Puzzles

Eric Pacuit

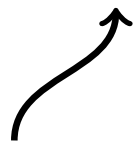
Department of Philosophy
University of Maryland

pacuit.org/esslli2019/puzzles

August 11, 2019

$$EU(A) = \sum_{o \in O} P_A(o) \times U(o)$$

$$EU(A) = \sum_{o \in O} P_A(o) \times U(o)$$



Expected utility of action A

$$EU(A) = \sum_{o \in O} P_A(o) \times U(o)$$

Expected utility of action A

Utility of outcome o

$$EU(A) = \sum_{o \in O} P_A(o) \times U(o)$$

$P_A(o)$: probability of o conditional on A — how likely it is that outcome o will occur, on the supposition that the agent chooses act A.

$P_A(o)$: probability of o conditional on A — how likely it is that outcome o will occur, on the supposition that the agent chooses act A.

Evidential: $\quad P_A(o) = P(o \mid A) = \frac{P(o \ \& \ A)}{P(A)}$

$P_A(o)$: probability of o conditional on A — how likely it is that outcome o will occur, on the supposition that the agent chooses act A.

Evidential: $P_A(o) = P(o \mid A) = \frac{P(o \ \& \ A)}{P(A)}$

Classical: $P_A(o) = \sum_{s \in S} P(s) f_{A,s}(o)$, where

$$f_{A,s}(o) = \begin{cases} 1 & A(s) = o \\ 0 & A(s) \neq o \end{cases}$$

$P_A(o)$: probability of o conditional on A — how likely it is that outcome o will occur, on the supposition that the agent chooses act A.

Evidential: $P_A(o) = P(o \mid A) = \frac{P(o \ \& \ A)}{P(A)}$

Classical: $P_A(o) = \sum_{s \in S} P(s) f_{A,s}(o)$, where

$$f_{A,s}(o) = \begin{cases} 1 & A(s) = o \\ 0 & A(s) \neq o \end{cases}$$

Causal: $P_A(o) = P(A \ \square\!\!\rightarrow o)$

P("if A were performed, outcome o would ensue")

(Lewis, 1981)

Dominance Reasoning and Act-State Dependence

	w_1	w_2
A	1	3
B	2	4

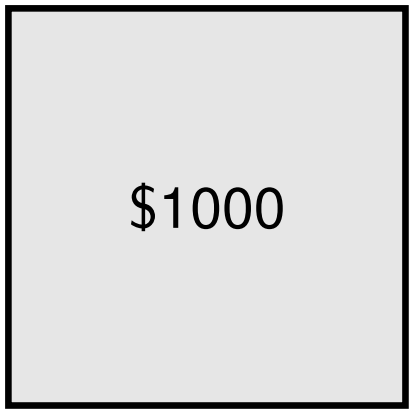
Dominance Reasoning and Act-State Dependence

	w_1	w_2
A	1	3
B	2	4

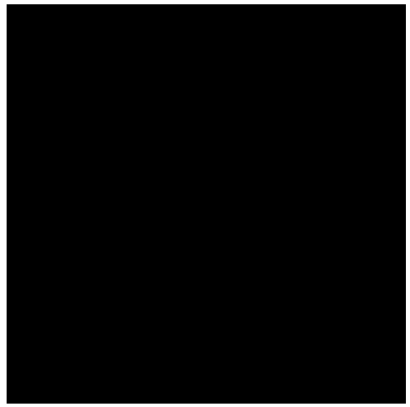
Dominance Reasoning and Act-State Dependence

Dominance reasoning is appropriate only when probability of outcome is *independent of choice*.

(A nasty nephew wants inheritance from his rich Aunt. The nephew wants the inheritance, but other things being equal, does not want to apologize. Does dominance give the nephew a reason to not apologize? *Whether or not the nephew is cut from the will may depend on whether or not he apologizes*.)



A



B

Choice:

one-box: choose box *B*

two-box: choose box *A* and *B*

Newcomb's Paradox

A very powerful being, who has been invariably accurate in his predictions about your behavior in the past, has already acted in the following way:

1. If he has predicted that you will open just box *B*, he has in addition put $1,000,000 in box *B*
2. If he has predicted you will open both boxes, he has put nothing in box *B*.

What should you do?

R. Nozick. *Newcomb's Problem and Two Principles of Choice*. 1969.

	$1 million in closed box	$0 in closed box
one-box	$1,000,000	$0
two-box	$1,001,000	$1,000

	$1 million in closed box	$0 in closed box
one-box	$1,000,000	$0
two-box	$1,001,000	$1,000

act-state dependence: $P(s) \neq P(s \mid A)$

Newcomb's Paradox

	B = 1M	B = 0
1 Box	1M	0
2 Boxes	1M + 1000	1000

Newcomb's Paradox

	B = 1M	B = 0
1 Box	1M	0
2 Boxes	1M + 1000	1000

	B = 1M	B = 0
1 Box	h	$1 - h$
2 Boxes	$1 - h$	h

Newcomb's Paradox

J. Collins. *Newcomb's Problem*. International Encyclopedia of Social and Behavorial Sciences, 1999.

Newcomb's Paradox

There is a conflict between maximizing your expected value (1-box choice) and dominance reasoning (2-box choice).

Newcomb's Paradox

There is a conflict between maximizing your expected value (1-box choice) and dominance reasoning (2-box choice).

What the Predictor did yesterday is *probabilistically dependent* on the choice today, but *causally independent* of today's choice.

$V(A) = \sum_w V(w) \cdot P_A(w)$

(the expected value of act A is a probability weighted average of the values of the ways w in which A might turn out to be true)

$V(A) = \sum_w V(w) \cdot P_A(w)$
(the expected value of act A is a probability weighted average of the values of the ways w in which A might turn out to be true)

EDT: $P_A(w) := P(w \mid A)$ (Probability of w given A is chosen)

CDT: $P_A(w) = P(A \mathrel{\square\!\!\rightarrow} w)$ (Probability of *if A were chosen then w would be true*)

Suppose 99% confidence in predictors reliability.

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

Suppose 99% confidence in predictors reliability.

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(M \mid B_1) + V(N)P(N \mid B_1)$$

Suppose 99% confidence in predictors reliability.

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(M \mid B_1) + V(N)P(N \mid B_1) =$$
$$1000000 \cdot 0.99 + 0 \cdot 0.01$$

Suppose 99% confidence in predictors reliability.

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(M \mid B_1) + V(N)P(N \mid B_1) =$$
$$1000000 \cdot 0.99 + 0 \cdot 0.01 = 990,000$$

Suppose 99% confidence in predictors reliability.

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(M \mid B_1) + V(N)P(N \mid B_1) =$$
$$1000000 \cdot 0.99 + 0 \cdot 0.01 = 990,000$$

$$V(B_2) = V(L)P(L \mid B_2) + V(K)P(K \mid B_2)$$

Suppose 99% confidence in predictors reliability.

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(M \mid B_1) + V(N)P(N \mid B_1) =$$
$$1000000 \cdot 0.99 + 0 \cdot 0.01 = 990,000$$

$$V(B_2) = V(L)P(L \mid B_2) + V(K)P(K \mid B_2) =$$
$$1001000 \cdot 0.01 + 1000 \cdot 0.99$$

Suppose 99% confidence in predictors reliability.

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(M \mid B_1) + V(N)P(N \mid B_1) =$$
$$1000000 \cdot 0.99 + 0 \cdot 0.01 = 990,000$$

$$V(B_2) = V(L)P(L \mid B_2) + V(K)P(K \mid B_2) =$$
$$1001000 \cdot 0.01 + 1000 \cdot 0.99 = 11,000$$

Let μ be the assigned to the conditional $B_1 \boxdot\!\!\rightarrow M$ (and $B_2 \boxdot\!\!\rightarrow L$) (both conditionals are true iff the Predictor put \$1,000,000 in box B yesterday).

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

Let μ be the assigned to the conditional $B_1 \mathrel{\Box\!\!\rightarrow} M$ (and $B_2 \mathrel{\Box\!\!\rightarrow} L$) (both conditionals are true iff the Predictor put \$1,000,000 in box B yesterday).

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(B_1 \mathrel{\Box\!\!\rightarrow} M) + V(N)P(B_1 \mathrel{\Box\!\!\rightarrow} N)$$

Let μ be the assigned to the conditional $B_1 \boxdot\!\!\rightarrow M$ (and $B_2 \boxdot\!\!\rightarrow L$) (both conditionals are true iff the Predictor put \$1,000,000 in box B yesterday).

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$V(B_1) = V(M)P(B_1 \boxdot\!\!\rightarrow M) + V(N)P(B_1 \boxdot\!\!\rightarrow N) =$
$1000000 \cdot \mu + 0 \cdot (1 - \mu)$

Let μ be the assigned to the conditional $B_1 \boxdot\!\!\rightarrow M$ (and $B_2 \boxdot\!\!\rightarrow L$) (both conditionals are true iff the Predictor put \$1,000,000 in box B yesterday).

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(B_1 \boxdot\!\!\rightarrow M) + V(N)P(B_1 \boxdot\!\!\rightarrow N) =$$
$$1000000 \cdot \mu + 0 \cdot (1 - \mu) = 1000000\mu$$

Let μ be the assigned to the conditional $B_1 \boxdot\!\!\rightarrow M$ (and $B_2 \boxdot\!\!\rightarrow L$) (both conditionals are true iff the Predictor put \$1,000,000 in box B yesterday).

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(B_1 \boxdot\!\!\rightarrow M) + V(N)P(B_1 \boxdot\!\!\rightarrow N) =$$
$$1000000 \cdot \mu + 0 \cdot (1 - \mu) = 1000000\mu$$

$$V(B_2) = V(L)P(B_2 \boxdot\!\!\rightarrow L) + V(K)P(B_2 \boxdot\!\!\rightarrow K)$$

Let μ be the assigned to the conditional $B_1 \boxdot\!\!\rightarrow M$ (and $B_2 \boxdot\!\!\rightarrow L$) (both conditionals are true iff the Predictor put \$1,000,000 in box B yesterday).

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(B_1 \boxdot\!\!\rightarrow M) + V(N)P(B_1 \boxdot\!\!\rightarrow N) =$$
$$1000000 \cdot \mu + 0 \cdot (1 - \mu) = 1000000\mu$$

$$V(B_2) = V(L)P(B_2 \boxdot\!\!\rightarrow L) + V(K)P(B_2 \boxdot\!\!\rightarrow K) =$$
$$1001000 \cdot \mu + 1000 \cdot (1 - \mu)$$

Let μ be the assigned to the conditional $B_1 \boxdot\!\!\rightarrow M$ (and $B_2 \boxdot\!\!\rightarrow L$) (both conditionals are true iff the Predictor put \$1,000,000 in box B yesterday).

B_1: one-box (open box B)
B_2: two-box choice (open both A and B)
N: receive nothing
K: receive \$1,000
M: receive \$1,000,000
L: receive \$1,001,000

$$V(B_1) = V(M)P(B_1 \boxdot\!\!\rightarrow M) + V(N)P(B_1 \boxdot\!\!\rightarrow N) =$$
$$1000000 \cdot \mu + 0 \cdot (1 - \mu) = 1000000\mu$$

$$V(B_2) = V(L)P(B_2 \boxdot\!\!\rightarrow L) + V(K)P(B_2 \boxdot\!\!\rightarrow K) =$$
$$1001000 \cdot \mu + 1000 \cdot (1 - \mu) = 1000000\mu + 1000$$

Causal Decision Theory

A. Egan. *Some Counterexamples to Causal Decision Theory*. Philosophical Review, 116(1), pgs. 93 - 114, 2007.

The Psychopath Button: Paul is debating whether to press the 'kill all psychopaths' button. It would, he thinks, be much better to live in a world with no psychopaths.

The Psychopath Button: Paul is debating whether to press the 'kill all psychopaths' button. It would, he thinks, be much better to live in a world with no psychopaths. Unfortunately, Paul is quite confident that only a psychopath would press such a button.

The Psychopath Button: Paul is debating whether to press the 'kill all psychopaths' button. It would, he thinks, be much better to live in a world with no psychopaths. Unfortunately, Paul is quite confident that only a psychopath would press such a button. Paul very strongly prefers living in a world with psychopaths to dying. Should Paul press the button?

(Set aside your theoretical commitments and put yourself in Paul's situation. Would you press the button? Would you take yourself to be irrational for not doing so?)

Death in Damascus

A man in Damascus knows that he has an appointment with Death at midnight. He will escape Death if he manages at midnight not to be at the place of his appointment. He can be in either Damascus or Aleppo at midnight.

Death in Damascus

A man in Damascus knows that he has an appointment with Death at midnight. He will escape Death if he manages at midnight not to be at the place of his appointment. He can be in either Damascus or Aleppo at midnight. As the man knows, Death is a good predictor of his whereabouts. If he stays in Damascus, he thereby has evidence that Death will look for him in Damascus. However, if he goes to Aleppo he thereby has evidence that Death will look for him in Aleppo.

Death in Damascus

A man in Damascus knows that he has an appointment with Death at midnight. He will escape Death if he manages at midnight not to be at the place of his appointment. He can be in either Damascus or Aleppo at midnight. As the man knows, Death is a good predictor of his whereabouts. If he stays in Damascus, he thereby has evidence that Death will look for him in Damascus. However, if he goes to Aleppo he thereby has evidence that Death will look for him in Aleppo. Wherever he decides to be at midnight, he has evidence that he would be better off at the other place. No decision is stable.

A. Gibbard and W. Harper. *Counterfactuals and Two Kinds of Expected Utility*. In Ifs: Conditionals, Belief, Decision, Chance, and Time, pp. 153 - 190, 1978.

- The crucial distinction is between an act and a decision to perform the act.

- Before performing an act, an agent may assess the act in light of a decision to perform it. Information the decision carries may affect the act's expected utility and its ranking with respect to other acts.

- Decision makers should make *self-ratifying*, or *ratifiable*, decisions.

H. Gaifman. *Self-reference and the acyclicity of rational choice*. Annals of Pure and Applied Logic, 96, pgs. 117 - 140, 1999.

The Irrational Choice

Mr. Z offers Adam two boxes, each containing \$10. Adam can choose either $S1$: to take the leftmost box and get \$10, or $S2$: to take the two boxes and get \$20.

The Irrational Choice

Mr. Z offers Adam two boxes, each containing \$10. Adam can choose either $S1$: to take the leftmost box and get \$10, or $S2$: to take the two boxes and get \$20. Before making his decision, Adam is informed by Mr. Z that if he acts irrationally, Mr. Z will give him a bonus of \$100.

(...to eliminate noise factors, assume that Adam believes that Mr. Z is serious, has the relevant knowledge, is a perfect reasoner and is completely trustworthy.)

"...the bonus condition in Z's statement has truth-conditions, and once Adam has chosen it can be evaluated...It is only from the perspective of *Adam* qua *deliberating rational agent* that the bonus condition must be excluded as meaningless."

"He could have chosen by whim, because of a feeling, a mood, or for no reason. The question how irrational choice is possible, what constitutes such a whim, impulse, temporary incoherence, weakness of will, or what have you, does not concern me here. I take it for granted that there will be cases which we shall characterize in this way (else 'rational' becomes a vacuous constraint).

"He could have chosen by whim, because of a feeling, a mood, or for no reason. The question how irrational choice is possible, what constitutes such a whim, impulse, temporary incoherence, weakness of will, or what have you, does not concern me here. I take it for granted that there will be cases which we shall characterize in this way (else 'rational' becomes a vacuous constraint). And if Adam chooses in this way he qualifies for the bonus, and will probably be surprised when he gets it. It is only from the perspective of *Adam* qua *deliberating rational agent* that the bonus condition must be excluded as meaningless." (Gaifman, pg. 123)

The Rational Choice

Mr. Z offers Adam two boxes, each containing \$10. Adam can choose either $S1$: to take the leftmost box and get \$10, or $S2$: to take the two boxes and get \$20. Before making his decision, Adam is informed by Mr. Z that if he acts **rationally**, Mr. Z will give him a bonus of \$100.

(...to eliminate noise factors, assume that Adam believes that Z. is serious, has the relevant knowledge, is a perfect reasoner and is completely trustworthy.)

(AC) The reason for choosing A can refer to each of the available options, but they cannot refer in an essential way to the *choosing* from these options (except through considerations of signaling).

Irrational Man

(straightforward reason) $20 is better than $10

(c) If Adam chooses $S2$ for the straightforward reason, then his choice is rational. Hence, he forfeits the bonus, which he could have received by choosing $S1$.

Irrational Man

(straightforward reason) $20 is better than $10

(c) If Adam chooses $S2$ for the straightforward reason, then his choice is rational. Hence, he forfeits the bonus, which he could have received by choosing $S1$.

(c) is ruled out by (AC)

Irrational Man

(straightforward reason) $20 is better than $10

(c) If Adam chooses $S2$ for the straightforward reason, then his choice is rational. Hence, he forfeits the bonus, which he could have received by choosing $S1$.

(c) is ruled out by (AC)

If Mr. Z is not assumed to be a perfect reasoner, Adam may rationally try to outsmart Z. (c) can be rephrased as a legitimate case of signaling: Adam signals (deceptively) to Mr. Z that choosing $S1$ he is behaving irrationally. Deceptive signaling is, of course, useless if you deal with a omniscient reasoner.

Newcomb's Paradox

(N1) Take one box for the reason: Given the evidence, if I take one box (make $B1$ true), I am likely to find there a very large sum; but if I take two I am likely to find the first empty, and the payoff from the second is comparatively paltry. The reasoning can be case in terms of expected utilities, where $P(E \mid B1)$ and $P(not - E \mid B2)$ are sufficiently high.

(N2) Take two boxes for the reason: Given the evidence, my doing does not influence in any way what the box already contains. Whatever is there, I do better by choosing $B2$.

- Orthodox Bayesian: It is a problem of act-state dependence (1-box)

- Orthodox Bayesian: It is a problem of act-state dependence (1-box)
- Causal Decision Theory: expected utility involves probabilities of causal counterfactuals (2-box)

- Orthodox Bayesian: It is a problem of act-state dependence (1-box)
- Causal Decision Theory: expected utility involves probabilities of causal counterfactuals (2-box)
- No Acyclic Reasons: reasoning cannot refer to the act of choice in an essential way (2-box)...plus some “mental gymnastics” (1-box)

- Orthodox Bayesian: It is a problem of act-state dependence (1-box)
- Causal Decision Theory: expected utility involves probabilities of causal counterfactuals (2-box)
- No Acyclic Reasons: reasoning cannot refer to the act of choice in an essential way (2-box)...plus some “mental gymnastics” (1-box)
- “Tickle”-defense (2-box)

- Orthodox Bayesian: It is a problem of act-state dependence (1-box)
- Causal Decision Theory: expected utility involves probabilities of causal counterfactuals (2-box)
- No Acyclic Reasons: reasoning cannot refer to the act of choice in an essential way (2-box)...plus some “mental gymnastics” (1-box)
- “Tickle”-defense (2-box)
- Evidential Decision Theory: decisions to act provides evidence for the consequences (1-box)

- Orthodox Bayesian: It is a problem of act-state dependence (1-box)
- Causal Decision Theory: expected utility involves probabilities of causal counterfactuals (2-box)
- No Acyclic Reasons: reasoning cannot refer to the act of choice in an essential way (2-box)...plus some “mental gymnastics” (1-box)
- “Tickle”-defense (2-box)
- Evidential Decision Theory: decisions to act provides evidence for the consequences (1-box)
- Ratifiability: decision makers must assess the act in light of the decision to perform it and only choose acts that are self-ratifiable (1-box)

Conditioning

Conditional Probability

The probability of *E given F*, dented $p(E|F)$, is defined to be

$$p(E|F) = \frac{p(E \cap F)}{p(F)}.$$

provided $P(F) > 0$.

Setting $p_t(\cdot) = p_0(\cdot \mid E)$ is demonstrably the correct thing to do just in case, for all propositions $H \in \Sigma$, both:

1. Certainty: $p_t(E) = 1$
2. Rigidity: $p_t(H \mid E) = p_0(H \mid E)$

People are often not aware of all that they have learnt or they fail to adequately represent it, and it is only the failure of the Rigidity condition that alerts us to this.

Three Prisoner's Problem

Three prisoners A, B and C have been tried for murder and their verdicts will told to them tomorrow morning. They know only that one of them will be declared guilty and will be executed while the others will be set free. The identity of the condemned prisoner is revealed to the very reliable prison guard, but not to the prisoners themselves. Prisoner A asks the guard "Please give this letter to one of my friends — to the one who is to be released. We both know that at least one of them will be released".

Three Prisoner's Problem

An hour later, *A* asks the guard "Can you tell me which of my friends you gave the letter to? It should give me no clue regarding my own status because, regardless of my fate, each of my friends had an equal chance of receiving my letter." The guard told him that *B* received his letter.

Prisoner *A* then concluded that the probability that he will be released is 1/2 (since the only people without a verdict are *A* and *C*).

Three Prisoner's Problem

But, *A* thinks to himself:

Three Prisoner's Problem

But, *A* thinks to himself:

> *Before I talked to the guard my chance of being executed was 1 in 3. Now that he told me B has been released, only C and I remain, so my chances of being executed have gone from 33.33% to 50%. What happened? I made certain not to ask for any information relevant to my own fate...*

Three Prisoner's Problem

But, *A* thinks to himself:

> *Before I talked to the guard my chance of being executed was 1 in 3. Now that he told me B has been released, only C and I remain, so my chances of being executed have gone from 33.33% to 50%. What happened? I made certain not to ask for any information relevant to my own fate...*

Explain what is wrong with *A*'s reasoning.

A's reasoning

Consider the following events:

G_A: "Prisoner A will be declared guilty" (we have $p(G_A) = 1/3$)

I_B: "Prisoner B will be declared innocent" (we have $p(I_B) = 2/3$)

A's reasoning

Consider the following events:

G_A: "Prisoner A will be declared guilty" (we have $p(G_A) = 1/3$)

I_B: "Prisoner B will be declared innocent" (we have $p(I_B) = 2/3$)

We have $p(I_B \mid G_A) = 1$: "If A is declared guilty then B will be declared innocent."

A's reasoning

Consider the following events:

G_A: "Prisoner A will be declared guilty" (we have $p(G_A) = 1/3$)

I_B: "Prisoner B will be declared innocent" (we have $p(I_B) = 2/3$)

We have $p(I_B \mid G_A) = 1$: "If A is declared guilty then B will be declared innocent."

Bayes Theorem:

$$p(G_A \mid I_B) =$$

A's reasoning

Consider the following events:

G_A: "Prisoner A will be declared guilty" (we have $p(G_A) = 1/3$)

I_B: "Prisoner B will be declared innocent" (we have $p(I_B) = 2/3$)

We have $p(I_B \mid G_A) = 1$: "If A is declared guilty then B will be declared innocent."

Bayes Theorem:

$$p(G_A \mid I_B) = p(I_B \mid G_A)\frac{p(G_A)}{p(I_B)} =$$

A's reasoning

Consider the following events:

G_A: "Prisoner A will be declared guilty" (we have $p(G_A) = 1/3$)

I_B: "Prisoner B will be declared innocent" (we have $p(I_B) = 2/3$)

We have $p(I_B \mid G_A) = 1$: "If A is declared guilty then B will be declared innocent."

Bayes Theorem:

$$p(G_A \mid I_B) = p(I_B \mid G_A)\frac{p(G_A)}{p(I_B)} = 1 \cdot \frac{1/3}{2/3} = 1/2$$

A's reasoning, corrected

But, A did not receive the information that B will be declared innocent, but rather that "the guard said that B will be declared innocent." So, A should have conditioned on the event:

I'_B: "The guard said that B will be declared innocent"

A's reasoning, corrected

But, A did not receive the information that B will be declared innocent, but rather that "the guard said that B will be declared innocent." So, A should have conditioned on the event:

I'_B: "The guard said that B will be declared innocent"

Given that $p(I'_B \mid G_A)$ is 1/2 (given that A is guilty, there is a 50-50 chance that the guard could have given the letter to B or C).

A's reasoning, corrected

But, A did not receive the information that B will be declared innocent, but rather that "the guard said that B will be declared innocent." So, A should have conditioned on the event:

I'_B: "The guard said that B will be declared innocent"

Given that $p(I'_B \mid G_A)$ is 1/2 (given that A is guilty, there is a 50-50 chance that the guard could have given the letter to B or C). This gives us the following correct calculation:

$$p(G_A \mid I'_B) = p(I'_B \mid G_A)\frac{p(G_A)}{p(I'_B)} = 1/2 \cdot \frac{1/3}{1/2} = 1/3$$

Monty Hall Dilemma

Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say number 1, and the host, who knows what's behind the doors, opens another door, say number 3, which has a goat. He says to you, "Do you want to pick door number 2?" Is it to your advantage to switch your choice of doors?

Monty Hall (1)

H_1: The care is behind door 1

H_2: The care is behind door 2

H_3: The care is behind door 3

Monty Hall (1)

H_1: The care is behind door 1

H_2: The care is behind door 2

H_3: The care is behind door 3

Monty Hall (2)

Reasoning 1: E: The car is not behind door 3 ($\neg H_3 \leftrightarrow H_1 \vee H_2$)

Monty Hall (2)

Reasoning 1: E: The car is not behind door 3 ($\neg H_3 \leftrightarrow H_1 \vee H_2$)

$$p(H_1 \mid E) \quad = \quad p(E \mid H_1) \frac{p(H_1)}{p(E)}$$

Monty Hall (2)

Reasoning 1: E: The car is not behind door 3 ($\neg H_3 \leftrightarrow H_1 \vee H_2$)

$$
\begin{aligned}
p(H_1 \mid E) &= p(E \mid H_1)\frac{p(H_1)}{p(E)} \\
&= p(E \mid H_1)\frac{p(H_1)}{p(E \mid H_1)p(H_1)+p(E \mid H_2)p(H_2)+p(E \mid H_3)p(H_3)}
\end{aligned}
$$

Monty Hall (2)

Reasoning 1: E: The car is not behind door 3 ($\neg H_3 \leftrightarrow H_1 \vee H_2$)

$$
\begin{aligned}
p(H_1 \mid E) &= p(E \mid H_1)\frac{p(H_1)}{p(E)} \\
&= p(E \mid H_1)\frac{p(H_1)}{p(E \mid H_1)p(H_1)+p(E \mid H_2)p(H_2)+p(E \mid H_3)p(H_3)} \\
&= 1 \cdot \frac{\frac{1}{3}}{1\cdot\frac{1}{3}+1\cdot\frac{1}{3}+0\cdot\frac{1}{3}}
\end{aligned}
$$

Monty Hall (2)

Reasoning 1: E: The car is not behind door 3 ($\neg H_3 \leftrightarrow H_1 \vee H_2$)

$$
\begin{aligned}
p(H_1 \mid E) &= p(E \mid H_1)\frac{p(H_1)}{p(E)} \\
&= p(E \mid H_1)\frac{p(H_1)}{p(E \mid H_1)p(H_1)+p(E \mid H_2)p(H_2)+p(E \mid H_3)p(H_3)} \\
&= 1 \cdot \frac{\frac{1}{3}}{1\cdot\frac{1}{3}+1\cdot\frac{1}{3}+0\cdot\frac{1}{3}} \\
&= 1 \cdot \frac{\frac{1}{3}}{\frac{2}{3}}
\end{aligned}
$$

Monty Hall (2)

Reasoning 1: E: The car is not behind door 3 ($\neg H_3 \leftrightarrow H_1 \vee H_2$)

$$
\begin{aligned}
p(H_1 \mid E) &= p(E \mid H_1)\frac{p(H_1)}{p(E)} \\
&= p(E \mid H_1)\frac{p(H_1)}{p(E \mid H_1)p(H_1)+p(E \mid H_2)p(H_2)+p(E \mid H_3)p(H_3)} \\
&= 1 \cdot \frac{\frac{1}{3}}{1\cdot\frac{1}{3}+1\cdot\frac{1}{3}+0\cdot\frac{1}{3}} \\
&= 1 \cdot \frac{\frac{1}{3}}{\frac{2}{3}} \\
&= \frac{1}{2}
\end{aligned}
$$

Monty Hall (2)

Reasoning 1: E: The car is not behind door 3 ($\neg H_3 \leftrightarrow H_1 \vee H_2$)

$$
\begin{aligned}
p(H_1 \mid E) &= p(E \mid H_1)\frac{p(H_1)}{p(E)} \\
&= p(E \mid H_1)\frac{p(H_1)}{p(E \mid H_1)p(H_1)+p(E \mid H_2)p(H_2)+p(E \mid H_3)p(H_3)} \\
&= 1 \cdot \frac{\frac{1}{3}}{1\cdot\frac{1}{3}+1\cdot\frac{1}{3}+0\cdot\frac{1}{3}} \\
&= 1 \cdot \frac{\frac{1}{3}}{\frac{2}{3}} \\
&= \frac{1}{2}
\end{aligned}
$$

Similarly for $p(H_2 \mid E)$, so do not switch.

Monty Hall (3)

Reasoning 2: F: Monty opened door number 3

$$
\begin{aligned}
p(H_2 \mid F) &= p(F \mid H_2)\frac{p(H_2)}{p(F)} \\
&= p(F \mid H_2)\frac{p(H_2)}{p(F \mid H_1)p(H_1)+p(F \mid H_2)p(H_2)+p(F \mid H_3)p(H_3)} \\
&= 1 \cdot \frac{\frac{1}{3}}{\frac{1}{2}\cdot\frac{1}{3}+1\cdot\frac{1}{3}+0\cdot\frac{1}{3}} \\
&= 1 \cdot \frac{\frac{1}{3}}{\frac{1}{2}} \\
&= \frac{2}{3}
\end{aligned}
$$

So, $p(H_1 \mid F) = \frac{1}{3}$ and $p(H_2 \mid F) = \frac{2}{3}$

Monty Hall (3)

Reasoning 2: F: Monty opened door number 3

$$
\begin{aligned}
p(H_2 \mid F) &= p(F \mid H_2)\frac{p(H_2)}{p(F)} \\
&= p(F \mid H_2)\frac{p(H_2)}{p(F \mid H_1)p(H_1)+p(F \mid H_2)p(H_2)+p(F \mid H_3)p(H_3)} \\
&= 1 \cdot \frac{\frac{1}{3}}{\frac{1}{2}\cdot\frac{1}{3}+1\cdot\frac{1}{3}+0\cdot\frac{1}{3}} \\
&= 1 \cdot \frac{\frac{1}{3}}{\frac{1}{2}} \\
&= \frac{2}{3}
\end{aligned}
$$

So, $p(H_1 \mid F) = \frac{1}{3}$ and $p(H_2 \mid F) = \frac{2}{3}$

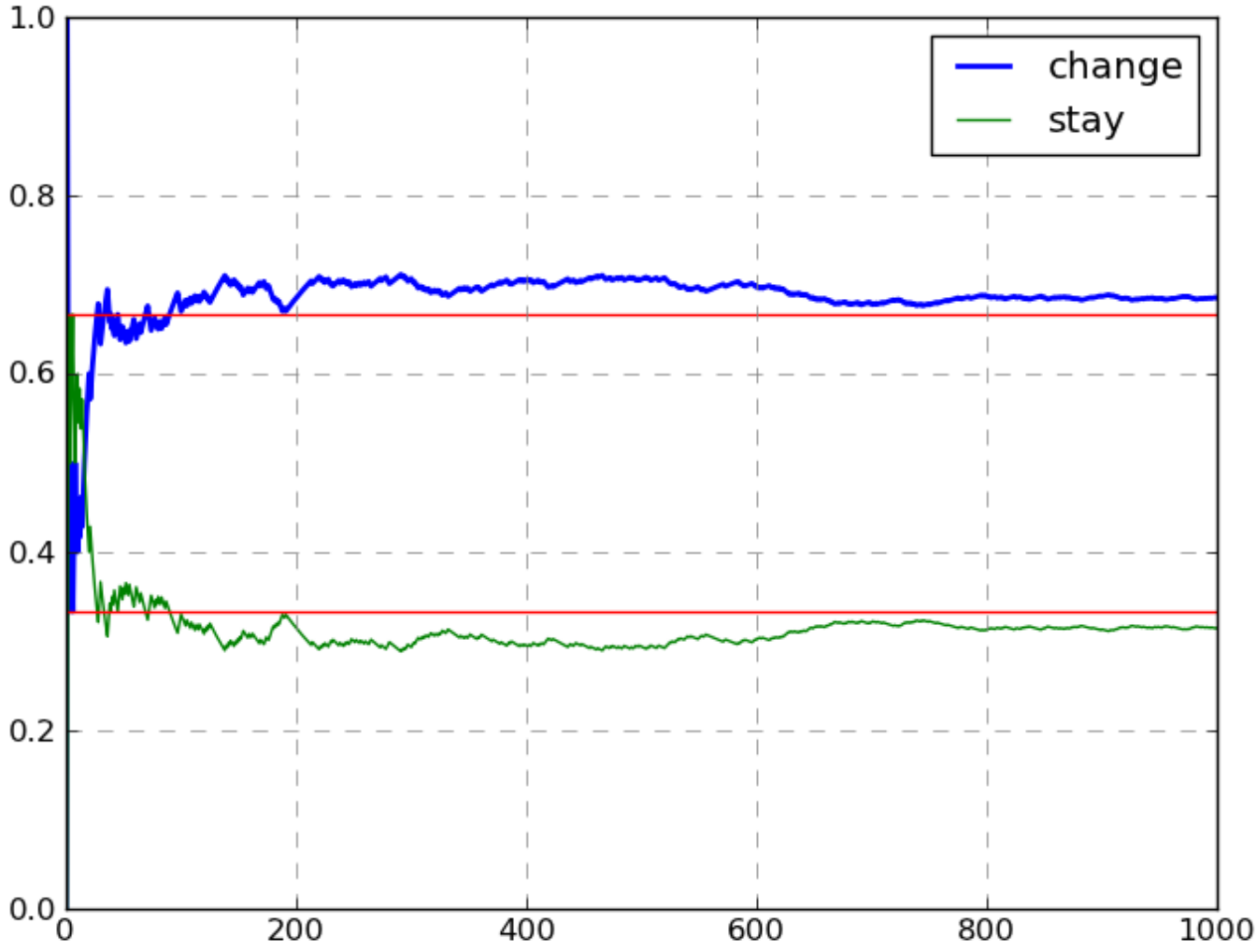
Monty Hall (3)

Reasoning 2: F: Monty opened door number 3

$$
\begin{aligned}
p(H_2 \mid F) &= p(F \mid H_2)\frac{p(H_2)}{p(F)} \\
&= p(F \mid H_2)\frac{p(H_2)}{p(F \mid H_1)p(H_1)+p(F \mid H_2)p(H_2)+p(F \mid H_3)p(H_3)} \\
&= 1 \cdot \frac{\frac{1}{3}}{\frac{1}{2}\cdot\frac{1}{3}+1\cdot\frac{1}{3}+0\cdot\frac{1}{3}} \\
&= 1 \cdot \frac{\frac{1}{3}}{\frac{1}{2}} \\
&= \frac{2}{3}
\end{aligned}
$$

So, $p(H_1 \mid F) = \frac{1}{3}$ and $p(H_2 \mid F) = \frac{2}{3}$, so you should switch

Monty Hall: Reasoning 1 vs. Reasoning 2



H. Leitgeb. *The Review Paradox: On the Diachronic Costs of Not Closing Rational Belief Under Conjunction*. Nous, 2013.

Bel_t is the set of propositions believed at time t

P_t is the agent's degree of belief function at time t

$t' > t$

P1 If the degrees of belief that the agents assigns to two propositions are identical then either the agent believes both of them or neither of them.

For all X, Y: if $P_t(X) = P_t(Y)$, then $Bel_t(X)$ iff $Bel_t(Y)$.

P2 If the agent already believes X, then updating on the piece of evidence X does not change her system of (all-or-nothing) beliefs at all.

For all X: if the evidence that the agent obtains between t and $t' > t$ is the proposition X, but it holds already that $Bel_t(X)$, then for all Y:

$$Bel_{t'}(Y) \text{ iff } Bel_t(Y)$$

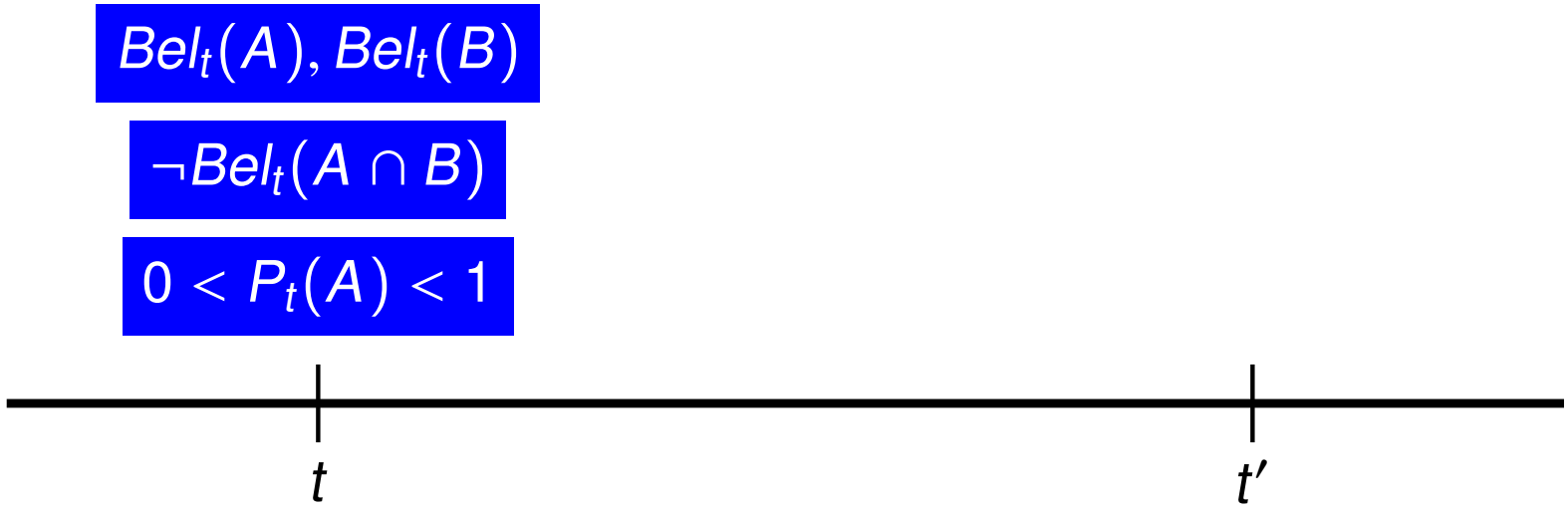
P3 When the agent learns, this is captured probabilistically by conditionalization.

For all X (with $P_t(X) > 0$): if the evidence that the agent obtains between t and $t' > t$ is the proposition X, but it holds already that $Bel_t(X)$, then for all Y:

$$P_{t'}(Y) = P_t(Y \mid X)$$

Assume $Bel_t(A), Bel_t(B)$ but not $Bel_t(A \cap B)$

- Suppose that the agent receive A as evidence.
- $P_{t'}(B) = P_t(B \mid A) = P_t(A \cap B \mid A) = P_{t'}(A \cap B)$.
- By P1, the agent must have the same doxastic attitude towards B and $A \cap B$.
- By P2, the agent's attitude towards B and $A \cap B$ must be the same at t' as at t.
- But, $Bel_t(B)$ and not $Bel_t(A \cap B)$



Assumption

$Bel_t(A), Bel_t(B)$

$\neg Bel_t(A \cap B)$

$0 < P_t(A) < 1$

$P_{t'}(B) = P_t(B \mid A)$

$P_{t'}(A \cap B) = P_t(A \cap B \mid A) = P_t(B \mid A)$

By P3

$Bel_t(A), Bel_t(B)$

$\neg Bel_t(A \cap B)$

$0 < P_t(A) < 1$

$P_{t'}(B) = P_t(B \mid A)$

$P_{t'}(A \cap B) = P_t(A \cap B \mid A) = P_t(B \mid A)$

$Bel_{t'}(B)$ iff $Bel_{t'}(A \cap B)$

By P1

$Bel_t(A), Bel_t(B)$

$\neg Bel_t(A \cap B)$

$0 < P_t(A) < 1$

$P_{t'}(B) = P_t(B \mid A)$

$P_{t'}(A \cap B) = P_t(A \cap B \mid A) = P_t(B \mid A)$

$Bel_{t'}(B)$ iff $Bel_{t'}(A \cap B)$

$Bel_t(A)$ iff $Bel_{t'}(A)$

$Bel_t(B)$ iff $Bel_{t'}(B)$

$Bel_t(A \cap B)$ iff $Bel_{t'}(A \cap B)$

By P2

$Bel_t(B)$

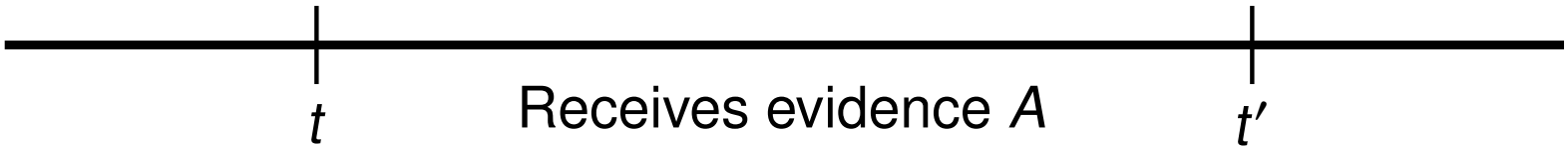
$\neg Bel_t(A \cap B)$

$0 < P_t(A) < 1$

$P_{t'}(B) = P_t(B \mid A)$

$P_{t'}(A \cap B) = P_t(A \cap B \mid A) = P_t(B \mid A)$

$Bel_{t'}(B)$ iff $Bel_{t'}(A \cap B)$



$Bel_t(A)$ iff $Bel_{t'}(A)$

$Bel_t(B)$ iff $Bel_{t'}(B)$

$Bel_t(A \cap B)$ iff $Bel_{t'}(A \cap B)$

$Bel_t(B)$ iff $Bel_{t'}(B)$ iff $Bel_{t'}(A \cap B)$ iff $Bel_t(A \cap B)$

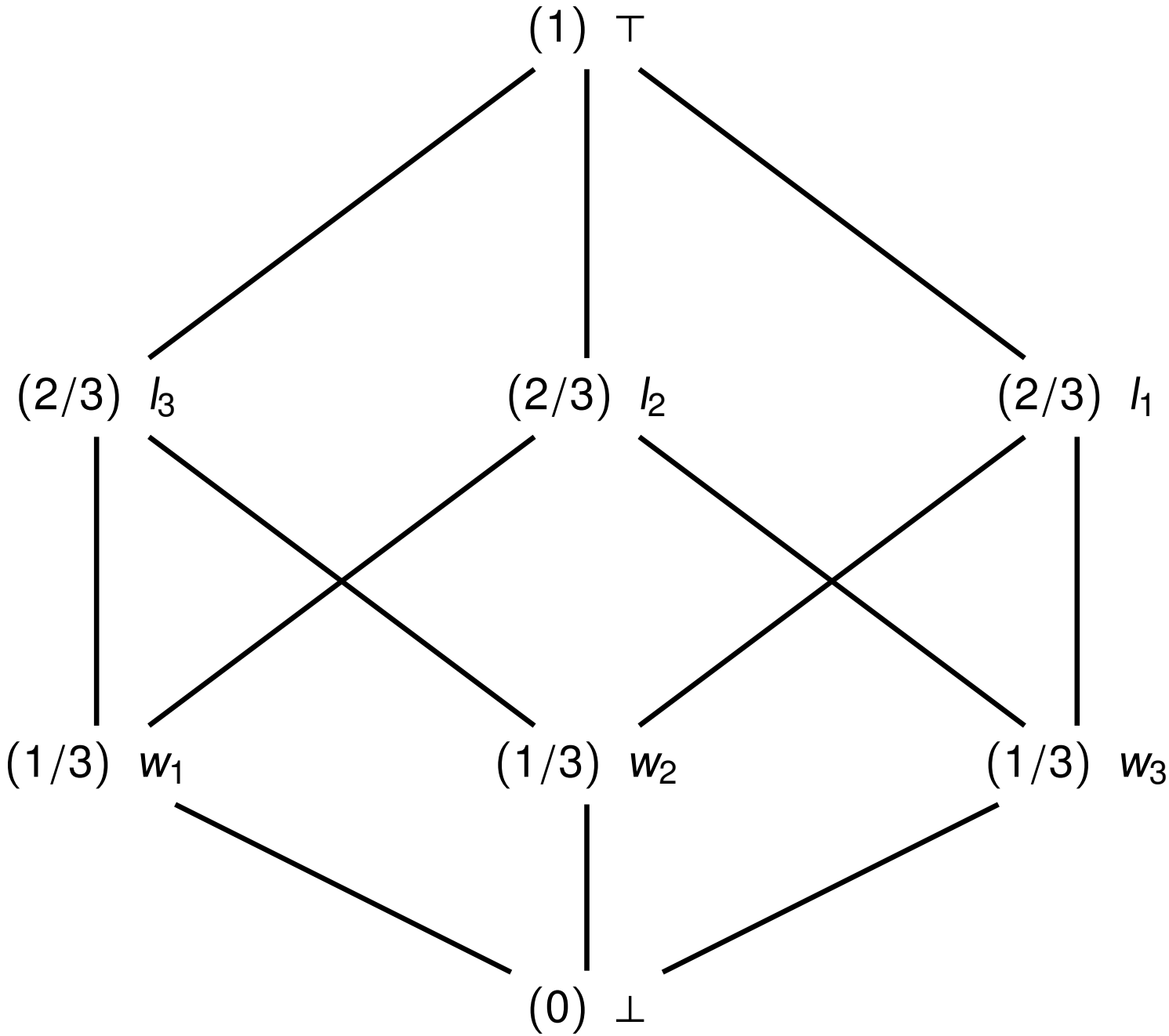
Lockean Thesis on Belief

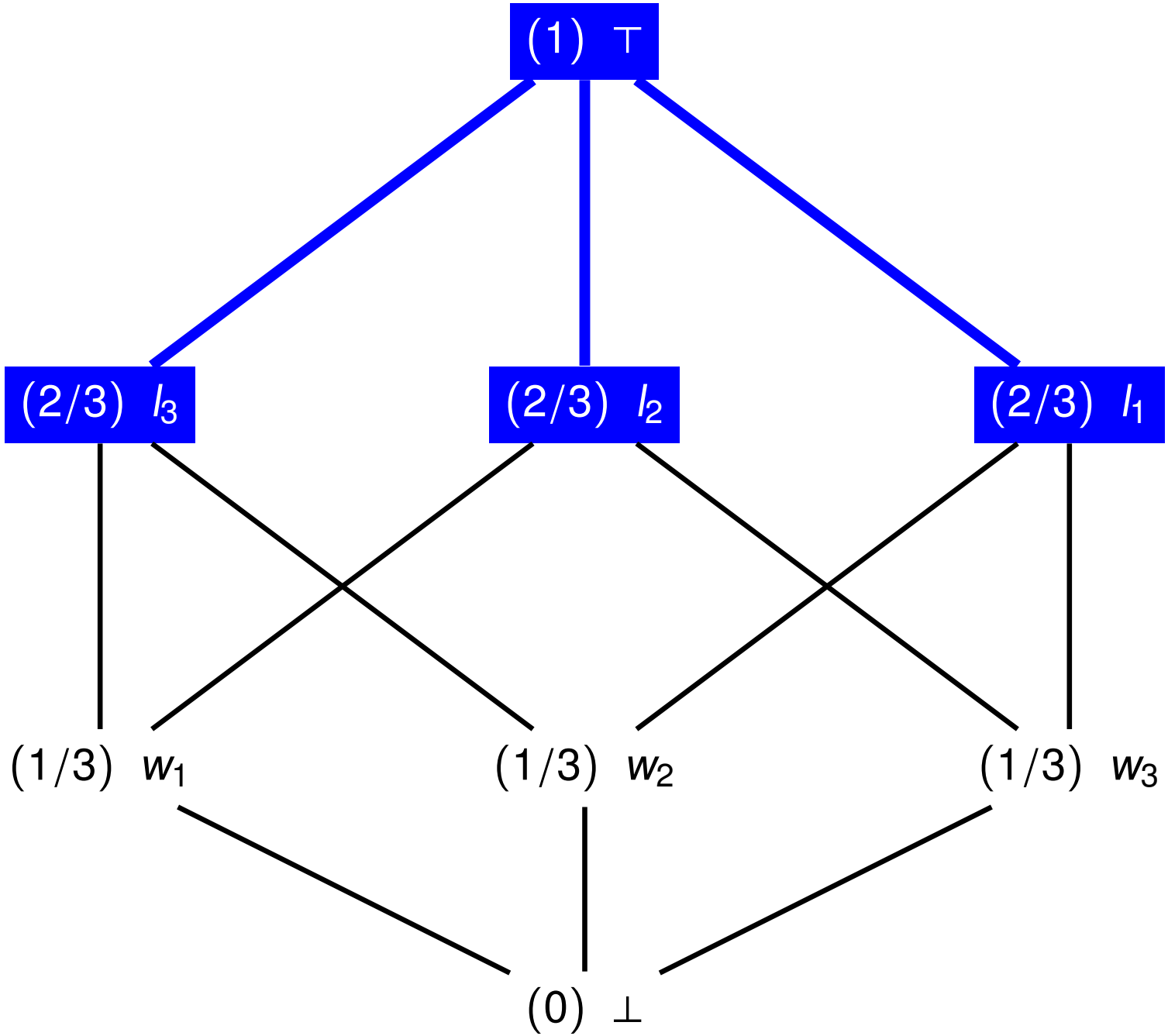
If our agent's beliefs are given by *Bel* and her credences by P, then if she is rational, P is a probability function and

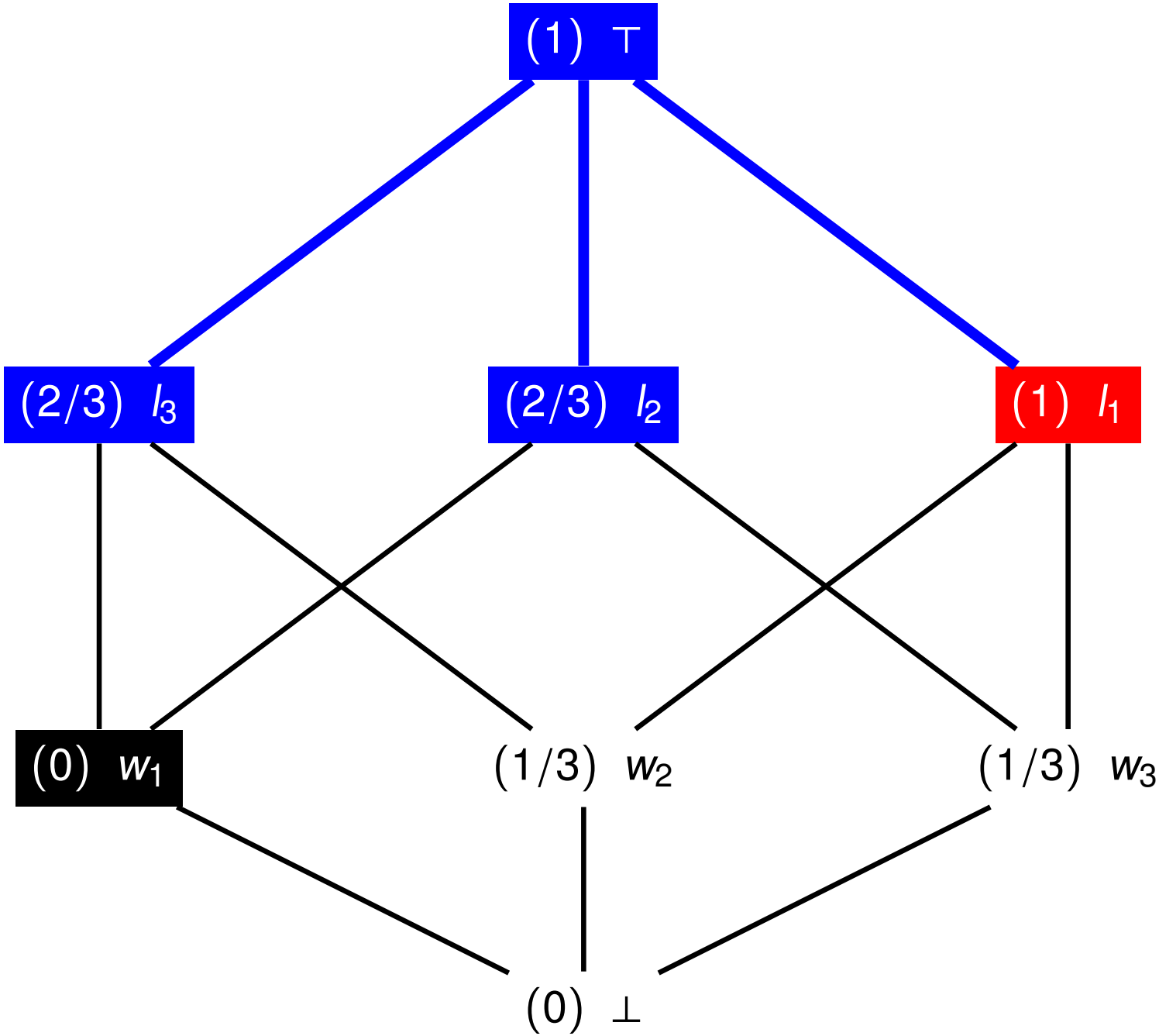
$$Bel(X) \text{ iff } P(X) \geq r$$

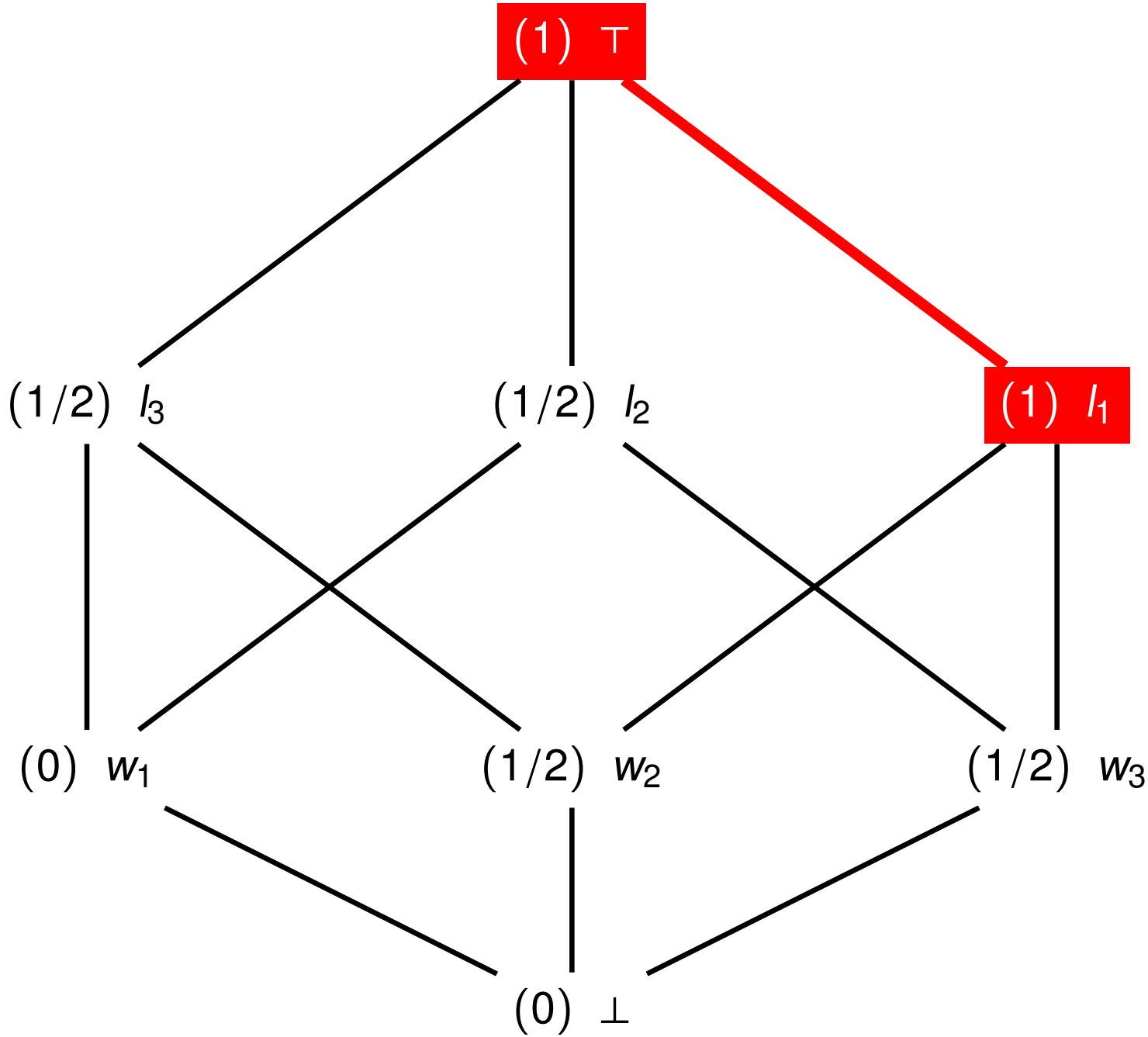
Beliefs that obey the Lockean thesis can be undermined by new evidence that is consistent with the agents current beliefs.

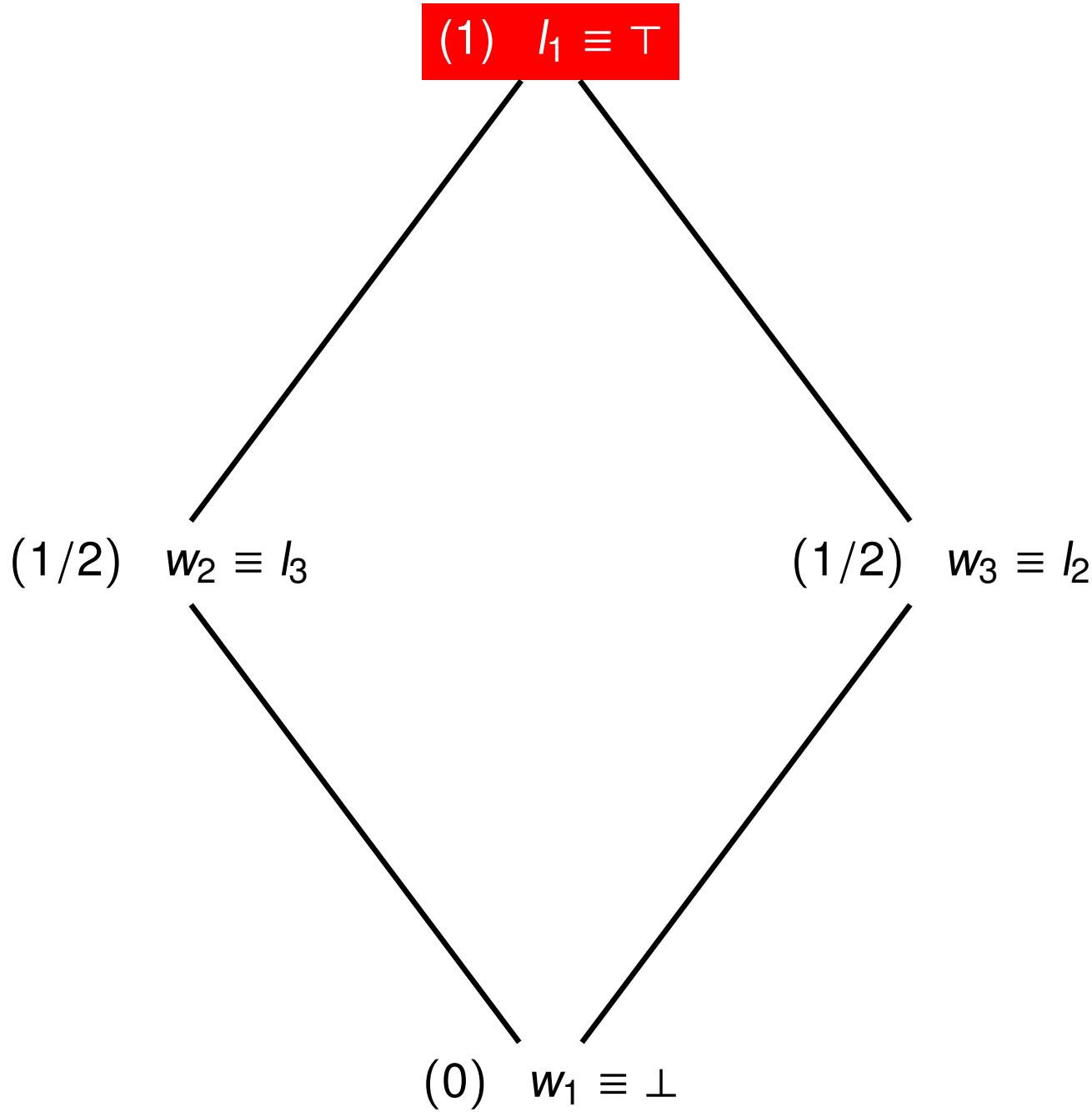
For each $i = 1, 2, 3$, let l_i be the proposition Ticket i won't win (and w_i is the proposition that "ticket i will win"). And let us set our threshold for Lockean belief at $r = 0.6$.


(1) ⊤
(2/3) l_3
(2/3) l_2
(2/3) l_1
(1/3) w_1
(1/3) w_2
(1/3) w_3
(0) ⊥


(1) ⊤
(2/3) l_3
(2/3) l_2
(2/3) l_1
(1/3) w_1
(1/3) w_2
(1/3) w_3
(0) ⊥


(1) ⊤
(2/3) l_3
(2/3) l_2
(1) l_1
(0) w_1
(1/3) w_2
(1/3) w_3
(0) ⊥


(1) ⊤
(1/2) l_3
(1/2) l_2
(1) l_1
(0) w_1
(1/2) w_2
(1/2) w_3
(0) ⊥


(1) $l_1 \equiv \top$
(1/2) $w_2 \equiv l_3$
(1/2) $w_3 \equiv l_2$
(0) $w_1 \equiv \bot$

Resiliency, Robust Belief, Stable Belief

B. Skyrms. *Resiliency, propensities, and causal necessity*. Journal of Philosophy, 74:11, pgs. 704 - 713, 1977.

A. Baltag and S. Smets. *Probabilistic Belief Revision*. Synthese, 2008.

H. Leitgeb. *Reducing belief simpliciter to degrees of belief*. Annals of Pure and Applied Logic, 16:4, pgs. 1338 - 1380, 2013.

R. Stalnaker. *Belief revision in games: forward and backward induction*. Mathematical Social Sciences, 36, pgs. 31 - 56, 1998.

Certainty: $P(H) = 1$

Absolute Certainty: for all E: $P(H \mid E) = 1$

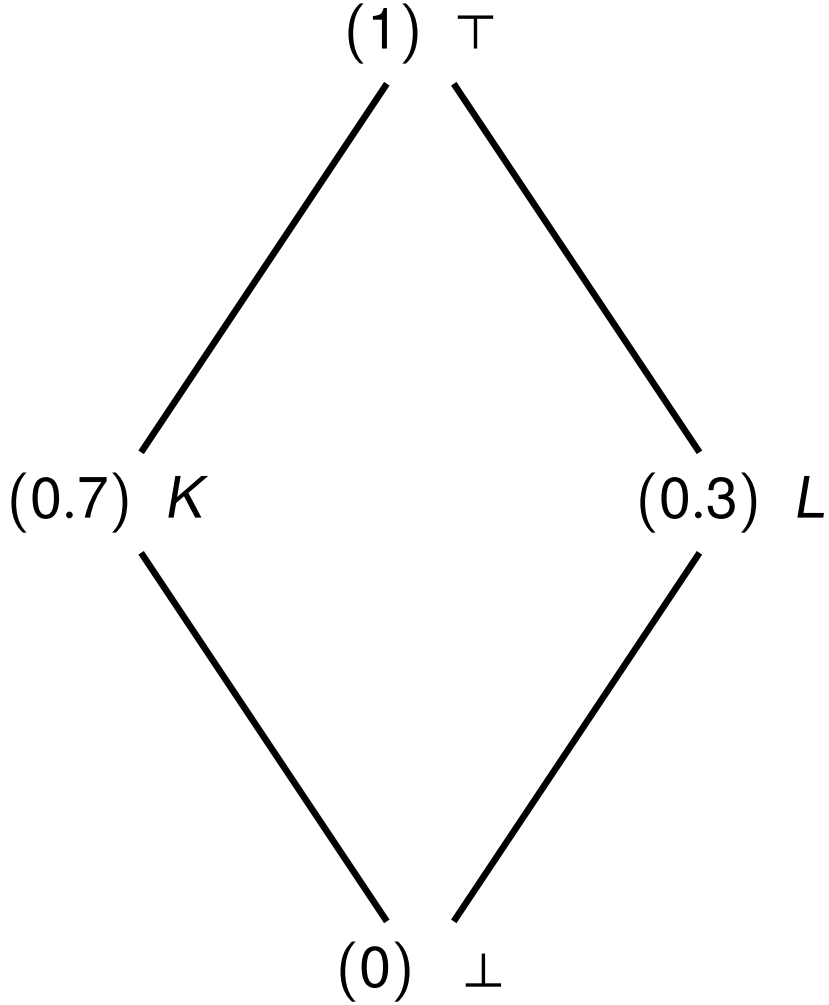
Certainty: $P(H) = 1$

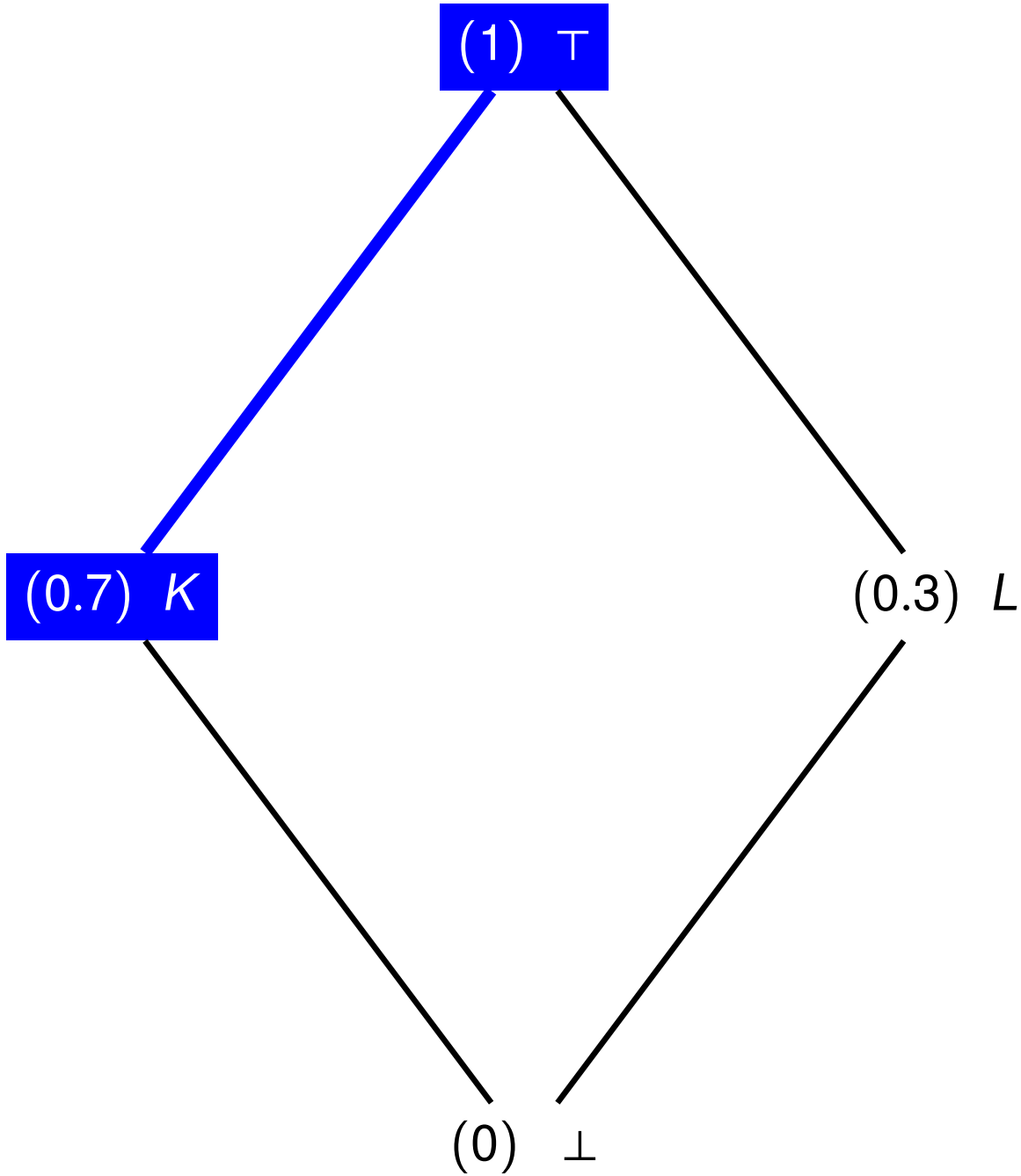
Absolute Certainty: for all E: $P(H \mid E) = 1$

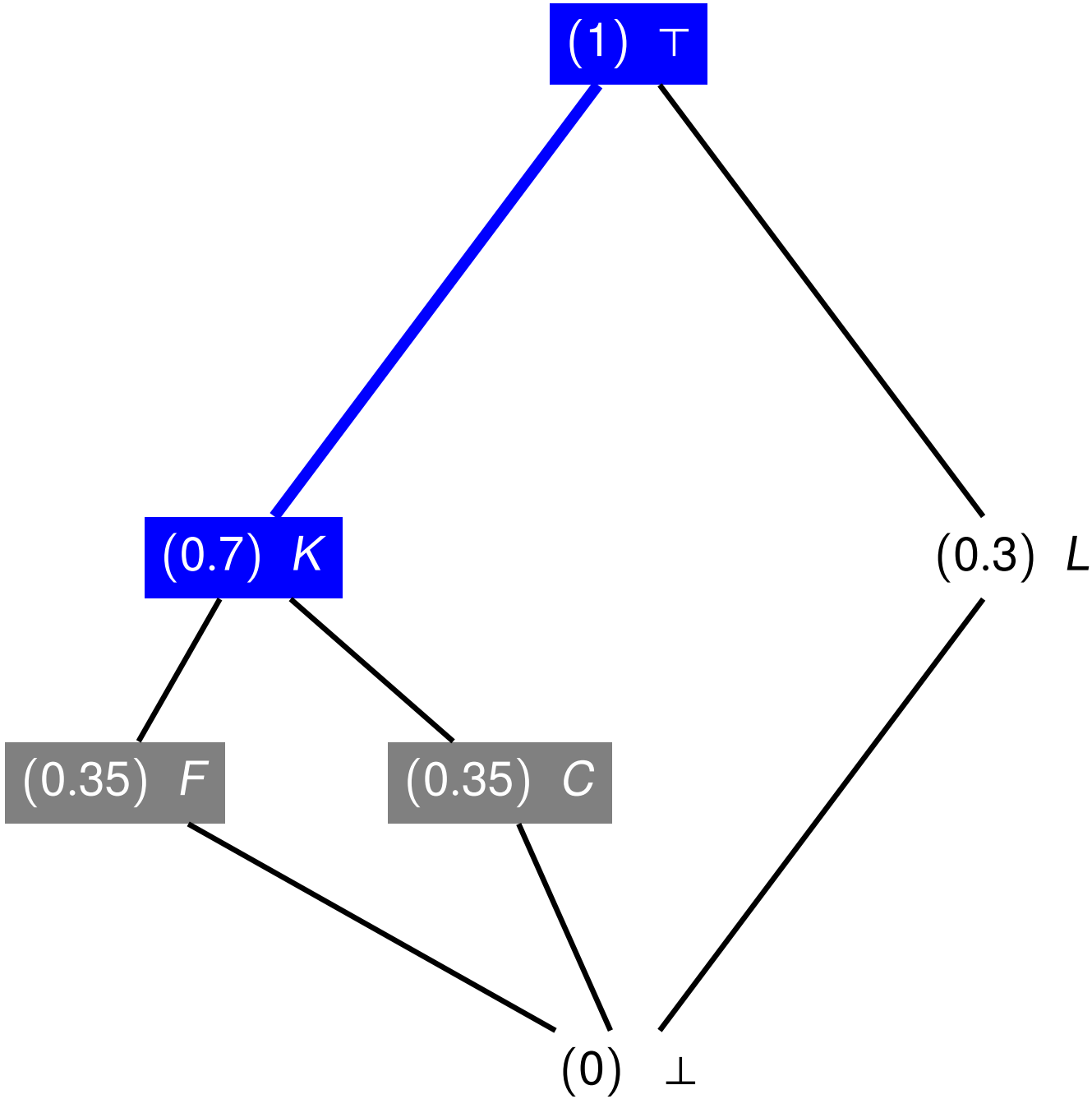
Stable Belief: for all $E \in \mathfrak{U}$ with $H \cap E \neq \emptyset$ and $P(E) \neq 0$:
$P(H \mid E) \geq t$

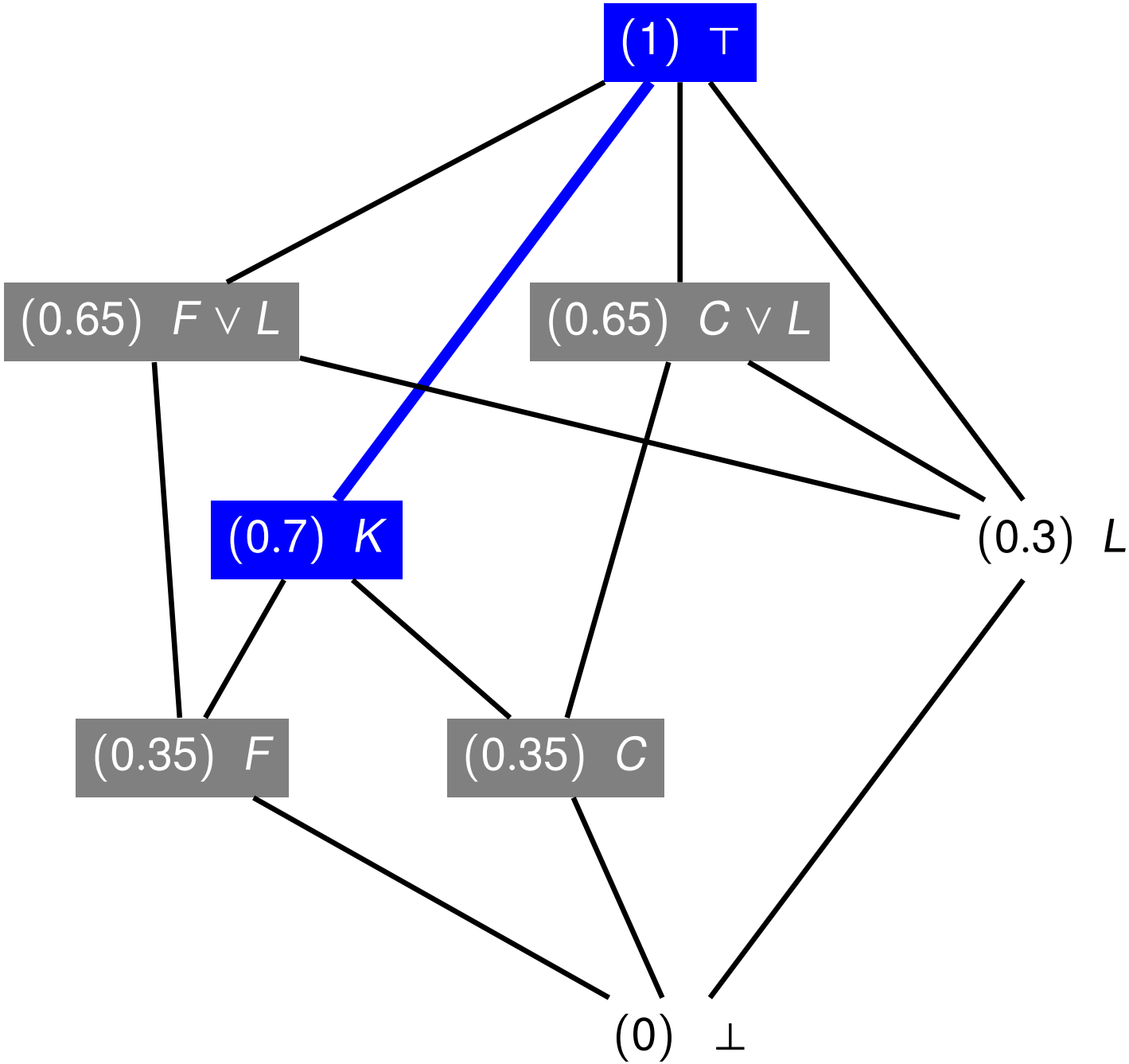
H. Leitgeb. *Reducing belief simpliciter to degrees of belief*. Annals of Pure and Applied Logic, 16:4, pgs. 1338 - 1380, 2013.

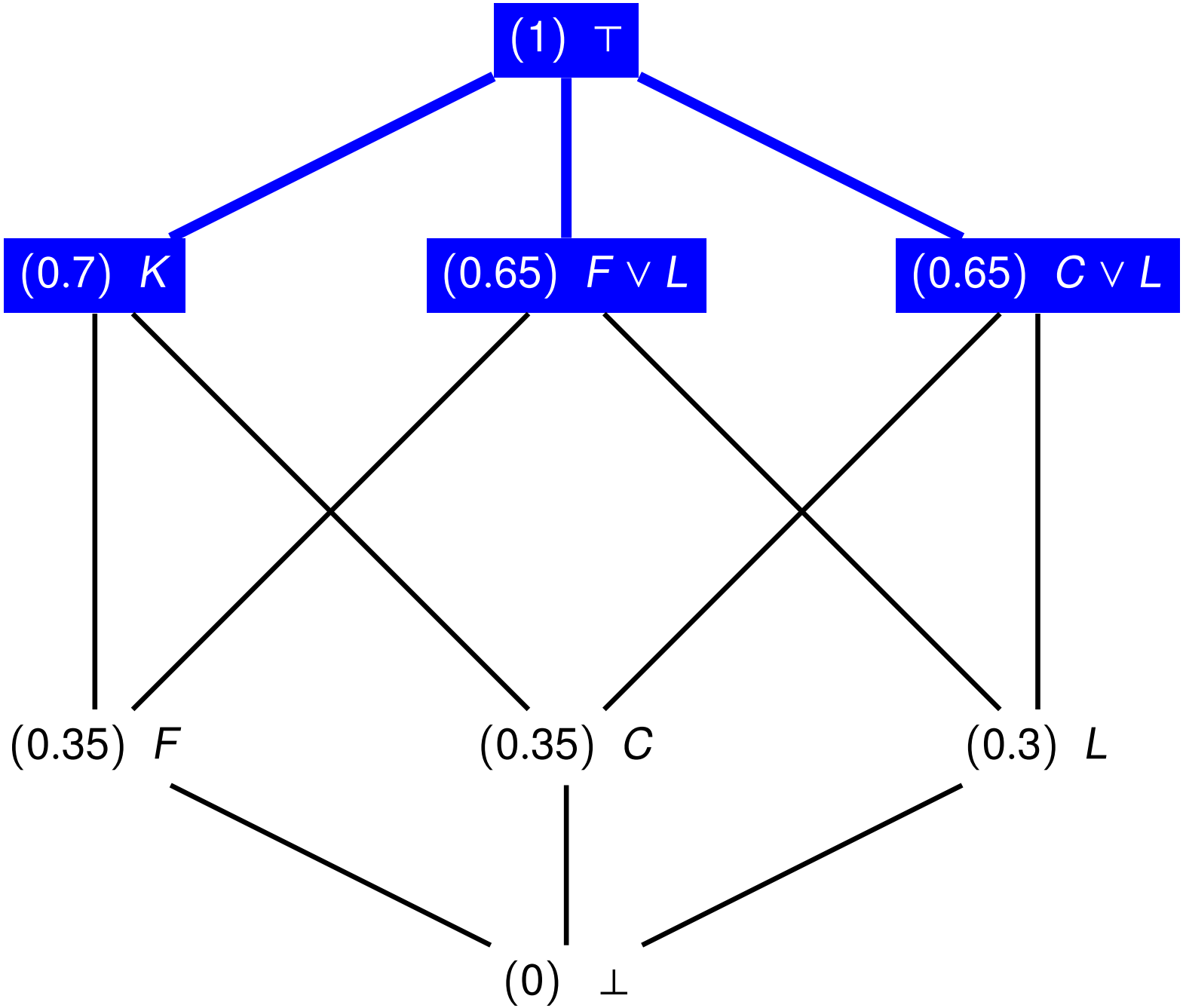
H. Leitgeb. *Humean Thesis of Belief*. Philosophical Review, 2015.

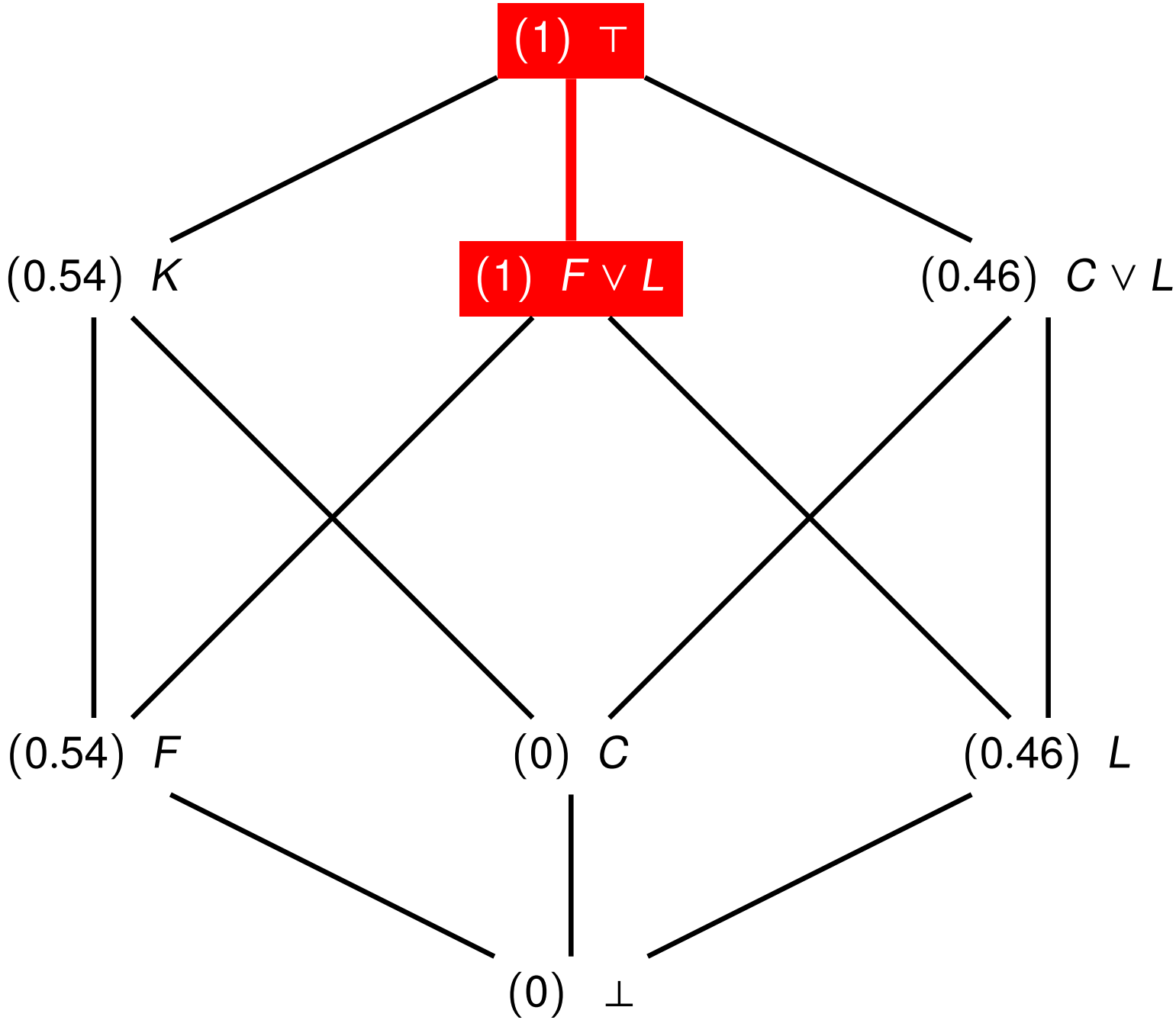

(1) ⊤
(0.7) K
(0.3) L
(0) ⊥


(1) ⊤
(0.7) K
(0.3) L
(0) ⊥


(1) ⊤
(0.7) K
(0.3) L
(0.35) F
(0.35) C
(0) ⊥


(1) ⊤
(0.65) F ∨ L
(0.65) C ∨ L
(0.7) K
(0.3) L
(0.35) F
(0.35) C
(0) ⊥


(1) ⊤
(0.7) K
(0.65) F ∨ L
(0.65) C ∨ L
(0.35) F
(0.35) C
(0.3) L
(0) ⊥


(1) ⊤
(0.54) K
(1) F ∨ L
(0.46) C ∨ L
(0.54) F
(0) C
(0.46) L
(0) ⊥

Thus, while stable belief is stable under acquisition of new (doxastically possible) evidence and Lockean belief is not, sstable belief is not stable under fine-graining of possibilities while Lockean belief is.

Leitgeb's Solution to the Lottery Paradox

In a context in which the agent is interested in *whether ticket i will be drawn*; for example, for $i = 1$: Let Π be the corresponding partition:

$$\{\{w_1\}, \{w_2, \ldots, w_{1,000,000}\}\}$$

The resulting probability measure P_Π is given so that P is given by P so that:

$$P_\Pi(\{\{w_1\}\}) = \frac{1}{1,000,000} \qquad P_\Pi(\{\{w_2, \ldots, w_{1,000,000}\}\}) = \frac{999,999}{1,000,000}$$

For example, this might be a context in which a single ticket holder—the person holding ticket 1—would be inclined to say of his or her ticket: “I believe it wont win.”

In a context in which the agent is interested in *which ticket will be drawn*: Let Π' be the corresponding partition that consists of all singleton subsets of W. The probability measure P^{Π} is the uniform probability on W.

The only P-stable set—and hence the only choice for the strongest believed proposition $B_W^{\Pi'}$—is W itself: our perfectly rational agent believes that some ticket will be drawn, but he or she does not believe of any ticket that it will not win

For example, this might be a context in which a salesperson of tickets in a lottery would be inclined to say of each ticket: “It might win” (that is, it is not the case that I believe that it won’t win).

In either of the two contexts from before, the theory avoids the absurd conclusion of the Lottery Paradox; in each context, it preserves the closure of belief under conjunction; and in each context, it preserves the Lockean thesis for some threshold ($r = \frac{999{,}999}{1{,}000{,}000}$ in the first case, $r = 1$ in the second case)-all of this follows from the theory of P-stability.