# From Degrees of Belief to Defeasible Knowledge

**Alexandru Baltag and Sonja Smets**

**(ILLC, University of Amsterdam)**

## The General Problem

**How can we model QUALITATIVE notions of "belief" and "knowledge", as well as their dynamics (belief revision, knowledge update) in a QUANTITATIVE probabilistic setting?**

## The Framework

We assume a probabilistic structure $(S, \mathcal{A}, \mathcal{E}, P, t)$:

- $S$: the set of all *possible worlds* (outcomes, states) consistent with the agent's information
- $\mathcal{A}$ an "algebra" of sets (=Boolean subalgebra of the powerset algebra $(\mathcal{P}(S), \cup, \cap, \neg, \emptyset)$), called *events* or *propositions*;
- $\mathcal{E} \subseteq \mathcal{A}$ the set of *observable events*: the potential evidence;
- $P : \mathcal{A} \to [0, 1]$ a probability measure giving the agent's *degrees of belief*;
- $t \in (0.5, 1)$ is the agent's *confidence "threshold"*.

## Cautious Agents: factive, plausible evidence

NOTE: In this talk, I take all actually observed "evidence" to be *factive* (true in the actual world).

However, we might want to require that the agent is cautious enough to **NOT assign probability** $0$ **to any potential evidence**:

$$P(E) \neq 0 \text{ for every } E \in \mathcal{E}.$$

Thus, the agent's credence after seeing new (true) evidence will always a true probability measure:

if the real world $w \in E$ then $P_E(\bullet) := P(\bullet|E)$ is a probability measure.

## Differences with Leitgeb's framework

So we essentially require that

$$\mathcal{E} \subseteq \{A \in \mathcal{A} | P(A) \neq 0\}.$$

Leitgeb's framework is **stronger**: it essentially amounts to taking

$$\mathcal{E} = \{A \in \mathcal{A} | P(A) \neq 0\}.$$

This is TOO strong: it requires that ALL non-null true propositions be observable.

This defeats the purpose: there is no need for a threshold $t < 1$ in Lockean belief!

In order to know/believe $A$, just wait till you observe evidence $A$ (if $A$ is actually true), or else evidence $\not\!A$ (if $A$ is false): you get true belief **with probability** 1 !

## Basic Notions

For a hypothesis $H \in \mathcal{A}$, we can consider the following:

**Infallibility**: $I(H)$ iff $H = W$.

**(Weak) Belief as "High Confidence"**: $B(H)$ iff $P(H) \geq t$.

**(Subjective) Certainty**: $C(H)$ iff $P(H) = 1$.

**Absolute Certainty**: $P(H|E) = 1$ for **every** possible evidence $E \in \mathcal{A}$ (for which $P(H|E)$ is defined)

*NOTE*: If we stick with Classical Probability Theory (as we will for now), then **certainty is the same as absolute certainty**.

This is **not** the case for extended probabilistic settings (Popper measure, lexicographic probabilities) that allow conditioning on evidence of probability 0.

## Belief: the Lockean Thesis

But what about (simple, qualitative) **belief**?

According to **Rickard Foley's interpretation**(1992) of Locke's **An Essay Concerning Human Understanding**, Locke seemed to propose the following equation:

**(qualitative) belief = degree of belief $\geq$ some (given) treshold**.

This identifies **simple "belief"** with "**a high degree of belief**", by putting e.g.

$$B(H) \text{ iff } \mu(H) \geq t,$$

or maybe

$$B(H) \text{ iff } \mu(H) > t,$$

for **some (high enough) threshold** $t > 0.5$.

## Conditional Belief

**High confidence in $H$ conditional on evidence $E$:**

$B(H|E)$ iff $P(H|E) \geq t$.

Let's call this "(weak) conditional belief".

Of course, the Lottery Paradox shows that this weak form of belief does not satisfy the axioms of doxastic logic: in particular, it does not satisfy Conjunctivity, because of the Lottery Parado:

$$B(H_1) \wedge B(H_2) \implies B(H_1 \wedge H_2).$$

## Leitgeb's Proposal: the Stability theory of Belief

Recently, Hannes Leitgeb (Leitgeb, APAL 2013) proposed a very interesting solution:

"the stability theory of belief".

Leitgeb gives extensive motivations to the above notions, based on what he calls the "Humean conception of belief" and on a modified Lockean thesis.

## Stable Belief

Essentially, he requires **stable high confidence in $H$ under updating with any plausible proposition (be it true or false) that is consistent with $H$**:

$H$ is a **stable belief** iff $P(H|E) \geq t$ for all $E \in \mathcal{E}$ such that $E \cap H \neq \emptyset$, (where recall that for Leitgeb $\mathcal{E} = \{A \in \mathcal{A} | P(A) \neq 0)$.

Let us put $Sb$ for **the family of all stsble beliefs**, and also write $Sb(H)$ iff $H \in Sb$.

Essentially, Leitgeb defines **"belief" as the closure of stable belief under logical consequence**:

$$Bel(H) \text{ holds} \quad \text{iff} \quad \text{there exists } J \in Sb \text{ such that } J \subseteq H.$$

"*Something is believed iff it is justifiable (i.e. entailed) by some stable belief.*"

## What if You Don't Buy All This?

This is a beautiful theory, but one based on doubtful motivations and making extremely strong and rather unrealistic requirements/

But... what if you don't buy these motivations and requirements?

A Bayesian could say that there is absolutely no reason to require stability of high confidence:

*almost nobody (– except for Leitgeb... and possibly Hume) ever claimed that belief has to be stable*!

On the contrary, belief (even when true) is deemed to be something evanescent, that can easily be lost, revised or suspended.

This is after all the **difference between belief and knowledge** (at least according to some philosophers).

"**True beliefs** too are a fine thing and altogether good in their effects so long as they stay with one, but they won't willingly stay long but instead run away from a person's soul, so they're not worth much until one **ties them down** by reasoning out the **explanation ('Logos')**.

(...)

And when they've been tied down, then for one thing they become items of **knowledge**, and for another, **permanent**.

And that's what makes knowledge more valuable than true belief, and the way knowledge differs from true belief is being tied down."

Plato, *Meno*.

## Back to Lockean belief

So... why not simply keep "belief" as determined by Lockean thesis?

Yes, it doesn't satisfy the traditional axioms of doxastic logic, but... many Bayesians would just say that this no problem:

just too bad for classical doxastic logic!

From now on, **"belief", denoted by $B$, is just Lockean belief**:

$$B(H) \text{ iff } P(H) \geq t.$$

As we'll see we WILL recover Leitgeb's "belief" $Bel$, but only by recognizing it as a **stronger** attitude: "**believing that you know**".

## Knowledge

**How can we model "knowledge" probabilistically?**

Various attempts to adapt classical accounts of knowledge to a Bayesian setting.

e.g. Roush 2007 (Tracking Truth) deveops a probabilistic version of the Sensitivity theory of knowledge, using ideas from confirmation theory.

Here I will present a probabilistic version of the Defeasibility theory of knowledge, formulated directly in terms of Lockean belief $B$ (NOT of Leitgeb's belief $Bel$).

Nevertheless, as we'll seem, this theory will provide an independent justification (and a new interpretation) for Leitgeb's "stability theory".

## Fallibilism

Is all knowledge infallible knowledge?

i.e., do we have

$$K(H) \Rightarrow H = W \ ?$$

*Fallibilism:* **knowledge without infallibility**

$\Rightarrow$ against the $S5$ semantics for "knowledge" (that uses universal quantification over all the possibilities consistent with one's knowledge, thus equating $K$ with $I$)

I will take fallibilism for granted, thus:

$$K \neq I$$

## Knowledge Without Certainty

But **can there be knowledge without certainty**?

Do we have

$$K(H) \Rightarrow P(H) = 1 \ ?$$

(Klein 1981) argued in favor of this implication.

(Meyer 1988) and many others argue against it.

I agree with the latest: any form of "inductive knowledge" would be practically impossible to ever achieve if the learner waits till she gets absolute certainty!

**Knowledge implies high confidence ($K(H) \Rightarrow P(H) \geq t$), but does not necessarily imply absolute certainty**.

## So, what is knowledge?

*Factivity*:

$$K(H) \Rightarrow H$$

*High Confidence* (="Weak Belief"):

$$K(H) \Rightarrow B(H)$$

**Are these two enough**? i.e. do we have

$$H \wedge B(H) \Rightarrow K(H)\,?$$

**Gettier-type Counterexamples**:

It seems to me that the answer is **no**.

The usual Gettier counterexamples can be easily adapted to disprove this.

## An Example: a No-False Lemma Gettier

I saw Tom Grabbit (whom I know very well) grabbing a rare book from a public library and running away with it. That same book was later announced on the TV to have been stolen. Let $A$ be the proposition "Tom Grabbit stole the book". I assign probability 0.9 to $A$, and my threshold is $t = 0.9$, so I believe $A$.

I am right (Tom indeed stole the book), and my justification for this belief involves no falsehoods (-I indeed saw Tom grabbing the book!).

But... unknown to me, Tom has a kleptomaniac twin brother John, looking just like him and having a long arrest and conviction record for stealing books from libraries.

If given this evidence $E$, my probability $P_E(A)$ would drop to 0.1. My justification is defeated! In fact, I come to believe the opposite: that Tom didn't do it!

## Knowledge should be more Resilient!

This is an example of a *true, but "un-safe" acceptance*: it can be lost after acquiring (new) true information.

According to many authors, **something so fragile cannot be called knowledge**.

"... 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)

## Defeasibility theory of Knowledge

The **"defeasibility theory" of knowledge** (Lehrer, Klein etc):

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

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

**Knowledge = "undefeated" true belief** (Lehrer)

## Undefeated Justification

In fact, Lehrer requires that, not only the belief is undefeated, but **its justification is also undefeated** (by any new true information):

> "If a person has knowledge, than that person's justification must be sufficiently strong that it is incapable of being defeated by evidence that he does not possess" (Pappas and Swain 1978).

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.*

## Objections to Defeasibility Theory

People argued that the defeasibility theory is *too strong*: there are counterexamples of intuitive "knowledge", that can nevertheless be defeated by some "deceiving" or "irrelevant" (but TRUE) new evidence.

But this means that the choice of of potential (non-deceiving, relevant) evidence $\mathcal{E}$ matters! "Real knowledge" is only undefeated by "non-deceiving", "relevant" evidence.

## Objections to Defeasibility Theory: "Misleading Defeaters"

We can safely ignore false defeaters (such as convincing lies), since the defeasibility theory insists that only (true) evidence can count as a defeater.

But suppose that, unknown to me, reliable witnesses testify that they have seen Tom's brother John in another place, nowhere near the library, at the exact time when I witnessed the crime.

Some authors claim that in this case the previous evidence $E$ is a "misleading defeater": my justification for thinking that it was Tom was in fact correct, and shouldn't have been affected by this additional (and true) information $E$. They claim that, contrary to defeasibility theory, in this case I **know** that Tom stole the book.

## I have my doubts

I am not convinced by this example.

It seems to me that the existence of the twin is in itself a piece of very relevant information that should naturally throw doubt over the identity of the thief.

Relevant evidence that lowers the probability of the hypothesis below the threshold **should** count as a defeater, even if **more** evidence might re-establish the original belief.

## Irrelevant, Improbable, Quirky Defeaters

However, there do exist misleading defeaters, typically involving true but highly improbable information of an unpredictable kind, about which I did not have any prior opinions (probabilistic or not), and which moreover were not in any way relevant

Suppose that there is no twin brother. But instead, my friend Cheryl is a psychic with very special powers of always guessing totally random, but always true, facts about the world, typically involving disjunctions between totally uncorrelated and independent statements. These facts come to her in dreams in an obvious random, unplanned and non-strategic manner, so there is nobody who cherry picks these facts with the intention to deceive me. (Cheryl is a woman of impeccable integrity, who never lies and never intends to deceive anybody).

Today, Cheryl will come to tell me

"Last night, I had a revelation in my dream: either Tom is not the thief or a black hole will swallow the universe tomorrow."

So I learn the evidence $\neg A \vee B$, where $B$ refers to the asteroid event.

Cheryl tells the truth, of course: no matter how unlikely, the black hole is coming! But I don't know this. After Cheryl's announcement, I will look in an astronomy book and I will find out that there a 1 chance in $10^{100}$ that a mini-black hole might suddenly swallow the Earth and most of the observable universe. I also correctly assume that $A$ and $B$ are independent: no correlation.

Given this, a brief calculation shows that $P(A|\neg A \vee B)$ is vanishingly small: once again I will have lost my belief in $A$.

Still, it seems clear that, in this case, my initial justification for believing *A was* correct, and *its correctness was in no way affected by any irrelevant facts about black holes*!

Intuitively, in this case I **knew** that Tom stole the book, period. The fact that I might lose my knowledge due to such a freakish accident of Nature (such as Cheryl's dreaming this particular disjunction) does not diminish the merits of my justification. **Any** uncertain belief, no matter how solid and well-justified, can be defeated by such quirky "evidence".

## Solution: the family of relevant evidence sets

There is an obvious solution. When discussing my current knowledge (from before Cheryl's announcement), we should **restrict** (AS WE DID) to a *subset* $\mathcal{E} \subseteq \mathcal{A}$ of the "total algebra" (of all possible events): **only the observable events that are related or relevant to my current state of knowledge and the issues at hand**.

Black holes engulfing (or not) the universe were not part of my knowledge before I talked to Cheryl. So (despite the probability of being hit being so small) I didn't know that the Earth will not be hit, nor I knew that it will be hit: I just didn't think about it, and assigned no probability to it.

AFTER Cheryl's announcement, I am forced to think of this unexpected, quirky possibility, no matter how unlikely and how irrelevant for $A$. I am thus forced to expand my algebra to a larger one $\mathcal{A}'$. In this new context, my old belief can be defeated: *now, I don't know A anymore*. But it still the case that I did know $A$ before!

BUT (you may say), I may not have ever thought of the possibility of Tom having a twin brother either.

Why was that a good defeater then?

## The Difference between the Two Stories

There is a difference between these two examples: Tom having or not a twin brother is a fact that is relevant to my current state of knowledge (about the identity of the thief), even if I never thought of it. I am ignoring this possibility at my peril. Since it is a relevant and true fact, my ignorance about this comes at a cost: my justification, though based on true facts only, is faulty.

But in the Cheryl story, the ignored facts were not relevant at all to my current state of knowledge. My knowledge was only defeated by the quirky event of Cheryl dreaming up this random disjunction of unrelated statements, one of which was totally irrelevant, very improbable and outside my thinking context. Such a "defeater" should not count against my justification.

## Undefeated (Lockean) Belief

Leaving justification on a side for a moment, let us try to develop a probabilistic version of these notions.

We say that the agent has **Undefeated (Lockean) belief** $U(H)$ in a hypothesis $H$ if *she has high confidence in* $H$ *no matter what (true) evidence she might learn*:

$$w \in U(H) \text{ iff } B(H|E) \text{ for all } E \in \mathcal{E} \text{ such that } w \in E.$$

## Negative Properties

Unfortunately, the operator $U$ does **not** satisfy the Conjunctivity of Knowledge:

$$U(A) \cap U(B) \not\subseteq U(A \cap B).$$

Another "negative" (?) feature is **lack of positive introspection**:

$$U(A) \not\subseteq U(U(A)).$$

So, *the operator $TU$ would give us a very "non-classical" (though interesting) notion of knowledge*!

A win for both Nozick and Williamson?! Knowledge without closure or introspection!

## Solution

The solution is simple:

*undefeated belief is not the same as knowledge*!

First, because **not every proposition is "knowable"**.

Second, because **reflective conscious knowledge requires at least believing that you know**:

$$K(H) \Longrightarrow BK(H).$$

Third, because the justification of one's knowledge needs to be more than undefeated:

**as more and more evidence accumulates, Lockean belief should track the justification's truth value, positively and negatively**.

## Knowledge, take 1

1. **truth**:

$$K_E(A) \;\Rightarrow\; E \wedge A$$

2. **conscious** ("belief that you know"):

$$K_E(A) \;\Rightarrow B(K_E(A)|E)$$

3. **undefeated**:

$$K_E(A) \wedge F \;\Rightarrow\; K_{E \wedge F}(A)$$

4. **maximality** with respect to (1), (2), (3):

$K_E(A)$ is the *weakest* proposition satisfying (1), (2), (3).

The last condition that knowledge is **nothing more** than what the first three conditions require, with no additional restrictions or qualifications.

## Equivalent non-circular definition

So $K(H)$ **is the weakest proposition (=largest set of worlds)** $A \subseteq W$ **satisfying** $A = H \cap U(A)$**.**

We can unfold this definition to produce an *equivalent non-circular one*:

$$K(H) \quad := \quad H \cap U(H) \cap U(U(H)) \cap U(U(U(H))) \cap \ldots$$

So $K$ is the **reflexive (=positively introspective) version of** $U$:

**$H$ is known iff if it is true, undefeated, the fact that it is so is also undefeated, etc ....**

## Second Take: True, Inherently undefeated Justification

$A$ is a **good justification** if it is *true*, believed and **inherently undefeated**:

*undefeated whenever it is true.*

We say that a hypothesis $H \subseteq W$ is **inherently undefeated** if *it can be defeated ONLY when it is false*; equivalently, if it is *undefeated whenever it is true.*

Formally:

$$A \subseteq U(A)$$

**Warning**

Strangely enough, **"inherently undefeated" does NOT imply "undefeated"**.

The first is a purely internal property, that can be discovered by introspection, while the second has to do with the truth in the real world.

However, **"true" plus "inherently undefeated" DO imply "undefeated"**:

$$w \in H \subseteq U(H) \Rightarrow w \in U(H).$$

## Knowledge, take 2

Knowing $H$ can now be defined as having a **true and inherently undefeated justification** for $H$ in terms of a priori knowledge and the evidence:

$$w \in K(H) \quad \text{iff} \quad \exists J \text{ such that } w \in J \subseteq H \cap U(J).$$

## Third Solution: Stable sensitivity

Being inherently undefeated looks very much like much like positive sensitivity: the belief in $J$ positively tracks the truth of $J$.

We might want to require negative sensitivity as well: whenever $J$ is false, it will be eventually defeated.

This gives

$$J = U(J).$$

We might in fact want to allow some exceptions, as long as they consist of "abnormal worlds", or more precisely if the set of exceptions is negligible:

$$P(J - U(J)) = P(U(J) - J) = 0$$

In fact, if we adopt both these conditions, we obtain an equivalent version.

## Sensitive Knowledge

$J$ is sensitive knowledge iff it is true, and undefeated belief tracks its truth value in almost all worlds:

$$w \in K^{sensitive}(J) \quad \text{iff} \quad w \in J \text{ and } P(J - U(J)) = P(U(J) - J) = 0$$

Equivalently, $J$ is a believed truth of which the agent is certain that his belief will be undefeated iff it is true:

$$K^{sensitive}(J) \iff J \wedge C(J \Rightarrow U(J)) \wedge C(\neg J \Rightarrow \neg U(J))$$

Equivalently

$$K^{sensitive}(J) \iff J \wedge P(U(J)|J) \wedge P(\neg U(J)|\neg J).$$

This is very much like Kevin Kelly's inductive knowledge: a truth that entails (with probability 1) that it will forever be believed (given any new evidence), while its falsehood entails (with probability 1) that the agent will eventually stop believing it (if given enough evidence).

Except that... we allow for exceptions: in some abnormal set of worlds, the agent might never stop believing it, even if false.

## Knowledge based on true, sensitive justifications

Sensitive knowledge, like Nozick's original notion, is not closed under logical entailment.

But we can close it under logic, and regain our same notion of knowledge:

$$w \in K(H) \quad \text{iff} \quad \exists J \text{ such that } w \in J \text{ and } K^{sensitive}(J).$$

## Equivalence of the above notions

*Our above definitions of knowledge are equivalent*:

**Proposition**. The following are equivalent, for all worlds $w \in W$ and propositions $H \in \mathcal{A}$:

- $K(H)$ in the sense of our first definition ("conscious, undefeated knowledge");
- $w \in A$ for some $A \in \mathcal{A}$ such that $A = H \cap U(A)$ (and hence $w$ belongs to the *largest* such fixed point $A$);
- $w \in H \cap U(H) \cap U(U(H)) \cap \ldots$;
- $H$ has a true, inherently undefeated justification: $\exists J$ such that $w \in J \subseteq H \cap U(J)$;
- $H$ has a true, sensitively known justification: $\exists J$ such that $w \in J \subseteq H$ and $K^{sensitive}(J)$.

## Example: Achieving Inductive Knowledge

Universal statements ("All ravens are black") can actually come to be **known** after finitely many experiments!

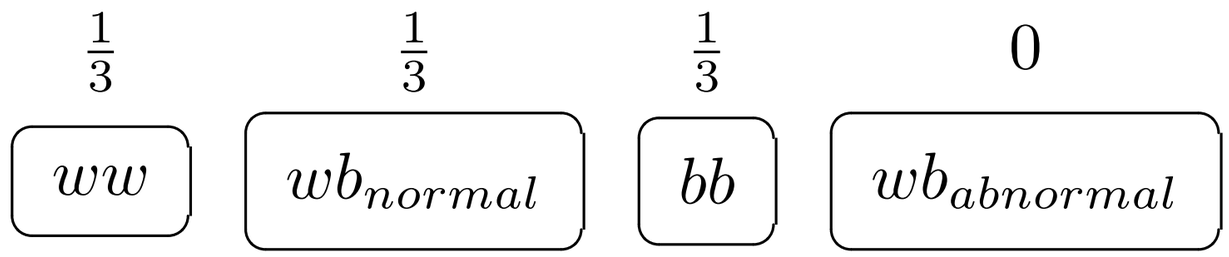
**Example**: Three urns are placed in front of me, and I inspect their contents: one ($ww$) contains two white balls, the second ($wb$) contains one white ball and one black ball, the third ($bb$) two black balls. The urns are then closed, and using a fair randomizing process only one of the urns is selected to be left in the room. The urns look identical on the outside, so I have no way to know which urn is left in the room.

In reality (unknown) to me, the first urn ($ww$) was selected.

I am allowed to make "experiments", by successively extracting one ball from the urn, noting its color and returning it back inside.

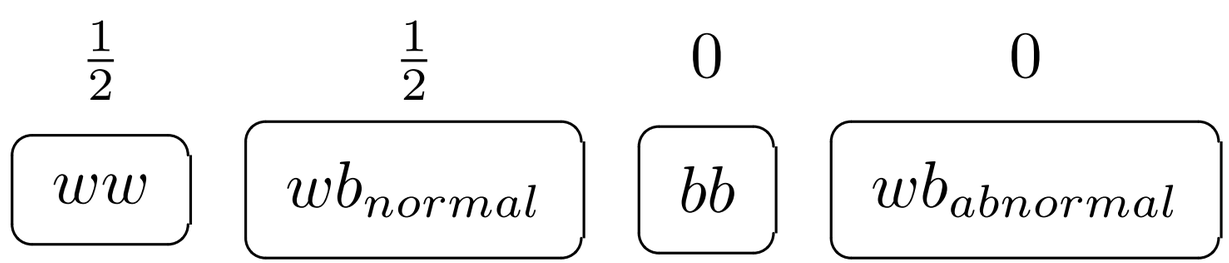
## Representation

Here, $t = 0.80$ and the real world is $ww$. The initial model:

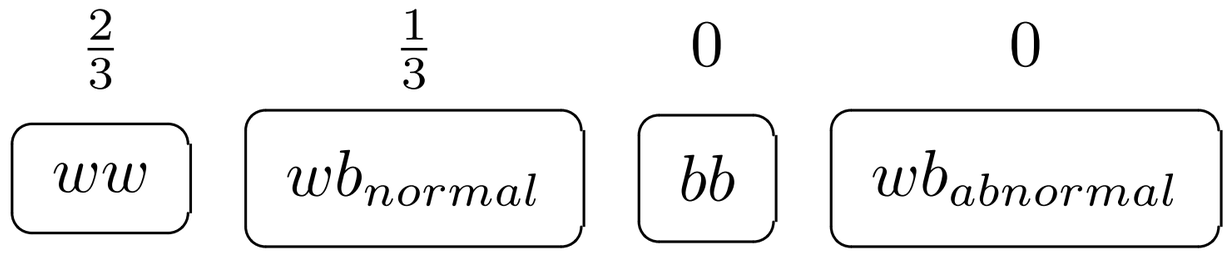



In world $wb_{normal}$ urn $wb$ was selected and I will eventually see a black ball; world $wb_{abnormal}$ is similar, except that I will forever happen to pick the white ball.

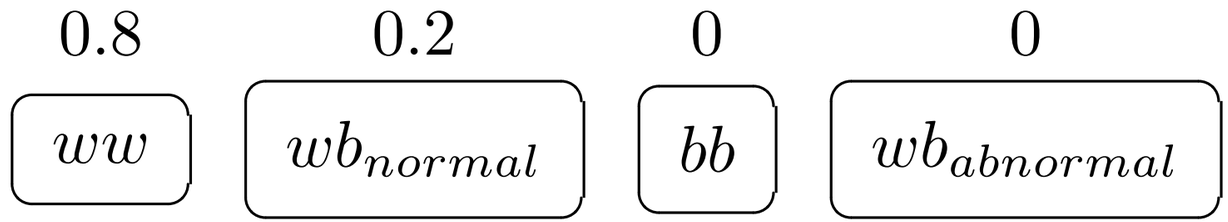
At this stage, I only know what I am certain of: $K\neg wb_{abnormal}$. After the first extraction (event $w_1$, in which I see a white ball), the new probability $P_{w_1}$ is:



After one more extraction (event $w_2$), the new probability $P_{w_1,w_2}$ is:




After a third extraction:




I reached my threshold!

Given the evidence $E = (w_1, w_2, w_3, w_4)$, I (correctly) believe that urn 1 is in the urn: if we put $H = \{(ww)\}$, then we have $B_E(H)$.

**Knowledge!**

But in fact I **know** it! Not with certainty, of course, but I know it!

Moreover, I **sensitively know it**: no need for any more justification!

$H$ is inherently undefeated: $U(H) = \{ww, wb_{abnormal}\}$, so $H \subseteq U(H)$, hence $P(U(H)|H) = 1$. I am certain that, if $H$ is true then I will continue to believe $E$ no matter what new evidence I may find.

Moreover, we have negative tracking with certainty, since

$$P(U(H) - H) = P(wb_{abnormal}) = 0$$

so that we have

$$P(\neg U(H)|\neg H) = 1.$$

I am certain that, if $H$ is false then I will eventually stop believing it (when seeing a black ball).

## More General Case

In the above we implicitly assumed (like Leitgeb) that

$$\mathcal{E} = \{A \in \mathcal{A} | P(A) \neq 0\}$$

But this doesn't work in most case: e.g. suppose that one of the urns has 50% white balls and 50% black balls, the second urn has 70% white and 30% black, and the third urn has 70% black and 30% white.

Will we ever come to know the true state of the urn, without opening it?

Well, not if we keep the above choice of $\mathcal{E}$: this amounts to having a way too strong standard of indefeasibility!

But the above choice of $\mathcal{E}$ is WRONG in this case: *we can only observe the sequence of colors of the extracted balls.*

So

$$\mathcal{E} = \{(o_1, \ldots, o_n) | n \geq 0, o_i \in \{w, b\}\}$$

Given this set of defeaters, we WILL eventually come to "know" the state of the urn!

The Strong Law of Large Numbers, combined with the Central Limit Theorem, ensure that.

## The Logic of Knowledge

**Proposition**. The *complete logic of the knowledge* operator $K$ is **the modal logic** $S4.3$:

$$K(A \Rightarrow B) \Rightarrow (K(A) \Rightarrow K(B))$$

$$K(A) \Rightarrow A$$

$$K(A) \Rightarrow K(K(A))$$

$$K(K(A) \Rightarrow B) \vee K(K(B) \Rightarrow A)$$

together with the usual **axioms of Propositional Logic** and the rules of **Modus Ponens** and **Necessitation** (*From* $A$ *infer* $K(A)$).

## The "Feeling of Knowledge": Believing that You Know

An agent has a *strong belief* in $H$ if **she believes that she knows** $H$.

Intuitively, *this notion of belief looks to the agent just like knowledge.*

From a **subjective** point of view, the two notions are *indistinguishable*:

$$BK(A) \Leftrightarrow BK(K(A)).$$

## Characterization

The following are equivalent:

- $BK(H)$;
- $\neg K \neg K(H)$;
- $H$ has a consistent, inherently undefeated (but possibly false) justification;
- $H$ has a consistent, sensitive (but possibly false) justification.
- $Bel(H)$ holds, in Leitgeb's sense.

So this gives us an interpretation of Leitgeb's belief as "the feeling of knowledge":

$$Bel = BK.$$

This can be extended to (Leitgeb's) (strong) *conditional beliefs*:

$$Bel(H|E) \;\Leftrightarrow\; B(K_E(H)|E).$$

## First Objection (Kevin Kelly)

K. Kelly came up with an objection against (the initial version of) Leitgeb's stability theory of belief, that could also be addressed against this version of defeasible knowledge.

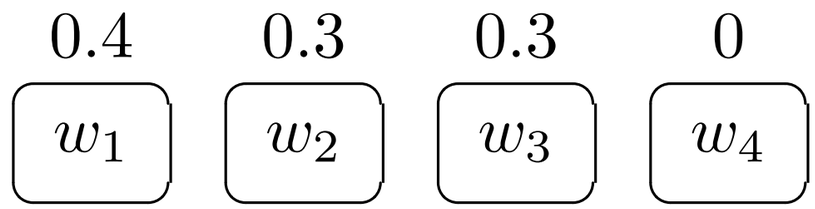
The problem is that, according to the logical accounts of knowledge change, the usual way of updating knowledge with new evidence is by **expansion**:
simply add the new evidence to all that was previously known, and close under logical entailment.

This means that $A$ is known after observing evidence $E$ iff $(E \Rightarrow P)$ was known before receiving the evidence:

$$K_E(A) \Leftrightarrow K(E \Rightarrow A)$$

But this will **not** match with Bayesian conditioning!

## Counterexample




with $t = \frac{4}{7}$.

Let $E = \{w_1, w_2, w_4\}$, $H = \{w_1, w_3\}$.

Suppose $w_3$ is the real world. Then we have

$$w_3 \models K_E H$$

but

$$w_3 \not\models K(E \Rightarrow H).$$

**Answer**

I used to think this was a problem, but not anymore.

It is only a problem for the AGM belief-revision theory.

According to my current view, probabilistic knowledge **should** track Bayesian conditioning:
this is exactly what allowed us to come to "know" universal statements after finitely many tests/

But this just means that knowledge update is simply different from expansion!

Something can come to be known (probabilistically) WITHOUT being entailed by the evidence plus things that were known before.

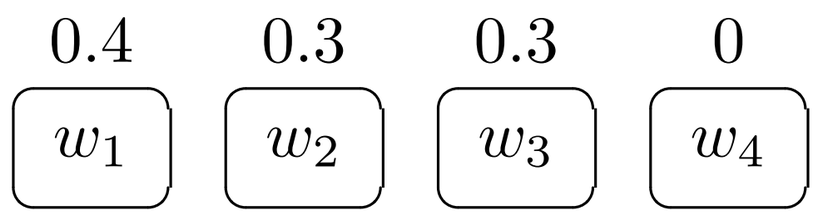
This also means that firm-belief revision does NOT satisfy the AGM postulates.

## Challenge for Logicians

Find the complete dynamic logic for (probabilistic) belief revision!

Funny enough, it is not Rational Monotonicity that fails: it's the OR rule, and the Sure Thing Principle.

Same model




with $t = \frac{4}{7}$, $E = \{w_1, w_2, w_4\}$, $H = \{w_1, w_3\}$. Then

$$Bel(w_1 \vee w_3 | w_3),$$

$$Bel(w_1 \vee w_3 | \neg w_3),$$

but

$$\neg Bel(w_1 \vee w_3).$$

## What is the dynamic logic of (probabilistic) knowledge updates?

This problem does NOT occur for knowledge, since knowledge has to be based only on (true) evidence, and in no world are both $E$ and $\neg E$ true. Hence, there is NO world satisfying both $K_E(H)$ and $K_{\neg E}(H)$: we have that

$$K_E(H) \Rightarrow \neg K_{\neg E}(H)$$

is a tautology, from which it follows that the Sure Thing Principle

$$K_E(H) \wedge K_{\neg E}(H) \Rightarrow K(H)$$

is vacuously true for $K$.

Still, what is the complete logic of (probabilistic, defeasible) knowledge updates?

## Puzzle Resolved!

It may look strange that the Sure Thing Principle fails for some reasonable notion of belief/acceptance.

But recall that *Bel* is nothing but "believing that you know" $BK$. Both $B$ and $K$ satisfy the OR rule, and so the Sure Thing Principle (though for different reasons).

However, one cannot use this to derive the Sure Thing Principle for $BK$, because of the failure of Conjunctivity for Lockean belief $B$.

## Objection 2: Context Sensitivity

The notions of "stability" and "being undefeated" (and hence also "being known") etc depend on the set $\mathcal{E}$ of observable evidence. If we go to a *more refined algebra*, in such a way that the old probabilities (of events in the old algebra) are *left the same*, and if we insist on identifying $\mathcal{E}$ with the family of all non-null events in the new algebra, them it can happen that something that was previously "known" is now defeated (by a new defeater, that only exists in the new algebra).

So **"knowledge" is not preserved when we change the context, by taking into account events that were previously disregarded**!

## ANSWER TO OBJECTION 2

This context sensitivity is a bug, not a feature!

In fact, it answers some of the objections against the Defeasibility Theory of Knowledge!

The choice of $\mathcal{E}$ is important": **only (and all)** the potential evidence that is **non-misleading and relevant** to the agent's question(s) should be included.

Every piece of fallible knowledge can be defeated, by some type of "misleading" evidence. But these kind of defeaters have to be excluded from $\mathcal{E}$.

So $\mathcal{E}$ can be "extended" only if the agent has previously missed relevant potential evidence! In which case, "knowledge" based on such wrong choice of $\mathcal{E}$ is not knowledge at all...