Logics for Evidence-Based Belief Revision

Sonja Smets

ILLC, University of Amsterdam

Based on joint work with Alexandru Baltag and Virginie Fiutek (ILLC)

Financial Support Acknowledgement:

European Research Council erc

Netherlands Organisation for Scientific Research

Plan

• K. Lehrer's informal account of knowledge as "undefeated justified acceptance" is based on the "(ultra-) justification game".

• Goal: a "game semantics for defeasible knowledge", as a formalization of Keith Lehrer's conception of knowledge

• Present a formal qualitative representation of an agent's information and justification: *"Justification models"*.

• Models, more general than the "evidence models" proposed by van Benthem and Pacuit, and "plausibility models" for belief.

The Controversy in Mainstream Epistemology

What is the "true" concept of knowledge, and how can one reliably acquire such knowledge?

From the pre-60's interpretation till now

The most common interpretation is:

"Knowledge" = true belief + "justification"

= "justified true belief"

In this form, the 'Platonic' thesis has been shattered by **Edmund Gettier's famous counterexamples** (Gettier, 1963).

Now we know that not just any justification will do!

Lehrer: Knowledge = **undefeated** justified acceptance (true belief).

What is Justification?

When is a belief "justified"?

• An "explanation" (justification) for a belief can be:

- a mathematical **proof**
- a *necessary truth* obtained by **introspection** (e.g. the belief that I exist).
- the result of an *action*, *direct observation* from an **absolutely reliable** source.

- some ideal form of *perfectly reliable, truthful communication from an* ***infallible*** *source* ("divine revelation").

We call this **"hard" information**.

Types of Justification: "hard versus soft" information

But most of common-day "knowledge" (including scientific knowledge, outside Mathematics) cannot be justified in this tight sense.

A justification can also be based on some "less than ideal" type of informational action, e.g.

- *an imprecise observation*
- a communication from a *fallible source*,
- a testimony of an *unreliable* witness etc.

We call this **"soft" information**.

Formalization: Justification Models

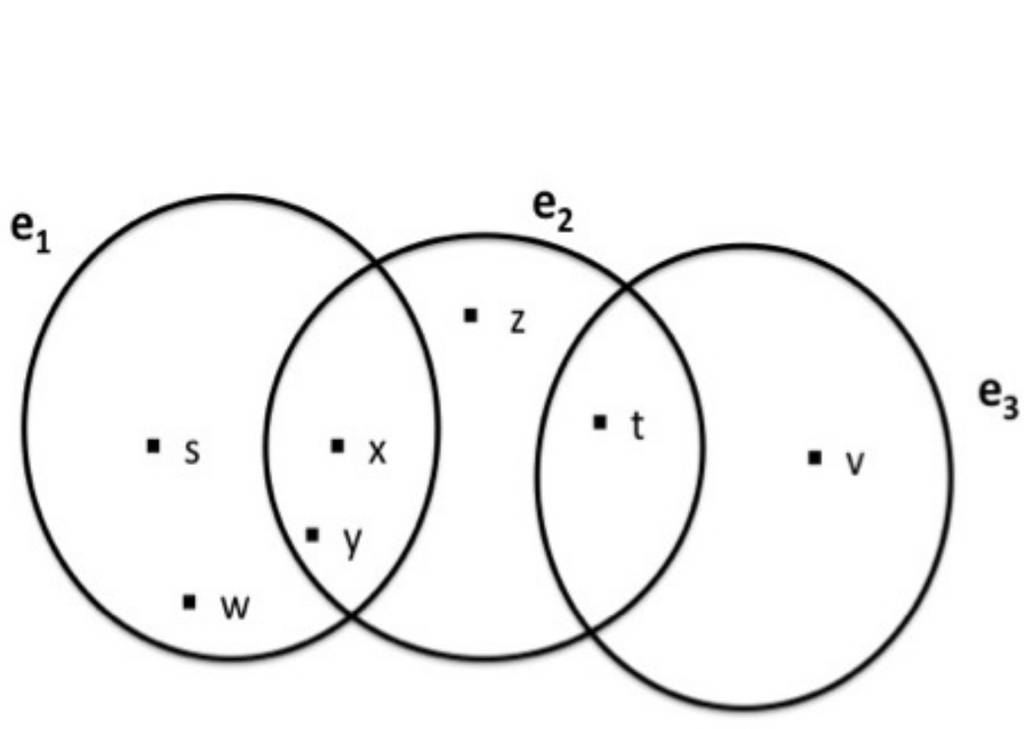
A **justification (or evidence)** model is a tuple $(S, E, \sim, \leq, \|\cdot\|)$, where

- S is a *finite* set of possible worlds, and $\|\cdot\|$ is a valuation map.
- $E \subseteq \mathcal{P}(S)$ is a family of subsets $e \subseteq S$, called **evidence (sets)**.

A **body of evidence** (or a "**justification**") is any consistent family of evidence sets, i.e. any $F \subseteq E$ such that $\bigcap F \neq \emptyset$.

We denote by $\mathcal{E} \subseteq \mathcal{P}(E)$ the family of all bodies of evidence.

Illustration



Evidence sets:

$e_1 = \{s,w,x,y\}$

$e_2 = \{x,y,z,t\}$

$e_3 = \{t,v\}$

Body of evidence:

$F = \{e_1\}$

$G = \{e_2\}$

$H = \{e_1 , e_2\}$

$I = \{e_2 , e_3\}$

BUT J is not a Body of evidence :

$J = \{e_1, e_3\}$

Justification Models Continued

A **justification** model is a tuple $(S, E, \sim, \leq, ||)$, where

- $\sim \subseteq S \times S$ is an *equivalence relation* ("**indistinguishability**").
- $\leq$ is a *partial preorder* on $\mathcal{E}$, satisfying the following constraints:

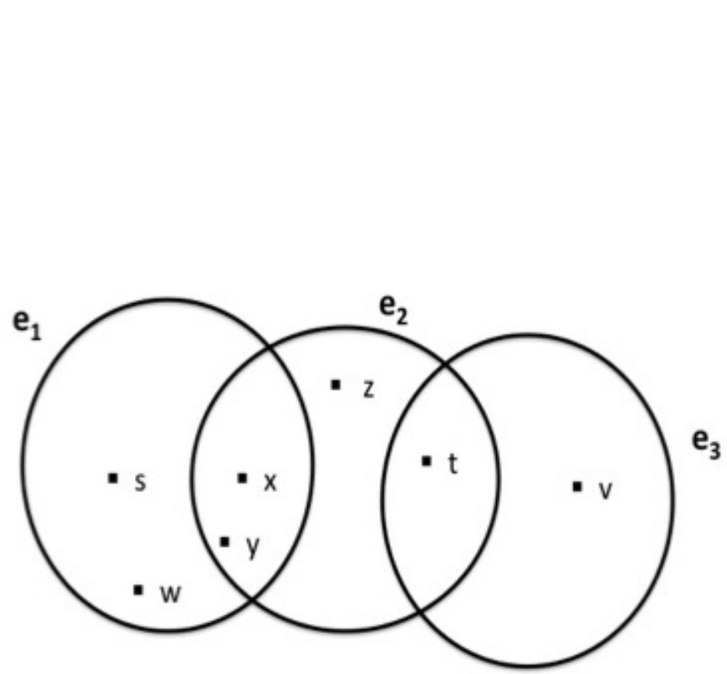
$$F \subseteq F' \Rightarrow F \leq F'$$

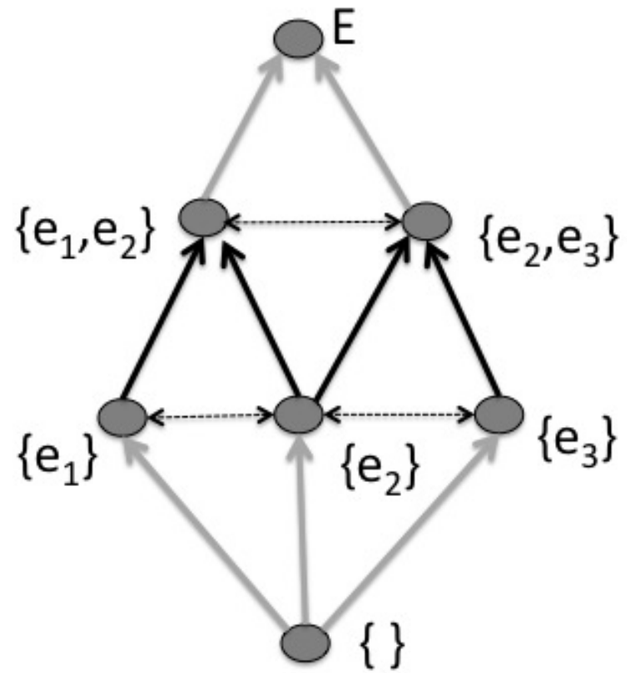
$$F \leq F', G \leq G' \text{ and } F' \cap G' = \emptyset \Rightarrow F \cup G \leq F' \cup G'$$

$$F < F', G \leq G' \text{ and } F' \cap G' = \emptyset \Rightarrow F \cup G < F' \cup G'$$

Illustration: Evidence sets and Hasse Diagram

Partial preorder compatible with the inclusion order on consistent evidence sets, + dashed arrows.





$\{e_1\} \subseteq \{e_1, e_2\}$

$\Downarrow$

$F \leq H$

Plausibility (Pre)Order

We read $F < G$ as "*the body of evidence G is (considered as) more plausible, easier to accept (by some implicit agent) than the body of evidence F "*.

For a state $s \in S$, put

$$E_s := \{e \in E | s \in E\}$$

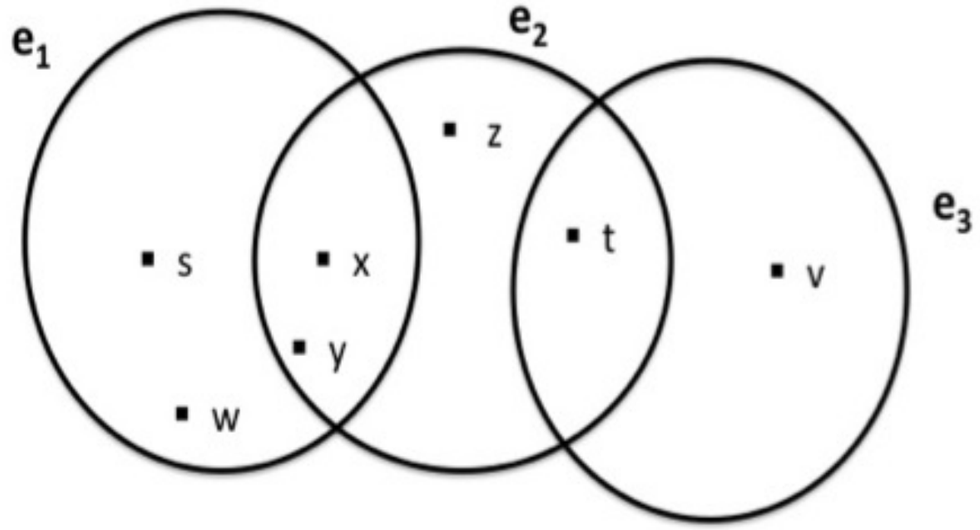
This is a body of evidence, namely the largest body of evidence consistent with the hypothesis s.

For two states $s, t \in S$, put

$$s \leq_E t \text{ iff } s \sim t \text{ and } E_s \leq E_t.$$

This makes any justification model $(S, E, \sim, \leq)$ into a plausibility model $(S, \sim, \leq)$.

Illustration



We write:
$E_x = \{e_1, e_2\}$
$E_s = \{e_1\}$

E_s is the largest body of evidence consistent with hypothesis s.

$$E_s \leq E_x$$

$$s \leq_E x$$

From Plausibility models back to justification models?

Conversely, every plausibility model $(S, \sim, \leq, \|\cdot\|)$ can be extended to a justification model $(S, E, \sim, \leq, \|\cdot\|)$, in such a way that

$$\leq = \leq_E$$

In general, this can be done in many different ways.

However, there is a **canonical** way: take E to be set of all **upwards-closed sets**

$$E = \{e \subseteq S | \forall s, t \in S\,(s \in e \wedge s \leq t \Rightarrow t \in e\}$$

Upwards-closed sets are also called **Grove spheres**, or "**strong beliefs**".

Other Examples

By taking $\leq$ to be inclusion

$$F \leq G \text{ iff } F \subseteq G,$$

we obtain the notion of belief given by van Benthem's and Pacuit's "evidence models".

Another natural choice is

$$F \leq G \text{ iff } |F| \leq |G|$$

(where $|F|$ is the cardinality of F).

Irrevocable Knowledge

This is just **Aumann (partitional) knowledge**, defined Kripke-style:

$$s \models_{\mathbf{S}} K\varphi \text{ iff } [s]_\sim \subseteq \|\varphi\|_{\mathbf{S}},$$

where

$$[s]_\sim = \{t \in S | s \sim t\}$$

is the **("information cell", i.e.) equivalence class** of s in the (partition determined by the) equivalence relation $\sim$.

K is truthful and both positively and negatively introspective.

Conditional Belief

We say that at a world $s \in S$, ψ **is believed given** φ (or "conditional on φ"), and write $B(\psi|\varphi)$, if ψ **is true in the most plausible φ-worlds**:

$$s \models_{\mathbf{S}} B(\psi|\varphi) \text{ iff } Max_{\leq}\|\varphi\|_{\mathbf{S}} \subseteq \|\psi\|_{\mathbf{S}},$$

where

$$Max_{\leq}P = \{s \in P | t \leq s \text{ for all } t \in S)\}$$

Local Connectedness

The plausibility relation $\leq_E$ is **locally connected** if it satisfies

$$s \sim t \Rightarrow s \leq_E t \vee t \leq_E s$$

Local connectedness is important: it validates the AGM principles of belief revision, stated in terms of conditional beliefs.

In particular, it validates "**Rational Monotonicity**" (which is the conditional-belief version of the AGM postulates of "Expansion").

An *example* of locally connected plausibility is the one given by the cardinality order.

A *counterexample* is the van Benthem-Pacuit order (inclusion): it is not locally connected, hence Rational Monotonicity fails.

Since we want to keep AGM, from now on we assume that $\leq_E$ is locally connected.

Updates on justification Models

New **hard** evidence φ is received.

This induces an update $!\varphi$, which changes the model to:

$$S' = \|\varphi\|\mathbf{s}$$

$$\sim'=\sim \cap (S' \times S')$$

$$E' = \{e \cap S' |\ e \in E,\ e \cap S' \neq \emptyset\}$$

i.e. the new evidence = old evidence consistent with the new states.

$$F' \leq' G' \text{ iff } \{e \in E | e \cap S' \in F'\} \leq \{e \in E | e \cap S' \in G'\}$$

i.e. iff the new evidence within G' is at least as strong as the new evidence in F'

There are other ways to change the model (by just *adding* the extra evidence set and giving it some degree of plausibility), which correspond to obtaining **soft** evidence.

For now, we restricted to the epistemic changes in models in which no new evidence is added.

So what's the missing ingredient?

"Knowledge" = "justified true belief" + what?

Plato: *"permanence" of belief.*

Hintikka: *"robustness" of belief.*

"... by saying "I know that p ", one makes a commitment stronger than one made by making a simple assertion; one proposes (it is part of one's proposition) to stick to this statement no matter what further information one expects to receive." (Hintikka, *Knowledge and Belief*, 1962)

An "absolute" interpretation

If by "further information" we mean *any further evidence* extracted from *any source, however unreliable or deceiving*, then this may include **misinformation**.

"Real knowledge", in this absolute sense, should be robust *even in the face of false evidence*. This gives us:

"Knowledge" = "absolutely unrevisable" belief

We will call this **irrevocable knowledge** and denote it by K. It is the standard notion used in Mathematics, Economics and Computer Science, and captures the *"hard information" possessed by an agent*.

"Stability" of belief

The "stability theory" of knowledge (Klein, Stalnaker, Rott, anticipated by Lehrer) takes a more "relative" interpretation: "information" means "**true** information".

"An agent knows that φ if and only if φ is true, she believes that φ, and she continues to believe φ if any *true* information is received" (Stalnaker 2006).

"A belief α is a piece of knowledge of the subject S iff α is not given up by S on the basis of any *true* information that S might receive" (Rott 2004).

Safe Belief

Let us denote the Stalnaker-Rott concept by $\Box$ and call it *"safe belief"*:
it is belief that is "safe", i.e. it is not defeasible by true evidence (though it can be defeated by false evidence).

Stalnaker had a first formalization of this concept in the setting of plausibility models.

He also noticed that belief is definable in terms of this notion:

$$B\varphi = \Diamond\Box\varphi$$

In my work with Alexandru, we rediscovered this notion (called "safe belief"), and considered it together with the usual (Aumann) notion of knowledge K, completely axiomatizing the resulting logic.

Definition of safe belief

$$s \models_{\mathbf{S}} \Box\varphi \text{ iff } \{t \in S | s \leq_E t\} \subseteq \|\varphi\|_{\mathbf{S}}$$

At least at a first approximation, safe belief is a formalization of the stability theory of knowledge:

$$s \models \Box P \text{ iff } s \models B(P|F) \text{ for all } F \in \mathcal{E} \text{ such that } s \models \bigwedge F.$$

Where F is a consistent body of true evidence

"Indefeasibility" of belief

In fact, the stability theory is a simplified version of Lehrer's *"defeasibility" theory of knowledge.*

Lehrer's Justification Game: the believing subject (Meno) in engaged in a dialogue with a *truthful and omniscient critic* (Socrates), who criticizes his justification for believing **P**, either be analyzing its consistency or by offering new true evidence. The subject *knows* **P** if he can always win the game, i.e. he *does not lose his justification for believing* **P** *when new evidence comes in.*

(In)defeasible Knowledge

We will call the Lehrer concept "(in)defeasible knowledge", formalize it in terms of a game semantics and prove it to be equivalent to the concept of safe belief.

So, from this perspective, the defeasibility theory is equivalent to the stability theory.

Moves of the Claimant

Agent (Believer, Proponent) :

Move	**Information Conveyed**
Claims P	BP
$F : P$	$F \in \mathcal{E} \wedge B(P\|F)$
Claims that F' is more plausible than F	$F < F'$

Moves of the Critic

Critic (Opponent) :

Move	Information Conveyed
Directly attacks P, or F	$!(\neg P)$ or $!(\neg e)$, for some $e \in F$
Indirect attack	$!Q$, s.t. $\neg B(P\|Q)$, or $\neg B(e\|Q)$, $e \in F$
Why? (More) Justification!?	$\neg K \neg (F \wedge \neg P)$
Attack against order $F < G$	$!(\neg G \wedge F)$

Rules of the Game

• Claimant makes the first move, claiming a proposition P to be true, based on his belief $B(P)$.

• Next the critic can pose an objection, attacking either the claim or its (lack of?) justification.

• The Claimant has to first update her model with the information conveyed by the Critic, then he has the choice whether to retract his claim, or else to stick with it.

• She can stick with a claim either by sticking with its previous justification and answering the objection of the Critic (thus offering further justification), or by retracting her previous justification and offering instead another one.

Rules of the Game continued

• **Every move for the believer is bound by the precondition that the information he conveys** to his opponent about herself holding certain beliefs, conditional beliefs or about the strength of her beliefs, **has to be truthful**, (i.e. the believer cannot make claims that go against the information she accepts in her own evaluation system, even if what he believes might actually be false in reality).

• The pre-condition for any move of the critic is that all the information he conveys has to be **true in the actual world, i.e. the critic cannot lie** (this is why we use public announcement operators ! in the information that the critic can convey).

Winning Conditions and Def. of (In)Defeasible Knowledge

At the end of the game Claimant wins iff her original belief $B(P)$ was undefeated and still justified; i.e. she offered at least one justification for P that was left undefeated.

Else, Critic wins.

Claimant has **(in)defeasible knowledge** of P iff she has a winning strategy in this game.

Proposition. There exists a winning strategy for the believer in the ultra justification game on the basis of the believer's claim that P iff $\Box P$.

Example of Justification Game

Claimant: There is a zebra here.

$$B(zebra)$$

Critic: Why do you think so?
(Justify!)

Claimant: I believe there is a zebra because I see a zebra.

$$see : zebra, \text{ i.e.: } see \in E \text{ and } B(see) \text{ and } B(zebra|see)$$

Critic: Maybe you are sleeping and dreaming that you see a zebra.
(Your evidence is consistent with no zebra! Need further justification!)

$$\neg K \neg (see \wedge dream \wedge \neg zebra)$$

Claimant: It is more reasonable for me to accept that I see the zebra because there is a zebra than to accept that I am dreaming a zebra!

$$(see \wedge dream) < (see \wedge zebra)$$

Critic: You are dreaming! You are asleep! You are only seeing the zebra in your dreams!

$$dream$$

Claimant: I still believe there is a zebra here, coincidental with my dreaming of it. I distinctly remember coming to the Zoo. Maybe I just fell asleep at the Zoo, right after seeing a real zebra? This would also explain why I am dreaming a zebra.

$$B(Zoo) \wedge B(zebra|Zoo \wedge dream \wedge see)$$

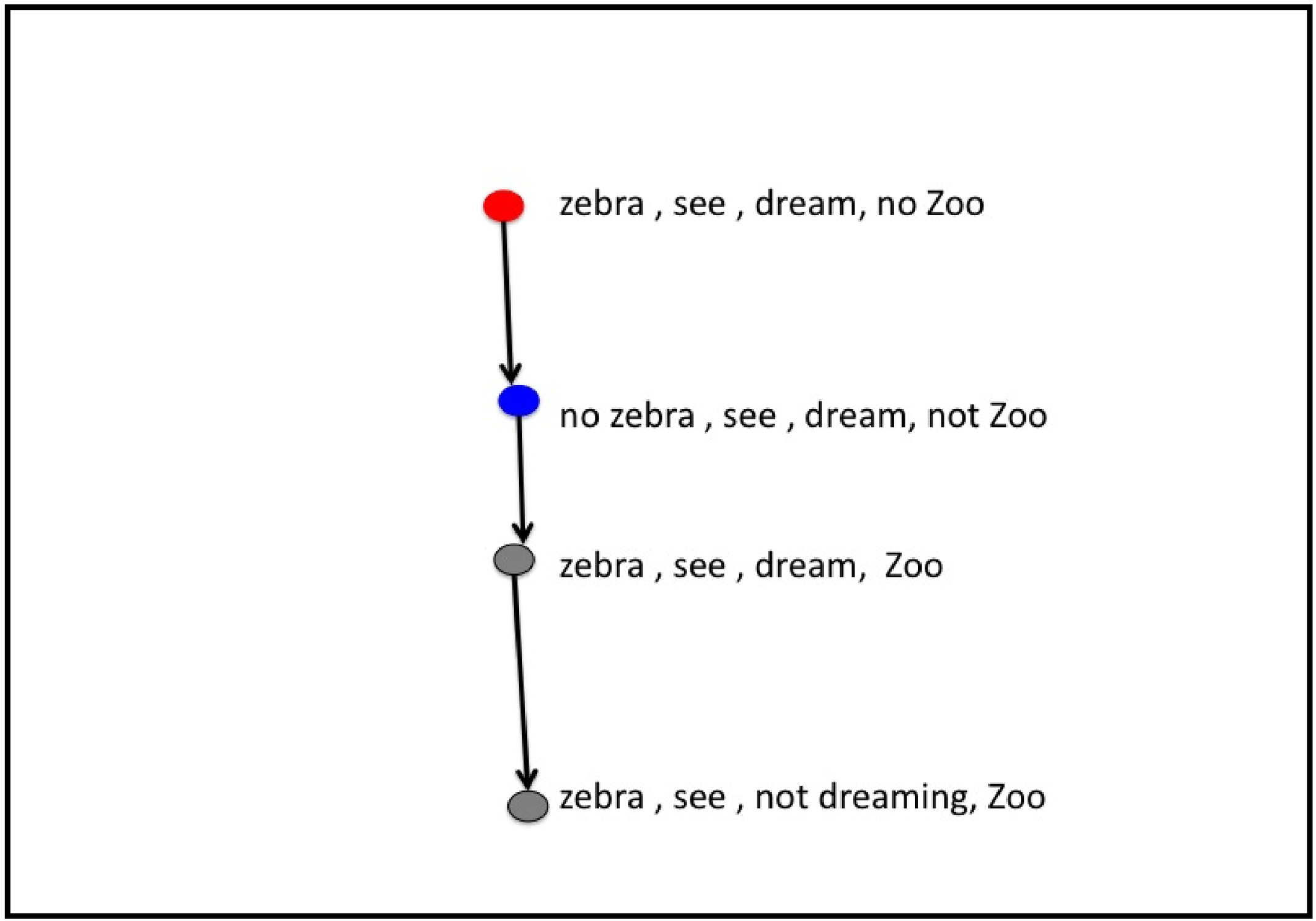
Critic: You are not at the Zoo. You are asleep in your bed, dreaming of zebras.

$$\neg Zoo$$

Claimant: Given that I am asleep in my bed, I have no justification left to believe there is a zebra here in the bedroom. So I give up: I no longer believe it!

The Claimant loses the game: he didn't really "know" that there was a zebra.

Sadly enough, his initial justified belief was in fact true: however implausible this might seem to him, there is a zebra in his bedroom!


zebra , see , dream, no Zoo
no zebra , see , dream, not Zoo
zebra , see , dream, Zoo
zebra , see , not dreaming, Zoo

Concluding Remarks

In order to formalize the personal justification game of Lehrer we need to adjust our logical setting. In our view, there exists a winning strategy for our agent in the personal justification game if and only if all his beliefs and conditional beliefs are internally consistent.

Because our plausibility models do not allow agents to have inconsistent beliefs, these games can be formalized only if we make the difference between an agents **belief system** and his **acceptance and reasoning system** explicit. What is needed is a model of an agents doxastic state in which inconsistent acceptances are allowed.

Formally, one can use Hintikka's models that allow impossible and possible worlds.

As a final remark, we are aware of only one other formal analysis of Lehrers work that relates close to our setting, and that is **Spohns analysis in terms of Ranking Theory**.

Even though Spohn does not offer a game semantics, an analysis of the differences between our qualitative approach (via Plausibility Models) and Spohns more quatitative setting in terms of grades of belief and disbelief would be interesting.