Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning
Main Track

The Topology of Surprise

Alexandru Baltag[1] , **Nick Bezhanishivili**[1] , **David Fernández-Duque**[2,3]
[1]ILLC, University of Amsterdam
[2]ICS of the Czech Academy of Sciences
[3]Department of Mathematics WE16, Ghent University
{A.Baltag,N.Bezhanishvili}@uva.nl, fernandez@cs.cas.cz

Abstract

In this paper we present a topological epistemic logic, with modalities for knowledge (modeled as the universal modality), knowability (represented by the topological interior operator), and unknowability of the actual world. The last notion has a non-self-referential reading (modeled by Cantor derivative: the set of limit points of a given set) and a self-referential one (modeled by Cantor's perfect core of a given set: its largest subset without isolated points). We completely axiomatize this logic, showing that it is decidable and PSPACE-complete, and we apply it to the analysis of a famous epistemic puzzle: the Surprise Exam Paradox.

1 Introduction

Epistemic logic has been formalized by Hintikka within the framework of possible-world semantics in relational (Kripke) models, and later rediscovered by game theorists (Aumann 1995) in the setting of partitional models (corresponding to the special case of S5 Kripke models, based on equivalence relations). In these forms, it has been used in Computer Science to reason about distributed systems, in AI to reason about agent-based knowledge representation, and in Philosophy to explore issues in formal epistemology.

An alternative interpretation of modal logic is based not on Kripke frames, but on topological spaces. This semantics can be traced back to McKinsey and Tarski (McKinsey and Tarski 1944). When the modal $\Diamond$ is interpreted as topological closure Cl and the modal $\Box$ as topological interior Int, one obtains a semantics for the modal logic S4. McKinsey and Tarski also suggested a second topological semantics, obtained by interpreting the modal $\Diamond$ as Cantor derivative (where recall that the derivative $d(A)$ of a set A consists of all limit points of A). The modal logic of Cantor derivative is semantically more expressive than the modal logic of the interior/closure operators: the latter can be defined in terms of derivative, but not vice-versa.

Since then, the usefulness of topological structures in Computer Science and Knowledge Representation has been well established. As noticed by Vickers (Vickers 1989) and Abramsky (Abramsky 1991), the notion of observability and its logic require a topological setting. Abstract notions of computability also involve topological structures, with a famous example being the Scott topology. Research on spatial reasoning, in both topological and metric incarnations, is also of significant interest for AI. More recently, developments in Formal Learning Theory (Kelly 1996; Brecht and Yamamoto 2010; Baltag, Gierasimczuk, and Smets 2015) and Distributed Computing (Goubault, Ledent, and Rajsbaum 2020) have taken a topological turn. Moreover, recent work in epistemic logic (Baltag et al. 2016; Özgün 2017; Baltag et al. 2019b) on modelling and reasoning about evidence and knowability uses topological structures.

These applications are based mostly on the notion of topological interior. Our paper builds on this existing work, but is the first to show the usefulness for Knowledge Representation of other topological notions, such as Cantor derivative. From a technical point of view, our formalism is obtained by adding to the logic of Cantor derivative a global modality (quantifying over all points), an operator capturing the perfect core $d^{\infty}(A)$ of a set A (defined as the largest subset of A that is equal to its own derivative) and a dynamic update modality (that goes from the original space to some definable subspace). Building on our previous work on topological μ-calculus (Baltag, Bezhanishvili, and Fernández-Duque 2021), we give a complete axiomatization, as well as decidability and complexity results. Our proof is natural and not difficult to grasp, due in large part to subtle technical innovations which allow for a much more direct approach than that of related results in the literature (see e.g. (Fernández-Duque 2011; Goldblatt and Hodkinson 2017)).

From a conceptual point of view, the key contribution of our paper is that we develop a logic of evidence-based knowledge, knowability, and (un)knowability of the actual world; and moreover, we apply it to the analysis of a famous epistemic paradox: the Surprise Examination paradox.

We start by adopting the learning-theoretic reading of topology (Kelly 1996; Baltag, Gierasimczuk, and Smets 2015; Baltag et al. 2019b), in which a topological space is a way to represent the actual and potential evidence that some (anonymous) agent may observe. The points of the space represent possible worlds (or possible states of the world): all the possibilities that are consistent with the agent's current knowledge. A proposition is known if it is true in all possible worlds. The potential evidence (that might be observed in the future) forms a topological basis $\mathcal{B}$: if a world x belongs to a basic open set $x \in U \in \mathcal{B}$, then the agent may observe proposition U in world x. The topological interior operator $\text{Int}(A)$ captures the *knowability* of proposition A through observations $U \in \mathcal{B}$. When the agent gains more knowledge, some possibilities are eliminated (being ruled out by the new information), and thus the space shrinks to a subspace: this corresponds to performing a knowledge update (described in our logic by dynamic update modalities).

While each of the above epistemic readings of standard topological notions are already known from the literature, the epistemic meaning of Cantor's derivative and the perfect core is not so obvious. In this paper, we propose a novel and very natural learning-theoretic interpretation of derivative. Essentially, the derivative $d(A)$ is the proposition asserting that *the actual world is unknowable (through observations), even if given (the additional information)* A.[1] Finally, the epistemic meaning of the perfect core $d^{\infty}(A)$ can be reconstructed from its fixed-point definition: essentially, $d^{\infty}(A)$ is the *self-referential version of Cantor's derivative*, i.e. the proposition asserting that "A is true, but the actual world is unknowable even given *this* information" (where 'this' refers to the very proposition that we are defining).

The main motivation for the introduction of the perfect core modality comes from our analysis of the Surprise Exam Paradox. The story says that a Student knows for sure that the date of the exam has been fixed in one of the five (working) days of next week. But he doesn't know in which day. The Teacher (who is known to be always truthful) announces that the *exam's date will be a surprise*: even in the evening before the exam, the Student will still not know for sure that the exam is tomorrow. Intuitively, the Student can then prove (by backward induction, starting with Friday) that *the exam cannot take place in any day of the week*: first, if the exam would be on Friday, then it wouldn't be a surprise (-since Friday is the last possible day, by Thursday evening the Student would know it); since the Teacher (truthfully) announced that the exam will be a surprise, it follows that the exam will not take place on Friday. But once Friday is eliminated, the Student can repeat the same argument, until all days are eliminated. But this is a contradiction (since we assumed the Student *knows* there will be an exam).

In some versions of the puzzle, there is an even more "paradoxical" follow-up: the assumption that the Teacher never lies is weakened, to allow the Student some way out. After deriving the above contradiction, he concludes that the Teacher lied: the exam will *not* be a surprise. Confident that, whenever the exams comes, he will somehow get to know it an evening in advance (and thus be able to study in that last evening), the Student parties every day. Then, when the exam comes (say, on Wednesday), it *will* indeed be a complete surprise! So the Teacher told the truth after all?!

[1]Indeed, prior to this paper, the dominant interpretation of derivative in the epistemological literature was Steinsvold's reading in terms of "belief" (Steinsvold 2007). That interpretation has been criticized as not correctly reflecting the intuitive properties of belief and its relations to knowledge (Baltag et al. 2019a). Though new, our interpretation is closer to an older work (Parikh 1992), based on a different framework: multi-agent $S5$ Kripke frames. In that setting, derivative is connected to ignorance (rather than to unknowability): the agent does not know the actual world.

In this paper, we give a full analysis of the paradox using our topological epistemic logic. We distinguish between non-self-referential interpretations of Teacher's announcement (which can be formalized using Cantor derivative) and self-referential interpretations (which are captured using the perfect core modality). The first interpretation was pursued (in a non-topological, and less transparent, setting) in (Gerbrandy 2007), and shown to be paradox-free: the only conclusion is that the exam cannot be on Friday, but the elimination process cannot be iterated. However, most logicians consider that the most natural (and intended) interpretation is the second, self-referential one. As in the above intuitive argumentation, this does lead to a contradiction. The correct conclusion is that a Teacher who is known to be always truthful *cannot make* such an announcement (since if she did, it would be a lie). In this, we agree with the verdict given in (Quine 1953). However, we also show that the above contradiction is only produced by the special evidential topology underlying the Surprise Exam Story. By changing the topology, we obtain "non-paradoxical" versions, in which the Teacher *can* truthfully make similar future-oriented self-referential "surprise" announcements. Our conclusion (against the opinions of many philosophical logicians) is that epistemic self-referentiality is *not* the cause of the apparent 'paradoxicality' of the Surprise Exam Paradox.

2 The Evidential Topology

As preliminaries, we recall here some notions from General Topology. In the view of our epistemic applications, we strengthen somewhat the standard notion of topological base, obtaining a concept that we call "strong base".

2.1 Topological Preliminaries

Definition 2.1 (Topology, strong base, open and closed sets, neighborhoods)**.** *A* ***strong (topological) base*** *on a set X (called a* space*, and whose elements $x \in X$ are called* points*) is a family $\mathcal{B} \subseteq \mathcal{P}(X)$ of subsets of X (called* basic open *sets), with the property that it is* closed under finite intersections*: if $\mathcal{U} \subseteq \mathcal{B}$ is any finite subfamily, then $\bigcap \mathcal{U} \in \mathcal{T}$. This is in fact equivalent to requiring that a base is closed only under binary intersections (if $U, V \in \mathcal{B}$, then $U \cap V \in \mathcal{B}$) and contains the whole space (i.e. $X \in \mathcal{B}$).*[2] *A* basic neighborhood *of a point $x \in X$ is a basic open set $U \in \mathcal{B}$ with $x \in U$.*

A ***topology*** *on a set X is a strong base $\mathcal{T} \subseteq \mathcal{P}(X)$, that satisfies the additional requirement that: it is* closed under arbitrary (possibly infinite) unions*: if $\mathcal{U} \subseteq \mathcal{T}$ is any subfamily, then $\bigcup \mathcal{U} \in \mathcal{T}$. The pair $(X, \mathcal{T})$ is a* ***topological space*** *and the sets $U \in \mathcal{T}$ are called* open *sets.*[3] *Their complements $X - U$ (with $U \in \mathcal{T}$) are called* closed*, and have dual closure properties to the opens. A* neighborhood *of a point $x \in X$ is an open set $U \in \mathcal{T}$ with $x \in U$.*

[2]This last condition follows from applying closure under finite intersections to the empty family $\mathcal{U} = \emptyset \subseteq \mathcal{B}$, since $\bigcap \emptyset = X$.

[3]By applying closure under unions to the empty family $\mathcal{U} = \emptyset$, it is easy to see that $\emptyset$ is open (as well as closed, being the complement $X - X$ of the open set X).

Operators in a Topological Space [Interior, closure, derivative] An *interior point* of a set $A \subseteq X$ is a point $x \in X$ s.t. there exists a neighborhood $U \in \mathcal{T}$ (of x) with $x \in U \subseteq A$. Given a strong basis $\mathcal{B}$ for the topology $\mathcal{T}$, it is easy to see that x is an interior point of A iff there exists a *basic* neighborhood $U \in \mathcal{B}$ (of x) s.t. $x \in U \subseteq A$. The ***interior*** $\mathrm{Int}(A)$ of a set $A \subseteq X$ is the set of all its interior points. A point $x \in X$ is *close* to a set $A \subseteq X$ if all its (basic) neighborhoods intersect A: for all $U \in \mathcal{T}$ (or equivalently, for all $U \in \mathcal{B}$) s.t. $x \in U$ we have $U \cap A \neq \emptyset$. The ***closure*** $\mathrm{Cl}(A)$ of the set A is the set of all points that are close to A. A *limit point* of a set $A \subseteq X$ is a point $x \in X$ s.t. every (basic) neighborhood U of x contains a point $y \in A$ with $y \neq x$; equivalently, x is a limit point of A iff $x \in \mathrm{Cl}(A - \{x\})$. The ***(Cantor) derivative*** $d(A)$ of a set A is the set of all the limit points of A. It is easy to see that $\mathrm{Cl}(A) = A \cup d(A)$. A non-limit point $x \in A - d(A)$ is called *isolated* in A.

It is important to note that operators Int, Cl and d are *monotonic* operators, e.g. in particular $A \subseteq B$ implies $d(A) \subseteq d(B)$.

Generated Topology The *topology generated by* a strong base $\mathcal{B} \subseteq \mathcal{P}(X)$ is the smallest topology $\mathcal{T} \subseteq \mathcal{P}(X)$ s.t. $\mathcal{B} \subseteq \mathcal{T}$. We then say that $\mathcal{B}$ is a *base for* $\mathcal{T}$. The generated topology can be explicitly characterized as consisting of all possible unions of basic opens: $\mathcal{T} = \{\bigcup \mathcal{U} : \mathcal{U} \subseteq \mathcal{B}\}$.

Subspace Topology Every subset $A \subseteq X$ of a topological space $(X, \mathcal{T})$ is a ***subspace*** of the original space, when endowed with the subspace topology $\mathcal{T}_A = \{A \cap U : U \in \mathcal{T}\}$. Every strong base $\mathcal{B}$ for $\mathcal{T}$ induces a corresponding strong base for $\mathcal{T}_A$, obtained by taking $\mathcal{B}_A = \{A \cap U : U \in \mathcal{B}\}$. All the above topological notions can be relativized to a subspace: e.g. for any subset $P \subseteq A$, we can define its relative interior $\mathrm{Int}_A(P)$ in A, closure $\mathrm{Cl}_A(P)$ in A and derivative $d_A(P)$ in A, by applying the above definitions in the subspace A. It is easy to see that $\mathrm{Int}_A(P) = A \cap \mathrm{Int}(P \cup (X - A))$, $\mathrm{Cl}_A(P) = A \cap \mathrm{Cl}(P)$, and $d_P(A) = A \cap d(P)$.

Perfect Sets and Perfect Core A set $A \subseteq X$ is said to be ***perfect*** if $A = d(A)$. The ***perfect core*** of a set A is a subset of A denoted by $d^\infty(A)$, and defined as the largest perfect subset of A.[4] The perfect core $d^\infty(A)$ is the largest fixed point of the relative derivative operator $d_A : \mathcal{P}(A) \to \mathcal{P}(A)$, that takes subsets $P \subseteq A$ into their relative derivative $d_A(P) = A \cap d(P)$ in A.[5] This fixed point exists (by the Knaster-Tarski fixed point theorem) because of the *monotonicity* of the relative derivative operator d_A (itself a consequence of the monotonicity of derivative and intersection). Using standard μ-calculus notation for this largest fixed point, we can thus write

$$d^\infty(A) = \nu P. A \cap d(P).$$

[4]Here, "largest" is used in the sense of inclusion: so the perfect core $d^\infty(A)$ is the unique set B satisfying the following three conditions: (1) $B \subseteq A$; (2) $B = d(B)$; (3) every set B' satisfying conditions (1) and (2) is included in B.

[5]Once again, "largest" is taken here in the sense of inclusion.

Cantor-Bendixson Rank For any set $A \subseteq X$, we define a transfinite sequence of subsets of A, by putting:

$$d^0(A) = A, \qquad d^{\alpha+1}(A) = d_A(d^\alpha(A)) = A \cap d^\alpha(A),$$

$$d^\lambda(A) = \bigcap_{\alpha<\lambda} d^\alpha(A) \text{ for limit ordinals } \lambda.$$

It is easy to check that this is a descending sequence

$$A = d^0(A) \supseteq d(A) = d^1(A) \supseteq \ldots \supseteq d^\alpha(A) \supseteq \ldots,$$

which thus must reach a *fixed point*; i.e. there must exist an ordinal α s.t. $d^{\alpha+1}(A) = d^\alpha(A)$. The smallest such ordinal is called the ***(Cantor-Bendixson) rank*** of A, denoted by $\mathrm{rank}(A)$. Moreover, *the fixed point of the above iterative process $d^\alpha(A)$ is the perfect core*:

$$d^{\mathrm{rank}(A)}(A) = d^\infty(A).$$

2.2 The Epistemic Interpretation of Topology

We proceed now to explain the intended epistemic interpretation of the above topological notions, in terms of observable evidence and information updates.

Possible Worlds, Knowledge, Observable Evidence, Evidential Topology We think of the points $x \in X$ as representing *possible worlds* (or possible states of the world): all the possibilities that are consistent with some (anonymous) agent's information. Only one of these points represents the *actual world* (the true state of affairs), but the agent may not know which one: all she knows for certain is that it belongs to the set X. Every subset $P \subseteq X$ represents a "proposition", which may be "true" (i.e., hold) in a given world or not. A proposition P is "known" for certain only if it is true in all possible worlds that are consistent with the agent's information, i.e. if $P = X$. A strong basis $\mathcal{B} \subseteq \mathcal{P}(X)$ represents our agent's *potential evidence*: the properties of the world that can in principle be *directly observed* by the agent. When $x \in U \in \mathcal{B}$, the agent may observe the truth of proposition U in world x. Note that only the observable properties that are *true* in a world x will be observed in x (i.e. we assume observations to be sound or "correct"). So in world x the observable evidence corresponds to basic neighborhoods of the point x. Note also that this is *not* yet "evidence in hand" (that the agent already possesses), but "evidence out there" (that might observed in the future). The two conditions that underlie our definition of strong basis have a clear epistemic meaning: closure under binary intersections says that our agent is *able to accumulate observations*: after observing propositions U and V, the agent will in effect have observed the truth of the conjunction $U \cap V$ (coming to know that $x \in U \cap V$); while the condition $X \in \mathcal{B}$ says that the agent can directly *observe the truth of a tautology*.

Knowability and Conditional Knowability Interior points $x \in \mathrm{Int}(P)$ represent worlds in which proposition P is *knowable* (or " verifiable") based on direct observations: P is true at x, and this fact can be known after some more evidence about x is observed. This interpretation follows directly from the definition: $x \in \mathrm{Int}(P)$ holds iff there exists some observable evidence that entails P (i.e. $U \in \mathcal{B}$ with

$x \in U \subseteq P$). So, as an epistemic proposition, $\mathrm{Int}(P)$ says that *proposition P can be known from observations*. More generally, the proposition $\mathrm{Int}(A \Rightarrow P) = \mathrm{Int}((X - A) \cup P)$ captures *conditional knowability*: it says that P *can be known (from observations) given* A.

Unknowability and Falsifiability The complement $X - \mathrm{Int}(P)$ thus corresponds to "unknowability" of P, while the closure $\mathrm{Cl}(P) = X - \mathrm{Int}(X - P)$ corresponds to unfalsifiability of P: $x \in \mathrm{Cl}(P)$ means that, no matter what more evidence about x will be observed, P will never be known to be false. Note though that *our notion of unknowability is not an absolute barrier to knowledge*: it only expresses the fact that P cannot be known by direct observations (of evidence observable by the agent). Such an 'unknowable' P may still become known based on information received from another source (e.g. another agent).

Verifiable and Falsifiable Propositions The open sets $U \in \mathcal{T}$ represent *(inherently) verifiable propositions*: the ones having the property that they are knowable/verifiable whenever they are true (cf. (Vickers 1989; Kelly 1996)). This interpretation is backed by the following equivalence:

$$P \in \mathcal{T} \text{ iff } P \subseteq \mathrm{Int}(P).$$

Similarly, the closed sets represent *(inherently) falsifiable propositions*: whenever they are false, they can become known to be false after some more evidence is observed.

Knowledge Updates The move from the original topology on X to the subspace topology on some subset $A \subseteq X$ corresponds to performing an *update* of the agent's knowledge base with the proposition A: the possible worlds not satisfying A are eliminated, so the agent comes to know A after that. The update can be the result of a direct observation $A \in \mathcal{B}$; but it can also be the result of some communication from some outside source of information (e.g. an announcement from some other agent), in which case it is quite possible that $A \notin \mathcal{B}$ (i.e. A is *not* observable by our agent). However, for this update-by-elimination to be justified, it is essential that our agent *knows for certain that the source of the new information is absolutely reliable* (e.g. the other, informing agent is telling the truth).[6] The relativized interior $\mathrm{Int}_A(P) = A \cap \mathrm{Int}(P \cup (X - A))$ in the subspace topology will then capture a notion of *updated knowability* (after updating with P, the agent can come to know A based on further observations).

Examples of Evidential Topologies

- *Complete ignorance*: the **trivial topology** $\mathcal{T} = \{\emptyset, X\}$ on a set X;
- *Omniscience (God's topology)*: the **discrete topology** $\mathcal{T} = \mathcal{P}(X) = \{Y | Y \subseteq X\}$ on X;
- Knowledge based on *measurements of a point on a line*: the **standard topology of real numbers** $X = \mathbb{R}$, with the topology $\mathcal{T}$ generated by the strong basis $\mathcal{B} = \{(a, b) : a, b \in \mathbb{Q}, a < b\}$ (open intervals with rational endpoints);
- Knowledge based on *measurements in space*: the **standard topology on** $\mathbb{R}^n$, with state space $X = \mathbb{R}^n$ and topology $\mathcal{T} = \{$ countable unions of rational open balls$\}$, i.e. $A \subseteq \mathbb{R}^n$ is open iff it is of the form $A = \bigcup_{i=1}^{\infty} \{x \in \mathbb{Q}^n | d(x, a_i) < b_i\}$, where $a_i, b_i \in \mathbb{Q}^n$ and d is the Euclidean distance in n-dimensional space $\mathbb{R}^n$. A strong basis for this topology consists of all finite intersections of rational open balls.

[6]When this is not the case, other forms of updating are to be considered (in which the non-A worlds are *not* eliminated, but only considered in some sense less plausible, or less probable, than the A-worlds).

Concrete Example: The Policeman and the Speeding Car A policeman may use radars with varying accuracy to determine whether a car is speeding in a 50 mph speed-limit zone. Then the *the set of possible worlds* is $X = (0, \infty)$ (since we assume the car is known to be *moving*). The strong base

$$\mathcal{B} = \{(a, b) : a, b \in \mathbb{Q}, 0 < a < b < \infty\}$$

consists of *all possible measurement results by arbitrarily accurate radars*. The topology $\mathcal{T}$ generated by $\mathcal{B}$ is the *standard topology on real numbers* (restricted to X). "*Speeding*" is the proposition $S = (50, \infty)$.

Suppose now that a radar with accuracy shows mph 51 ± 2 mph. This induces an *update*: the original space X shrinks to the subspace $A = (49, 53)$. In this updated space, "*Speeding*" becomes $S_A = (50, 53)$. Still, even now (in the subspace A, i.e. after the radar reading), the policeman does *not know* that the car is speeding (since $S_A \neq A$). However, the property "the car is speeding" *is in principle verifiable* (by the policeman): *if* the car is indeed speeding, then its velocity must be some $x \in S_A = (50, 53)$. Given a more accurate radar, the policeman can obtain a better measurement (a, b) with $x \in (a, b) \subseteq S_A$. This is reflected in the fact that $S_A = (50, 53)$ is *open* in the standard topology.

In contrast, *Not-Speeding* $NS = (0, 50]$ is *in general not verifiable* (not open). This means that *whether NS is knowable or not depends on the actual speed*! For instance, NS is knowable in the world in which the speed is $x = 49$. But it is *not* knowable in the world $x = 50$. On the other hand, not-speeding NS is in general falsifiable (closed in X): whenever it is false, it can be disproved by a sufficiently accurate measurement of the speed.

The Epistemic Interpretation of Cantor Derivative To understand the derivative, recall the equivalence:

$$x \in d(A) \text{ iff } x \in \mathrm{Cl}(A - \{x\}).$$

But note that $\mathrm{Cl}(A - \{x\}) = X - \mathrm{Int}(X - (A - \{x\})) = X - \mathrm{Int}((X - A) \cup \{x\}) = X - Int(A \Rightarrow \{x\})$. Using our interpretation of $X - P$ as negation of the proposition P, and of $\mathrm{Int}(A \Rightarrow P)$ as conditional knowability (of P given A), we conclude that

$$x \in d(A) \text{ iff } x \text{ is not knowable given } A.$$

So, as an epistemic proposition, Cantor's derivative $d(A)$ says that "*the actual world is unknowable given A*".

The Epistemic Meaning of the Perfect Core Looking now at the perfect core $d^{\infty}(A)$, we can infer its epistemic meaning from the above fixed-point identity:

$$d^{\infty}(A) = \nu P. A \cap d(P).$$

The perfect core can thus be understood as the *self-referential version of Cantor's derivative*: $d^\infty(A)$ captures the epistemic proposition "A is true, but the actual world is unknowable given *this* information" (where 'this' refers to the very proposition that we are defining). As we'll see, this is precisely the kind of self-referential statement that plays a key role in the Surprise Examination Paradox.

3 The Logic of Derivative and Perfect Core

In this section we introduce the formal syntax and semantics of our logic. We begin by defining the formal languages $\mathcal{L}_{\langle\cdot\rangle}$ and $\mathcal{L}$ we will work with:

Syntax. The language $\mathcal{L}_{\langle\cdot\rangle}$ of dynamic-epistemic logic of derivative and perfect core consists of formulas recursively defined by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid \varphi\wedge\varphi \mid \Diamond\varphi \mid \odot\varphi \mid \widehat{K}\varphi \mid \langle\varphi\rangle\varphi$$

The language $\mathcal{L}$ of (static) epistemic logic of derivative and perfect core is the fragment of $\mathcal{L}_{\langle\cdot\rangle}$ obtained by eliminating all dynamic modalities $\langle\varphi\rangle$.

Semantics. We interpret this language on **epistemic topo-models** $\mathbf{M} = (X, \mathcal{T}, \|\cdot\|)$: topological spaces $(X,\mathcal{T})$ with a valuation function (mapping every atomic sentence p into a subset $\|p\| \subseteq X$). The semantics is given by extending this valuation recursively to all of $\mathcal{L}_{\langle\cdot\rangle}$, defining $\|\varphi\|_{\mathbf{M}}$ using the usual clauses for Booleans, while

$$\|\Diamond\varphi\|_{\mathbf{M}} = d(\|\varphi\|_{\mathbf{M}})$$

is the Cantor derivative of $\|\varphi\|_{\mathbf{M}}$ wrt the topology $\mathcal{T}$, and

$$\|\odot\varphi\|_{\mathbf{M}} = d^\infty\|\varphi\|_{\mathbf{M}} = \nu P.\big(\|\varphi\|_{\mathbf{M}}\cap d(P)\big)$$

is the perfect core of $\|\varphi\|_{\mathbf{M}}$. The operator $\widehat{K}$ is just the global existential modality, quantifying existentially over all possible worlds: $\|\widehat{K}\varphi\|_{\mathbf{M}} = X$ if $\|\varphi\|_{\mathbf{M}} \neq \emptyset$, otherwise $\|\widehat{K}\varphi\|_{\mathbf{M}} = \varnothing$. Finally, $\langle\varphi\rangle\psi$ is the dynamic modality for epistemic updates, whose semantics is given by evaluating ψ in the updated model: if, for any subset $A\subseteq X$, we put $\mathbf{M} = (A, \mathcal{T}_A, \|\cdot\|_A)$ for the updated model, with the subspace topology $\mathcal{T}_A = \{U\cap A : U\in\mathcal{T}\}$ and relativized valuation $\|p\|_A = \|p\|\cap A$, then we set

$$\|\langle\varphi\rangle\psi\|_{\mathbf{M}} = \|\psi\|_{\|\varphi\|},$$

where $\|\varphi\| = \|\varphi\|_{\mathbf{M}}$ is the valuation of φ in the original model. As usual, we may write $(\mathbf{M},x)\models\varphi$ iff $x\in\|\varphi\|_{\mathbf{M}}$. When the model $\mathbf{M}$ is clear from the context, we may skip it, writing e.g. $\|\varphi\|$ and $x\models\varphi$.

In an epistemic context, we read $\widehat{K}$ as *epistemic possibility*: $\widehat{K}\varphi$ says that "as far our agent knows, φ *may* be true", in the sense that φ is consistent with the agent's information. We read $\Diamond\varphi$ as saying that "the actual world is unknowable (through observations) given φ"; we read $\odot\varphi$ as a self-referential statement, saying that "φ is true, but the actual world is unknowable (through observation) given *this* information" (where 'this information' refers to the very proposition we are defining); finally, we read $\langle\varphi\rangle\psi$ as saying that "φ holds, and ψ will also hold after updating with φ".

Abbreviations: We will use the standard abbreviations for propositional connectives $\varphi\vee\psi$, $\varphi\Rightarrow\psi$, $\varphi\Leftrightarrow\psi$, $\top$ and $\bot$, as well as the following additional ones: $\Box\varphi := \neg\Diamond\neg\varphi$, $K\varphi := \neg\widehat{K}\neg\varphi$, $\widehat{\mathcal{K}}\varphi := \varphi\vee\Diamond\varphi$, and $\mathcal{K}\varphi := \neg\widehat{\mathcal{K}}\neg\varphi$. To justify these notations, note that K is just the universal modality (quantifying universally over all worlds that are possible according to our agent), $\Diamond$ is just the closure modality and $\mathcal{K}$ is just the interior modality: $\|K\varphi\| = X$ iff $\|\varphi\| = X$, and $\|K\varphi\| = \emptyset$ otherwise; $\|\widehat{\mathcal{K}}\varphi\| = \mathrm{Cl}(\|\varphi\|)$; and $\|\mathcal{K}\varphi\| = \|\varphi\wedge\Box\varphi\| = \mathrm{Int}(\|\varphi\|)$. So, given our interpretation of possible worlds, closure and interior, we can read $K\varphi$ as "φ is known" (to our agent), $\mathcal{K}\varphi$ as "φ is knowable" (through observations by our agent), and read $\widehat{\mathcal{K}}\varphi$ as "φ cannot be falsified" (by any observations by the agent).

Theorem 3.1. *[**Completeness for** $\mathcal{L}_{\langle\cdot\rangle}$] The following system is a sound and complete axiomatization of the dynamic-epistemic logic of Cantor derivative and perfect core $\mathcal{L}_{\langle\cdot\rangle}$:*

- *Axioms and Rules of Propositional Logic.*
- *Necessitation Rule, and Distribution (=Kripke's Axiom), for the modalities K, $\Box$ and $[\varphi]$.*[7]
- *Positive and negative introspection for knowledge:*
$$K\varphi\Rightarrow KK\varphi \qquad \neg K\varphi\Rightarrow K\neg K\varphi$$
- *Positive Introspection of Knowability* (if φ is knowable, then it is knowable to be knowable)*:* $\mathcal{K}\varphi \Rightarrow \mathcal{K}\mathcal{K}\varphi$
- *Knowledge implies knowability:* $K\varphi\Rightarrow\mathcal{K}\varphi$
- *Monotonicity rule for the perfect core:* $\dfrac{\varphi\to\psi}{\odot\varphi\to\odot\psi}$
- *Fixed Point Axiom:* $\odot\varphi \Rightarrow (\varphi\wedge\Diamond\odot\varphi)$
- *Induction Axiom:* $\mathcal{K}(\varphi\Rightarrow\Diamond\varphi) \Rightarrow (\varphi\Rightarrow\odot\varphi)$
- *Reduction axioms for update modalities:*
$$\langle\varphi\rangle p \Leftrightarrow (\varphi\wedge p)$$
$$\langle\varphi\rangle\neg\theta \Leftrightarrow (\varphi\wedge\neg\langle\varphi\rangle\theta)$$
$$\langle\varphi\rangle\widehat{K}\theta \Leftrightarrow (\varphi\wedge\widehat{K}\langle\varphi\rangle\theta)$$
$$\langle\varphi\rangle\Diamond\theta \Leftrightarrow (\varphi\wedge\Diamond\langle\varphi\rangle\theta)$$
$$\langle\varphi\rangle\odot\theta \Leftrightarrow \odot\langle\varphi\rangle\theta$$

Proving soundness is an easy verification. Completeness follows immediately from putting together the following two results:

Theorem 3.2. *[**Provable Co-expressivity of** $\mathcal{L}_{\langle\cdot\rangle}$ **and** $\mathcal{L}$] Every formula in the language $\mathcal{L}_{\langle\cdot\rangle}$ is provably equivalent*[8] *to some formula in the static fragment $\mathcal{L}$. Hence, the two logics $\mathcal{L}_{\langle\cdot\rangle}$ and $\mathcal{L}$ have the same expressivity.*[9]

[7]In fact, Necessitation for $\Box$ follows from Necessitation for K and the axiom "Knowledge implies knowability".

[8]This means that the equivalence is provable in the above axiomatic system for $\mathcal{L}_{\langle\cdot\rangle}$.

[9]But they differ in succinctness: formulas in $\mathcal{L}_{\langle\cdot\rangle}$ can be in general exponentially more succinct than their translations in $\mathcal{L}$. In addition, they can capture the desired dynamic-epistemic scenarios in a much more transparent and direct way than their translations. This makes dynamic modalities very useful for applications, and justifies our choice of the larger language $\mathcal{L}_{\langle\cdot\rangle}$.

Theorem 3.3. ***[Completeness for $\mathcal{L}$]*** *The system obtained from the above axiomatic system for $\mathcal{L}_{\langle\cdot\rangle}$ by eliminating all axioms and rules that refer to dynamic modalities (specifically: eliminating Necessitation and Distribution for $[\varphi]$, as well as all the reduction axioms) is a sound and complete axiomatization of the static epistemic logic of Cantor derivative and perfect core $\mathcal{L}$.*

Proof Summary While the proof of Theorem 3.2 is an easy induction (using the reduction axioms to gradually push the dynamic modalities past other operators and then eliminate them), the proof of Theorem 3.3 is highly non-trivial, and uses methods that we developed in our recent work on topological μ-calculus (Baltag, Bezhanishvili, and Fernández-Duque 2021). Hence, we only give here a bird's eye overview of this proof. Essentially, we start from the canonical model Ω (comprising all maximally consistent theories accessible from some fixed theory), a standard construction in modal logic. But we should stress that Ω is *not* our intended model.[10] Indeed, the usual Truth Lemma fails for our logic $\mathcal{L}$ in the canonical model: formulas are not necessarily satisfied in Ω by the theories that contain them. Next, for any given finite set of formulas Σ, we select a special submodel of the canonical model Ω^Σ (called the Σ-final model), consists of "Σ-final theories": essentially, these are the ones whose cluster is locally definable by some formula in Σ. Our goal will be to show that the Truth Lemma does hold in Ω^Σ for Σ-formulas. It is easy to show that Ω^Σ satisfies the usual Existential Witness Lemma for modalities $\Diamond$ and $\widehat{K}$ (and formulas in Σ), but extending this to the perfect core modality $\odot$ requires some work. Another key ingredient in our proof is the fact that Ω^Σ is "essentially" a finite object: though possibly infinite in size, it has finite 'depth', and moreover it contains only finitely many bisimilarity classes. As a consequence, the largest fixed points of the operators $P \mapsto d_{\|\varphi\|}(P)$ (that define $\|\odot\varphi\|$) are all attained in Ω^Σ below some fixed *finite* stage of the Cantor-Bendixson process. We then use these ingredients to prove our Truth Lemma for the final model Ω^Σ.

The full details are in the Appendix to the extended version (see Supplementary Material), where we also use the selection method to obtain a finite submodel of Ω^Σ that satisfies the same relevant formulas, and then analyzing the complexity of the selection algorithm, thus proving:

Theorem 3.4. ***[FMP, Decidability and Complexity]*** *The (static and dynamic) logics of Cantor derivative and perfect core have the strong finite model property (and hence they are decidable). Moreover, the satisfiability problem for the static logic $\mathcal{L}$ is* PSPACE*-complete.*

Some Technical-Historical Connections. As mentioned in the Introduction, McKinsey and Tarski (McKinsey and Tarski 1944) were the first to look at the modal logic of topological closure and topological interior. In our notations, these are captured by the knowability modalities $\widehat{\mathcal{K}}$ and $\mathcal{K}$. They showed that this is the same as the modal logic of reflexive-transitive frames, more precisely the modal logic S4. In our formalism, the axiom 4 corresponds to our axiom of Positive Introspection for knowability: $\mathcal{K}\varphi \Rightarrow \mathcal{K}\mathcal{K}\varphi$. We refer to (van Benthem and Bezhanishvili 2007) for an overview of results on topological completeness of modal logics above S4.

As also mentioned in the Introduction, McKinsey and Tarski also considered the modal logic of Cantor derivative. Esakia (Esakia 2001; Esakia 2004) showed that the derivative logic of all topological spaces is the same as the logic of weakly-transitive frames, namely the modal logic wK4 = K + w4, where w4 is the weak transitivity axiom: $\Diamond\Diamond p \rightarrow \Diamond p \vee p$. In our formalism, this is easily seen to be *equivalent* to the above-mentioned axiom of Positive Introspection for knowability. Indeed, given our definition of knowability, the axiom $\mathcal{K}\varphi \Rightarrow \mathcal{K}\mathcal{K}\varphi$ can be unfolded into

$$(\varphi \wedge \Box\varphi) \Rightarrow \Box\Box\varphi.$$

This is a Sahlqvist formula (see e.g. (Chagrov and Zakharyaschev 1997)) corresponding to the weak-transitivity condition on relational models, whose equivalent dual form is Esakia's weak transitivity axiom w4.

4 Surprise: Non-Self-Referential Version

There are many 'solutions' to the Surprise Exam Paradox in the literature (Quine 1953; McLelland and Chihara 1975; Wright and Sudbury 1977; Sorensen 1984; Chow 1998; Hall 1999; Gerbrandy 2007; Levi 2009). Some of them concern different versions of the puzzle, in which some of the assumptions are suspended , e.g. the Student may *not know for sure* (but only believe) that there will be an exam next week, or that the Teacher always tells the truth. Though interesting, these provide "easy" ways to avoid the contradiction, so we will ignore these weakened versions, focusing on the version in which these assumptions are granted. Even so, most of the solutions proposed in the literature are unfortunately informal, or only half formalized. Gerbrandy's approach (Gerbrandy 2007) is one of the few exceptions. We hereby briefly summarize his approach.

Gerbrandy's Solution The setting uses the older, non-topological version of our dynamic-epistemic logic, more precisely the so-called Public Announcement Logic: an epistemic model $\mathbf{M} = (X, \|\cdot\|)$ is simply given by a set of possible worlds X together with a valuation map; the logic is restricted to the fragment generated by atomic sentences, Boolean connectives, the knowledge operator $K\varphi$ (modeled as universal modality) and the dynamic update operators $[\varphi]\theta$ (also called 'public announcement', and modeled by relativization to the *subset* $\|\varphi\|$, with no subspace topological structure). Like our logic, this logic is single-agent: the Teacher is only treated as an infallible *source* of truthful information, not as an agent. So the knowledge operator K refers to the Student's knowledge. Knowability $\mathcal{K}\varphi$, derivative modality $\Diamond\varphi$ and perfect core $\odot\varphi$ do not belong to this language. But the update modalities are still eliminable, via the reduction laws for Booleans and knowledge.

More specifically, the set $X = \{x_1, x_2, x_3, x_4, x_5\}$ consists of five possible worlds, with the obvious meaning: for

[10] In fact, the notion of truth in the canonical model will play no role in this paper: we never evaluate our formulas in it. Instead, we only use a few basic syntactic properties of this model.

each $1 \leq i \leq 5$, x_i is the world in which the exam will come in the corresponding i^{th} day of the week. The language has 5 atomic sentences $\{p_i : 1 \leq i \leq 5\}$, where p_i means "the exam will be in the i^{th} day". The valuation is again obvious: $\|p_i\| = \{x_i\}$. Clearly, this model satisfies $K(\bigvee_{i=1}^5 p_i)$, which captures one of the main assumptions of the puzzle: the Student knows for sure there will be an exam in the next week. Furthermore, for each $1 \leq i \leq 5$, the passage of the previous days without any exam can be 'simulated' in this logic by an update with the sentence $\bigwedge_{j=1}^{i-1} \neg p_j$: indeed, this is the information gained by the Student by the evening of day $i-1$. Hence, Gerbrandy formalizes Teacher's announcement as the sentence

$$\text{SURPRISE} \;:=\; \bigwedge_{i=1}^{5} [\bigwedge_{j=1}^{i-1} \neg p_j] \neg K p_i.$$

This sentence says that, no matter in which day i will the exam come, by the evening of day $i-1$ the Student will not know for sure that the exam will be the next day. Using the reduction axioms, this formula can be simplified to

$$\text{SURPRISE} \;\Leftrightarrow\; \bigwedge_{i=1}^{5} \neg K(\bigvee_{j=1}^{i} p_j).$$

Finally, the assumption that the Student knows for sure that the Teacher never lies is implemented by performing an update with the sentence SURPRISE: all worlds in which the sentence is false are eliminated, and the model shrinks to $\|\text{SURPRISE}\|$. But, using the above static equivalent, it is easy to see that, in the model X the sentence SURPRISE is false only in world w_5 (in which the exam is on Friday) and true in all the others. Hence, the model shrinks to $\|\text{SURPRISE}\| = \{x_1, x_2, x_3, x_4\}$.

Thus, according to Gerbrandy, *the only valid conclusion is that the exam cannot be on Friday*: the first elimination step in the informal reasoning underlying the 'paradox' is the only correct one. All further elimination steps are *not* justified: e.g., the second step (eliminating Thursday) would require performing *a second update* with the sentence SURPRISE. But the Teacher only announced the sentence once! The sentence SURPRISE was true before being announced (assuming the exam won't be on Friday), but *nothing guarantees that the sentence will still be true after this announcement*: the Teacher did not claim *that*! If say, the exam will be on Thursday, then the sentence SURPRISE changes its truth value (from true to false) after the Teacher's announcement: this does not in any way contradict the truthfulness of Teacher's announcement (since it *was true* at the moment when it was announced). So the apparent 'paradox' only points to the existence of sentences that change their truth value after being announced.[11]

A first objection to the above approach is that it gives a very "low level" formalization of the sentence SURPRISE, that is highly dependent on irrelevant details (such as the number of days in the week, the linear temporal order of the observable evidence in the form of day-passing, etc). If we change the story to cover 2 weeks, the sentence SURPRISE changes. Even worse: we can build similar stories, to which the above approach simply cannot be applied, since e.g. the number of worlds is infinite, the potential observations are also infinitely many, and they cannot be arranged in any salient linear order. Let us look now at such an example.

Infinite Surprise Let us denote the set of positive integers by $\mathbb{N}$. It is known that the Teacher chose a point x belonging to the set

$$A = \{0\} \cup \{1/n : n \in \mathbb{N}\} \cup \{1/n(n+1) : n \in \mathbb{N}\}$$

and marked it on the real line drawn on a board. The Student can perform observations, measuring the position of the point, with any arbitrary precision $\epsilon > 0$ (by building better and better measurement devices); but obviously, he can never measure the position with infinite precision ($\epsilon = 0$)! But the Teacher (who is known to be always truthful) tells the Student: "*No matter how good your measurement is, you will never know the exact position of the point*!"

Intuitively, the Student can reproduce the Surprise Exam argument to conclude that $x \notin A$, obtaining a contradiction (since he *knows* that $x \in A$). First, if the point is of the form $x = \frac{1}{n(n+1)}$ for some $n \in \mathbb{N}$, then he eventually be able to know its location exactly, if he continues increasing the precision of his measurements: indeed, whenever he will reach a precision $\epsilon < |\frac{1}{n(n+1)} - \frac{1}{(n+1)(n+2)}| = \frac{1}{n(n+1)(n+2)}$, then his measurement will yield an open interval of the form $(a - \epsilon, a + \epsilon) \ni x$, whose only intersection with A is the singleton $\{x\} = \{\frac{1}{n(n+1)}\}$ consisting of the exact position. But this contradicts the Teacher's announcement (that he will never know the exact position); this contradiction rules out all points of the form $\frac{1}{n(n+1)}$, so x must belong to the set $\{0\} \cup \{1/n : n \in \mathbb{N}\}$. By repeating the argument, the Student can rule out next all points of the form $x = \frac{1}{n}$ (with $n \in \mathbb{N}$), since in any such case he will eventually be able to know its location exactly (whenever he reaches a precision $\epsilon < |\frac{1}{n} - \frac{1}{(n)(n+1)}| = \frac{1}{n(n+1)}$), concluding that *$x$ must belong to the singleton set* $\{0\}$. So now the Student *knows* the exact location $x = 0$ (without even having had to do any measurement), again contradicting Teacher's announcement!

Though the argument is essentially identical to the Surprise Exam, it cannot be treated using the above approach, due to the fact that both the possible worlds and the possible observations (measurement intervals) are infinite.

This is where the topological approach comes to the rescue. By abstracting away from day-passing or measurements, and considering them to be just special cases of families of observable evidence, given in the form of strong topological bases, we can see the sentence SURPRISE simply says that "*the actual world is not knowable through observations*". Using our semantics, this is captured by the formula

$$\text{SURPRISE} \;:=\; \Diamond\top,$$

where $\Diamond$ is the derivative modality wrt the evidential topology (generated by the basis $\mathcal{B}$). In the case of our Infinite

[11] Such examples are called 'Moore sentences' and are by now well-understood as non-paradoxical utterings, easily dealt with in the framework of Dynamic Epistemic Logic.

Surprise, it is clear what the evidential topology is: the *standard* topology on the set A, generated by the family $\mathcal{B} = \{(a, b) \cap A : a, b \in \mathbb{Q}, a < b\}$ of (relativized) open intervals with rational endpoints. Applying Gerbrandy's analysis to this topological version, we see that $\|\text{SURPRISE}\|_A = \|\Diamond\top\|_A = d_A(A) = d(A) = \{0\} \cup \{\frac{1}{n} : n \in \mathbb{N}\}$ (since all other points are isolated in A), and we can thus conclude that *only this first elimination step is correct*: the only information that can be extracted from Teacher's announcement is that $x \in \{\frac{1}{n} : n \in \mathbb{N}\}$. Further elimination steps are not justified: though true when it was announced, the sentence $\Diamond\top$ may have changed its truth value after the announcement.

Going back to the original Surprise Exam story, what is the evidential topology in that case? Since "observations" correspond there to the passing of days without exams, the relevant strong base is

$$\mathcal{B} = \{O_1, O_2, O_3, O_4, O_5\},$$

where $O_i = X - \{x_j : j < i\} = \{x_j : i \leq j\}$. Here, $O_1 = X$ corresponds to the trivial tautological observation (before Monday) that the exam will be in one of the 5 days of next week; O_2 corresponds to the negative observation after Monday morning: that the exam was not on Monday (hence it will be in one of the remaining four days); etc. The generated evidential topology is $\mathcal{T} = \{\emptyset\} \cup \mathcal{B}$. Once again, as in Gerbrandy's analysis, $\|\Diamond\top\| = X - \{x_5\} = \{x_1, x_2, x_3, x_4\}$ (since x_5 is the only isolated point in this topology).

We have thus obtained a uniform treatment of the puzzle, that simplifies and generalizes Gerbrandy's solution.

5 Surprise: Self-Referential Version

While the above formalization of the sentence SURPRISE seems natural at first sight, there is something profoundly odd about it. The teacher announced that the *exam's date **will** be a surprise*: this seemed to point to the *actual future*, as it will unfold *after* this announcement is made. However, the above formalization allows for the possibility that the announcement was meant to be true only before the announcement (or counterfactually: if no such announcement was made), but to possibly change its truth value to false after the announcement is made. In that case, in what sense can one still claim that the Teacher was truthful in her announcement about "will" happen?

Looking at the sentence $\Diamond\top$ (or at Gerbrandy's more complicated non-topological counterpart), we can see that the best way to describe it in natural language is a counterfactual statement of the type: "*the exam's date **would** have been a surprise, if I didn't make this very announcement*". Moreover, this interpretation in terms of a counterfactual (instead of the actual) future seems to be crucial for Gerbrandy's 'solution' of the paradox.

However, this is *not* what the Teacher said, and it does *not* sound like the most natural interpretation of her statement. When referring to the future in an announcement, it is typically implicitly assumed that the speaker factors in her own announcement action: thus, she is expected to use the word "will" to refer to what will happen after she makes the announcement. "It will be a surprise" means that it *will* be so, not that it would have been so in some other possible future.

Thus, to understand the Teacher's statement we need to make explicit its implicit self-referentiality, reading it as "*You will not know in advance the exam day (i.e. after hearing **this** very announcement)*". Most authors who wrote about the paradox agree that *this self-referential interpretation is the intended one*.

Gerbrandy was aware of this interpretation (without formalizing it), but like many other logicians he thought that it leads to a genuine, Liar-like paradox, because of its circularity. In contrast, other logicians, such as Quine, argued in older work (Quine 1953) that there is no real paradox, but only an impossible assumption: the conclusion should only be that a *source who is known to always tell the truth cannot make such a (future-oriented, implicitly self-referential) announcement* (since that would be a lie).

Using our derivative and dynamic modalities, we can formalize the self-referential announcement as a 'circular' proposition P satisfying the equation

$$P = \langle P\rangle\Diamond\top.$$

Moreover, this is *all* that is claimed in the Teacher's announcement: there is no other implicit information in it. This means that we are looking at the *most general statement* satisfying the equation, i.e. the *largest fixed point* of the operator $P \mapsto \langle P\rangle\Diamond\top$. Using standard μ-calculus notation, we can write the statement as

$$\text{SURPRISE}^{\infty} := \nu P.\langle P\rangle\Diamond\top,$$

and call it the *self-referential surprise announcement*. Although the above formalization is not in our language $\mathcal{L}_{\langle\cdot\rangle}$ (but only in its fixed-point extension), it can be given an equivalent formulation. Using our reduction laws, we can see that $\langle P\rangle\Diamond\top$ is equivalent to $P \wedge \Diamond\langle P\rangle\top$, which in turn is equivalent to $P \wedge \Diamond P$. So the sentence SURPRISE^{∞} is equivalent to any of the following formulas:

$$\nu P.P \wedge \Diamond P = \nu P.\Diamond P = \nu P.(\top \wedge \Diamond P) = \odot\top.$$

Thus, the formula $\odot\top$, denoting the perfect core of our space $\|\odot\top\|_X = d^{\infty}(X)$, captures the full self-referential meaning of the surprise announcement SURPRISE^{∞}. There is nothing paradoxical with this type of self-referentiality: the monotonicity of the derivative operator ensures the existence of the fixed point. If a Teacher who is known never to lie made this announcement, that would induce an update that shrinks the original space X to its perfect core X^{∞}.

We can now recognize the successive eliminative steps in the Student's reasoning as corresponding to the Cantor-Bendisxon process of calculating the perfect core: the first step eliminates the isolated point x_5, calculating the Cantor derivative $d^1(X) = X - \{x_5\}$; the next step calculates $d^2(X) = X - \{x_4, x_5\}$; etc. After five steps, we reach a fixed point $d^5(X) = d^{\infty}(X) = \emptyset$. A similar remark applies to our above Infinite Surprise example: the first step yields $d^1(A) = \{0\} \cup \{\frac{1}{n} : n \in \mathbb{N}\}$; the next step yields $d^2(A) = \{0\}$; finally, the third step reaches the fixed point

$d^3(A) = d^\infty(A) = \emptyset$. And since in both cases the perfect core is empty, a contradiction is actually reached!

But, in this interpretation, *all the elimination steps are justified* (unlike in Gerbrandy's counterfactual interpretation). The conclusion is that, in the self-referential version of the story, *the Student's entire inductive eliminative reasoning is entirely correct*! The contradiction obtained in the end ($\|\text{SURPRISE}^\infty\| = d^\infty(X) = \emptyset$) only shows that *the update with* SURPRISE$^\infty$ *cannot be truthfully performed in this case*: if it is known that the Teacher never lies, then *the statement* SURPRISE$^\infty$ *is false, and in fact known to be false*, regardless of the day of the exam.

Liar-like paradox? Not really. The sentence SURPRISE$^\infty$ has in any case some definite truth value, unlike the Liar sentences. As already mentioned, one of the assumptions of the story must simply be false: either it is *not known for sure that the Teacher always tells the truth*, or else the Teacher *never makes this self-referential announcement* in these particular situations (since it would be a lie). The *appearance* of paradox is due to the fact in this specific example the only fixed point is the empty set. However, a proposition with empty extension is by definition *not* paradoxical, but just false (in all possible worlds).

But this doesn't validate the Students' ultimate conclusion (in the follow-up story): partying every day is *not* justified. *That last follow-up step is the Student's only mistake*. If the Student gives up the first assumption (that he knew that the Teacher never lies), then the whole iterative elimination reasoning is *blocked*: even the first step is no longer justified! So, in that case, the Student can no longer be sure that the Teacher lies: she may be lying, or she may be telling the truth. All bets are off, the exam might come any day. Studying every day, instead of partying, is now the only safe option.

Our diagnosis agrees with Quine's: a Teacher who is known not to lie cannot truthfully make the announcement SURPRISE$^\infty$ in our two examples. But, contrary to Quine, Gerbrandy and other philosophical logicians, we claim that this impossibility result is *not* due to the self-referential character of the announcement. Self-referentiality is only dangerous when applied to non-monotonic operators (such as negation, e.g. the Liar). But derivative is monotonic, so *the type of self-referentiality involved in the Surprise story is innocuous*.[12] In fact, the sentence SURPRISE$^\infty$ can even be *true* in some situations! To see this, let us consider a modified version of the above Infinite Surprise example.

Infinite Surprise with a Twist Everything goes as in the Infinite Surprise story, except that this time the Teacher choses a point x belonging to the set $B = A \cup [1, 2]$, where $A = \{0\} \cup \{\frac{1}{n} : n \in \mathbb{N}\} \cup \{\frac{1}{n(n+1)} : n \in \mathbb{N}\}$ is the set in the previous (untwisted) version of Infinite Surprise. The same Cantor-Bendixson inductive process of elimination can be now used to show that the perfect core is $d^4(B) = d^\infty(B) = [1, 2]$. In this situation, an update with the same self-referential sentence SURPRISE$^\infty$ shrinks the set of possible points to the subspace $[1, 2]$. In other words, an announcement of this sentence by a Teacher known to lie simply conveys the information that the actual points satisfies $x \in [1, 2]$. A smart Student should be able to correctly infer this information, by applying the same type of "paradoxical" reasoning as in the above examples. But no contradiction is reached now: this scenario *can* happen, and if the point really is in $[1, 2]$ then the Teacher told the truth!

The conclusion of our analysis is that, in any Surprise-like paradox, *the appearance of "paradoxicality" is not due to self-referentiality, but only to the fact that the perfect core happens to be empty*. The existence of non-empty perfect sets is a topological fact, that has important epistemic consequences: the self-referential sentence involved in Surprise-like scenarios *can* in fact be *true* (even if in the standard story it turns out to be false). The Surprise Exam 'Paradox' is not a paradox at all, and the Students' inductive process of elimination is a correct logical argument[13]: just a special case of the inductive Cantor-Bendixson process of calculating the perfect core! Thus, our analysis reveals deep connections between the apparent paradox and classical work in Analysis and Topology.

6 Concluding Remarks

In this paper, we developed a unified topological interpretation of knowledge, observable evidence, knowability and knowledge updates, and studied a notion of "epistemic surprise" (expressing the unknowability of the actual world), that comes in two flavors: a non-self-referential version (described by Cantor derivative) and a self-referential one (described by the perfect core). We applied these notions to the analysis of the Surprise Exam Paradox, gave a complete axiomatization of the associated logic, and proved that it is decidable and that its static fragment is PSPACE-complete.

Some outstanding open questions still remain. First, what is the *complexity of our dynamic logic* $\mathcal{L}_{\langle\cdot\rangle}$? Although the reduction to $\mathcal{L}$ is exponential, we conjecture that $\mathcal{L}_{\langle\cdot\rangle}$ is still PSPACE-complete. Second: developing a *multi-agent version* of our logic would be of great value for studying epistemic dialogues, security protocols and other multi-agent epistemic scenarios and puzzles. In future work, we plan to tackle these open problems and their applications.

[12] In contrast, the Liar sentence requires a fixed point for negation/complementation, which doesn't exist in a Boolean algebra. Another possible source of the feeling of paradox given by the Surprise Exam story might be the *negative* form of the Surprise sentence, as expressed in natural language, which makes it superficially similar to the Liar sentence. Thus, its self-referentiality may *look* dangerous at first sight. But looks are deceiving: in the expression "the actual world can be known, given P", the proposition P appears conditionally, and thus in a negative position; hence, when we negate this expression (saying "the world cannot be known, given P"), P reverts to a positive position. This explains the monotonicity of Cantor's derivative (and relative derivative), and thus the non-paradoxical nature of SURPRISE$^\infty$.

[13] With the obvious exception of the follow-up story: as we explained above, going to party every day (after giving up on the initial assumption that it was known that the Teacher never lies) is the Student's only mistake.

Acknowledgments

The third author was supported by the FWO-FWF Lead Agency grant G030620N (FWO)/I4513N (FWF).

References

Abramsky, S. 1991. Domain theory in logical form. *Ann. Pure Appl. Logic* 51(1-2):1–77.

Aumann, R. 1995. Backward induction and common knowledge of rationality. *Games and Economic Behavior* 8:6–19.

Baltag, A.; Bezhanishvili, N.; and Fernández-Duque, D. 2021. The topological mu-calculus: completeness and decidability. In *Proc. of LICS 36*, 1–13. IEEE Press.

Baltag, A.; Bezhanishvili, N.; Özgün, A.; and Smets, S. 2016. Justified belief and the topology of evidence. In *Proc. of WoLLIC 2016*, volume 9803 of *LNCS*, 83–103. Springer.

Baltag, A.; Bezhanishvili, N.; Ozgun, A.; and Smets, S. 2019a. A topological approach to full belief. *JPL* 48:205–244.

Baltag, A.; Gierasimczuk, N.; Özgün, A.; Sandoval, A. L. V.; and Smets, S. 2019b. A dynamic logic for learning theory. *Journal of Logical and Algebraic Methods in Programming* 109:100485.

Baltag, A.; Gierasimczuk, N.; and Smets, S. 2015. On the solvability of inductive problems: A study in epistemic topology. In *Proc. of TARK 2015*, 81–98. ENTCS.

Brecht, M., and Yamamoto, A. 2010. Topological properties of concept spaces. *Information & Computation* 208(4):327–340.

Chagrov, A., and Zakharyaschev, M. 1997. *Modal Logic*. New York: The Clarendon Press.

Chow, T. Y. 1998. The surprise examination or unexpected hanging paradox. *American Mathematical Monthly* 105:41—-51.

Esakia, L. 2001. Weak transitivity—a restitution. In *Logical investigations, No. 8 (Russian) (Moscow, 2001)*. Moscow: “Nauka”. 244–255.

Esakia, L. 2004. Intuitionistic logic and modality via topology. *Annals of Pure and Applied Logic* 127(1-3):155–170. Provinces of logic determined.

Fernández-Duque, D. 2011. Tangled modal logic for spatial reasoning. In Walsh, T., ed., *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, 857–862. IJCAI/AAAI.

Gerbrandy, J. 2007. The surprise examination in dynamic epistemic logic. *Synthese* 155:21–33.

Goldblatt, R., and Hodkinson, I. 2017. Spatial logic of tangled closure operators and modal mu-calculus. *Ann. Pure Appl. Log.* 168(5):1032–1090.

Goubault, E.; Ledent, J.; and Rajsbaum, S. 2020. A simplicial complex model for dynamic epistemic logic to study distributed task computability. *Information & Computation*.

Hall, N. 1999. How to set a surprise exam. *Mind* 108(432):647–703.

Kelly, K. T. 1996. *The Logic of Reliable Inquiry*. Oxford University Press.

Levi, K. 2009. The solution to the surprise exam paradox. *Southern Journal of Philosophy* 47(2):131–158.

McKinsey, J. C. C., and Tarski, A. 1944. The algebra of topology. *Ann. of Math.* 45:141–191.

McLelland, J., and Chihara, C. 1975. The surprise examination paradox. *JPL* 4(1):71–89.

Özgün, A. 2017. *Evidence in Epistemic Logic : A Topological Perspective*. Ph.D. Dissertation, ILLC, Univ. of Amsterdam and Univ. of Lorraine.

Parikh, R. 1992. Finite and infinite dialogues. In *Logic from Computer Science*, 481–497. Springer.

Quine, M. V. 1953. On a so-called paradox. *Mind* 62(245):65–67.

Sorensen, R. 1984. Recalcitrant variations of the prediction paradox. *Australasian Journal of Philosophy* 69(4):355–362.

Steinsvold, C. 2007. *Topological models of belief logics*. Ph.D. Dissertation, City University of New York.

van Benthem, J., and Bezhanishvili, G. 2007. Modal logics of space. In *Handbook of spatial logics*. Dordrecht: Springer. 217–298.

Vickers, S. 1989. *Topology via Logic*. Cambridge: Cambridge University Press.

Wright, C., and Sudbury, A. 1977. The paradox of the unexpected examination. *Australasian Journal of Philosophy* 55(1):41—58.