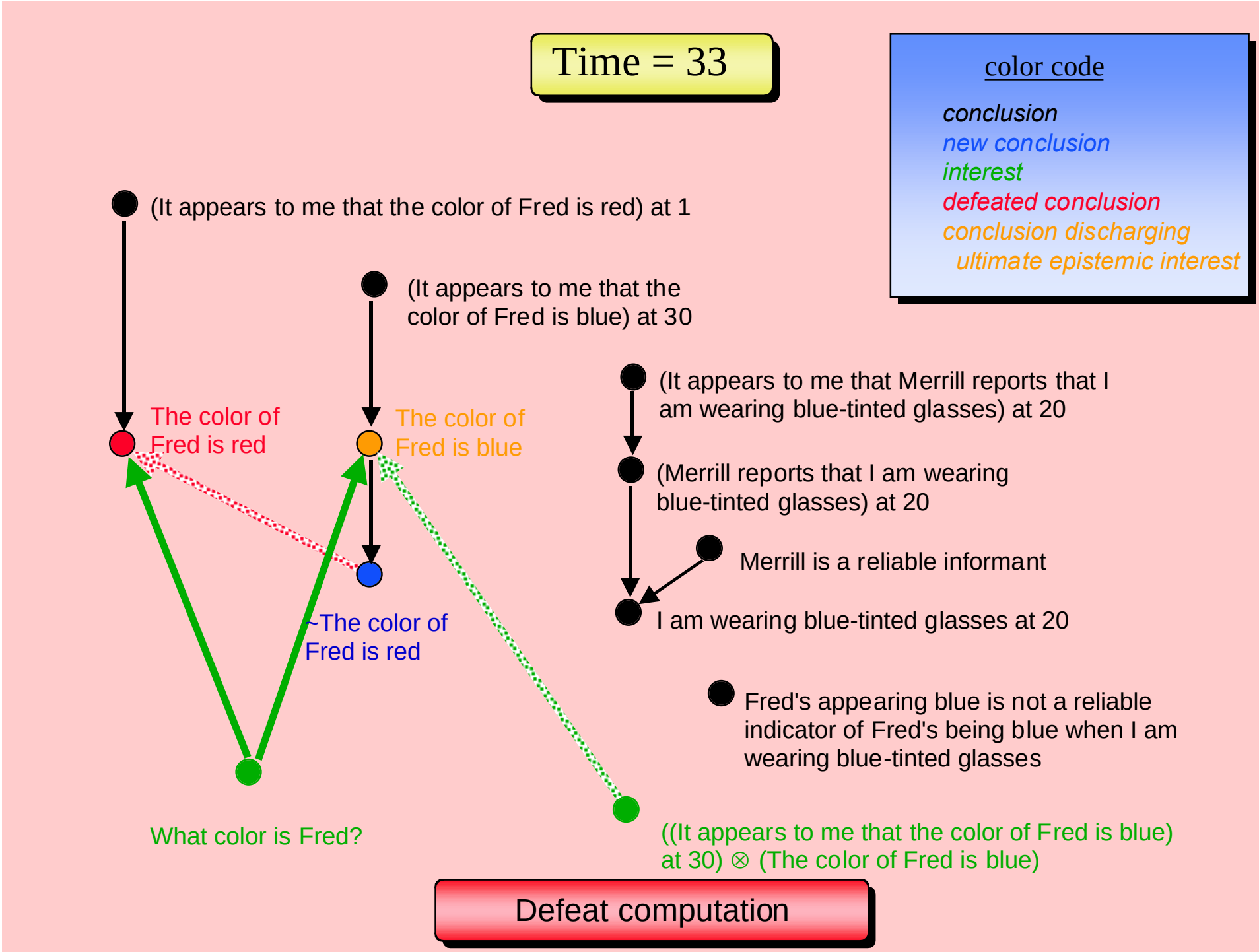


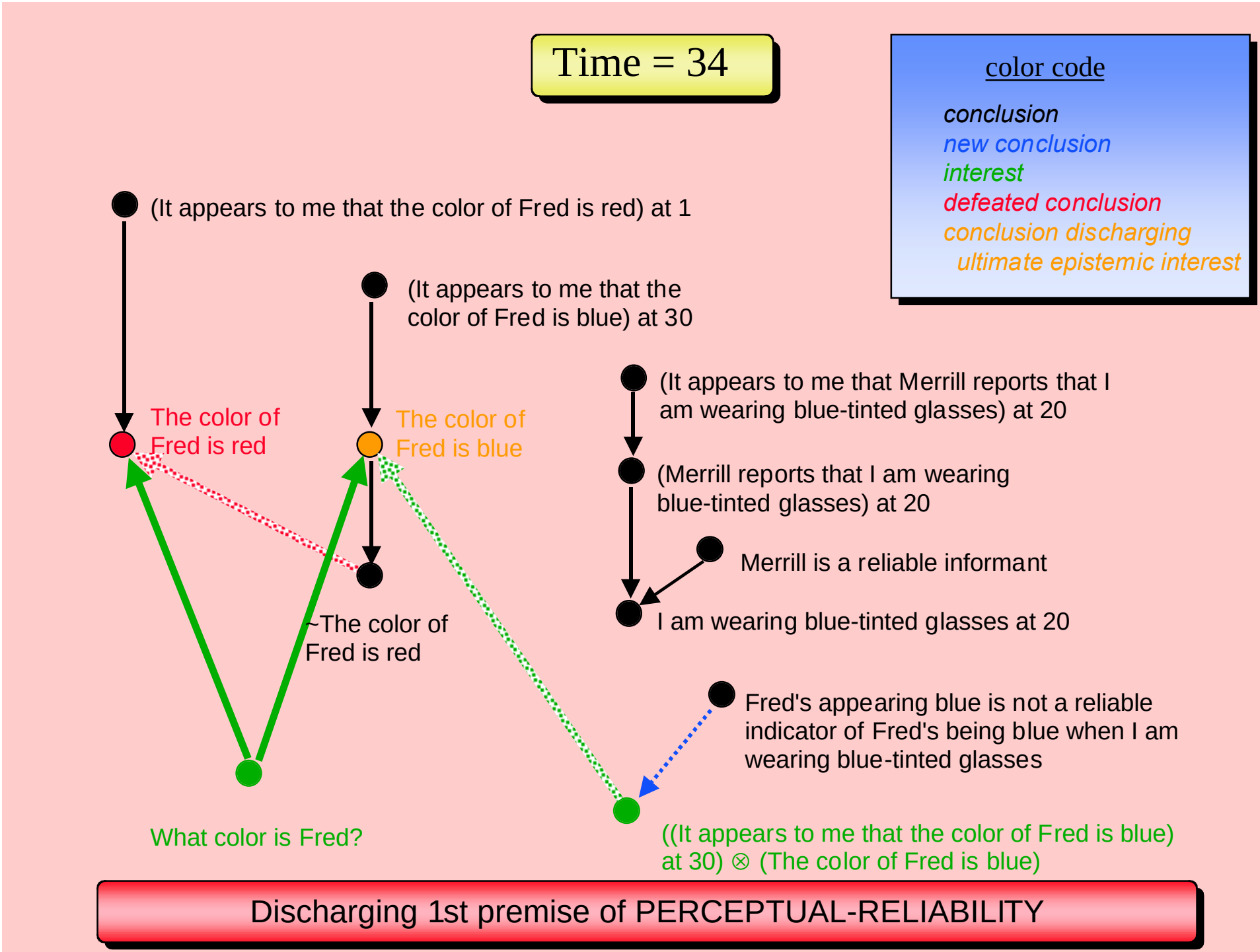
Fred's appearing blue is not a reliable indicator of Fred's being blue when I am wearing blue-tinted glasses

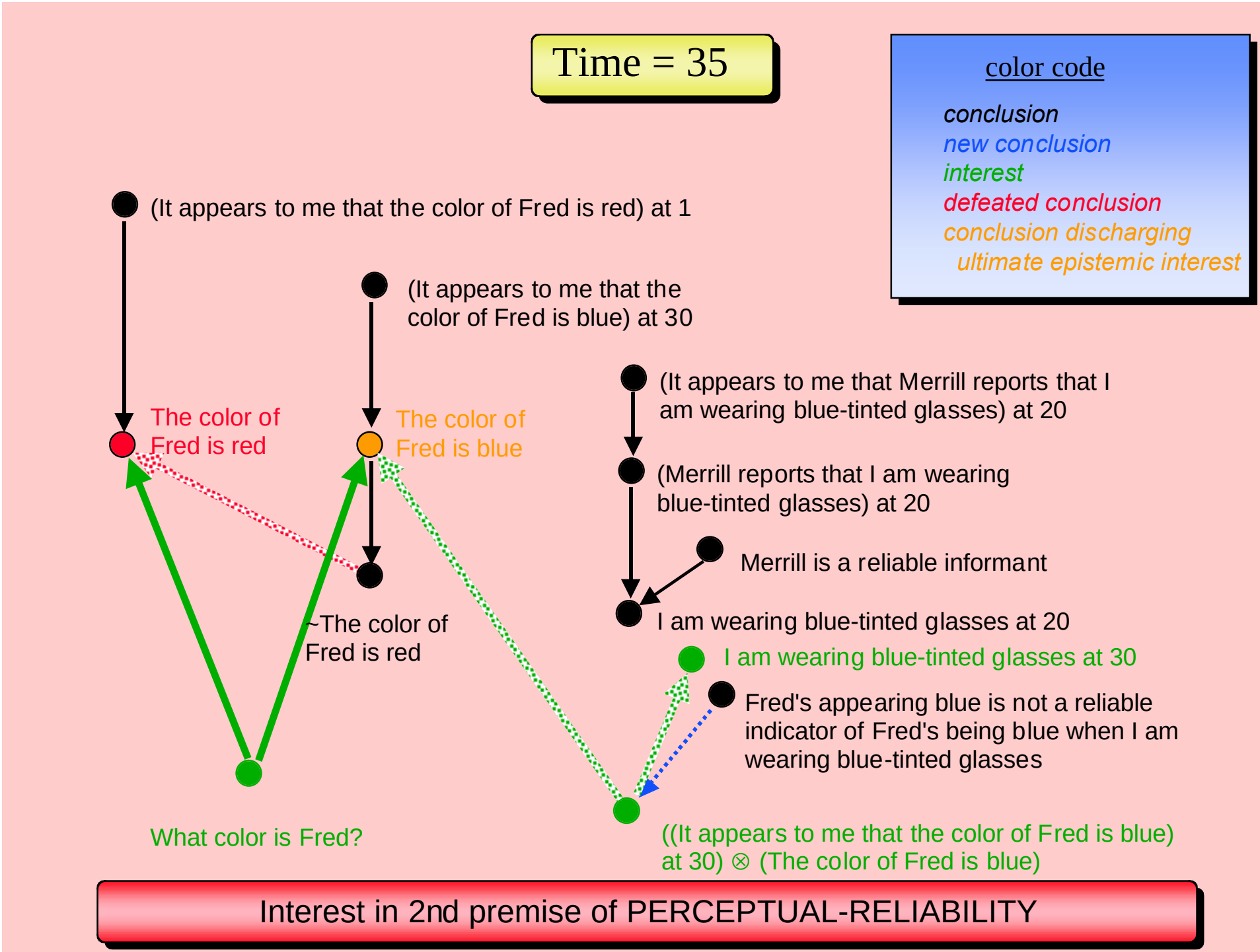
by PERCEPTION

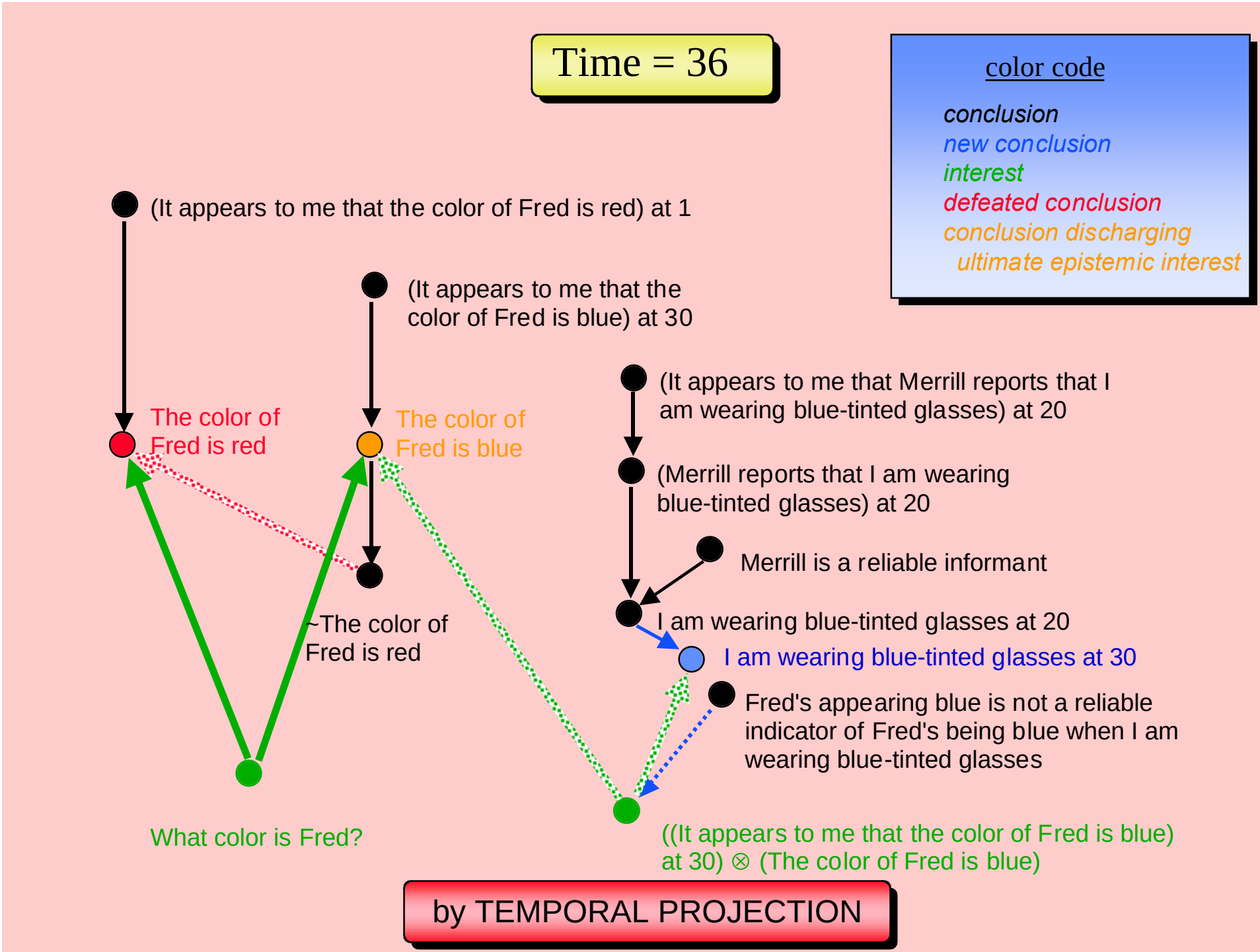

Time = 32
color code
conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest
(It appears to me that the color of Fred is red) at 1
(It appears to me that the color of Fred is blue) at 30
(It appears to me that Merrill reports that I am wearing blue-tinted glasses) at 20
The color of Fred is red
The color of Fred is blue
(Merrill reports that I am wearing blue-tinted glasses) at 20
Merrill is a reliable informant
~The color of Fred is red
I am wearing blue-tinted glasses at 20
Fred's appearing blue is not a reliable indicator of Fred's being blue when I am wearing blue-tinted glasses
What color is Fred?
((It appears to me that the color of Fred is blue) at 30) ⊗ (The color of Fred is blue)
Interest in undercutter

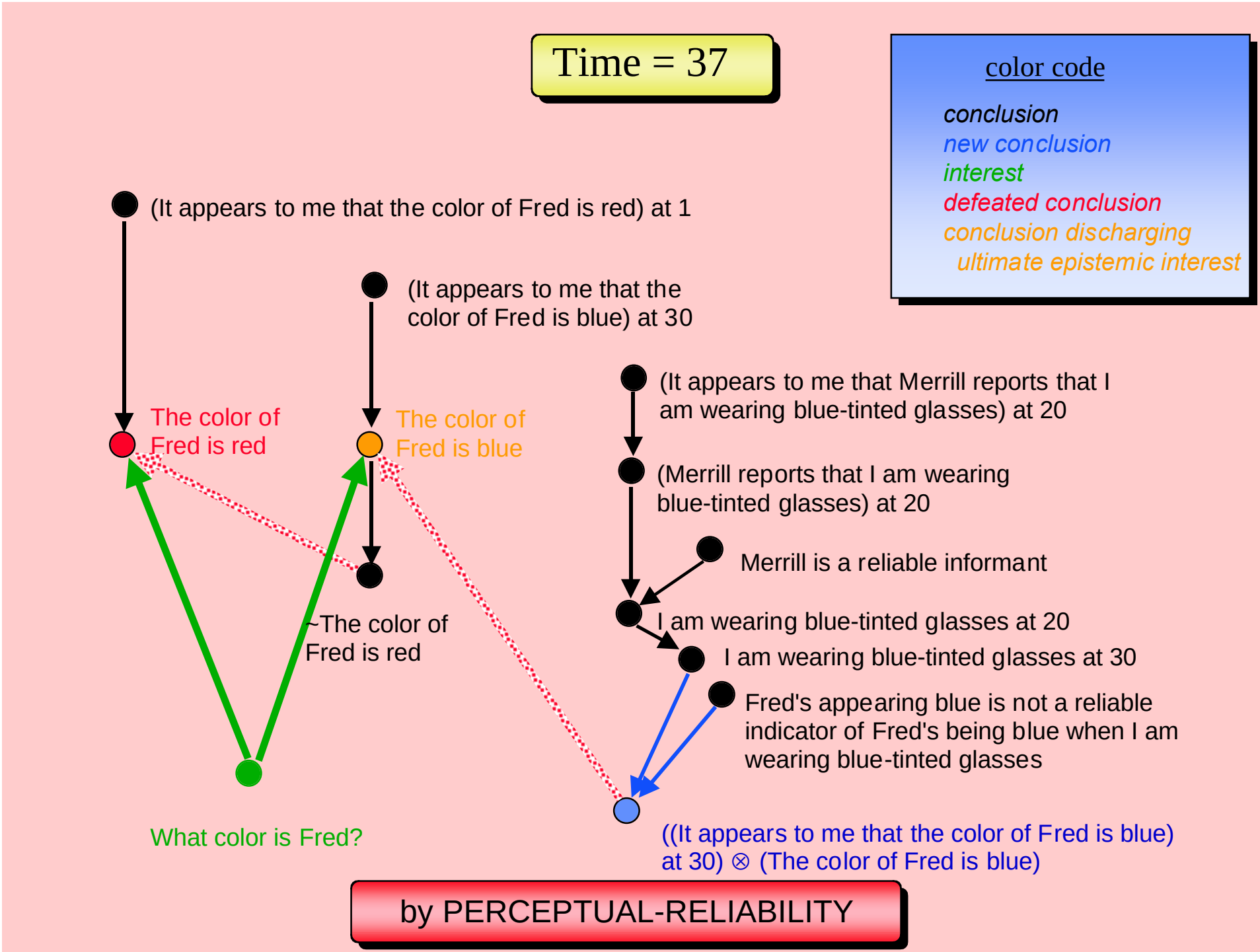

Time = 33
color code
conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest
(It appears to me that the color of Fred is red) at 1
(It appears to me that the color of Fred is blue) at 30
(It appears to me that Merrill reports that I am wearing blue-tinted glasses) at 20
The color of Fred is red
The color of Fred is blue
(Merrill reports that I am wearing blue-tinted glasses) at 20
Merrill is a reliable informant
~The color of Fred is red
I am wearing blue-tinted glasses at 20
Fred's appearing blue is not a reliable indicator of Fred's being blue when I am wearing blue-tinted glasses
What color is Fred?
((It appears to me that the color of Fred is blue) at 30) ⊗ (The color of Fred is blue)
by INCOMPATIBLE COLORS

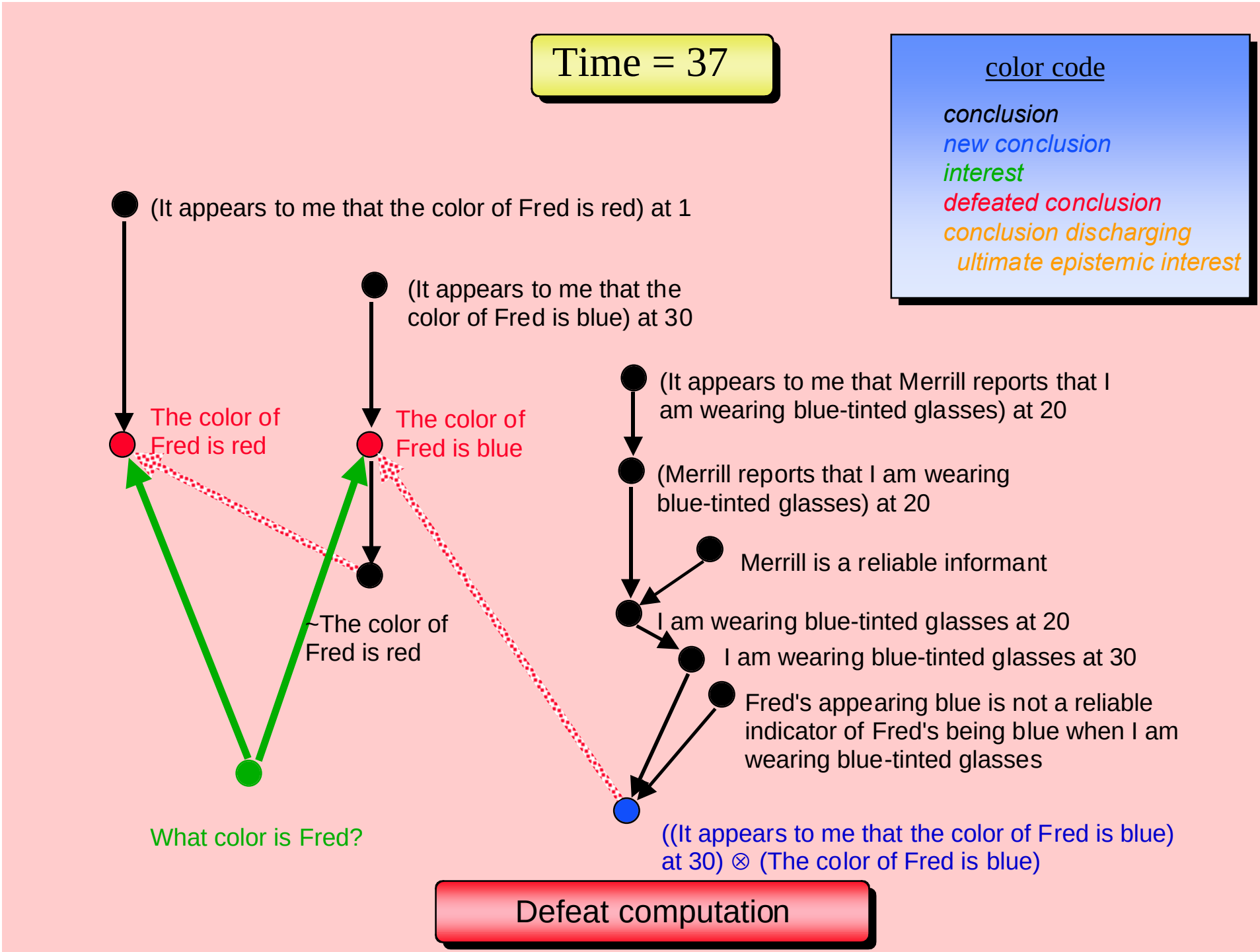

Time = 33
color code
conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest
(It appears to me that the color of Fred is red) at 1
(It appears to me that the color of Fred is blue) at 30
(It appears to me that Merrill reports that I am wearing blue-tinted glasses) at 20
The color of Fred is red
The color of Fred is blue
(Merrill reports that I am wearing blue-tinted glasses) at 20
Merrill is a reliable informant
~The color of Fred is red
I am wearing blue-tinted glasses at 20
Fred's appearing blue is not a reliable indicator of Fred's being blue when I am wearing blue-tinted glasses
What color is Fred?
((It appears to me that the color of Fred is blue) at 30) ⊗ (The color of Fred is blue)
Defeat computation

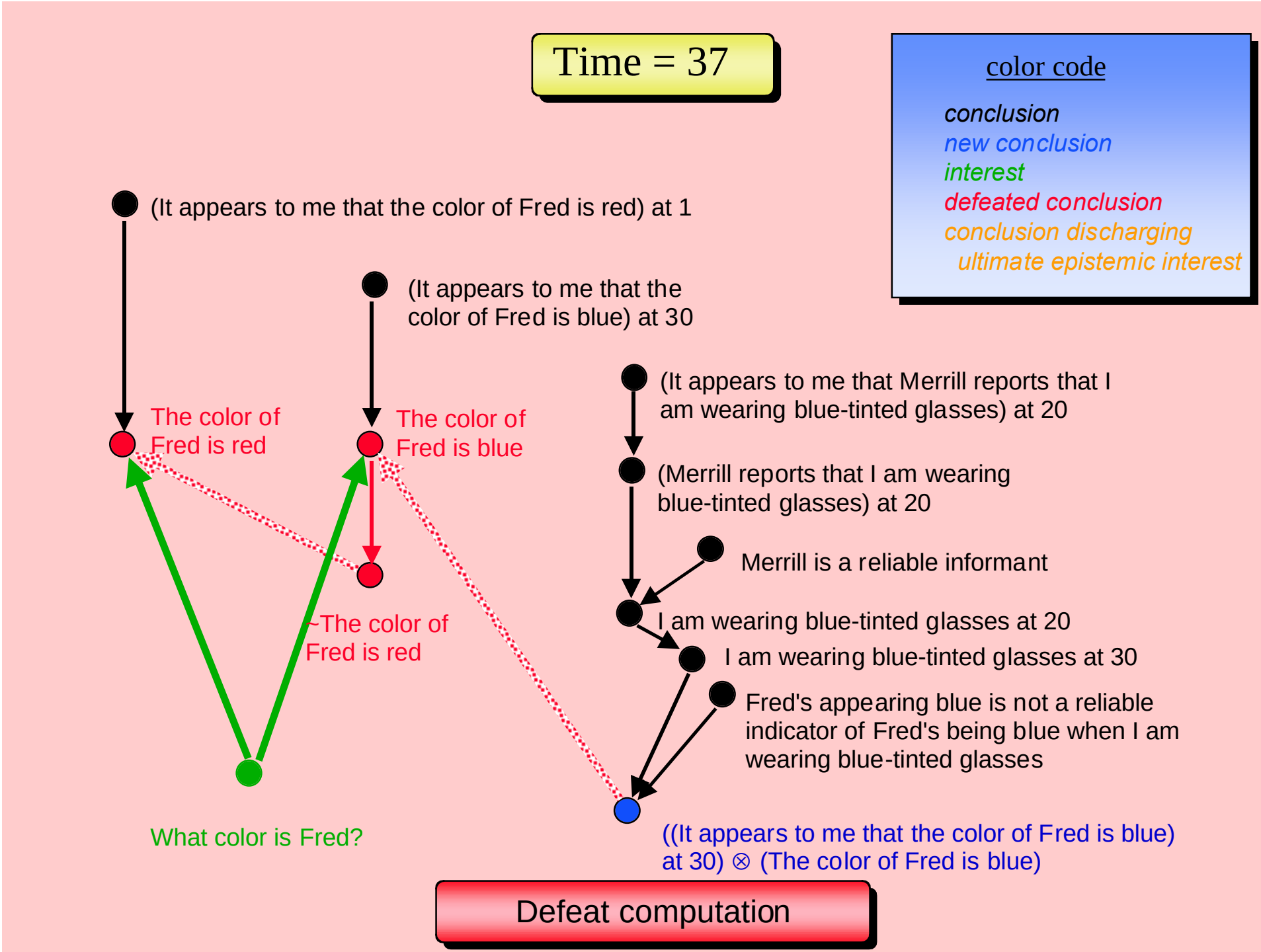

Time = 34
color code
conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest
(It appears to me that the color of Fred is red) at 1
(It appears to me that the color of Fred is blue) at 30
(It appears to me that Merrill reports that I am wearing blue-tinted glasses) at 20
The color of Fred is red
The color of Fred is blue
(Merrill reports that I am wearing blue-tinted glasses) at 20
Merrill is a reliable informant
~The color of Fred is red
I am wearing blue-tinted glasses at 20
Fred's appearing blue is not a reliable indicator of Fred's being blue when I am wearing blue-tinted glasses
What color is Fred?
((It appears to me that the color of Fred is blue) at 30) ⊗ (The color of Fred is blue)
Discharging 1st premise of PERCEPTUAL-RELIABILITY

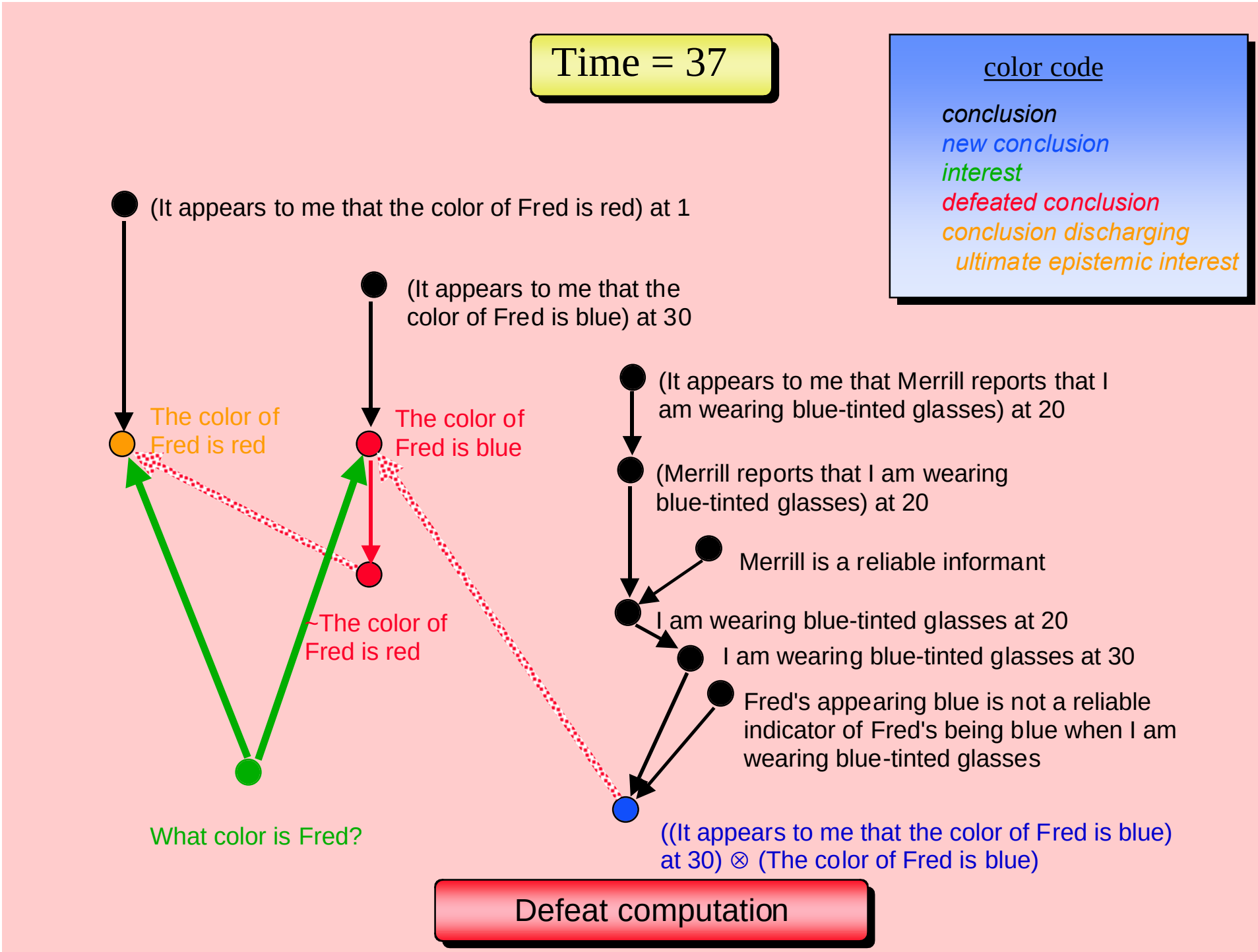

Time = 35
color code
conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest
(It appears to me that the color of Fred is red) at 1
(It appears to me that the color of Fred is blue) at 30
(It appears to me that Merrill reports that I am wearing blue-tinted glasses) at 20
The color of Fred is red
The color of Fred is blue
(Merrill reports that I am wearing blue-tinted glasses) at 20
Merrill is a reliable informant
~The color of Fred is red
I am wearing blue-tinted glasses at 20
I am wearing blue-tinted glasses at 30
Fred's appearing blue is not a reliable indicator of Fred's being blue when I am wearing blue-tinted glasses
What color is Fred?
((It appears to me that the color of Fred is blue) at 30) ⊗ (The color of Fred is blue)
Interest in 2nd premise of PERCEPTUAL-RELIABILITY

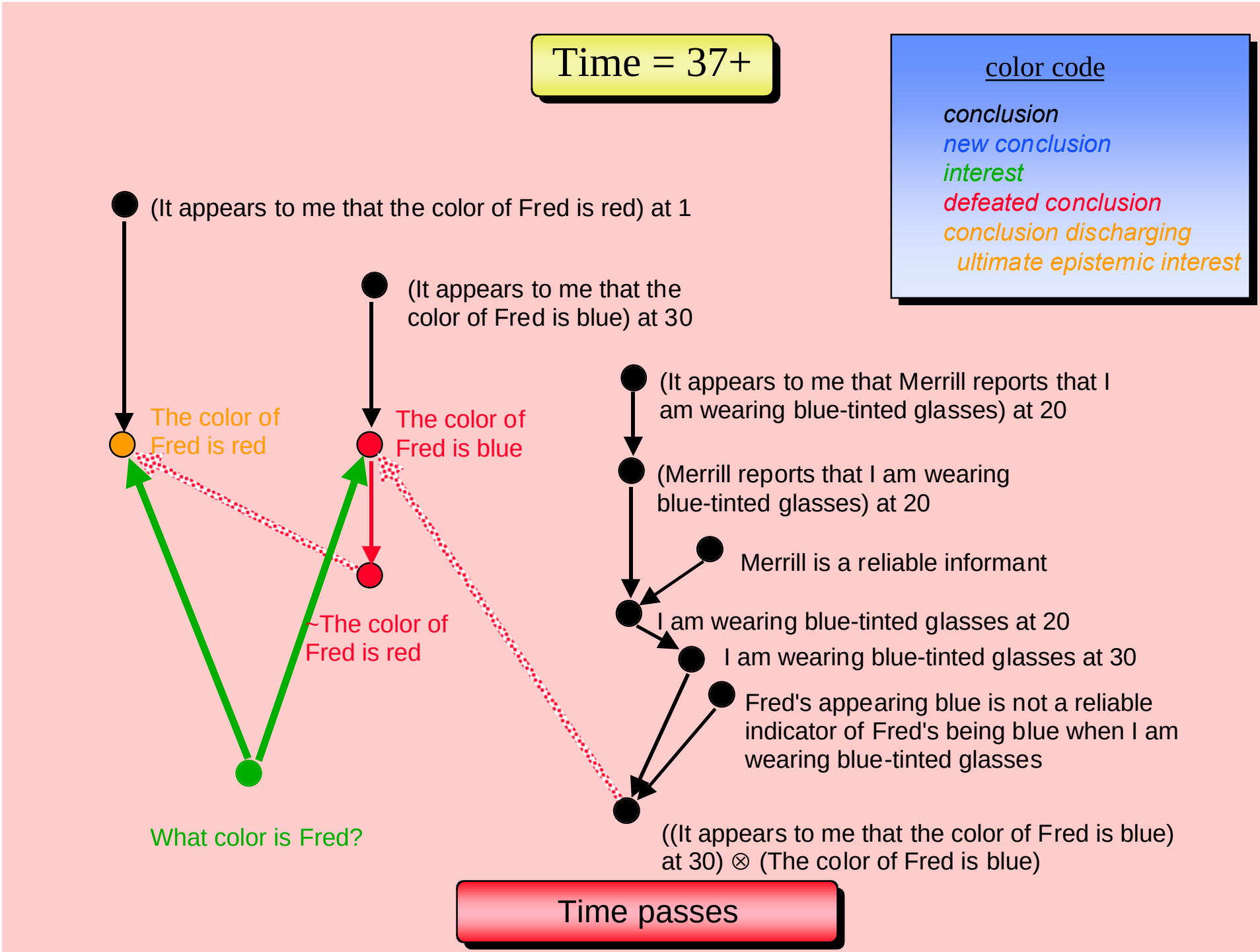

Time = 36
color code
conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest
(It appears to me that the color of Fred is red) at 1
(It appears to me that the color of Fred is blue) at 30
(It appears to me that Merrill reports that I am wearing blue-tinted glasses) at 20
The color of Fred is red
The color of Fred is blue
(Merrill reports that I am wearing blue-tinted glasses) at 20
Merrill is a reliable informant
~The color of Fred is red
I am wearing blue-tinted glasses at 20
I am wearing blue-tinted glasses at 30
Fred's appearing blue is not a reliable indicator of Fred's being blue when I am wearing blue-tinted glasses
What color is Fred?
((It appears to me that the color of Fred is blue) at 30) ⊗ (The color of Fred is blue)
by TEMPORAL PROJECTION


Time = 37
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
(It appears to me that the color of Fred is blue) at 30
(It appears to me that Merrill reports that I am wearing blue-tinted glasses) at 20
The color of Fred is red
The color of Fred is blue
(Merrill reports that I am wearing blue-tinted glasses) at 20
Merrill is a reliable informant
~The color of Fred is red
I am wearing blue-tinted glasses at 20
I am wearing blue-tinted glasses at 30
Fred's appearing blue is not a reliable indicator of Fred's being blue when I am wearing blue-tinted glasses
What color is Fred?
((It appears to me that the color of Fred is blue) at 30) ⊗ (The color of Fred is blue)
by PERCEPTUAL-RELIABILITY


Time = 37
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
(It appears to me that the color of Fred is blue) at 30
(It appears to me that Merrill reports that I am wearing blue-tinted glasses) at 20
The color of Fred is red
The color of Fred is blue
(Merrill reports that I am wearing blue-tinted glasses) at 20
Merrill is a reliable informant
~The color of Fred is red
I am wearing blue-tinted glasses at 20
I am wearing blue-tinted glasses at 30
Fred's appearing blue is not a reliable indicator of Fred's being blue when I am wearing blue-tinted glasses
What color is Fred?
((It appears to me that the color of Fred is blue) at 30) ⊗ (The color of Fred is blue)
Defeat computation


Time = 37
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
(It appears to me that the color of Fred is blue) at 30
(It appears to me that Merrill reports that I am wearing blue-tinted glasses) at 20
The color of Fred is red
The color of Fred is blue
(Merrill reports that I am wearing blue-tinted glasses) at 20
Merrill is a reliable informant
~The color of Fred is red
I am wearing blue-tinted glasses at 20
I am wearing blue-tinted glasses at 30
Fred's appearing blue is not a reliable indicator of Fred's being blue when I am wearing blue-tinted glasses
What color is Fred?
((It appears to me that the color of Fred is blue) at 30) ⊗ (The color of Fred is blue)
Defeat computation


Time = 37
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
(It appears to me that the color of Fred is blue) at 30
(It appears to me that Merrill reports that I am wearing blue-tinted glasses) at 20
The color of Fred is red
The color of Fred is blue
(Merrill reports that I am wearing blue-tinted glasses) at 20
Merrill is a reliable informant
~The color of Fred is red
I am wearing blue-tinted glasses at 20
I am wearing blue-tinted glasses at 30
Fred's appearing blue is not a reliable indicator of Fred's being blue when I am wearing blue-tinted glasses
What color is Fred?
((It appears to me that the color of Fred is blue) at 30) ⊗ (The color of Fred is blue)
Defeat computation


Time = 37+
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
(It appears to me that the color of Fred is blue) at 30
(It appears to me that Merrill reports that I am wearing blue-tinted glasses) at 20
The color of Fred is red
The color of Fred is blue
(Merrill reports that I am wearing blue-tinted glasses) at 20
Merrill is a reliable informant
~The color of Fred is red
I am wearing blue-tinted glasses at 20
I am wearing blue-tinted glasses at 30
Fred's appearing blue is not a reliable indicator of Fred's being blue when I am wearing blue-tinted glasses
What color is Fred?
((It appears to me that the color of Fred is blue) at 30) ⊗ (The color of Fred is blue)
Time passes

Causal Reasoning

- For a rational agent to be able to construct plans for making the environment more to its liking, it must be able to reason causally.
- In particular, it must be able to reason its way through the frame problem.
- OSCAR implements a solution to the frame problem. It has three constituents:
  - TEMPORAL-PROJECTION
  - CAUSAL-IMPLICATION

    If t* > t, "A-at-t and P-at-t and (A when P is causally-sufficient for Q)" is a defeasible reason for "Q-at-t*".
  - CAUSAL-UNDERCUTTER

    If $t_0 < t < t^*$, "A-at-t and P-at-t and (A when P is causally-sufficient for ~Q)" is an undercutting defeater for the inference from Q-at- t_0 to Q-at-t by TEMPORAL-PROJECTION.

The Yale Shooting Problem

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.

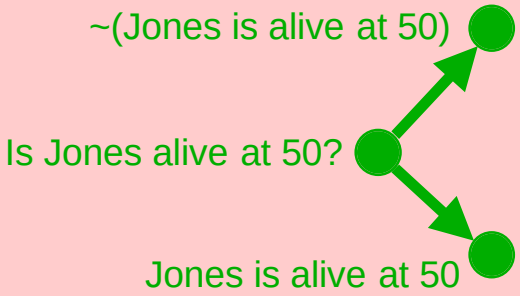
Should I conclude that Jones becomes dead?

Time = 0

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

((The trigger is pulled when the gun is loaded) is causally sufficient for ~(Jones is alive) after an interval 10)



given

Time = 20

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

The gun is loaded at 20

((The trigger is pulled when the gun is loaded) is causally sufficient for ~(Jones is alive) after an interval 10)

~(Jones is alive at 50)

Is Jones alive at 50?

Jones is alive at 50

Jones is alive at 20

given

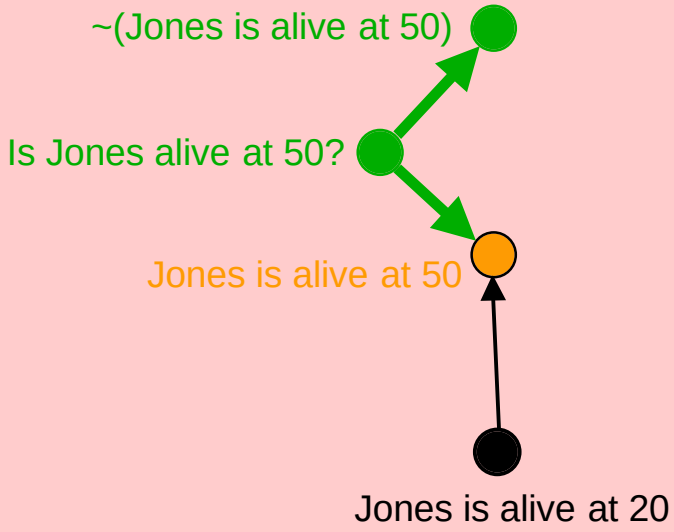
Time = 21

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

The gun is loaded at 20

((The trigger is pulled when the gun is loaded) is causally sufficient for ~(Jones is alive) after an interval 10)



by TEMPORAL PROJECTION

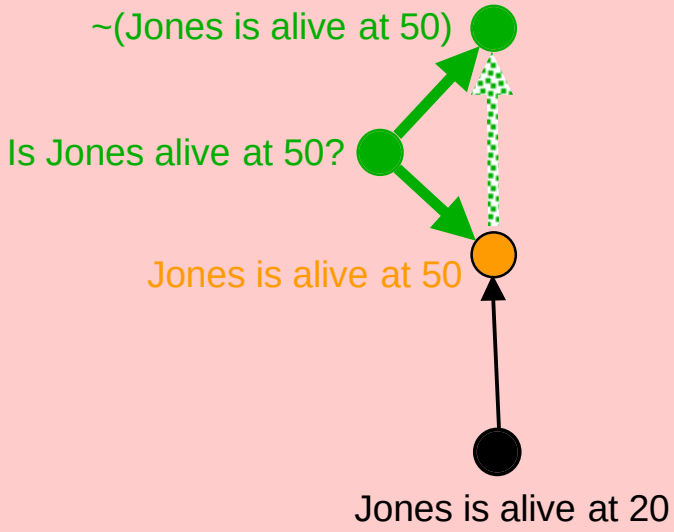
Time = 22

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

The gun is loaded at 20

((The trigger is pulled when the gun is loaded) is causally sufficient for ~(Jones is alive) after an interval 10)



Interest in rebutter

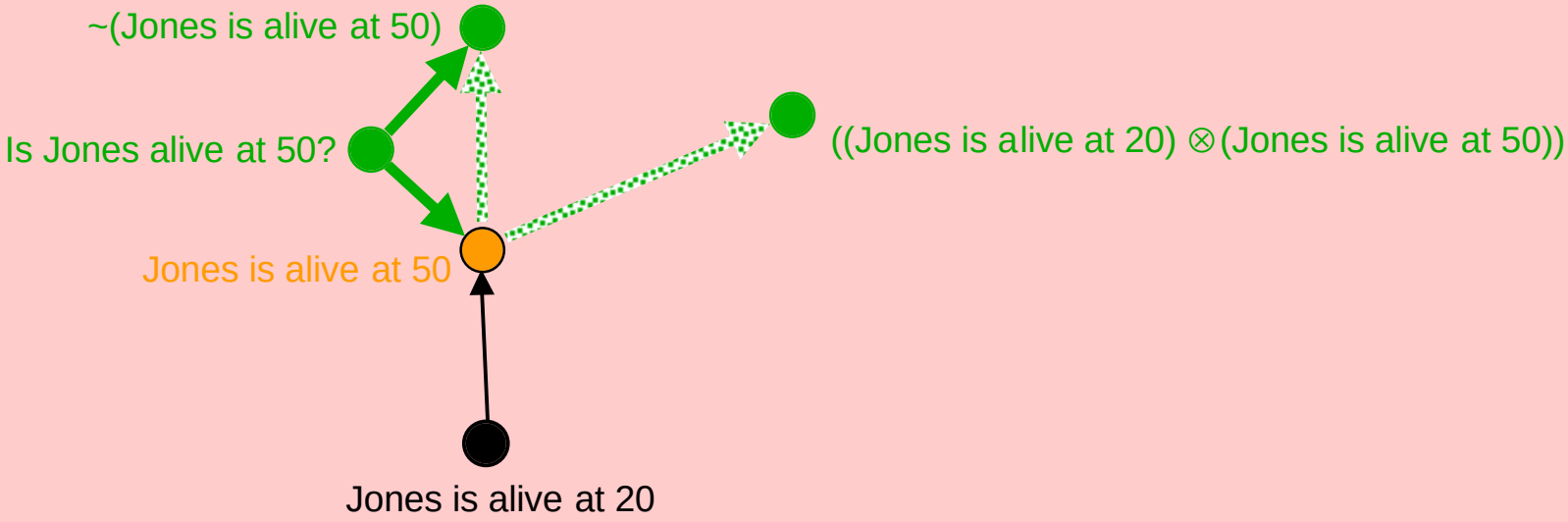
Time = 23

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

The gun is loaded at 20

((The trigger is pulled when the gun is loaded) is causally sufficient for ~(Jones is alive) after an interval 10)

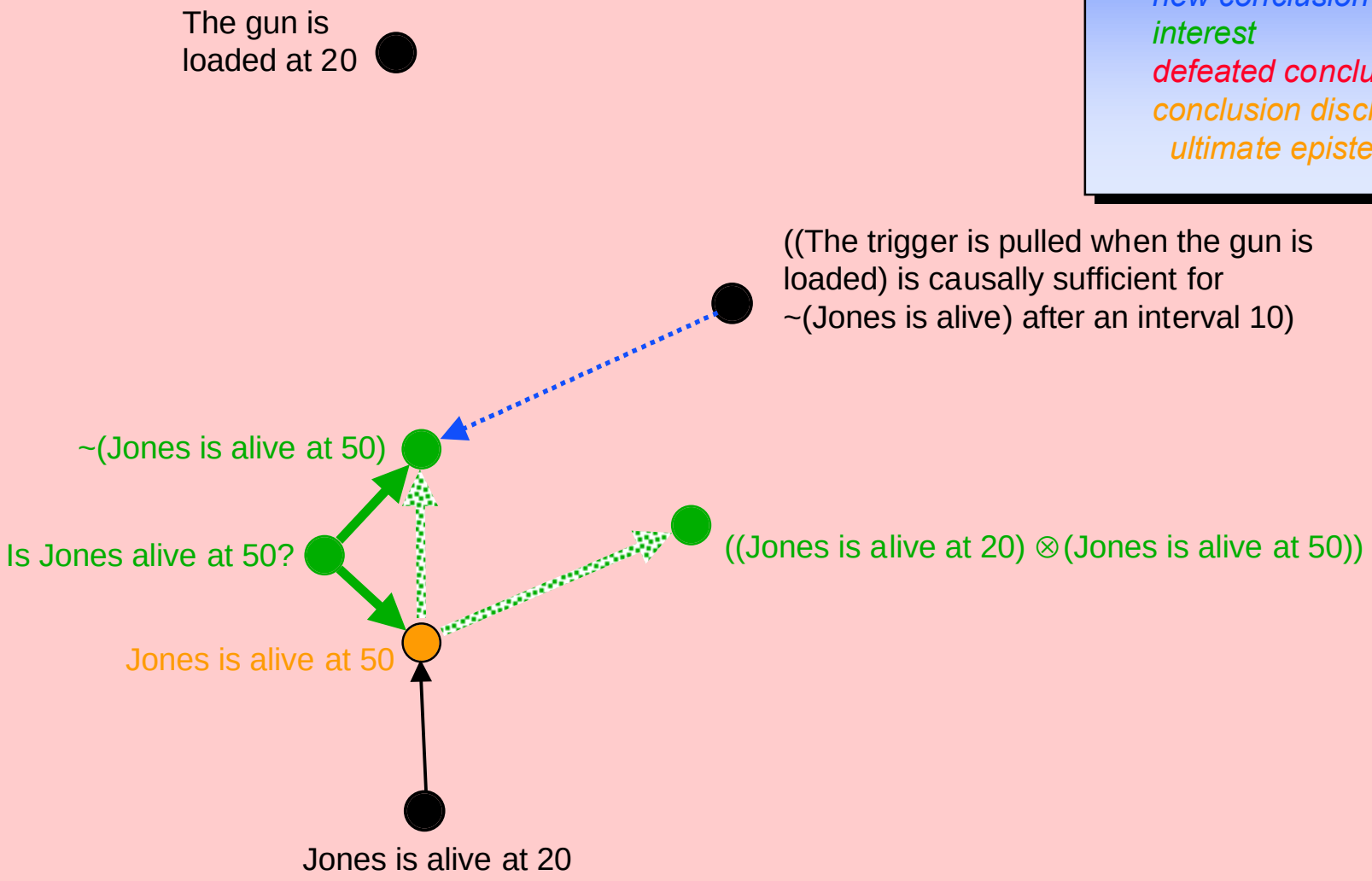


Interest in undercutter

Time = 24

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

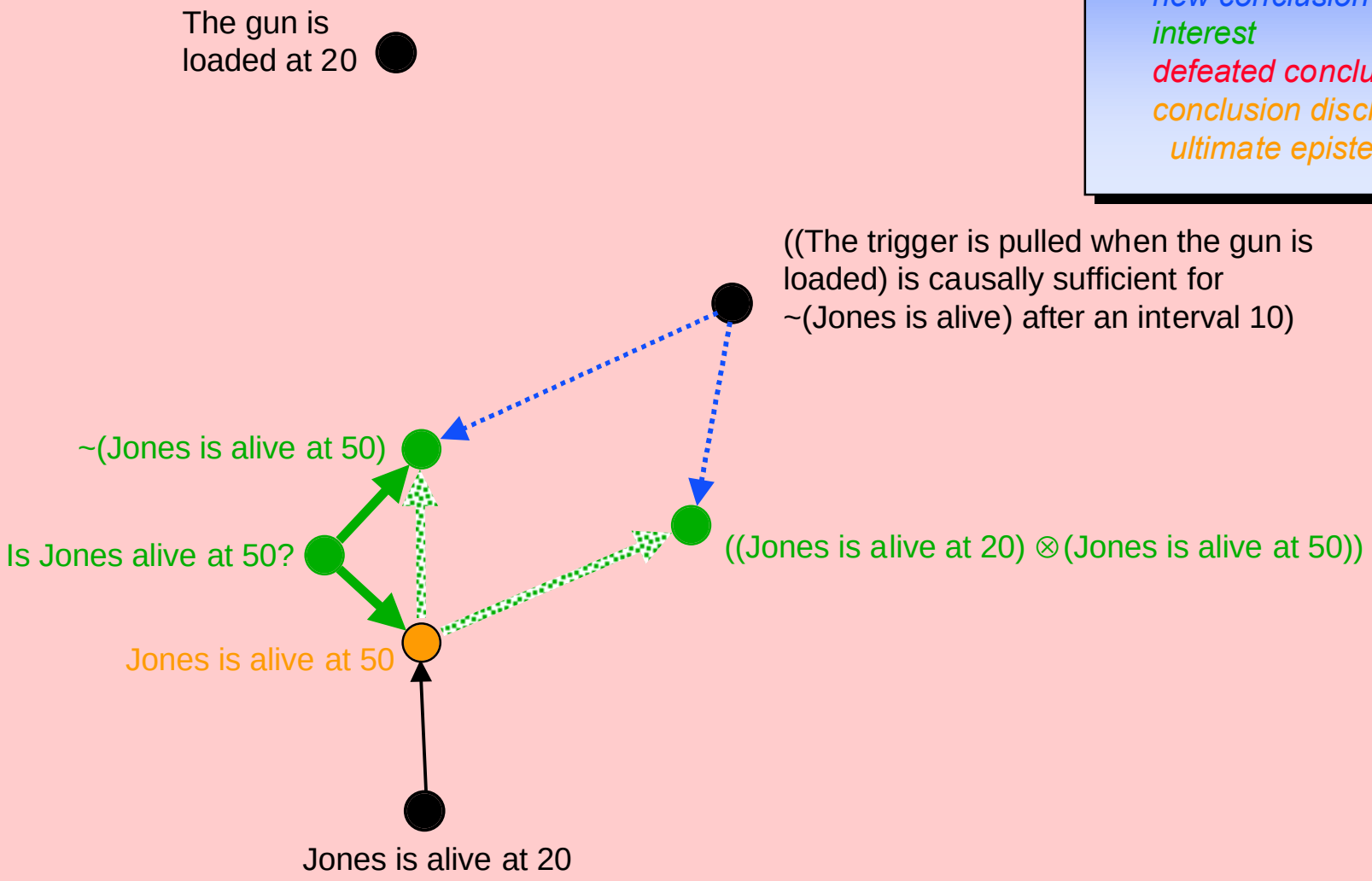


Discharging 1st premise of CAUSAL-IMPLICATION

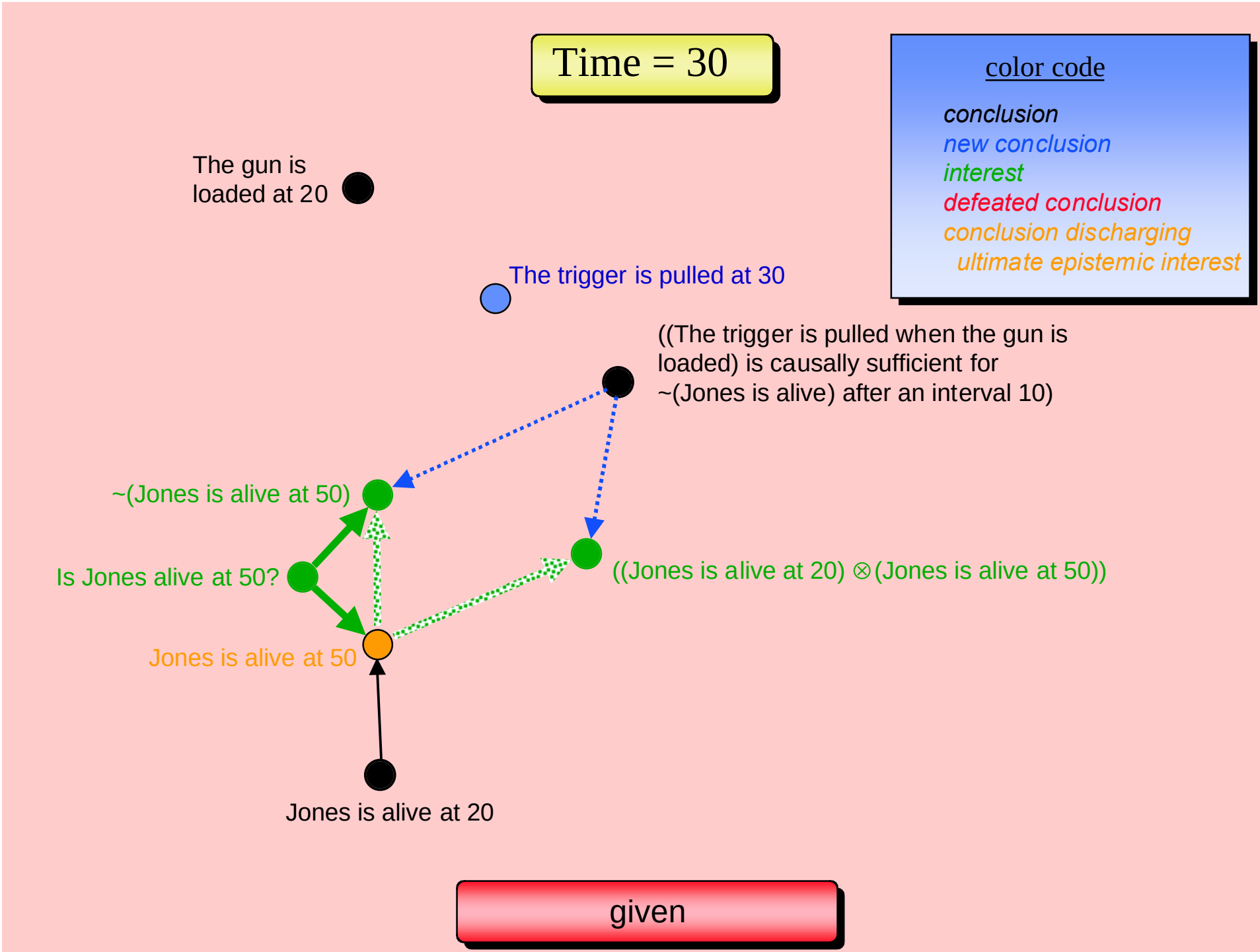
Time = 25

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest



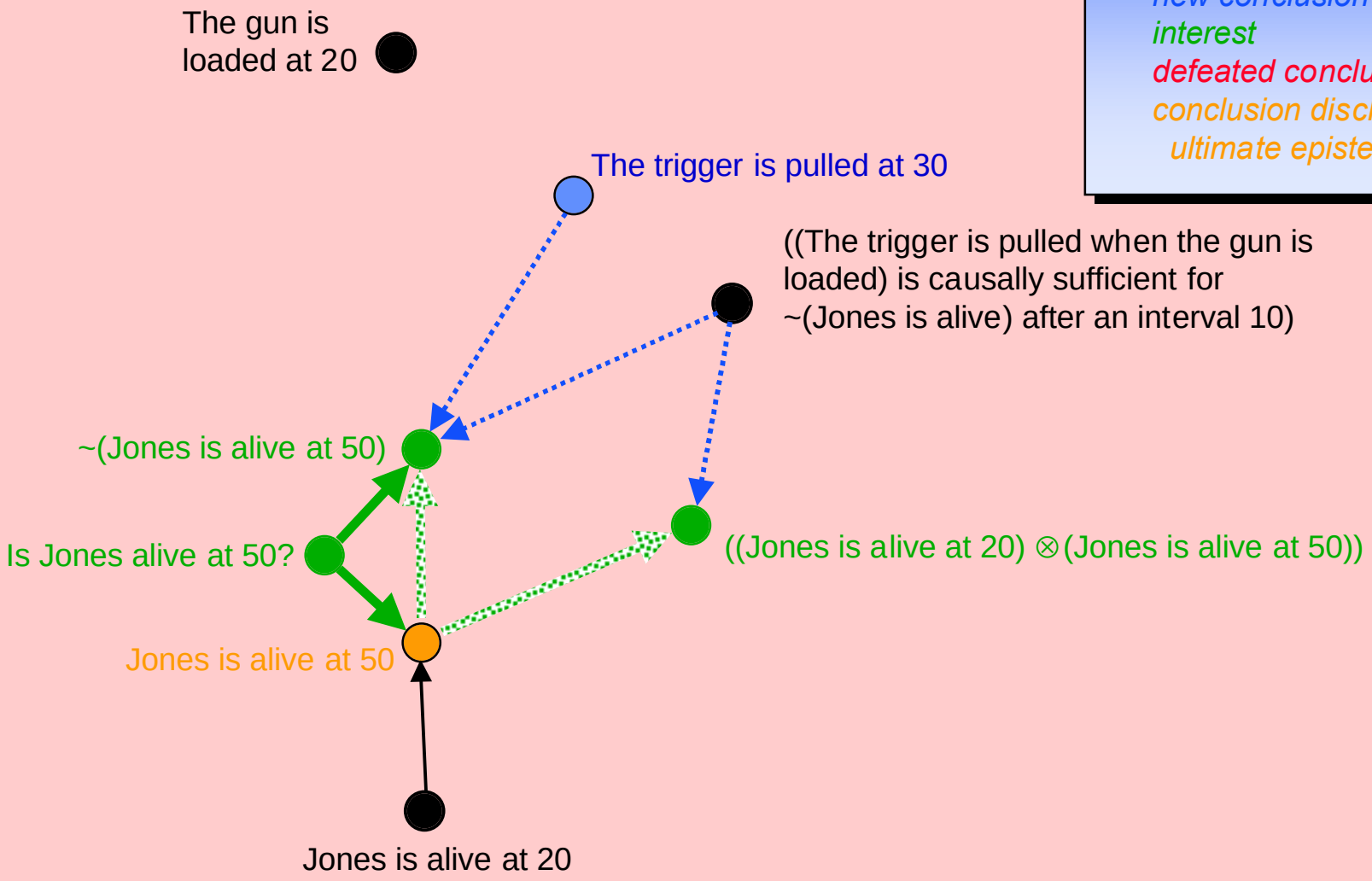
Discharging 1st premise of CAUSAL-UNDERCUTTER


Time = 30
color code
conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest
The gun is loaded at 20
The trigger is pulled at 30
((The trigger is pulled when the gun is loaded) is causally sufficient for ~(Jones is alive) after an interval 10)
~(Jones is alive at 50)
Is Jones alive at 50?
((Jones is alive at 20) ⊗(Jones is alive at 50))
Jones is alive at 50
Jones is alive at 20
given

Time = 31

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

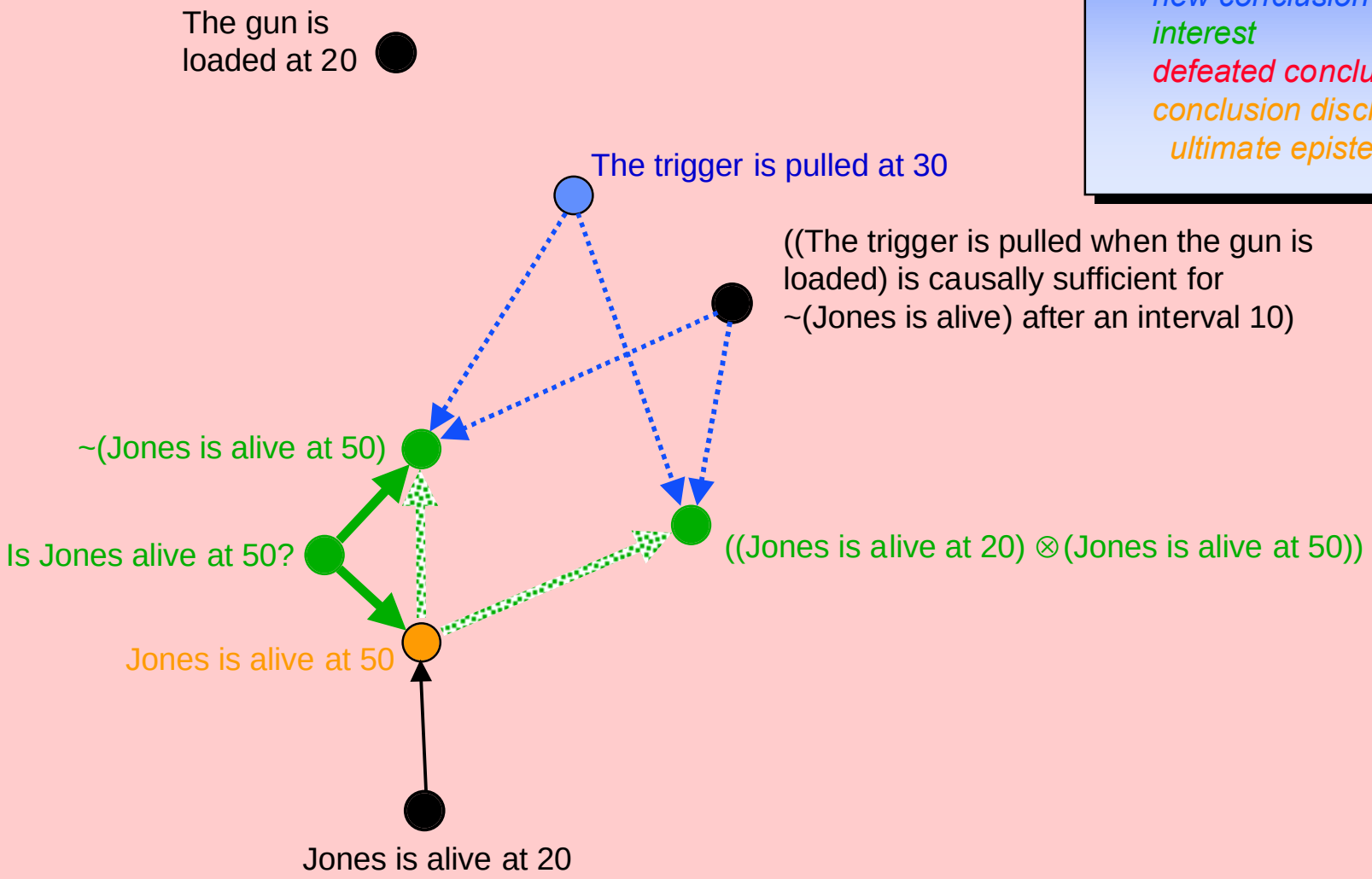


Discharging 2nd premise of CAUSAL-IMPLICATION

Time = 32

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

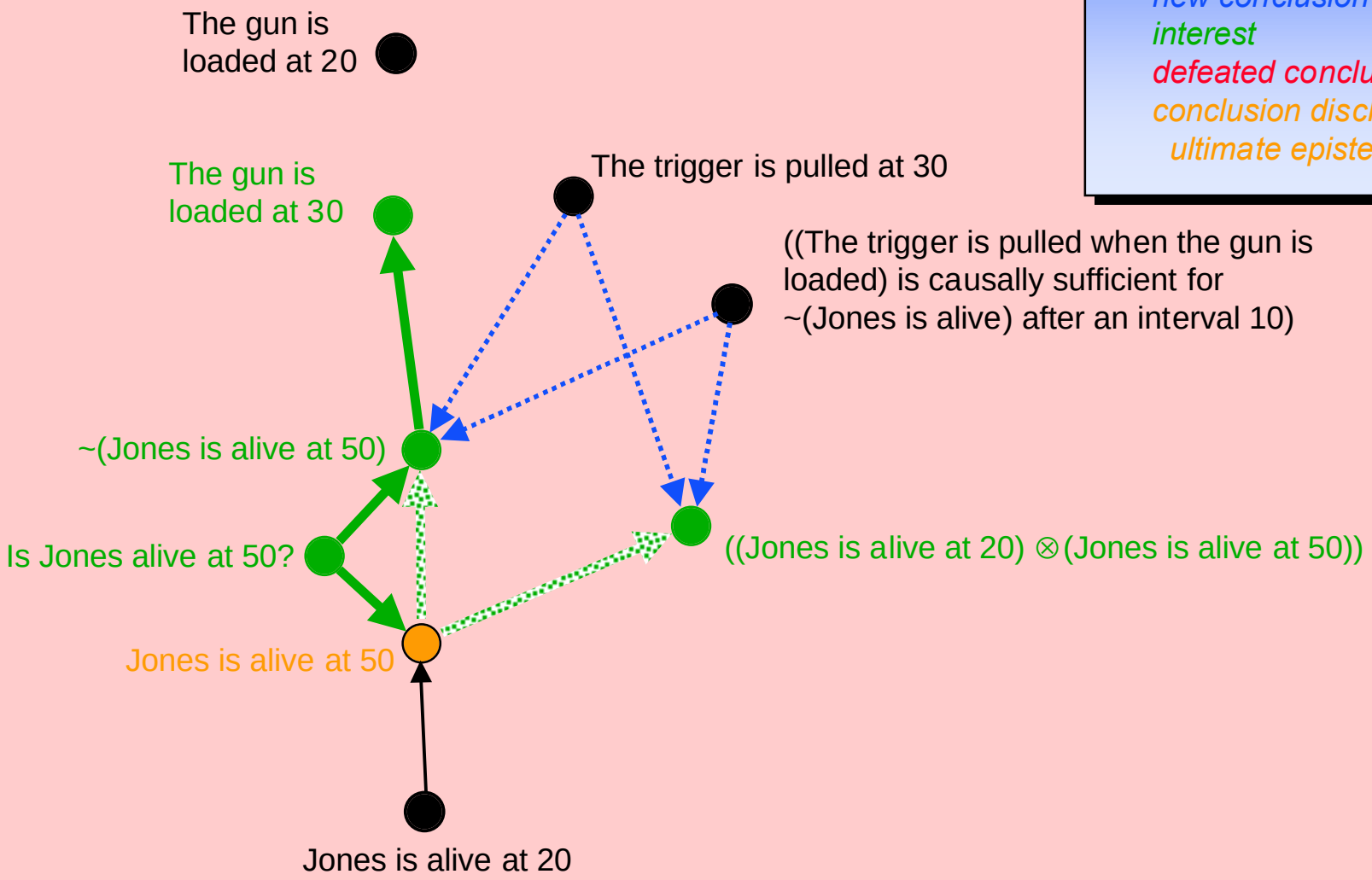


Discharging 2nd premise of CAUSAL-UNDERCUTTER

Time = 33

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

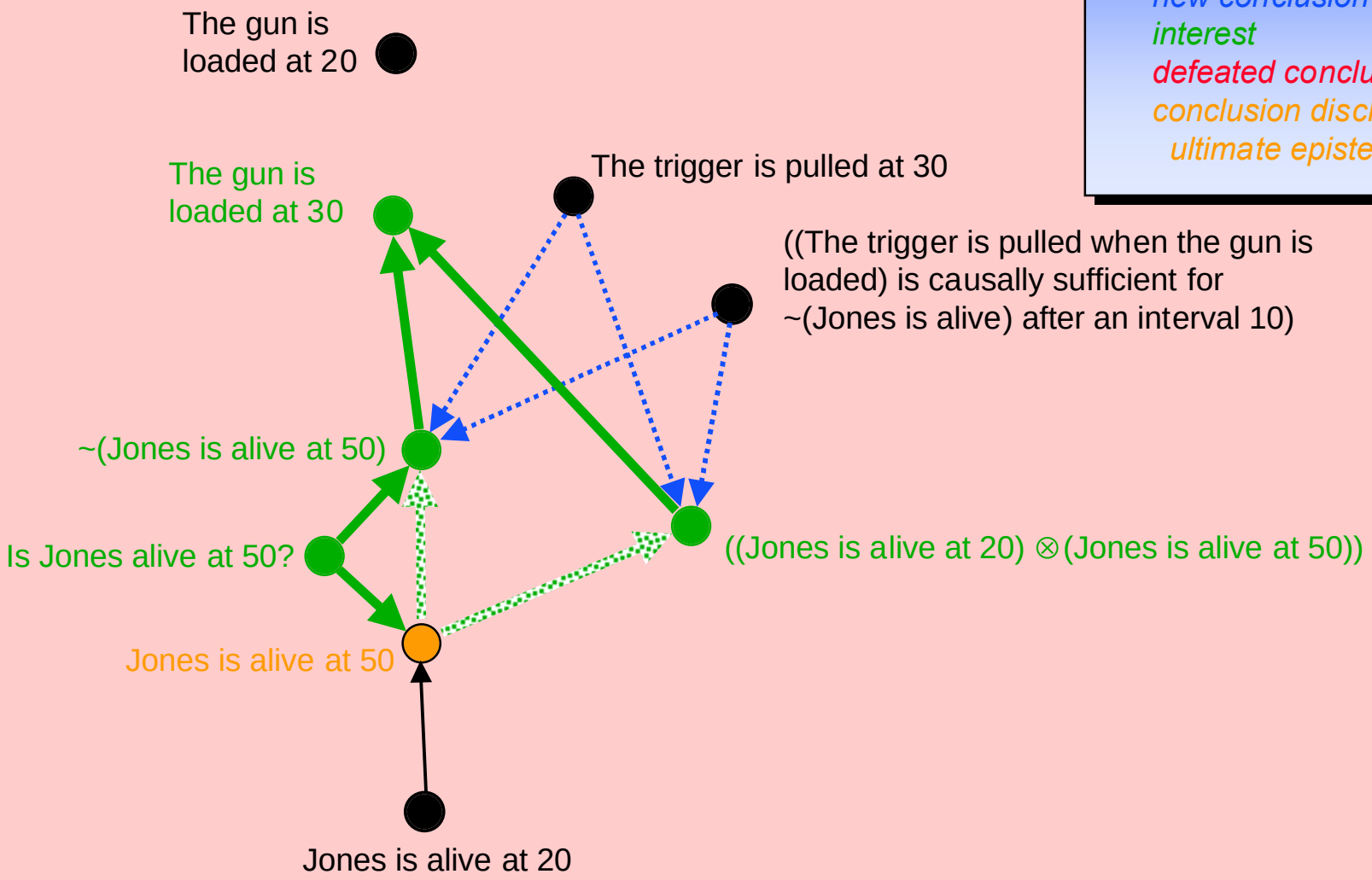


Interest in 3rd premise of CAUSAL-IMPLICATION

Time = 34

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

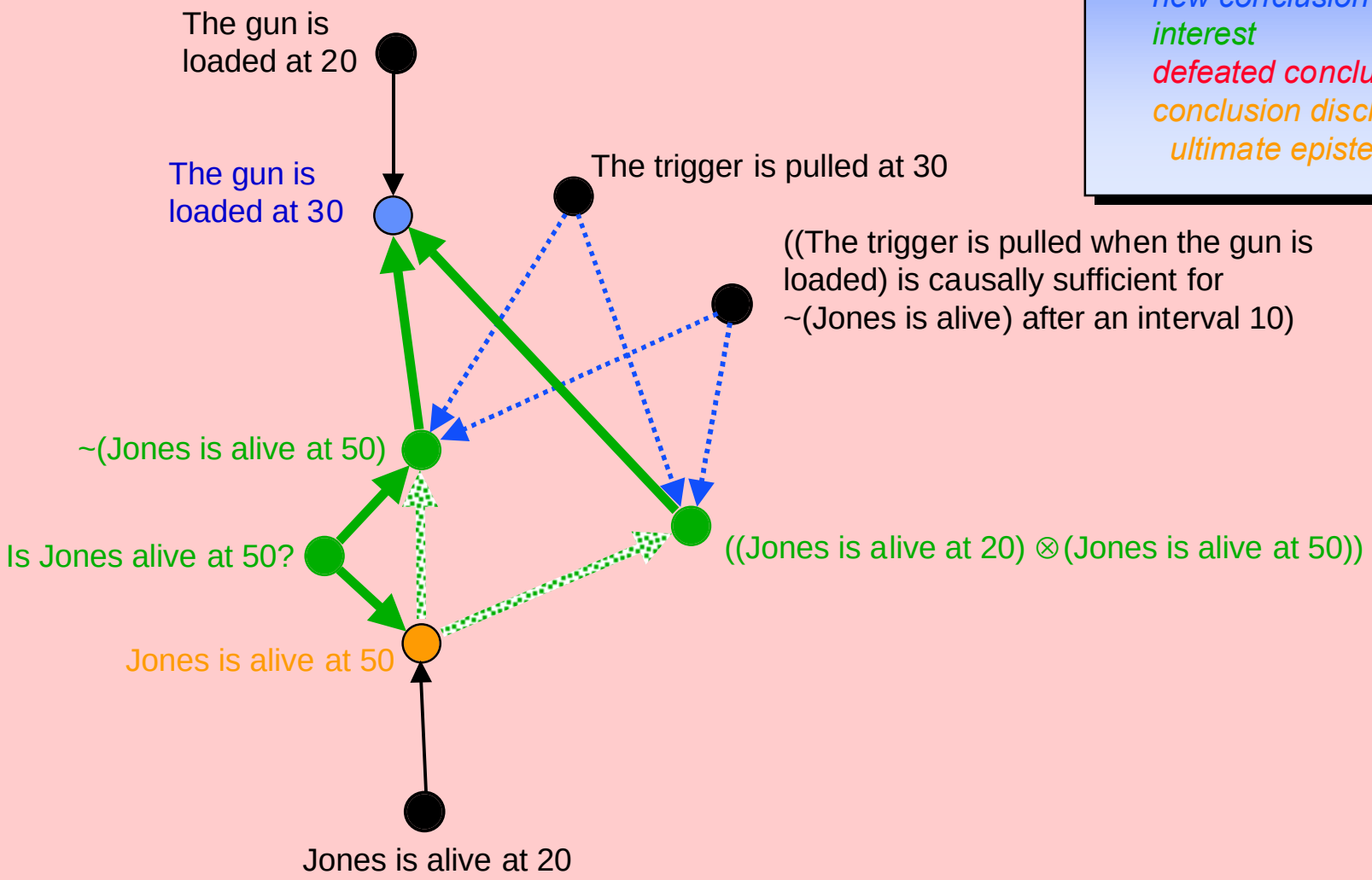


Interest in 3rd premise of CAUSAL-UNDERCUTTER

Time = 35

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

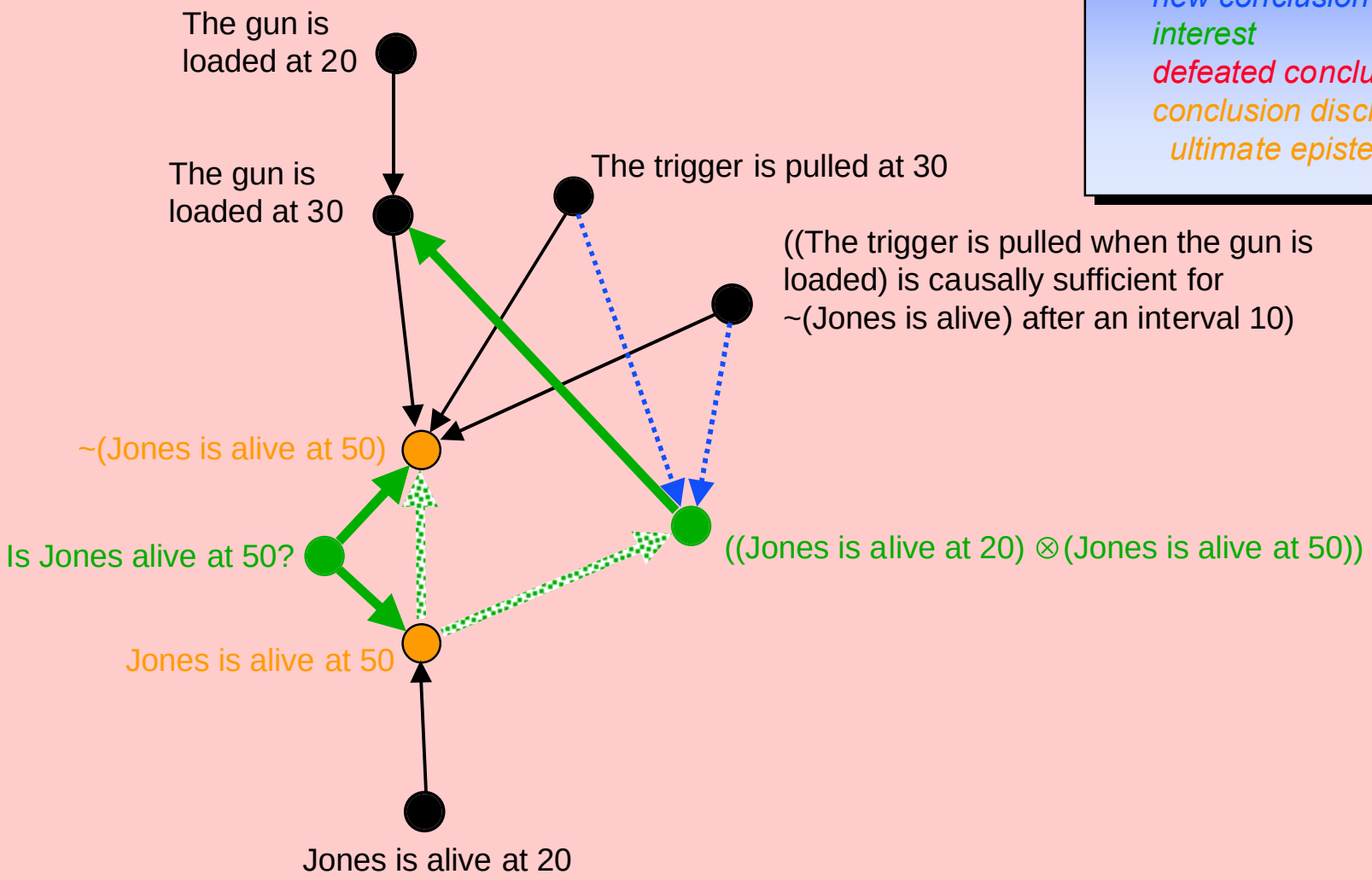


by TEMPORAL PROJECTION

Time = 36

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

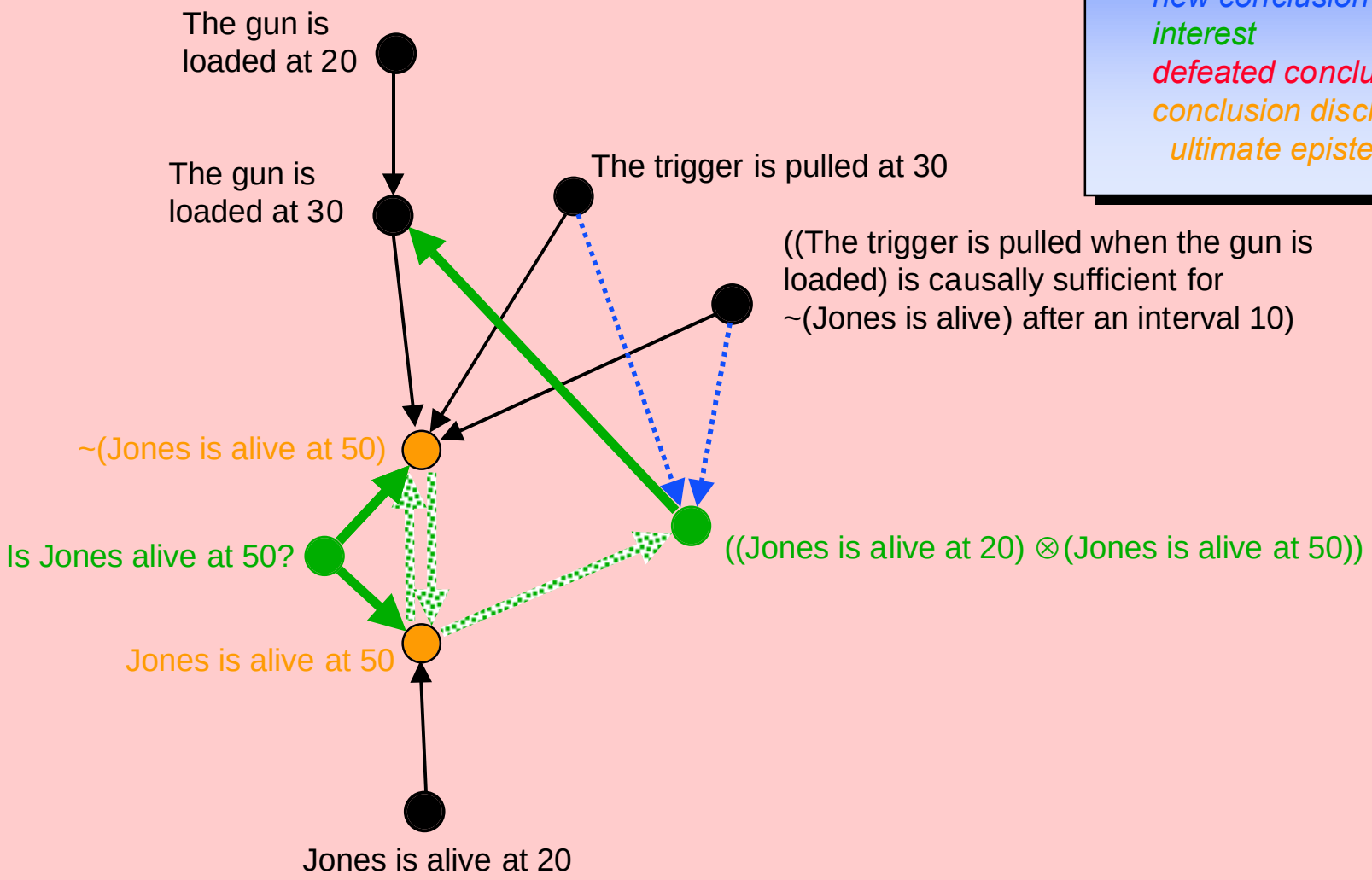


by CAUSAL-UNDERCUTTER

Time = 36

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

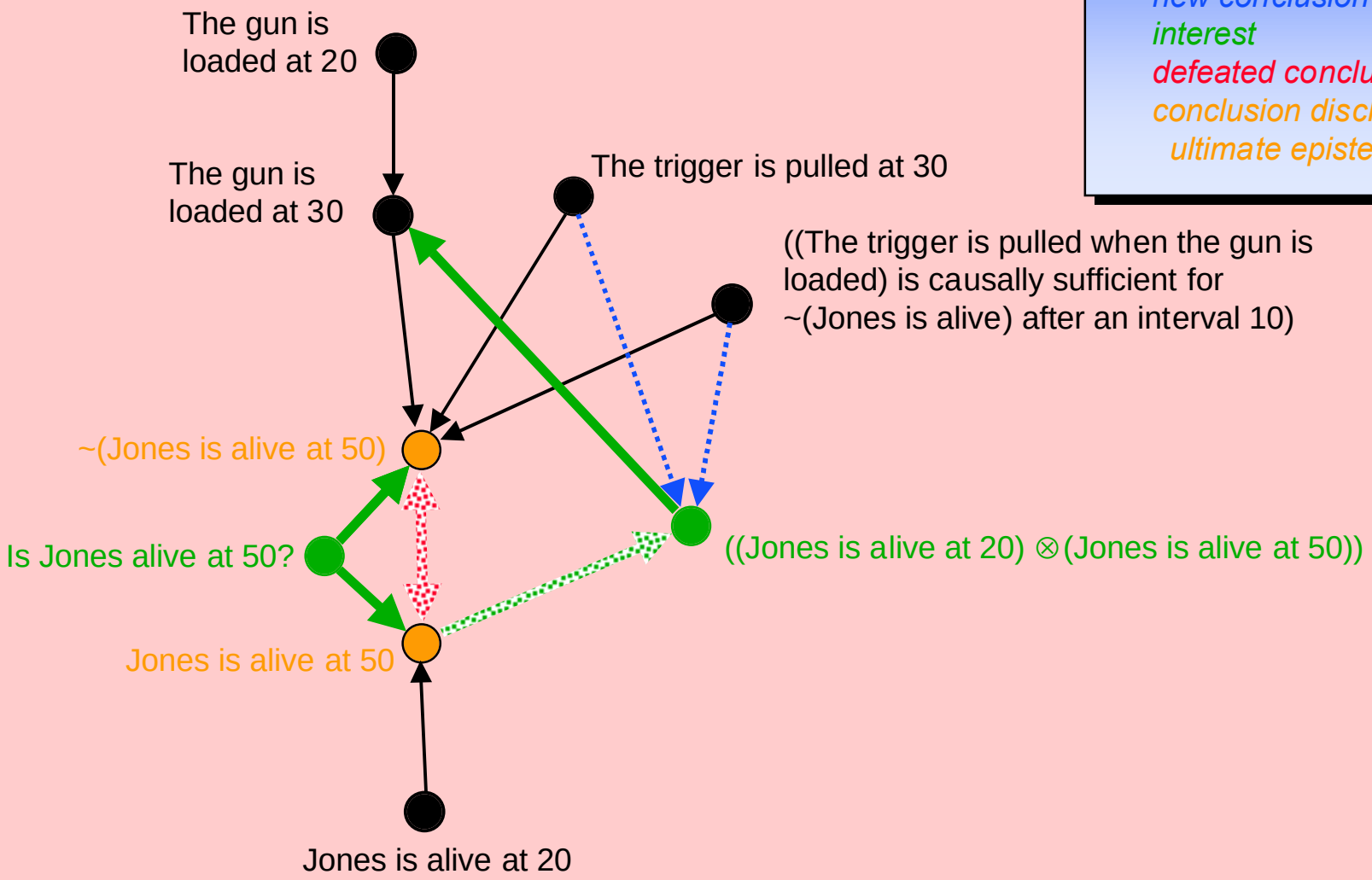


Interest in rebutter

Time = 37

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

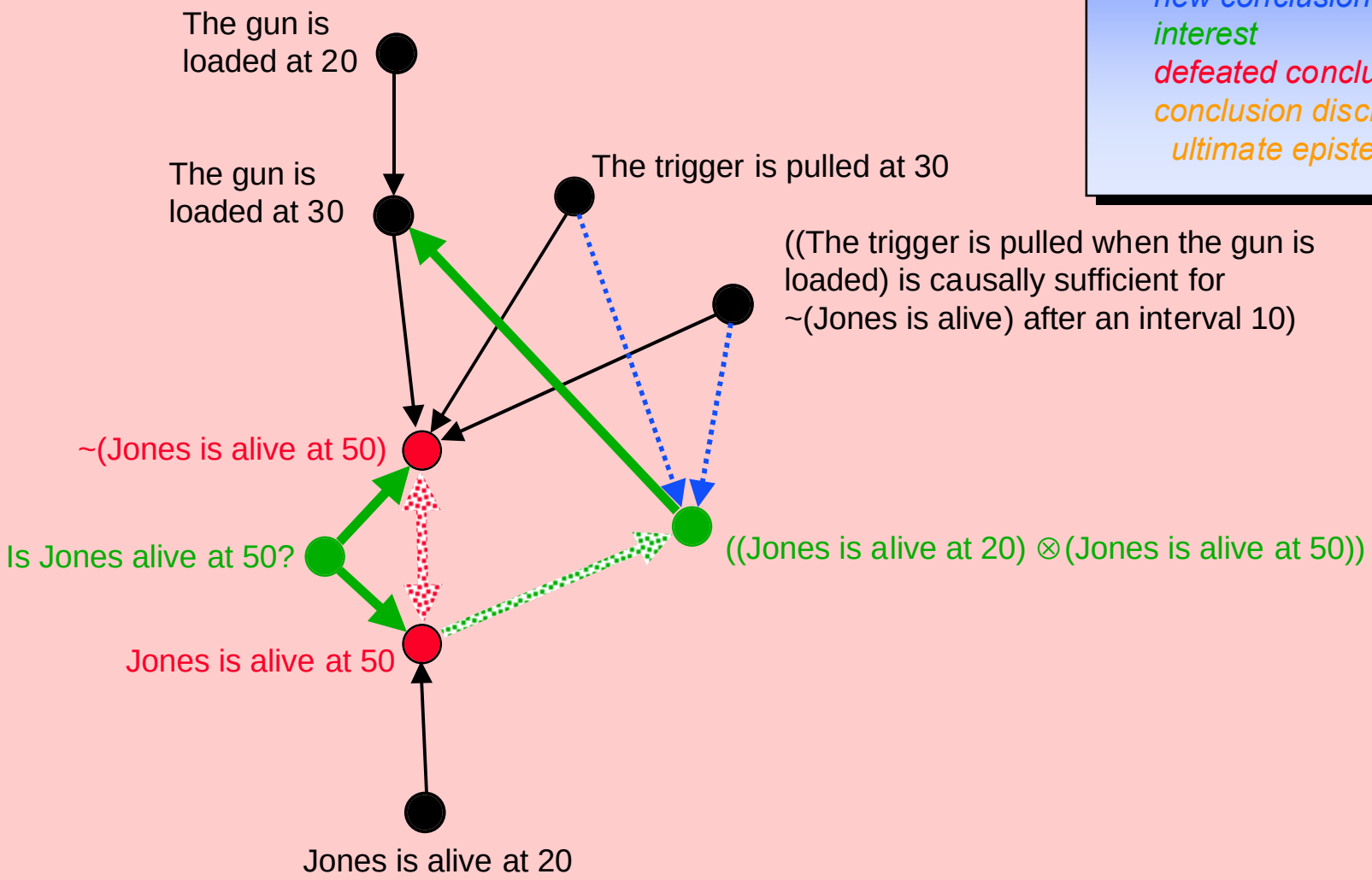


Discharge interests

Time = 38

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

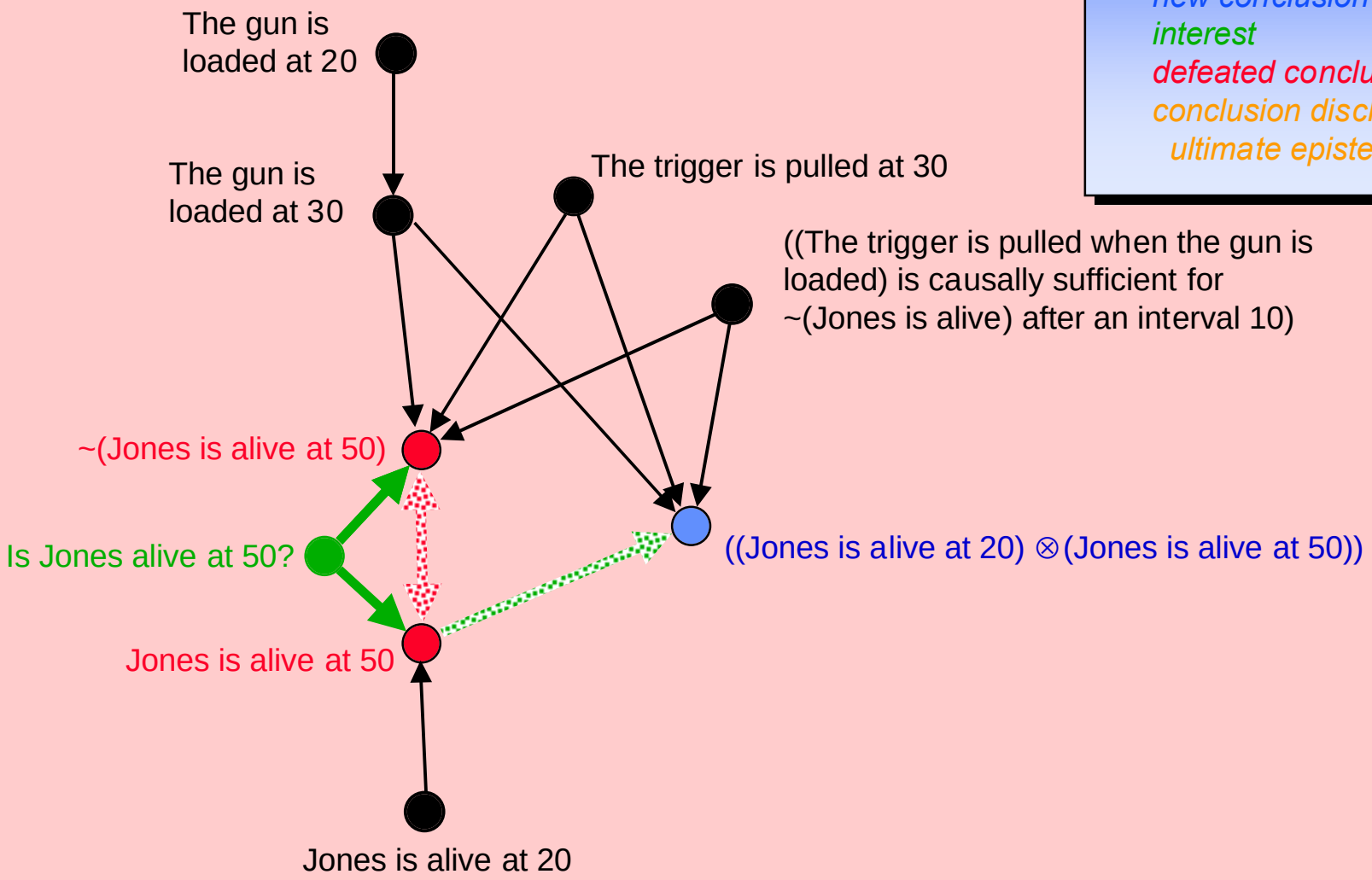


Defeat status computation

Time = 39

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

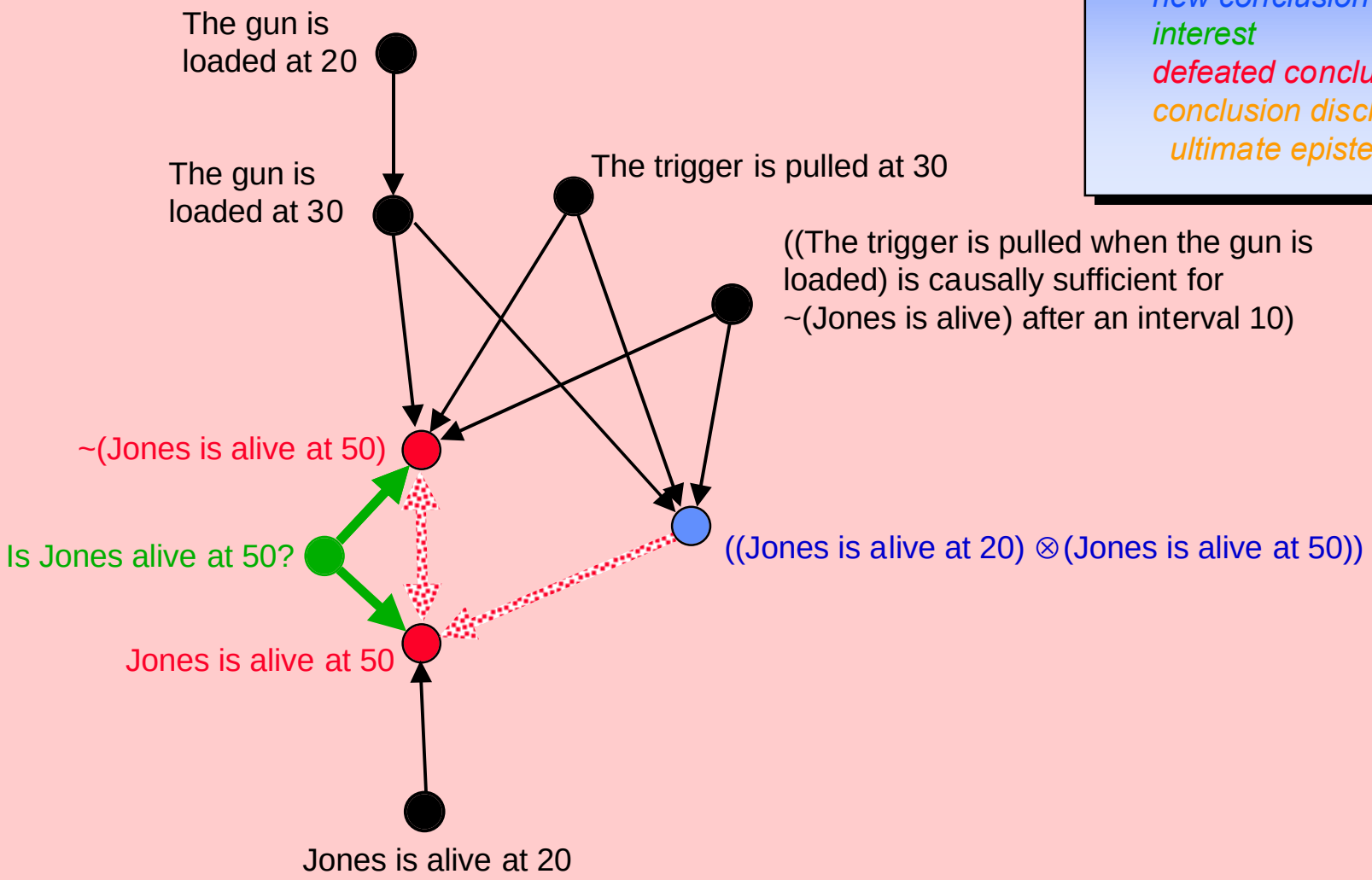


By CAUSAL-UNDERCUTTER

Time = 39

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

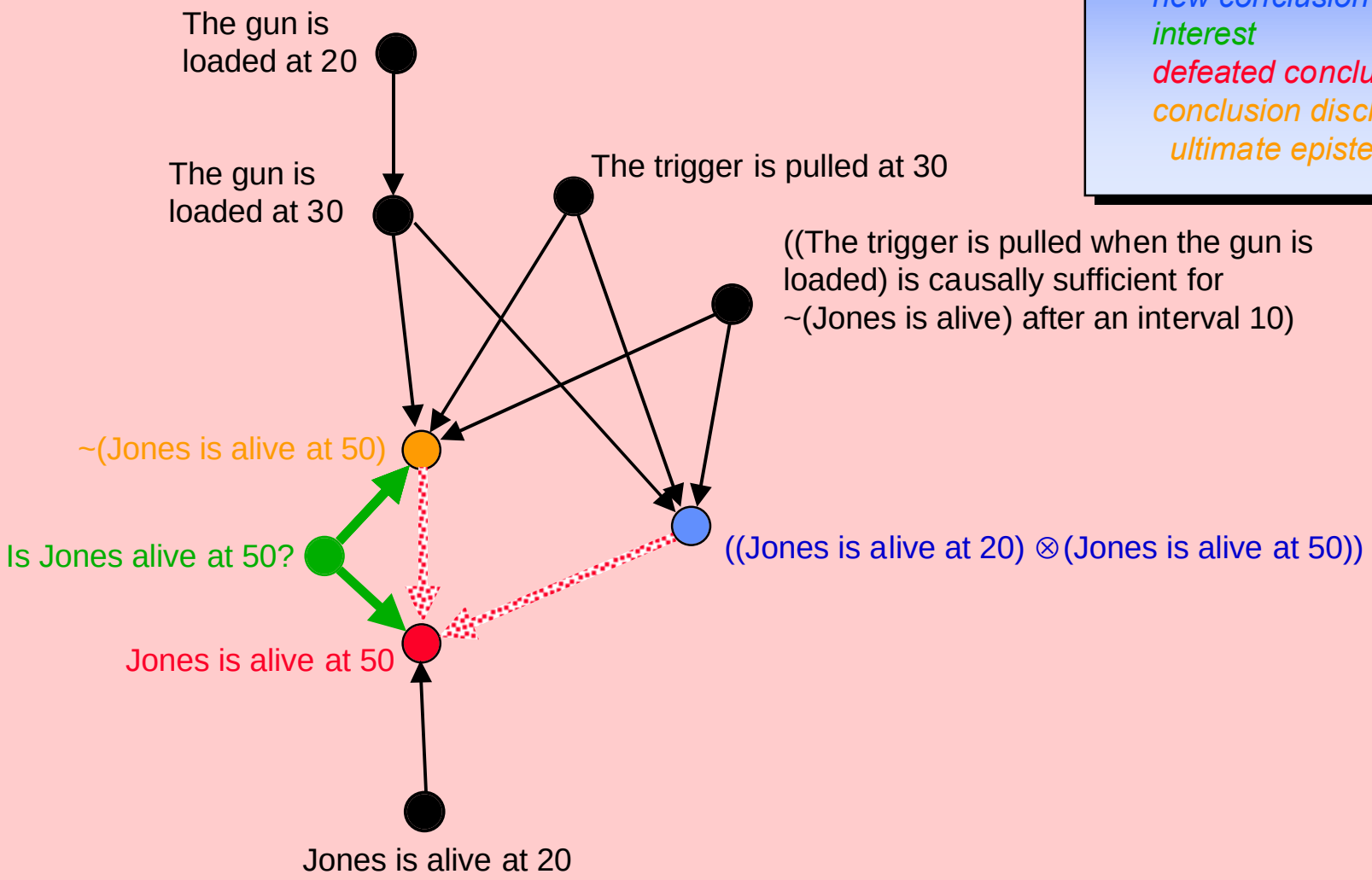


Defeat status computation

Time = 39

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest

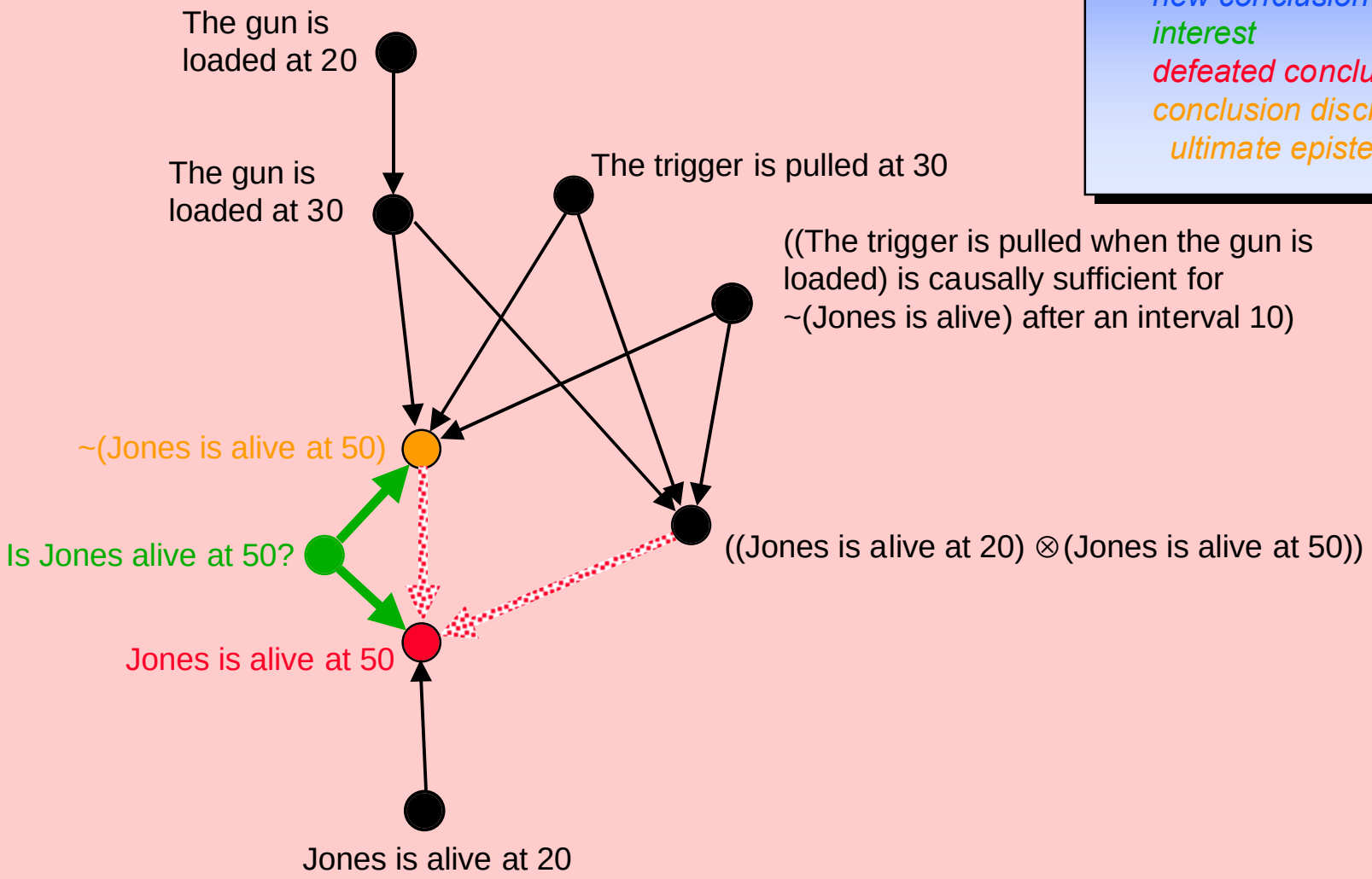


Defeat status computation

Time = 39+

color code

conclusion
new conclusion
interest
defeated conclusion
conclusion discharging
ultimate epistemic interest



Defeat status computation

Part Four: Practical Cognition

- Given an agent capable of sophisticated epistemic cognition, how can it make use of that in practical cognition?
- We can regard practical cognition as having four components:
  - goal selection
  - plan-construction
  - plan-selection
  - plan-execution
- Although it is natural to think of these as components of practical cognition, most of the work will be carried out by epistemic cognition.

Plan Construction

- Standard planning algorithms assume that we come to the planning problem with all the knowledge needed to solve it.
- This assumption fails for autonomous rational agents.
  - The more complex the environment, the more the agent will have to be self-sufficient for knowledge acquisition.
  - The principal function of epistemic cognition is to provide the information needed for practical cognition.
  - As such, the course of epistemic cognition is driven by practical interests.
  - Rather than coming to the planning problem equipped with all the knowledge required for its solution, the planning problem itself directs epistemic cognition, focusing epistemic endeavors on the pursuit of information that will be helpful in solving current planning problems.

Plan Construction

- Paramount among the information an agent may have to acquire in the course of planning is knowledge about the consequences of actions under various circumstances.
- Sometimes this knowledge can be acquired by reasoning from what is already known.
- Often it will require empirical investigation.
  - Empirical investigation involves acting, and figuring out what actions to perform requires further planning.
- So planning drives epistemic investigation, which may in turn drive further planning.
- In autonomous rational agents operating in a complex environment, planning and epistemic investigation must be interleaved.

Goal Regression Planning

- I assume that rational agents will engage in some form of goal-regression planning.
- This involves reasoning backwards from goals to subgoals whose achievement will enable an action to achieve a goal.
  - Such reasoning proceeds in terms of causal knowledge of the form "performing action A under circumstances C is causally sufficient for achieving goal G". *Planning-conditional* (A/C) $\Rightarrow$ G.
- Subgoals are typically conjunctions.
- We usually lack causal knowledge pertaining directly to conjunctions, and must instead use causal knowledge pertaining to the individual conjuncts.
  - We plan separately for the conjuncts of a conjunctive subgoal.
  - When we merge the plans for the conjuncts, we must ensure that the separate plans do not destructively interfere with each other. (We must "resolve threats".)

Avoiding Destructive Interference

- **Conventional planners assume that the planner already knows the consequences of actions under all circumstances, and so destructive interference can be detected by just checking the consequences.**
- **An autonomous rational agent may have to engage in arbitrarily much epistemic investigation to detect destructive interference.**
  - **The set of "threats" will not be recursive.**
- **Theorem: If the set of threats is not recursive, then the set of planning (problem,solution) pairs is not recursively enumerable.**
- **Corollary: A planner that insists upon ruling out threats before merging plans for the conjuncts of a conjunctive goal may never terminate.**

"The logical foundations of goal-regression planning in autonomous agents", *Artificial Intelligence,* 1998.

Defeasible Planning

- If the set of threats is not recursive, a planner must operate defeasibly, *assuming* that there are no threats unless it has reason to believe otherwise.
- That a plan will achieve a goal is a factual matter, of the sort normally addressed by epistemic cognition.
- So we can perform plan-search by adopting a set of defeasible reason-schemas for reasoning about plans.

Principles of Defeasible Planning

GOAL-REGRESSION

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

SPLIT-CONJUNCTIVE-GOAL

Given an interest in finding a plan for achieving (G_1-at-t_1 & G_2-at-t_2), adopt interest in finding plans $plan_1$ for G_1-at-t_1 and $plan_2$ for G_2-at-t_2. Given such plans, infer defeasibly that the result of merging them will achieve (G_1-at-t_1 & G_2-at-t_2).

— a number of additional reason-schemas are also required —

"The logical foundations of goal-regression planning in autonomous agents", *Artificial Intelligence,* 1998.

Example — Pednault's Briefcase

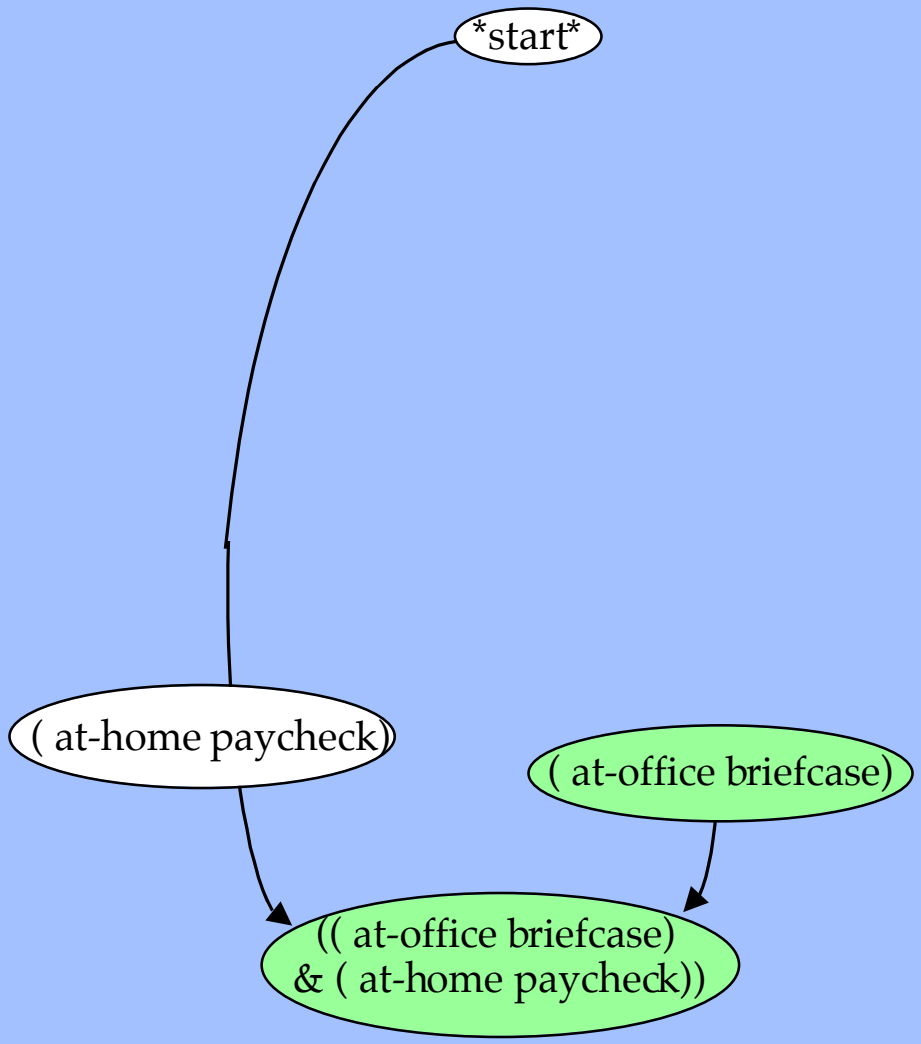
- (at-home briefcase)
- (at-home paycheck)
- (in-briefcase paycheck)
- $(\forall x)[((\text{in-briefcase } x) \,\&\, (\text{remove-from-briefcase } x)) \Rightarrow \sim(\text{in-briefcase } x)]$
- $[(\text{at-home briefcase}) \,\&\, \text{take-briefcase-to-office}] \Rightarrow (\text{at-office briefcase})$
- $(\forall x)[((\text{at-home briefcase}) \,\&\, (\text{in-briefcase } x) \,\&\, \text{take-briefcase-to-office}) \Rightarrow \sim(\text{at-home } x)]$

Goal: (at-home paycheck) & (at-office briefcase)

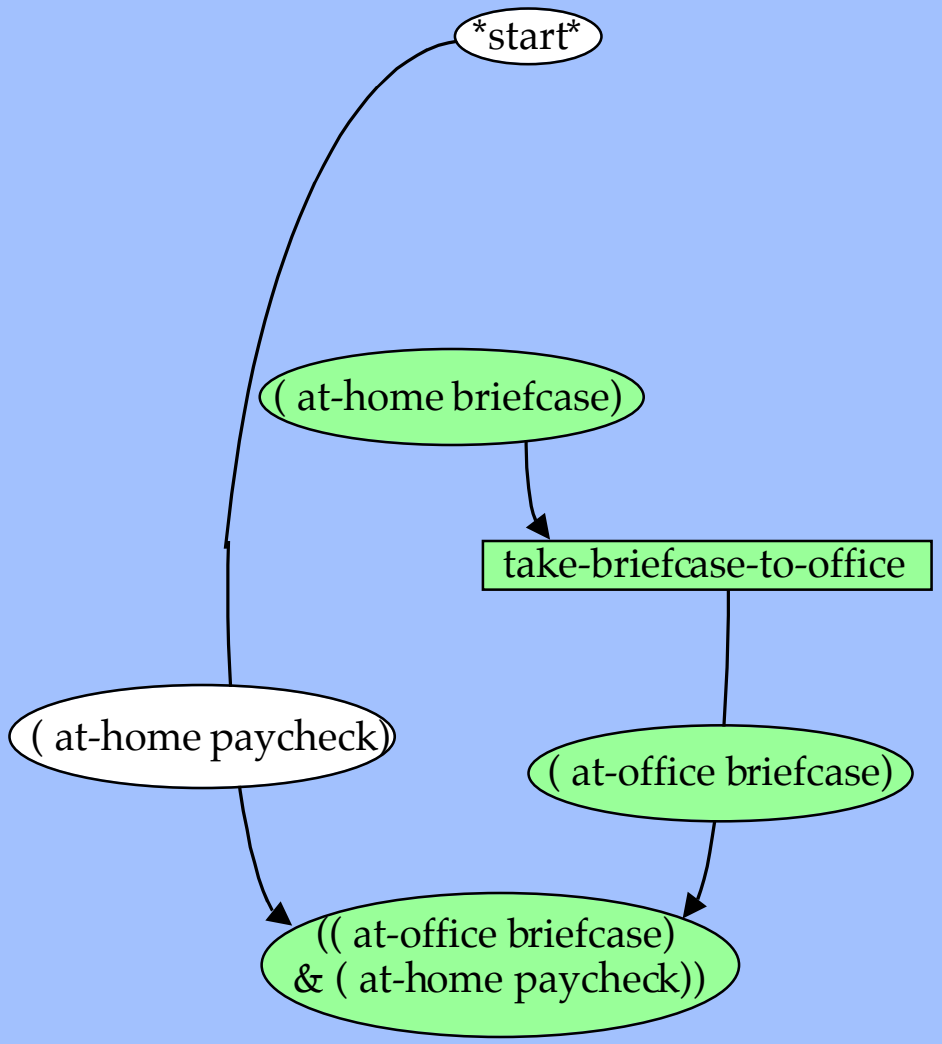
Example — Pednault's Briefcase

Example — Pednault's Briefcase

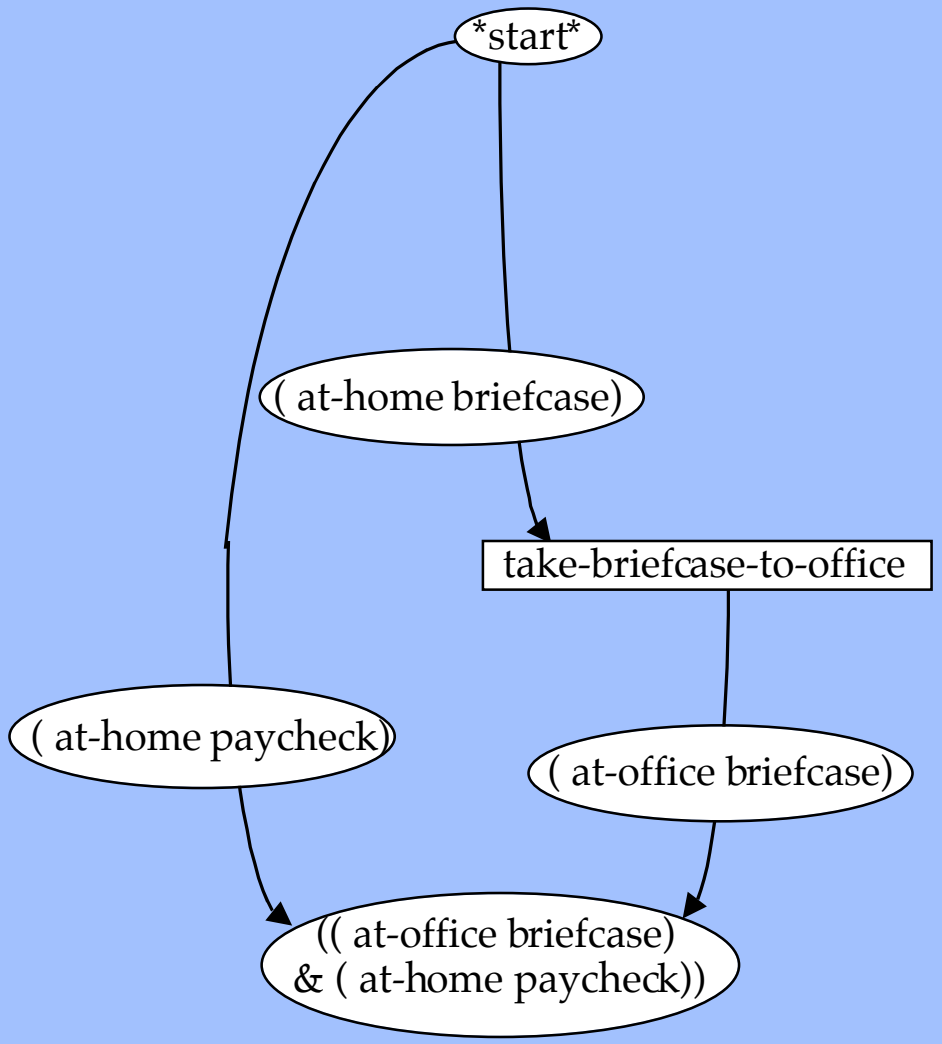
Example — Pednault's Briefcase



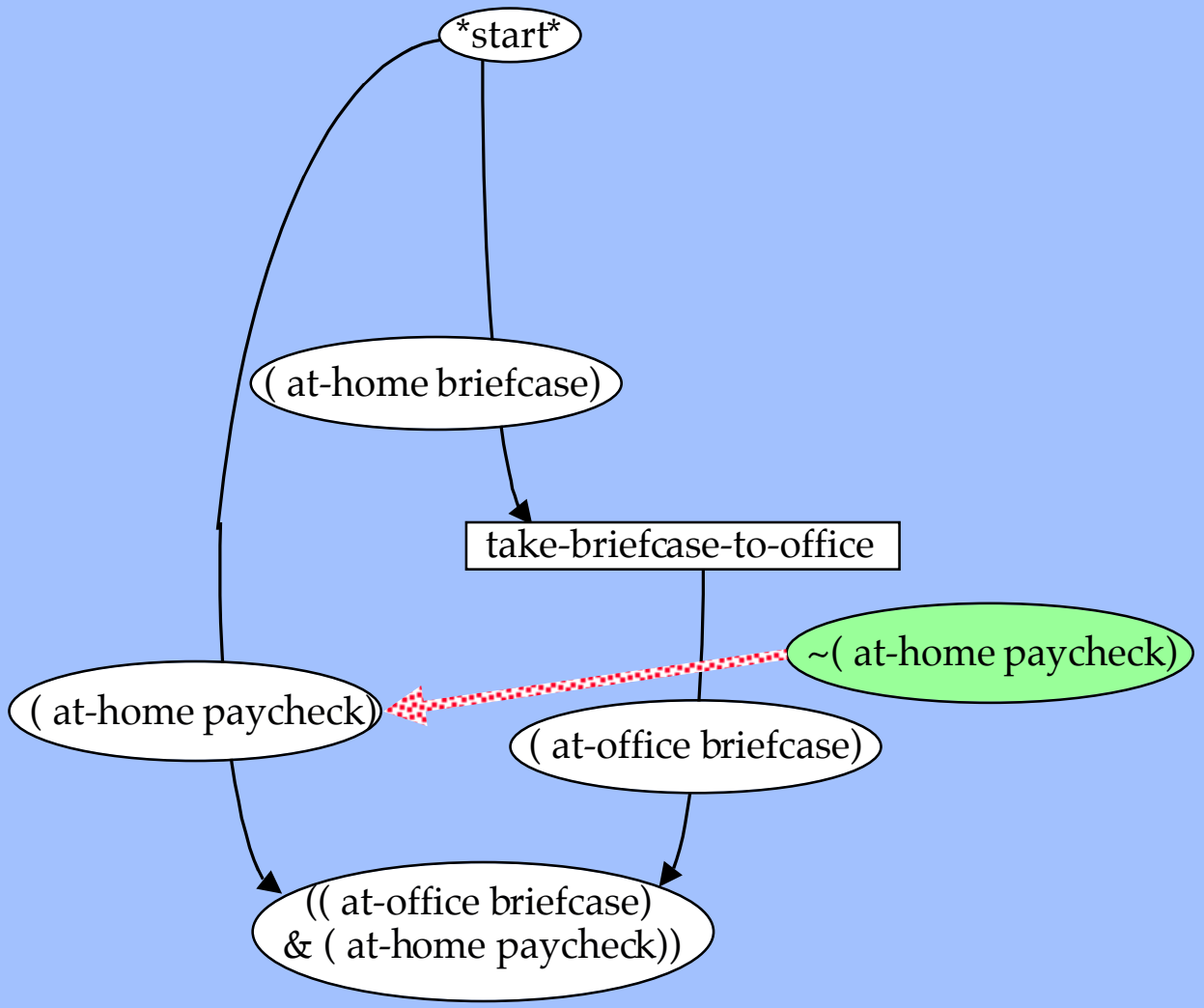
Example — Pednault's Briefcase



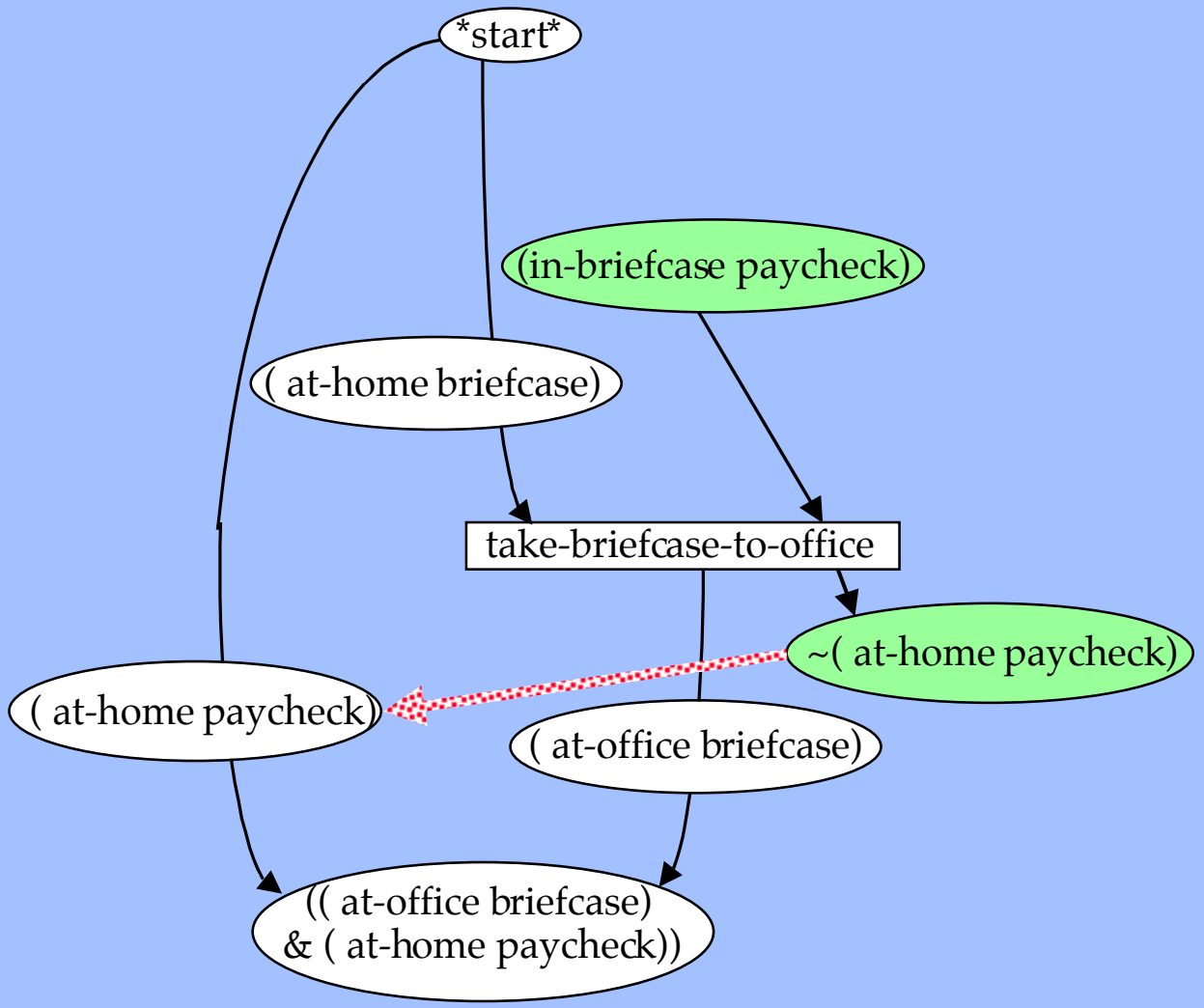
Example — Pednault's Briefcase



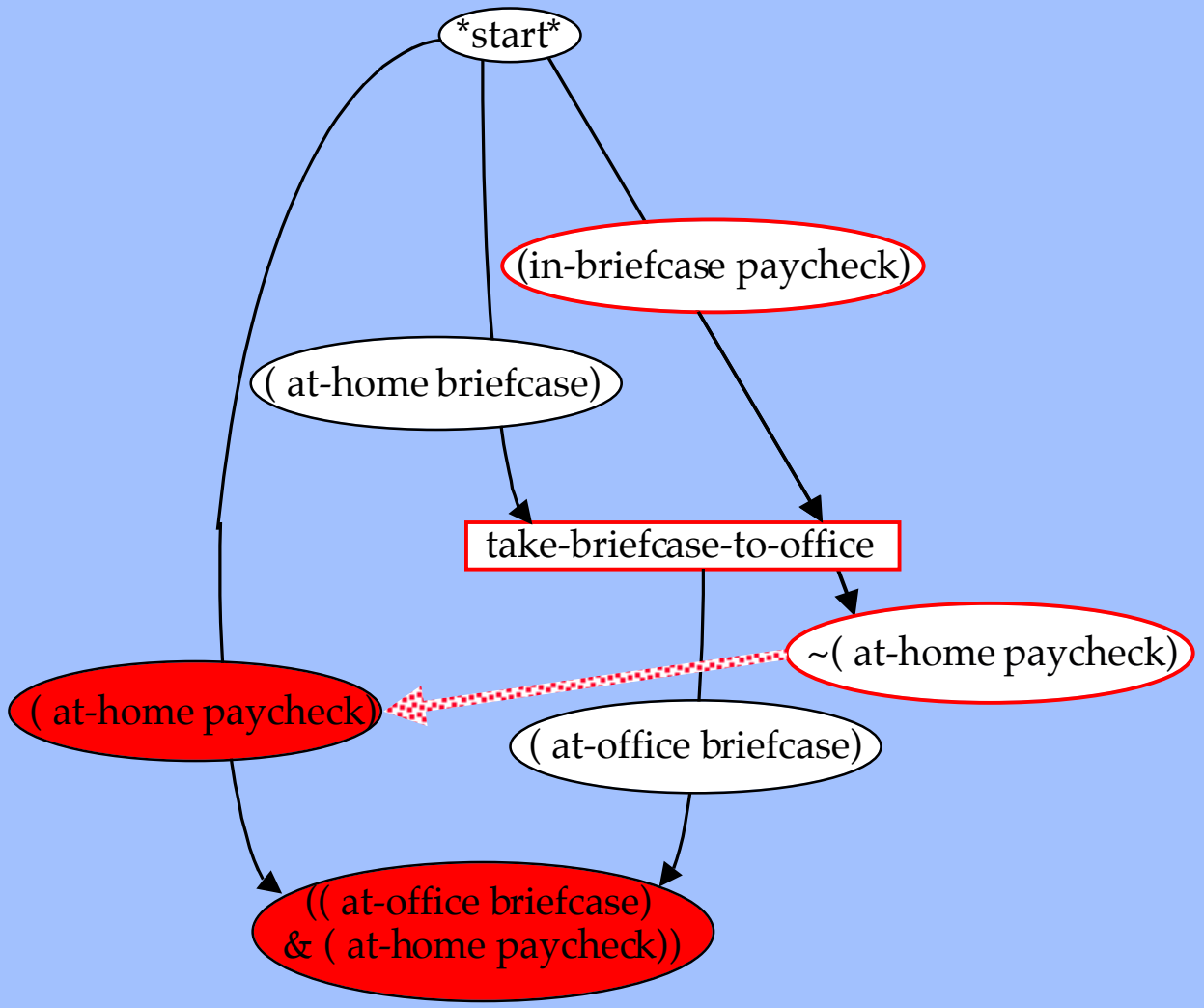
Example — Pednault's Briefcase



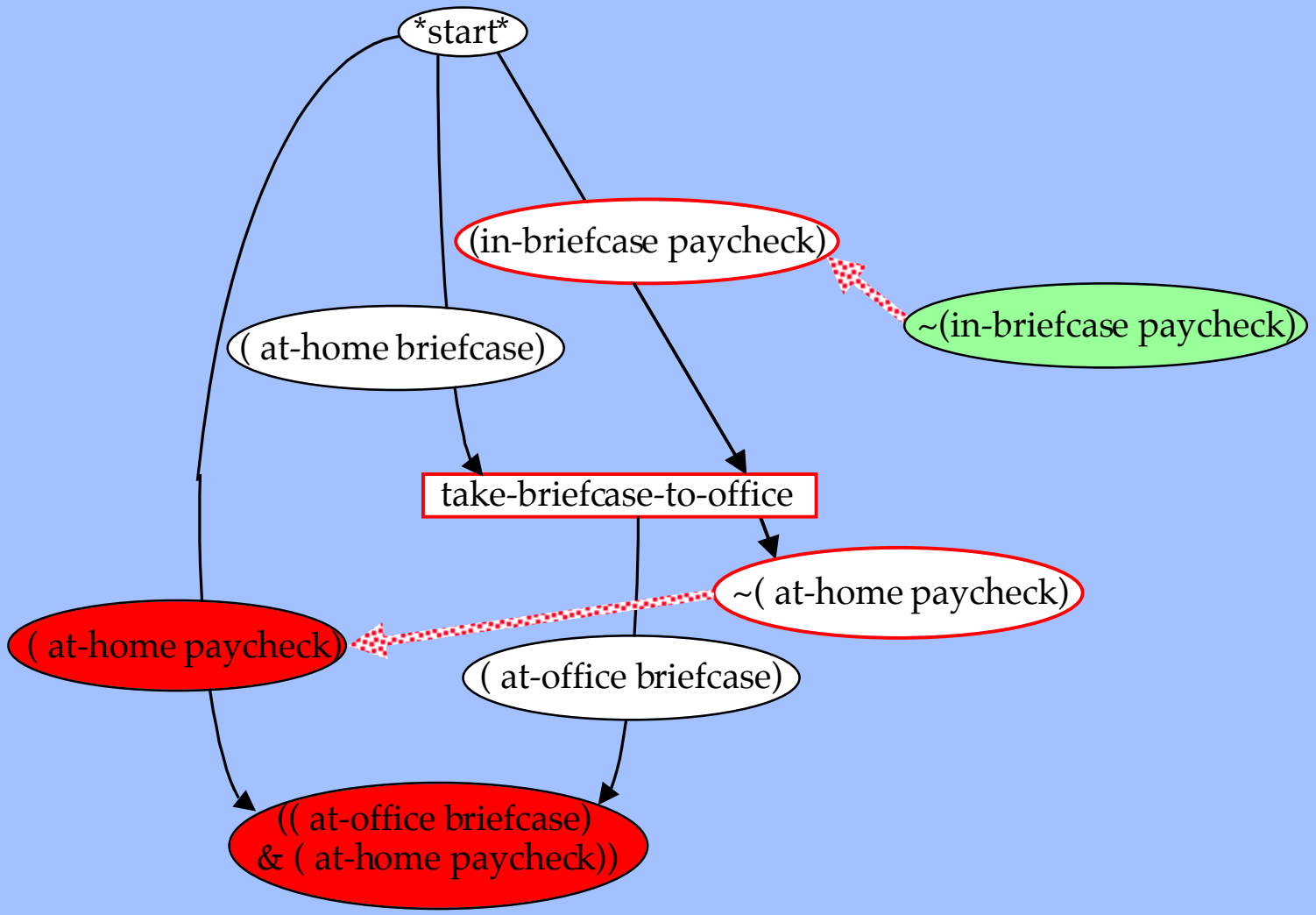
Example — Pednault's Briefcase



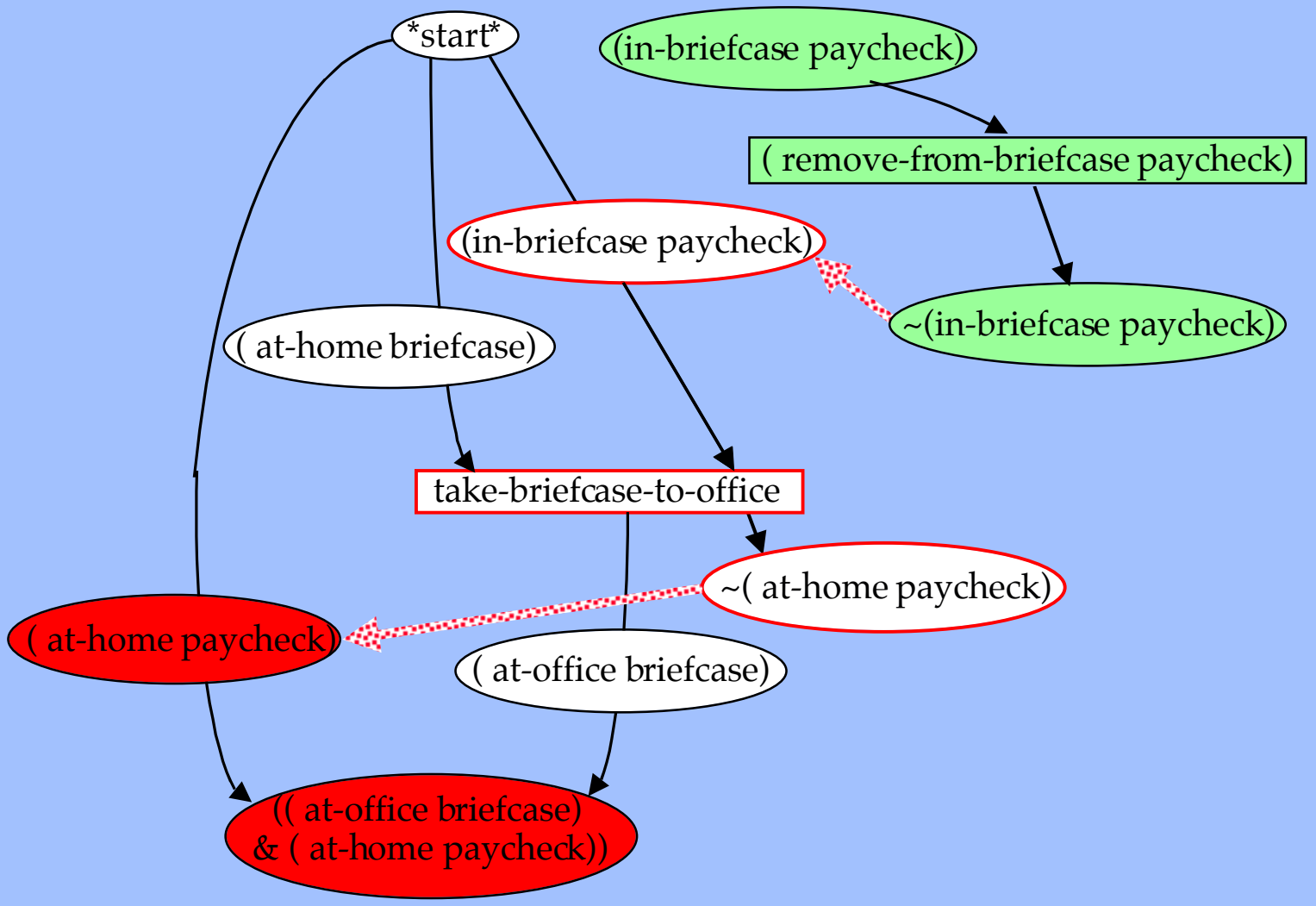
Example — Pednault's Briefcase



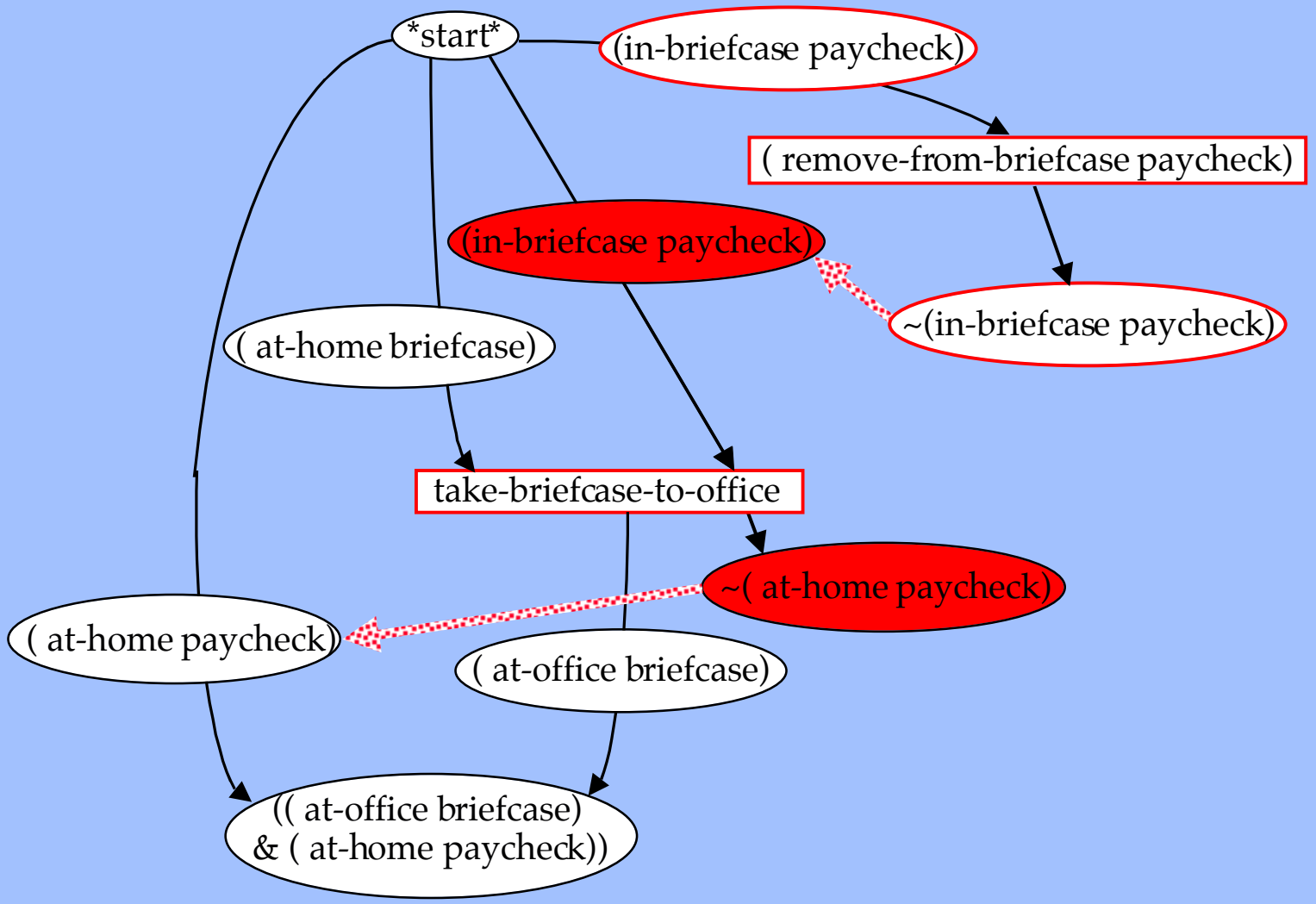
Example — Pednault's Briefcase



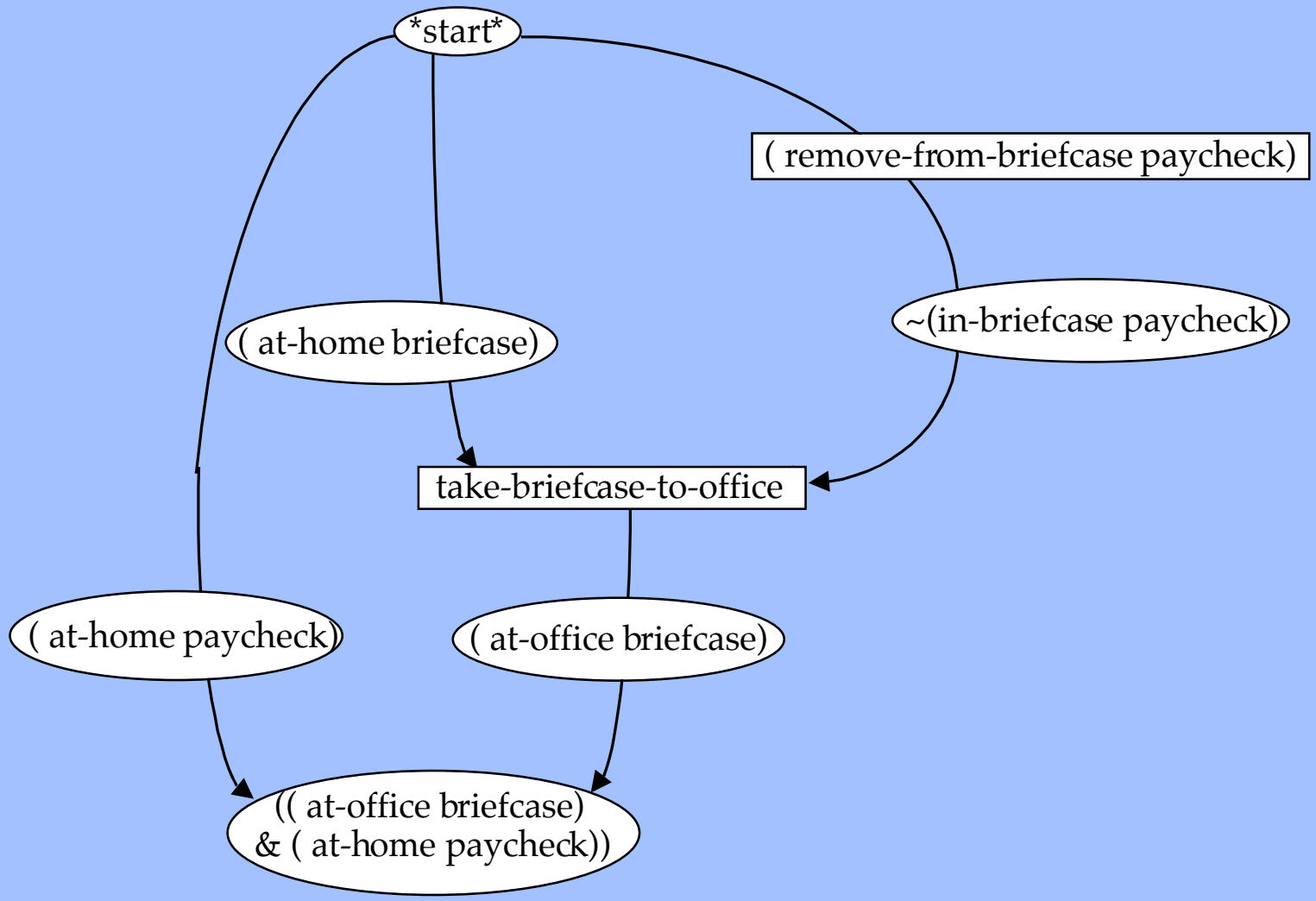
Example — Pednault's Briefcase



Example — Pednault's Briefcase



Example — Pednault's Briefcase



Adopting Plans (work in progress)

- Plan construction produces plans that purport to achieve their goals, but *adopting* such a plan requires a further cognitive step.
  - Such a plan is not automatically adoptable:
    - » Its execution costs might be greater than the value of the goal achieved.
    - » It might interact adversely with other plans already adopted, increasing their execution costs or lowering the value of their goals.
      - Sometimes the impacted plan should be rejected.
      - Sometimes the new plan should be rejected.
- In deciding whether to adopt a plan, we must evaluate it in a roughly decision-theoretic manner.
- This might suggest the use of Markov-decision planning (POMDP's).
  - Assertion — such planning will always be computationally infeasible in planning domains of real-world complexity.

Decision-Theoretic Goal-Regression

- Proposal — It is possible to perform feasible decision-theoretic planning by modifying conventional goal-regression planning in certain ways.
- Goal-regression planning can be performed by applying classical planning algorithms but appealing to probabilistic connections rather than exceptionless causal connections.
  - This is computationally easier than "probabilistic planning" (e.g., BURIDAN)
  - It isn't really the probability of the plan achieving its goals that is important — it is the expected value. The expected value can be high even with a low probability if the goals are sufficiently valuable.

Computing Expected Values

- Once a plan is constructed, an expected value can be computed. This computation is defeasible.
  - An initial computation can be made using just the probabilities of plan-steps having their desired outcomes in isolation.
  - Then a search can be undertaken for conditions established by other steps of the plan that alter the probabilities. This is analogous to the search for threats in deterministic causal-link planning.
  - Similarly, the initial computation uses default values for the execution costs of plan steps and the values of goals.
  - These values can be different in the context of conditions established by other steps of the plan, so a search can be undertaken for such conditions. This is also analogous to the search for threats.
- Expected values can be increased by adding conditional steps.
- Expected values can be increased by planning hierarchically.

Computing Expected Values

- The preceding computation computes the expected value of the plan *in isolation*.
- But that is not the relevant expected value. The agent may have adopted other plans whose execution will change the context and hence change both the probabilities and values used in computing expected values.
- Let the agent's *master plan* be the result of merging all of the agent's local plans into a single plan.
- In deciding whether to adopt a new plan, what is really at issue is the effect that will have on the expected value of the master plan.

Computing Expected Values

- Changes to the master plan may consist of simultaneously adopting and withdrawing several plans. It is *changes* that must be evaluated decision-theoretically.
- The value of a change is the difference between the expected value of the master plan after the change and its expected value before the change. This is the *differential expected value of the change*.

Computing Expected Values

- In a realistic agent in a complex environment, the master plan may grow very large.
- It is important to be able to employ simple computations of expected value defeasibly.
  - It can be assumed defeasibly that different local plans are *evaluatively independent*, in the sense that the expected value of the combined plan is the sum of the expected values of the individual plans.
  - This makes it easy to compute differential expected values defeasibly.
  - The search for considerations that would make this defeasible assumption incorrect is precisely the same as the search described above for considerations *within a plan* that would change the defeasible computation of its expected value. The only difference is that we look for considerations established by other constituents of the master plan.

Maximizing vs. Satisficing

- Conventional decision theory would tell us to choose a master plan having a maximal expected value.
- That is at least computationally infeasible in complex domains.
- There may not even be a maximally good plan.
- We should instead *satisfice* — seek plans with positive expected values, and always maintain an interest in finding better plans.
  - A plan is defeasibly adoptable if it has a positive expected value, or if its addition to the master plan increases the value of the latter.
  - The adoption is defeated by finding another plan that can be added to the master plan in its place and will increase the value of the master plan further.
- So we are always on the lookout for better plans, but we are not searching for a single “best” plan.

Plan Execution

- The master plan also provides a useful database for plan execution.
- Execution requires epistemic monitoring to
  1. ensure that things are going as planned
  2. determine when or whether the antecedents of conditional steps are satisfied.
- We must update the master plan as steps are performed, and update its evaluation.
- It is defeasibly reasonable to expect the master plan to remain adoptable as steps are performed.
  - However, new information may alter this defeasible expectation, forcing the abandonment or modification of constituent plans in the master plan.

Conclusions

- An architecture for "anthropomorphic agents" must mimic (but not necessarily duplicate) human rational cognition.
- Practical cognition makes choices based upon information supplied by epistemic cognition.
- Most of the work in rational cognition is carried out by epistemic cognition, *and must be done defeasibly*.
- OSCAR implements a sophisticated system of defeasible reasoning that enables it to deal defeasibly with:
  - perception
  - change and persistence
  - causation
  - probabilities

Conclusions

- **Sophisticated agents operating in complex environments cannot plan by using conventional planning algorithms that produce r.e. sets of solutions.**
- **However, the ideas underlying conventional planning algorithms can be resurrected as *defeasible* principles for reasoning about plans.**
- **Defeasible principles of deterministic planning can be generalized to produce defeasible principles of decision-theoretic planning.**
- **In decision-theoretic planning, decisions about whether to adopt new plans (and perhaps to reject previously adopted plans) must be made on the basis of the effect that has on the expected value of the master plan.**
- **An efficient computation of the expected value of the master plan can be done defeasibly.**

THE END

**(This talk, and related material,
can be downloaded from
http://www.u.arizona.edu/~pollock)**