Learning-Theoretic Models for Temporal Doxastic Logic: With An Application to the Surprise Exam Paradox

Hanti Lin

March 16, 2012

Abstract

The present paper introduces learning-theoretic models for temporal doxastic logic, which represent how belief responses to information in the flow of time. The assumption about how belief should change is extremely weak, allowing us to characterize what kinds of belief revision procedures can do what kinds of jobs. An application is given to the surprise exam paradox, which establishes three results. (1) It is impossible to believe that there is a surprise exam, if one requires the preservation axiom in the AGM theory of belief revision. (2) It is possible to believe the surprise proposition, if one only requires that belief revision procedures should satisfy the well-known system P for nonmonotonic logic. (3) It is impossible to believe that one knows the surprise proposition, if the student's class only has three days and knowledge implies stable belief.

1 Introduction

The present paper introduces learning-theoretic models for temporal doxastic logic, which represent how belief responses to information in the flow of time. Each of the models is obtained, roughly speaking, by taking the Cartesian product of a temporal dimension and a Kripke model for information and belief. The assumption about belief revision is very weak: one's belief is always consistent, one always believes the information, and one's belief is a function of information. As a consequence, the models form a Kripke semantic framework for formal learning theory.[1] That enables us to characterize what kinds of belief revision procedures can do what kinds of jobs, and that is important for the main application in the present paper: the surprise exam paradox.

What concerns me in the surprise exam paradox is not how the student should accommodate the teacher's announcement that there will be a surprise example. There

[1]For an exposition of formal learning theory that concerns issues in philosophy of science, see Kelly (1996).

are separate, equally fundamental questions, regardless of what the teacher says or whether she says anything.

Q1 Is it possible for the student to believe that there will be a surprise exam?

My answer is yes, only if the student violates the preservation axiom in the AGM theory of belief revision.[2] Then:

Q2 Is it possible for the student to believe the existence of the surprise exam with a reasonable belief revision procedure?

My answer is yes, and one candidate belief revision procedure is the well-known system P in nonmonotonic logic,[3] which corresponds to the now-standard, Adams' conditional logic.[4] Furthermore:

Q3 Is it possible that the student believes that she knows the existence of the surprise exam?

My answer is no, assuming that the student's class only has three days and that, following Plato's *Meno*, knowledge entails stable belief.

I will focus only on the models, leaving aside the language for temporal doxastic logic, because I want to get to the application as quickly as possible. The proofs are given in the appendix.

2 Information Frame

Let H be a set of *possible histories*, T a totally order set of *times*. Let $H \times T$ be the Cartesian product of H and T, each of whose elements (h, t) is called the *moment* of history h at time t. A *proposition* A is a set of moments, understood as expressing that the true, current moment is in A. An *information dynamics* I is a function that assigns to each moment (h, t) a proposition $I(h, t)$ that represents the agent's total information at (h, t). Call $I(h, t)$ the *information state* that the agent has at moment (h, t). The agent is informed at moment (h, t) that A is true iff $I(h, t) \subseteq A$. A triple of the form $(H \times T, I)$ is called an *information frame*.

For an example, suppose that today is Sunday (t_0), and the agent is now thinking about the exam to be given in a class. The class has three days in a week: Monday (t_1), Tuesday (t_2), and Wednesday (t_3). Suppose, further, that the student knows for sure on Sunday that the class will have one and only one exam in the coming week, but not sure

[2]The axioms are given in Harper (1975), whose representation theorem is given in Alchourrón, Gärdenfors, and Makinson (1985).

[3]Kraus, Lehmann, Magidor (1990).

[4]Adams (1975).

about when exactly. So there are three possible histories: the exam occurs on Monday (h_1), on Tuesday (h_2), or on Wednesday (h_3). The information frame that represents the scenario is depicted in figure 1. The dark grey dots (h_i, t_i) are the possible moments

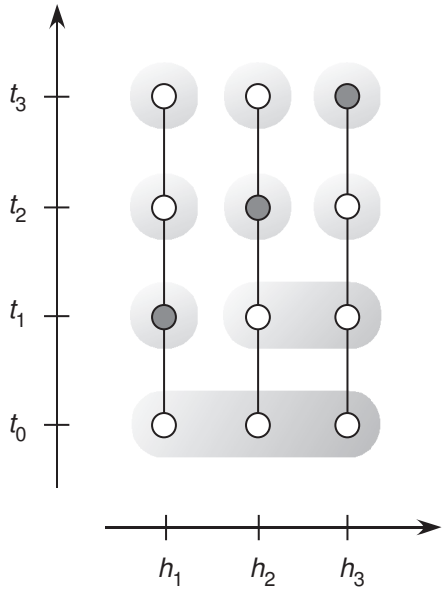

Figure 1: The information frame for the exam scenario

at which the exam occurs. When the student is at moment (h_i, t_j), her information state $I(h_i, t_j)$ is the proposition represented by the "grey cloud" that surrounds (h_i, t_j). For example, suppose that the student is now at the moment (h_2, t_1), i.e., the day before the Tuesday exam. At that moment, her total information is $I(h_2, t_1) = \{(h_2, t_1), (h_3, t_1)\}$. Hence, she has complete information about the current time t_1, because her information state entails that the current time is t_1. Her information state excludes history h_1 and is compatible with both h_2 and h_3, because she knows that today is Monday, observes that there is no exam today, and infers that the exam day is either Tuesday or Wednesday.

3 Learning Method

Relative to an information frame $(H{\times}T, I)$, a *belief dynamics* B assigns to each moment (h, t) a proposition $B(h, t)$ that represents the agent's belief state. So, at moment (h, t), the agent believes proposition A iff $B(h, t) \subseteq A$. Following the standard Kripke semantics,

$$\text{the proposition "the agent believes } A\text{"} \ = \{(h, t) \in H{\times}T : B(h, t) \subseteq A\}.$$

A belief dynamics B is a *learning method* iff it satisfies the following extra conditions:

- belief is always consistent: $B(h,t) \neq \varnothing$;
- information is always believed: $B(h,t) \subseteq I(h,t)$;
- belief depends only on information: $I(h,t) = I(h',t') \implies B(h,t) = B(h',t')$.

An example is given in the next section.

4 Who Believes That There Will Be A Surprise Exam?

Consider the information frame of the exam scenario (figure 1), relative to which we constructs a learning method as follows. Since belief depends only on information, assigning belief states to moments reduces to assigning belief states to information states. In figure 2, each information state (i.e., each cloud) is assigned a belief state, which is represented by the nonempty circle contained in that cloud. Each circle is required to

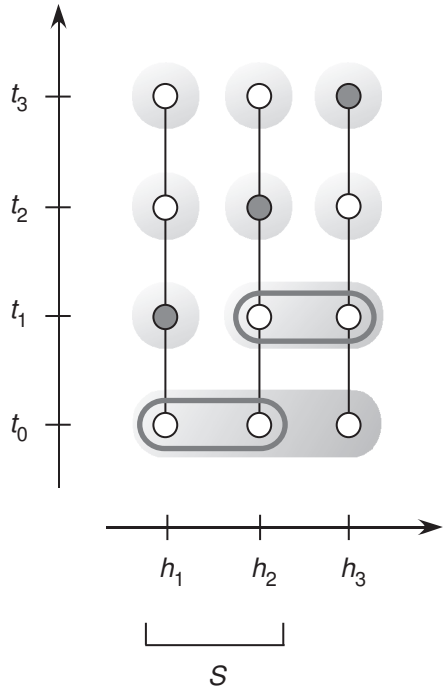

Figure 2: A learning method for the exam scenario

be contained in a cloud because the agent always believes the information. When an

information state is a singleton, the assigned belief state must be that singleton (by consistency of belief), so it will not be drawn for the sake to simplicity. So figure 2 completely specified a learning method. It is only one of the totally 21 learning methods for the information frame of the exam scenario.[5]

With an appropriate definition of surprise, the student represented by figure 2 believes *on Sunday* that there is a surprise exam. A formal treatment is given in the next section. Here is quick, informal exposition. Let S be the proposition that there is a surprise exam. Note that S should be a timeless proposition, i.e., if it is true at a moment, it is true at each moment with the same history. So we can simply talk about the histories in which S is true. S is true in history h_i iff $B(h_i, t_{i-1}) \not\subseteq \{h_i\}\times T$, namely, at the moment (h_i, t_{i-1}) right before the exam moment (h_i, t_i), the student does not believe that the exam will occur at the next moment. So, in figure 2, $S = \{h_1, h_2\}\times T$. Hence, on Sunday (t_0), the student believes S.

5 The Exam Scenario Formally Defined

A formal treatment of the exam scenario is given as follows. Suppose that the class has n days in a week and that, on Sunday, the subject knows for sure that there is exactly one exam in this week. Let the set of times be $T = \{t_0, t_1, \ldots, t_n\}$, where t_0 denotes Sunday and t_i denotes the i-th day of the class for $i \geq 1$. So $t_i \leq t_j$ if and only if $i \leq j$. Let the set of histories be $H = \{h_1, \ldots, h_n\}$, where h_i denotes the history in which the exam takes place on the i-th day of the class. The proposition E that *there is an exam today* is represented by the diagonal set of moments:

$$E \quad = \quad \{(h_i, t_i) : i = 1, \ldots, n\}.$$

When it is day t_j and no exam has occurred, the agent has information that the exam must occur in day $t_{j+1}, \ldots,$ or t_n, and thus the actual history must be $h_{j+1}, \ldots,$ or h_n. In general, the information states are defined as follows.

$$I(h_i, t_j) \quad = \quad \begin{cases} \{h_i\} \times \{t_j\}, & \text{if } i \leq j; \\ \{h_{j+1}, \ldots, h_n\} \times \{t_j\}, & \text{otherwise.} \end{cases}$$

The proposition S that *there is a surprise exam in this week* can only be defined when one specifies a belief dynamics, because one's surprise about an exam depends on one's belief on the day before the exam day. Relative to each learning method B,

[5]The number 21 is calculated as follows: we have exactly one 2-element cloud that admits of $2^2 - 1 = 3$ nonempty inner circles, and exactly one 3-element cloud that admits of $2^3 - 1 = 7$ nonempty inner circles, and all the other clouds are singletons; so there are totally $3 \times 7 = 21$ ways to specify a learning method.

let S_B denote the proposition that there is a surprise exam. So S_B is true in history h_i if and only if, at the moment (h_i, t_{i-1}) right before the exam day, the agent does not believe that the exam will occur tomorrow. In symbols:

$$S_B \quad = \quad \{(h_i, t_j) : B(h_i, t_{i-1}) \not\subseteq \{h_i\} \times T\}. \tag{1}$$

We will be interested in whether the agent believes on Sunday that there is a surprise exam in this week, i.e. whether the following is the case:

$$B(h_i, t_0) \quad \subseteq \quad S_B.$$

The subscript B in S_B will be omitted when there is no danger of confusion. The learning method in figure 2 witnesses the following possibility result:

Proposition 1. *Let the class have three days in a week, i.e. $n = 3$. Then there exists a learning method B such that the agent with B believes on Sunday that there is a surprise exam in this week; namely, for each history h_i,*

$$B(h_i, t_0) \quad \subseteq \quad S_B.$$

In general:

Corollary 1. *The above proposition holds for each finite $n \geq 3$.*

6 Who Cannot Believe That There Will Be A Surprise Exam?

One feature of the learning method in figure 2 is that it violates one of the AGM axioms for belief revision, called *preservation*. The present section shows that it is no accident.

The preservation axiom says: *if the new information is compatible with your old beliefs (about the actual history is), then retain those old beliefs in the new belief state.* That idea can be formulated in the present framework as follows. Since we are only interested in the cases where agent has complete information about the current time, the only important contents of information and belief are about the possible histories, i.e., about the *projections* to the history dimension. Let $\bar{A}$ be the projection of proposition A to the history dimension:

$$\bar{A} \quad = \quad \{h : \exists t \in T, (h, t) \in A\}.$$

Say that learning method B is *preservative* if and only if:[6]

$$\bar{I}(h_i, t_{j+1}) \cap \bar{B}(h_i, t_j) \neq \varnothing \quad \Longrightarrow \quad \bar{B}(h_i, t_{j+1}) \subseteq \bar{B}(h_i, t_j). \tag{2}$$

Then we have:

Proposition 2. *Whenever the agent adopts a preservative learning method B, she does not believe on Sunday that there is a surprise exam in this week; namely, for each history h_i,*

$$B(h_i, t_0) \quad \not\subseteq \quad S_B.$$

The above proposition suggests that if one can rationally believe that there is a surprise exam, then one can rationally violate the preservation axiom. But the result itself leaves open whether we should apply modus ponens or modus tollens, and I will not make a final decision in the present paper. It is worth noting that, after abandoning the preservation axiom, we still have a well-controlled belief procedure, namely the system P for nonmonotonic logic (Kraus, Lehmann, Magidor 1990). The learning method in figure 2 have the agent believe in the surprise exam, and it does satisfy each of the axioms in system P if we express the axioms the way we express preservation in formula (2).

7 Impossibility of Believing That One Knows

This section establishes an impossibility result for believing that one knows. Following the ideas in Plato's *Meno*, knowledge seems to entail stable belief. Define stable belief as follows. Let A be a timeless proposition, i.e. it can be expressed as the Cartesian product of a subset of histories and the set T of all times. Say that the agent *stably believes* A from moment (h_i, t_j) if and only if, for each $j' \geq j$, she believes that A at moment $(h_i, t_{j'})$. Then we have:

Proposition 3. *Suppose that the class has exactly three days in a week, i.e., $n = 3$. Then no learning method have the student believe on Sunday that she stably believes from Sunday that there will be a surprise exam.*

Corollary 2. *Continuing from the preceding proposition. If knowledge entails stable belief, then no learning method have the student believe on Sunday that she knows on Sunday that there will be a surprise exam.*

[6]Note that the antecedent concerns the compatibility between the new information and the old belief only about histories, not about times. Without restricting to the histories, the new information and the old belief themselves are always incompatible/disjoint. That is because today I get up with the new information state that contains only the moments on Monday, but my old belief state yesterday contains only the moments on Sunday.

References

Adams, E.W. (1975) *The Logic of Conditionals*, Dordrecht: D. Reidel.

Alchourrón, C.E., P. Gärdenfors, and D. Makinson (1985), "On the Logic of Theory Change: Partial Meet Contraction and Revision Functions", *The Journal of Symbolic Logic*, 50: 510-530.

Harper, W. (1975) "Rational Belief Change, Popper Functions and Counterfactuals", *Synthese*, 30(1-2): 221-262.

Kelly, K. (1996) *The Logic of Reliable Inquiry*, Oxford: Oxford University Press.

Kraus, S., Lehmann, D. and Magidor, M. (1990) "Nonmonotonic Reasoning, Preferential Models and Cumulative Logics", *Artificial Intelligence* 44: 167-207.

A Proofs

Proof of Proposition 1. Construct the learning method in figure 2. Behold! □

Proof of Corollary 1. Construct a learning method that includes the learning method in figure 2 as a sub-method. □

Proof of Proposition 2. Consider history h_n, in which the exam is held on the last day. Let t_j be the earliest day on which the new information contradicts the old belief state, i.e.:

$$\bar{I}(h_n, t_j) \cap \bar{B}(h_n, t_{j-1}) \;=\; \varnothing. \tag{3}$$

But:

$$\bar{I}(h_n, t_j) \;=\; \{h_{j+1}, \ldots, h_n\}; \tag{4}$$

$$\bar{I}(h_n, t_{j-1}) \;=\; \{h_j, h_{j+1}, \ldots, h_n\} \supseteq \bar{B}(h_n, t_{j-1}). \tag{5}$$

By (3), (4), and (5), we have:

$$\bar{B}(h_n, t_{j-1}) \;=\; \{h_j\}, \tag{6}$$

which implies that:

$$\bar{B}(h_j, t_{j-1}) \;=\; \{h_j\}, \tag{7}$$

because moments $(h_n, t_{j-1}), (h_j, t_{j-1})$ shares the same information state and, thus, share the same belief state. But (7) and the definition (1) of $\bar{S}_B$ implies that:

$$h_j \quad \notin \quad \bar{S}_B. \tag{8}$$

So, to prove the proposition, it suffices to how that for each history h_i,

$$h_j \quad \in \quad \bar{B}(h_i, t_0), \tag{9}$$

which is established as follows. By the construction of day t_j, there is no new information that contradicts an old belief state before day t_j in history h_n. So we have that:

$$\bar{B}(h_n, t_{j-1}) \quad \subseteq \quad \bar{B}(h_n, t_0), \tag{10}$$

by applying preservation $j-1$ times along history h_n. Then, by (6) and (10), we have that:

$$h_j \quad \in \quad \bar{B}(h_n, t_0), \tag{11}$$

which implies (9) because all Sunday moments share the same information state and, thus, share the same belief state. □

The proof of proposition 3 proceeds by the following lemmas.

Lemma 1. *Let the class have* 3 *days in a week, i.e.* $n = 3$. *Then one believes on Sunday that there will be a surprise exam if and only if her learning method is one of the four in figures 3 and 4. Namely, those four methods exhaust the methods* B *such that for every* $i \leq 3$,

$$B(h_i, t_0) \quad \subseteq \quad S_B.$$

Proof. Keep in mind that history h_3 must have no surprise exam. First, determine the range of the belief states that can be assigned to the two-element information state. The assigned belief state must be one of the two that appear figures 3 and 4. That is because if the two-element information state is assigned the singleton belief state $\{(h_2, t_1)\}$, then both histories h_2 and h_3 would have no surprise exam and, thus, the only way to believe on Sunday in a surprise exam is to believe proposition $\{h_1\}$, which makes h_1 have no surprise exam, and hence every history has no surprise exam—contradiction. Second, determine the range of the belief states that can be assigned to the three-element information state. The assigned belief state must be one of the two that appear figures 3 and 4. That is because the assigned belief state must exclude (h_3, t_0) and must not be identical to the singleton $\{h_1, t_0\}$, in order to believe in a surprise exam at all. Now,

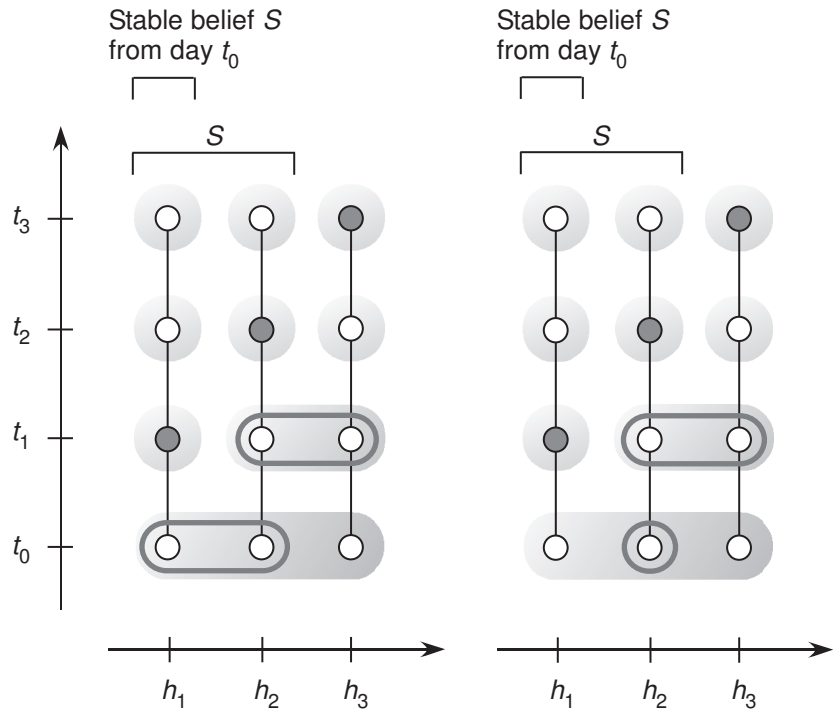

Figure 3: Learning methods for a Sunday belief that there is a surprise exam

we have established that there are two non-singleton information state and each of them can be assigned two belief states, which gives rise to the four learning methods in figures 3 and 4. Then it is routine to verify that all those four learning methods have the agent believe in a surprise exam on Sunday. □

Lemma 2. *Continuing from the above lemma. For all the four learning methods B in figures 3 and 4, the agent does not believe on Sunday that she stably believes from Sunday that S_B is true.*

Proof. The proposition that the agent stably believes S_B from Sunday t_0 is depicted for each of the four belief dynamics. Then, behold! □

Proof of Proposition 3. Let B be a learning method over the frame in figure **??**. There are two exhaustive cases. Case 1: the agent does not believe at moment (h_i, t_0) that S_B is true. Then, since the beliefs determined by a learning method is introspective, the agent believes at (h_i, t_0) that she does not believe t_0 that S_B is true. Then, since beliefs are assumed to be consistent, the agent does not believe at (h_i, t_0) that she believes at t_0 that S_B is true. Then, since knowledge entails belief, the agent does not believe

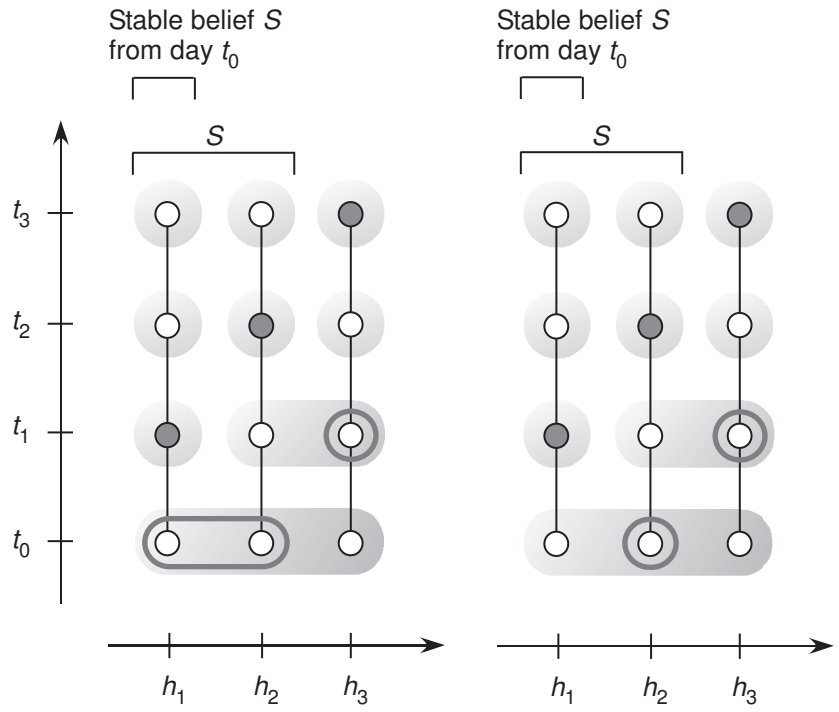

Figure 4: Learning methods for a Sunday belief that there is a surprise exam

at (h_i, t_0) that she knows at t_0 that S_B is true. Case 2: on Sunday the agent with B believes S_B. Then, by lemma 1, B is one of the four learning methods in figures 3 and 4, and thus lemma 2 applies. □