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 1944). When the modal $\Diamond$ is interpreted as topological closure $\mathrm{Cl}(A)$ and the modal $\Box$ as topological interior $\mathrm{Int}(A)$, 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 $d(A)$ (consisting 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. The notion of observability and its logic require a topological setting, cf. (Vickers 1989) and (Abramsky 1991). 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 Rausbaum 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 extending the logic of Cantor derivative (Esakia 2001) with a global modality (Goranko and Passy 1992) that quantifies over all points, as well as 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 $\mathrm{Int}(A)$ 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).[1]

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, though somewhat related to an older work (Parikh 1992) (which assumes a different framework: multi-agent $S5$ Kripke frames, seen as a special case of topological spaces). Essentially, the derivative $d(A)$ is the proposition asserting that *the actual world is unknowable (through observations), even if given (the additional information) A*. We derive this interpretation in a straightforward way, by putting together the semantics of derivative and the knowability reading of the interior operator.[2] Finally, we infer the epistemic meaning of the perfect core $d^{\infty}(A)$ 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 Student knows for sure that there will be an exam in one of the five (working) days of next week. But he doesn't know in which day. The Teacher (who is 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*: if the exam is on Friday, then it wouldn't be a surprise (-since Friday is the last possible day, by Thursday evening the Student would know it), contrary to the Teacher's (true) announcement; so the exam cannot take place on Friday. But once Friday is eliminated, the Student can repeat the same argument, until all days are eliminated. This is a contradiction (since 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?!

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 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. $X \in \mathcal{B}$).*[3] *A* basic neighborhood *of a point $x \in X$ is a basic open set $U \in \mathcal{B}$ with $x \in U$.*

[1] These dynamic operators are topological versions of the update modalities of Public Announcement Logic and Dynamic Epistemic Logic (Plaza 1989), (Baltag, Moss, and Solecki 1998), (van Ditmarsch, van der Hoek, and Kooi 2007), (van Benthem 2011).

[2] 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 (Parikh 1992), where derivative $d(A)$ is connected to *ignorance* (rather than to unknowability): the agent does *not know* the actual world (even if given A). Still, our treatment of the Surprise Exam Paradox using the perfect core and the Cantor-Bendixson process was partially inspired from Parikh's treatment of classical epistemic scenarios (such as the Two Numbers' dialogue) using the same topological tools.

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

*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 sets $U \in \mathcal{T}$ are called* open *sets.*[4] *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 $x \in U$.*

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 $P \subseteq X$ 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*** 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.[5] 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.[6] 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

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

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

[4] 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).

[5] 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.

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

$x \in U \cap V$); 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).[7] The relativized interior $\mathrm{Int}_A(P) = A \cap \mathrm{Int}(P \cup (X - A))$ in the subspace topology captures 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.

Concrete Example: the policeman and the speeding car (Parikh, Moss, and Steinsvold 2007) 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, "textitSpeeding" 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

[7] 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).

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 $[\varphi]$.*[8]
- *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:*

$$\langle\varphi\rangle p \;\Leftrightarrow\; (\varphi \wedge p)$$

$$\langle\varphi\rangle\neg\theta \;\Leftrightarrow\; (\varphi \wedge \neg\langle\varphi\rangle\theta)$$

$$\langle\varphi\rangle\widehat{K}\theta \;\Leftrightarrow\; (\varphi \wedge \widehat{K}\langle\varphi\rangle\theta)$$

$$\langle\varphi\rangle\Diamond\theta \;\Leftrightarrow\; (\varphi \wedge \Diamond\langle\varphi\rangle\theta)$$

$$\langle\varphi\rangle\odot\theta \;\Leftrightarrow\; \odot\langle\varphi\rangle\theta$$

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

Proving soundness is an easy verification. Completeness follows immediately from 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*[9] *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.*[10]

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 nontrivial, 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 (relegating the details ti the Appendix). 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.[11] 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 strategy is to show that the Truth Lemma does hold in Ω^{Σ} for Σ-formulas. It is easy to check 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 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. These ingredients are used to prove our Truth Lemma for the final model Ω^{Σ}.

The full details are in the Appendix, 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 were the first to look at the modal logic of topological closure and topological interior (McKinsey and Tarski 1944). 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 S4 of reflexive-transitive frames. 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 showed that the derivative logic of all topological spaces is the same as the logic of weakly-transitive frames (Esakia 2001; Esakia 2004), namely the modal logic wK4 $=$ K $+$ w4, where w4 is the weak transitivity axiom: $\Diamond\Diamond p \to \Diamond p \vee p$. 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. The approach in (Gerbrandy 2007) is one of the few exceptions, and we hereby briefly summarize it.

Gerbrandy's Solution The setting used by Gerbrandy to treat the paradox is the one of (non-topological) Public Announcement Logic (Plaza 1989): 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

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

[10] 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 $\mathcal{L}_{\langle\cdot\rangle}$.

[11] 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.

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 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.[12]

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 will eventually be able to know its location exactly, if given precise enough measurements: indeed, whenever he will reach a precision $\epsilon < |\frac{1}{n(n+1)} - \frac{1}{(n+1)(n+2)}| = \frac{1}{n(n+1)(n+2)}$, his measurement will yield an open interval of the form $(a-\epsilon, a+\epsilon) \ni x$, whose intersection with A is the singleton $\{x\} = \{\frac{1}{n(n+1)}\}$ consisting of the exact position. Since this contradicts the Teacher's announcement, all points of the form $\frac{1}{n(n+1)}$ are ruled out, 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 $x = \frac{1}{n}$ (since in any such case he will eventually be able to know its location exactly, when 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,

[12] 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 (van Ditmarsch, van der Hoek, and Kooi 2007; van Benthem 2011).

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 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 scenario? Since "observations" correspond in that case 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 background observation 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; 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-Bendixson 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 self-referential interpretation, *all the elimination steps are justified* (unlike in Gerbrandy's counterfactual interpretation): *the Student's entire inductive eliminative reasoning is 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 a 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 *cannot make this self-referential announcement* (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).

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 the only safe option.

Our diagnosis thus 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*.[13] 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![14]

In conclusion, *the appearance of "paradoxicality" in the Surprise Exam story 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 it is false in the standard version). The Surprise Exam 'Paradox' is not a paradox at all, and the Students' inductive process of elimination is a correct logical argument[15]: just a special case of the inductive Cantor-Bendixson process of calculating the perfect core! Thus, our solution 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

[13] 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$^\infty$.

[14] Similar non-paradoxical processes of iterated elimination ending in a non-empty fixed point occur elsewhere in epistemic logic: knowledge dialogues (Parikh 1992), converting implicit knowledge into common knowledge by publicly sharing information within a group (van Benthem 2002), reaching equilibria in epistemic game theory by repeated public announcements of substantive rationality (van Benthem 2007).

[15] 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.

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. Third, our concepts of knowability and unknowability are closely related to the topo-logical account of learning given in (Dabrowski, Moss, and Parikh 1996) in terms of *observational effort*. It would be interesting to elucidate this relationship in more depth.

In future work, we plan to tackle these open problems and their applications.

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.

Baltag, A.; Moss, L.; and Solecki, S. 1998. The logic of public announcements, common knowledge, and private suspicions. 43–56.

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.

Dabrowski, A.; Moss, L.; and Parikh, R. 1996. Topological reasoning and the logic of knowledge. *Annals of Pure and Applied Logic* 78(1):73–110.

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.

Goranko, V., and Passy, S. 1992. Using the universal modality: gains and questions. *J. Logic Comput.* 2(1):5–30.

Goubault, E.; Ledent, J.; and Rausbaum, 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.; Moss, L.; and Steinsvold, C. 2007. Topology and epistemic logic. In *Handbook of spatial logics*. Dordrecht: Springer. 299—–341.

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

Plaza, J. 1989. Logics of public communication. In *Proceedings 4th International Symposium on Methodologies for Intelligent Systems*, 201–216.

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.

van Benthem, J. 2002. One is a lonely number. 96–129. ASL and A.K. Peters, Wellesley MA.

van Benthem, J. 2007. Rational dynamics and epistemic logic in games. *International Game Theory Review* 9.

van Benthem, J. 2011. *Logical Dynamics of Information and Interaction*. Cambridge University Press.

van Ditmarsch, H.; van der Hoek, W.; and Kooi, B. 2007. *Dynamic Epistemic Logic*. Springer,.

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.

Appendix

A Relational semantics

We start from a special case of the above topological semantics: Alexandroff spaces $(X, \mathcal{T})$ are the ones in which the topology $\mathcal{T}$ is *closed under arbitrary intersections*. It is well-known such spaces admit a presentation as relational Kripke frames.

Special Case: Standard Relational Models If we restrict to the class of *Alexandroff spaces*, then we obtain as a special case a *relational semantics* for the above logic. It is well-known that Alexandroff spaces are the same as standard *relational models* $(X, R, \|\cdot\|)$, with R *irreflexive and weakly transitive*: i.e. if $wRsRv$, then either $w = v$ or wRv. The above topological semantics for $\Diamond$ simply corresponds in this case to the standard clause for existential Kripke modalities, while the above topological semantics for $\odot$ amounts to putting $w \models \odot\varphi$ iff there is an infinite chain of (not necessarily distinct) worlds

$$w = w_0 \; R \; w_1 \; R \; w_2 \; R \ldots R \; w_n \; R \ldots$$

with $w_n \models \varphi$ for all n. Moreover, one can easily see that $\mathcal{K}$ is in this case the (universal) Kripke modality for the reflexive closure $\mathrm{Id} \cup R$ of R, which (due to weak transitivity) coincides with its reflexive-transitive closure R^*.[16]

Non-standard Relational Models The above relational clauses can give an interpretation of our syntax in any relational model $(X, R, \|\cdot\|)$ (not necessarily associated to an Alexandroff topo-model). In particular, we'll be interested in dropping the irreflexivity condition, and thus interpreting our syntax in models in which R is only required to be weakly transitive.

Lemma A.1. *The logic of weakly transitive relational models (for our syntax) is the same as the logic of irreflexive and weakly transitive models.*

Proof. Given any weakly transitive model $\mathbf{M} = (X, R, \|\cdot\|)$, we associate to it an irreflexive and weakly transitive model $\tilde{\mathbf{M}} = (\tilde{X}, \tilde{R}, \|\cdot\|^{\sim})$, by letting W^{i} and W^{r} be the set of irreflexive and reflexive points of $\mathbf{M}$, respectively, and setting

$$\tilde{X} := \big(W^{\mathrm{i}} \times \{0\}\big) \cup \big(W^{\mathrm{r}} \times \{0, 1\}\big).$$

It is useful to consider a map $\pi : \tilde{X} \to X$, given by $\pi(x, i) := x$. Using this, we can define the accessibility relation on $\tilde{X}$ by putting

$$\tilde{x} \; R \; \tilde{y} \quad \text{if} \quad \pi(\tilde{x}) \; R \; \pi(\tilde{y}) \text{ and } \tilde{x} \neq \tilde{y},$$

for all $\tilde{x}, \tilde{y} \in \tilde{X}$; and we define the valuation on $\tilde{X}$ by

$$\|p\|^{\sim} := \{\tilde{x} \in \tilde{X} : \pi(\tilde{x}) \in \|p\|\}.$$

It is easy to see that $\tilde{\mathbf{M}}$ is an irreflexive and weakly transitive relational model, and that the map $\pi : \tilde{X} \to X$ is a p-morphism with respect to both modalities $\Diamond, \widehat{\mathcal{K}}$ of our syntax. So the two models are modally equivalent wrt our syntax. □

[16] Here, Id is the identity relation on X.

B Proof of Completeness

We prove here our main completeness result (Theorem 3.1). For this, we need to prove Theorem 3.2 (on the fact that $\mathcal{L}_{\langle\cdot\rangle}$ and $\mathcal{L}$ are provably co-expressive) and Theorem 3.3 (completeness for the static logic $\mathcal{L}$).

Proof of Theorem 3.2 Let φ be any formula in $\mathcal{L}_{\langle\cdot\rangle}$. We need to show that there exists some formula $\varphi' \in \mathcal{L}$, such that $\vdash \varphi \Leftrightarrow \varphi'$ is a theorem in our axiom system. The proof uses the reduction axioms we have given, as well as the following two derivable reduction laws:

$$\langle\varphi\rangle(\psi \wedge \theta) \;\Leftrightarrow\; (\langle\varphi\rangle\psi \wedge \langle\varphi\rangle\theta),$$

$$\langle\varphi\rangle\top \;\Leftrightarrow\; \varphi.$$

The first follows from Necessitation and Distribution for $[\varphi]$, together with the laws of propositional logic. The second follows the reduction axioms, together with the definition of $\top := q \vee \neg q$ (for some chosen atom q) and propositional logic. Using all the reduction axioms and laws, we can gradually push any innermost dynamic modality $[\psi]$ occurring in a subformula of φ of the form $[\psi]\theta$ (where θ contains no dynamic modalities) past each next connective, until it is pushed to the bottom and then eliminated, obtaining a formula θ' that is provably equivalent to $[\psi]\theta$. This formula θ' may still have some occurrences of dynamic modalities (coming from ψ, which now occurs 'online' in θ', rather than inside $[\psi]$), but their number will be at most one less than in the original subformula $[\psi]\theta$. By propositional logic, Necessitation and Distribution (for the modalities K, $\Box$ and $[\varphi]$), together with the Mononoticity rule for $\odot$, it follows that we can replace the subformula $[\psi]\theta$ with its equivalent θ' within our original formula φ, obtaining an equivalent formula φ'_1 that has fewer dynamic modalities than φ. Repeating this procedure n times (where n is at most the number of dynamic modalities occurring in the original formula φ), we obtain a provably equivalent formula $\varphi' = \varphi'_n$, that has no dynamic modalities, i.e. belongs to the fragment $\mathcal{L}$.

The rest of this section will be dedicated to proving Theorem 3.3 (completeness for the static logic $\mathcal{L}$).

Canonical Model The standard 'canonical model' construction provides an (infinite) weakly-transitive model. This is a non-standard relational model (since irreflexivity is not guaranteed). A *theory* is a maximally consistent set of formulas (i.e. a set T that is consistent and has no proper consistent extension). We can define a *canonical equivalence relation* between theories, by putting

$$T \sim T' \;\text{ iff }\; \forall\varphi \;(\text{ if } K\varphi \in T \text{ then } \varphi \in T')\,.$$

Similarly, the *canonical accessibility relation* $\longrightarrow$ between two theories T, T' is given as usual, by putting

$$T \longrightarrow T' \;\text{ iff }\; \forall\varphi \;(\text{ if } \Box\varphi \in T \text{ then } \varphi \in T')\,.$$

The axioms for K are Sahlquist, so that $\sim$ is an equivalence relation and $T \longrightarrow S$ implies that $T \sim S$. To ensure that the knowledge modality really quantifies over all possible worlds, we need to restrict our model so that the relation $\sim$ becomes the universal relation. For this, we now *fix a theory* T_0, and we will restrict our canonical construction to the generated submodel. Let Ω be the family of all theories T s.t. $T_0 \sim T$. The *canonical model for* T_0 is the structure $\mathbf{\Omega} = (\Omega, \longrightarrow, \|\cdot\|)$, where the canonical accessibility relations $\longrightarrow$ are restricted here to Ω, and $\|\cdot\|$ is the *canonical valuation* on Ω, given by

$$\|p\| \;:=\; \{T \in \Omega : p \in T\}.$$

Since the weak-transitivity condition is Sahlquist, it immediately follows that *the canonical model* $(\Omega, \longrightarrow, \|\cdot\|)$ *is indeed weakly-transitive* (though not irreflexive). As a consequence, *the reflexive-transitive closure* $\longrightarrow^*$ *of the canonical relation coincides with its reflexive closure* $\longrightarrow \cup\, \mathrm{Id}_\Omega$.

We will make use of two other well-known properties of the canonical model, given by the next two lemmas.

Lemma B.1 (Lindenbaum Lemma)**.** *Every consistent set* Φ *of formulas has a maximally consistent extension (*$T \in \Omega$ *s.t.* $\Phi \subseteq T$*).*

Lemma B.2 (Canonical Witness Lemma)**.** *For every theory* $T \in \Omega$ *and formula* φ*, we have:*

1. $\Diamond\varphi \in T$ *iff there exists some theory* $T' \in \Omega$ *s.t.* $T \longrightarrow T' \ni \varphi$.
   We also have an equivalent statement in $\Box$*-form:*
   $$\Box\varphi \in T \;\textit{ iff }\; \forall T' \in \Omega\,(\textit{ if } T \longrightarrow T' \textit{ then } \varphi \in T')\,.$$
2. $\widehat{K}\varphi \in T$ *iff there exists some theory* $T' \in \Omega$ *s.t.* $T \sim T' \ni \varphi$.
   The statement in K*-form is:*
   $$K\varphi \in T \;\textit{ iff }\; \forall T' \in \Omega\,(\textit{ if } T \sim T' \textit{ then } \varphi \in T')\,.$$

The left-to-right implication in the first statement above is known as the (Canonical) $\Diamond$-Existence Lemma. The proofs are well-known, and these results imply that the so-called Truth Lemma holds in the canonical model for the $\odot$-free fragment of our logic.

We similarly obtain a Canonical $\widehat{\mathcal{K}}$-Witness Lemma, using the following result.

Lemma B.3. *For theories* $T, T' \in \Omega$*, we have:*

$$T \longrightarrow^* T' \;\textit{ iff }\; \forall\varphi(\textit{ if } \mathcal{K}\varphi \in T \textit{ then } \varphi \in T'),$$

where $\longrightarrow^* = \longrightarrow \cup\, \mathrm{Id}_\Omega$ *is the reflexive closure of* $\longrightarrow$*.*

Proof. The *left-to-right implication*: Assume that $T \longrightarrow^* T'$. If $T = T'$, then $\mathcal{K}\varphi \in T$ implies by definition that $\varphi \in T = T'$, as desired. If $T \neq T'$, then we must have $T \longrightarrow T'$, and then $\mathcal{K}\varphi \in T$ implies by definition that $\Box\varphi \in T$, which implies that $\varphi \in T'$ (by the Canonical $\Diamond$-Witness Lemma), as desired.

The *right-to-left implication*: Assume that we have $\forall\varphi(\mathcal{K}\varphi \in T \Longrightarrow \varphi \in T')$. To show that $T \longrightarrow^* T'$, we assume that $T \neq T'$, and we need to prove that $T \longrightarrow T'$. Since $T \neq T'$, there exists some formula $\theta \in T$ with $\theta \notin T'$. To show the desired conclusion, let φ be any arbitrary formula s.t. $\mathcal{K}\varphi \in T$, and we need to prove that $\varphi \in T'$. From $\mathcal{K}\varphi \in T$ we infer that $\varphi \in T$, hence $(\varphi \vee \theta) \in T$; similarly, from $\mathcal{K}\varphi \in T$ we infer that $\Box\varphi \in T$, hence $\Box(\varphi \vee \theta) \in T$. Putting these together, we obtain $\mathcal{K}(\varphi \vee \theta) \in T$. By our assumption, this implies that $(\varphi \vee \theta) \in T'$, and since $\theta \notin T'$, we conclude that $\varphi \in T'$, as desired. □

As a consequence, we immediately get:

Lemma B.4 (Canonical $\widehat{\mathcal{K}}$-Witness Lemma)**.** *For every formula φ and theory $T \in \Omega$, we have $\widehat{\mathcal{K}}\varphi \in T$ iff there exists some theory $T' \in \Omega$ s.t. $T \longrightarrow^* T' \ni \varphi$.*

The proof is immediate, given Lemma B.3.

Interestingly enough, the analogue of the Existence Lemma for $\odot$ also holds in the canonical model:

> If $\odot\varphi \in T \in \Omega$, then there is an infinite chain $T = T_0 \longrightarrow T_1 \longrightarrow \ldots \longrightarrow T_n \longrightarrow \ldots$, with $\varphi \in T_n \in \Omega$ (and hence $\varphi \in T_n$) for all n.

We state this fact without proof, since we will not need it. Unfortunately, *the converse fails*: there exist theories T which are part of an infinite φ-chain as above, but $\odot\varphi \notin T$.

Example 1. *Consider atoms $(p_n)_{n<\omega}$ and check that for every n, the set $\Phi_n := \{p_n, \neg\odot\top\} \cup \{\mathcal{K}(p_i \Rightarrow \Diamond p_{i+1}) : i < \omega\}$ is consistent (since all finite subsets are satisfiable). Use the Canonical Truth Lemma for Basic Modal Logic (and the fact that $\mathcal{K}$ is definable in it) to construct $(T_n)_{n<\omega}$ with $\Phi_n \subseteq T_n$ and $T_0 \to T_1 \to \ldots \to T_n \to \ldots$. Thus, $T_0 \models \odot\top$ although $(\neg\odot\top) \in T_0$.*

So we don't have a full Canonical $\odot$-Witness Lemma, and as a consequence the Truth Lemma fails in the canonical model for the full language (with the perfect core modality $\odot$). Moreover, the filtration method (standardly used to deal with this problem in the case of PDL) does not seem to work here either. Surprisingly though, the older and simpler 'selection' method works: we will look at submodels of the canonical model, obtained by selecting only a special kind of theories, called 'final' theories.

Canonical Submodels Any subset $X' \subseteq X$ of the set of worlds of a relational model $M = (X, \longrightarrow, \|\cdot\|)$ determines a unique *submodel*, obtained by taking: X' as its set of worlds; the restriction of $\longrightarrow$ to X' as its accessibility relation; and the valuation given by $\|p\| \cap X'$. A *canonical submodel* is a submodel of the canonical model.

Final Theories Given a formula θ, a theory $T \in \Omega$ is *θ-final* if we have: $\theta \in T$, and for all theories $S \in \Omega$, if $T \longrightarrow S$ and $\theta \in S$ then $S \longrightarrow T$ (hence $T \longleftrightarrow S$). Given a set Σ of formulas, a theory $T \in \Omega$ is *Σ-final* if it is θ-final for some formula $\theta \in \Sigma$.

Final Model Let Σ be any set of formulas. The *Σ-final model* is the canonical submodel determined by the set $\Omega^{\Sigma} := \{T \in \Omega : T \text{ is } \Sigma\text{-final}\}$ of all Σ-final theories.

The final model may be infinite, but we can show that it has finite 'depth' whenever Σ is finite. For this, we need the following definition:

Depth of a point in a model Given a (weakly transitive, not necessarily irreflexive) relational model $\mathbf{M} = (X, \longrightarrow, \sim, \|\cdot\|)$, and a point $x \in X$, a *strict (finite) x-chain* is a finite sequence of points of the form $x = x_0 \longrightarrow x_1 \longrightarrow \ldots x_n$ with $x_{i+1} \not\longrightarrow x_i$ for all $i < n$. The number n is called the *length* of our finite chain. The *depth* $\mathrm{dpt}(x)$ *of the point* $x \in X$ is the supremum of the lengths of all x-chains:

$$\mathrm{dpt}(x) := \sup\{n \in \mathbb{N} : \exists \text{ a strict } x\text{-chain of length } n\}.$$

In general, we have $\mathrm{dpt}(x) \geq 0$, with $\mathrm{dpt}(x) = 0$ iff for every $y \in X$, $x \longrightarrow y$ implies $y \longrightarrow x$; and $\mathrm{dpt}(x) = \omega$ iff there exist x-chains of every length $n \in \mathbb{N}$. The *depth* $\mathrm{dpt}(\mathbf{M})$ *of the model* $\mathbf{M}$ is the supremum of the depths of all points of the model:

$$\mathrm{dpt}(\mathbf{M}) := \sup\{\mathrm{dpt}(x) : x \in X\}.$$

Lemma B.5. *Let $\mathbf{M} = (X, \longrightarrow, \sim, \|\cdot\|)$ be a relational model, and $x, y \in X$ be two points. Then we have the following:*

1. *if $x \longrightarrow^* y$, then $\mathrm{dpt}(x) \geq \mathrm{dpt}(y)$;*
2. *if $x \longleftrightarrow y$, then $\mathrm{dpt}(x) = \mathrm{dpt}(y)$;*
3. *if $x \longrightarrow y$ and $\mathrm{dpt}(x) = \mathrm{dpt}(y)$, then $x \longleftrightarrow y$;*
4. *if $x \longrightarrow y$ and $y \not\longrightarrow x$, then $\mathrm{dpt}(x) > \mathrm{dpt}(y)$.*

Proof. Easy verification. □

Lemma B.6 (Finite Depth Lemma)**.** *Assume that Σ is a finite set of formulas of size $|\Sigma|$. Then the Σ-final model Ω^{Σ} has a finite depth bounded by $|\Sigma| - 1$:*

$$\mathrm{dpt}(\mathbf{M}) \leq |\Sigma| - 1.$$

In other words: for every strict chain of Σ-final theories $T_0 \longrightarrow T_1 \longrightarrow \ldots T_n$ (satisfying $T_{i+1} \not\longrightarrow T_i$ for all $i < n$), we have that $n \leq |\Sigma| - 1$.

Proof. Suppose, towards a contradiction, that $T_0 \longrightarrow T_1 \longrightarrow \ldots T_n$ is a strict chain of Σ-final theories of length $n \geq |\Sigma|$. Since all T_i are Σ-final, there exist formulas $\theta_0, \ldots, \theta_n \in \Sigma$ s.t. T_i is θ_i-final (and hence $\theta_i \in T$) for all $i \leq n$. But this is a sequence of $n + 1 \geq |\Sigma| + 1 > |\Sigma|$ formulas in Σ, so some formula θ must be repeated. Let θ be such a repeating formula in the enumeration, and let i and j be indices such that $i < j$ and $\theta_i = \theta_j = \theta$.

So we have $T_i \longrightarrow T_{i+1} \longrightarrow^* T_j$, with both T_i and T_j being θ-final, and so also $T_i \longrightarrow^* T_j$. We have two cases: either $T_i \longrightarrow T_j$ or $T_i = T_j$. We claim that in both cases we have $T_{i+1} \longrightarrow^* T_i$. To show this, consider first the case $T_i \longrightarrow T_j$. By θ-finality we get $T_i \longleftrightarrow T_j$, hence $T_i \longrightarrow T_{i+1} \longrightarrow^* T_j \longleftrightarrow T_i$, and thus $T_i \longrightarrow T_{i+1} \longrightarrow^* T_i$, as desired. In the second case, we assume $T_i = T_j$, so we immediately obtain $T_{i+1} \longrightarrow^* T_j = T_i$, as desired.

So we showed that we have $T_i \longrightarrow T_{i+1} \longrightarrow^* T_i$. There are again two cases: either $T_i \longrightarrow T_{i+1} \longrightarrow T_i$, or $T_i \longrightarrow T_{i+1} = T_i$. In the first case, we immediately conclude that $T_i \longleftrightarrow T_{i+1}$, which contradicts the 'strictness' of our chain. In the second case, we have $T_{i+1} = T_i \longrightarrow T_{i+1} = T_i$, so we again conclude that $T_i \longleftrightarrow T_{i+1}$, in contradiction with our 'strictness' assumption. □

In order to prove completeness with respect to the final model, we first need to show that every consistent formula belongs to some final theory. This is achieved by combining the Lindenbaum Lemma with the following

Lemma B.7 (Final Lemma)**.** *If $\varphi \in T \in \Omega$, then there exists some φ-final theory $T^* \in \Omega$ such that $T \longrightarrow^* T^*$ (and obviously, $\varphi \in T^*$, by finality).*

Proof. We will use a well-known variant of Zorn's Lemma, stated for preorders: a preordered set $(\mathcal{S}, \leq)$ has a maximal element if every chain has an upper bound. (Here, being maximal in a preordered set means that there is no strictly larger element.)

Let $\varphi \in T \in \Omega$. Take $\mathcal{S} := \{T' \in \Omega : T \longrightarrow^* T' \ni \varphi\}$, with the relation $\longrightarrow^*$ as its preorder. Let $\mathcal{S}' \subseteq \mathcal{S}$ be a chain of theories in $\mathcal{S}$. To show that it has an upper bound, take the set

$$\Phi := \{\varphi\} \cup \{\mathcal{K}\theta : \mathcal{K}\theta \in T' \text{ for some } T' \in \mathcal{S}'\}$$

We show that Φ *is consistent*: suppose this is not the case. Then there exists some *finite* such inconsistent subset $\Phi' = \{\varphi\} \cup \{\mathcal{K}\theta_1, \ldots, \mathcal{K}\theta_n\}$, with $\mathcal{K}\theta_1 \in T_1, \ldots, \mathcal{K}\theta_n \in T_n$ for some theories $T_1, T_2, \ldots, T_n \in \mathcal{S}'$. Since $\mathcal{S}'$ is a chain, we can assume that $T_1, T_2, \ldots T_{n-1} \longrightarrow^* T_n$, and thus $\mathcal{K}\theta_1, \ldots, \mathcal{K}\theta_n \in T_n$. Since $T_n \in \mathcal{S}$, we also have $\varphi \in T_n$, so $\Phi' \subseteq T_n$, which contradicts the consistency of T_n.

Applying now Lindenbaum's Lemma, there exists some maximally consistent extension $S \in \Omega$ with $\Phi \subseteq S$. By construction (and using Lemma B.3), we have $T' \longrightarrow^* S$ for all $T' \in \mathcal{S}'$, so S is an upper bound for the chain $\mathcal{S}$. Applying Zorn's lemma, we obtain a $\longrightarrow^*$-maximal element $T^* \in \mathcal{S}$. In particular, this means that $\varphi \in T^*$ and $T \longrightarrow^* T^*$, as desired. To prove that T^* is φ-final, suppose that $T^* \longrightarrow S \ni \varphi$; we have to show that $S \longrightarrow T^*$. By the $\longrightarrow^*$-maximality of T^*, we must have $S \longrightarrow^* T^*$, i.e. either $S \longrightarrow T^*$ or $S = T^*$. If $S \longrightarrow T^*$, then we are done. If $S = T^*$, then $S = T^* \longrightarrow S = T^*$, so we get again $S \longrightarrow T^*$, as desired. □

The next step is to establish an analogue of the $\Diamond$-Witness Lemma for final theories:

Lemma B.8 (Final Witness Lemma)**.** *For any theory $T \in \Omega$ and formula φ, we have:*

1. *$\Diamond\varphi \in T$ iff there exists some φ-final theory T' such that $T \longrightarrow T'$.*
2. *$\widehat{K}\varphi \in T$ iff there exists some φ-final theory T' such that $T \sim T'$.*

(Obviously, we have $\varphi \in T'$ in both cases, by finality.)

Proof. We prove the first claim, as the second is analogous. The *left-to-right implication*: by the Canonical $\Diamond$-Witness Lemma B.2, $\Diamond\varphi \in T$ implies the existence of some theory S with $T \longrightarrow S$ and $\varphi \in S$. By the Final Lemma B.7, there exists some φ-final theory S^* with $S \longrightarrow S^*$ and $\varphi \in S^*$. If $T \longrightarrow S^*$, then we can take $T' := S^*$ and we are done (since S^* is φ-final and $T \longrightarrow S \ni \varphi$, as desired). If $T \not\longrightarrow S^*$, then from this and $T \longrightarrow S \longrightarrow S^*$ we get by weak transitivity that $T = S^*$, and so $T \longrightarrow S \longrightarrow S^* = T$. In this case, we can take $T' := S$. Indeed, since we already know that $T \longrightarrow S \ni \varphi$, to finish the proof we only need to check that S is φ-final. For this, let $U \in \Omega$ be any theory with $S \longrightarrow U \ni \varphi$; we need to show that $U \longrightarrow S$. From $S^* = T \longrightarrow S \longrightarrow U$, we obtain by weak transitivity that either $U = S^* = T \longrightarrow S$ (and we are done), or $S^* \longrightarrow U \ni \varphi$. In the second case, by the φ-finality of S^*, we have $U \longrightarrow S^* = T \longrightarrow S$; by weak transitivity, we obtain either $U \longrightarrow S$ (and we are done) or $U = S \longrightarrow U = S$. So, in all cases, we concluded that $U \longrightarrow S$, as desired.

The converse follows directly from the Canonical $\Diamond$-Witness Lemma B.2, as a special case. □

Next, we will show that an analogue of the Witness Lemma for $\odot$ does hold in the Σ-final model (unlike in the canonical model). In fact, for finite Σ, we will prove a strong version of this lemma, in which we replace the infinite chain of ϕ-theories witnessing a formula of the form $\odot\phi \in T$ (according to the semantic clause for $\odot$) with a *very special kind of infinite chain*: a "witnessing cluster" $T \longrightarrow T' \longleftrightarrow T''$ with $\varphi \in T \cap T' \cap T''$. Our goal is to prove a $\odot$-Witness Lemma for final theories, that uses the witnessing-cluster condition. For this, we first need two following preliminary results.

Lemma B.9. *If $T \in \Omega$ is θ-final, then it is also $\widehat{\mathcal{K}}\theta$-final.*

Proof. Assume T is θ-final. To show that it is also $\widehat{\mathcal{K}}\theta$-final, observe that we have $\widehat{\mathcal{K}}\theta \in T$ (since $\theta \Longrightarrow \widehat{\mathcal{K}}\theta$ is a theorem in our logic). Second, let $S \in \Omega$ be s.t. $T \longrightarrow S$ and $\widehat{\mathcal{K}}\theta \in S$, and we need to prove that $S \longrightarrow T$. Since $\widehat{\mathcal{K}}\theta \in S$, we have either $\theta \in S$ or $\Diamond\theta \in S$. In the first case, from $T \longrightarrow S \ni \theta$ and the fact that T is θ-final, we conclude that $S \longrightarrow T$, as desired. In the second case, from $\Diamond\theta \in S$ we infer (by the Canonical $\Diamond$-Witness Lemma) that there exists $S' \in \Omega$, with $S \longrightarrow S' \ni \theta$. Since $T \longrightarrow S \longrightarrow S'$, by weak transitivity we have either $T = S'$ or $T \longrightarrow S'$. If $T = S'$, then we conclude $S \longrightarrow S' = T$, and we are done. If $T \longrightarrow S'$, then since T is θ-final and $\theta \in S'$, we get $S' \longrightarrow T$. Thus we have $S \longrightarrow S' \longrightarrow T$, hence by weak transitivity we get that either $S \longrightarrow T$ (and we are done) or $S = T$ (in which case $S = T \longrightarrow S = T$, so we again obtain $S \longrightarrow T$, as desired). □

Lemma B.10. *Let T, T', T'' be theories, and φ, θ be formulas, such that: $T \longrightarrow T' \longleftrightarrow T''$, T' is θ-final, and $\varphi \in T \cap T' \cap T''$. Then $\odot\varphi \in T$.*

Proof. Since T' is θ-final, we have $\theta \in T'$, and so also $\widehat{\mathcal{K}}\theta \in T'$. Note also that, by the Canonical $\Diamond$-Witness Lemma B.2, $T'' \longrightarrow T' \ni \theta$ implies that $\Diamond\theta \in T''$, hence also $\widehat{\mathcal{K}}\theta \in T''$. Putting these facts together with $\varphi \in T \cap T' \cap T''$, we conclude that $(\widehat{\mathcal{K}}\theta \wedge \varphi) \in T', T''$.

To prove our lemma, we first show the following

Claim: $\mathcal{K}((\widehat{\mathcal{K}}\theta \wedge \varphi) \Rightarrow \Diamond(\widehat{\mathcal{K}}\theta \wedge \varphi)) \in T'$.

To prove this claim, we need to show two facts: (1) $((\widehat{\mathcal{K}}\theta \wedge \varphi) \Rightarrow \Diamond(\widehat{\mathcal{K}}\theta \wedge \varphi)) \in T'$; and (2) $\Box((\widehat{\mathcal{K}}\theta \wedge \varphi) \Rightarrow \Diamond(\widehat{\mathcal{K}}\theta \wedge \varphi)) \in T'$.

Proof of fact (1): From $(\widehat{\mathcal{K}}\theta \wedge \varphi) \in T''$ and $T' \longrightarrow T''$, we obtain $\Diamond(\widehat{\mathcal{K}}\theta \wedge \varphi)) \in T'$ (by Lemma B.2), and the desired conclusion follows by basic laws of propositional logic.

Proof of fact (2): by the Canonical $\Diamond$-Witness Lemma B.2, it is enough to show that $\forall S \in \Omega$, if $T' \longrightarrow S$ and $\widehat{\mathcal{K}}(\theta \wedge \varphi) \in S$, then $\Diamond(\widehat{\mathcal{K}}\theta \wedge \varphi) \in S$. To check this, let S be s.t. $T' \longrightarrow S \ni (\widehat{\mathcal{K}}\theta \wedge \varphi)$. Since T' is θ-final, by Lemma B.9 it is also $\widehat{\mathcal{K}}\theta$-final; from this, together with $T' \longrightarrow S \ni \widehat{\mathcal{K}}\theta$, we obtain that $S \longrightarrow T'$. Using this together with the fact that $(\widehat{\mathcal{K}}\theta \wedge \varphi) \in T'$, and applying again the Canonical $\Diamond$-Witness Lemma B.2, we conclude that $\Diamond(\widehat{\mathcal{K}}\theta \wedge \varphi) \in S$, as desired.

Using now the above Claim and the Induction Axiom, we conclude that $\odot(\widehat{\mathcal{K}}\theta \wedge \varphi) \in T'$. Using the Montonicity rule we see that $\odot\varphi \in T'$. Since $T \longrightarrow T'$, we get $\Diamond\odot\varphi \in T$ (again by Lemma B.2), and since $\varphi \in T$, we have $(\varphi \wedge \Diamond\odot\varphi) \in T$. Finally, using Induction and Monotonicity we see that $\vdash (\varphi \wedge \Diamond\odot\varphi) \Rightarrow \odot\varphi$ is a theorem in our system. We conclude that $\odot\varphi \in T$, as desired. □

Now we can prove the following (strong version of) $\odot$-Witness Lemma:

Lemma B.11 (Final $\odot$-Witness Lemma). *Let Σ be a finite set of formulas, φ be a formula with $\varphi \in \Sigma$, and T be a Σ-final theory such that $\varphi \in T$. The following are equivalent:*

(1) $\odot\varphi \in T$;

(2) *there exist $\odot\varphi$-final theories T', T'', with $T \longrightarrow T' \longleftrightarrow T''$;*

(3) *there exist Σ-final theories T', T'', with $T \longrightarrow T' \longleftrightarrow T''$ and $\varphi \in T', T''$;*

(4) *there exists an infinite chain of Σ-final theories $T = T_0 \longrightarrow T_1 \longrightarrow \ldots T_n \longrightarrow \ldots$, such that $\varphi \in T_n$ for all n.*

Proof. (1) $\Rightarrow$ (2): Assume $\odot\varphi \in T$. By the Final Lemma B.7, there exists some $\odot\varphi$-final theory T' such that $T \longrightarrow T'$ and $\odot\varphi \in T'$. Since $\vdash \odot\varphi \implies \Diamond\odot\varphi$ is a theorem of our logic, we must have $\Diamond\odot\varphi \in T'$. By the Final $\Diamond$-Witness Lemma B.8, there exist some $\odot\varphi$-final theory T'' such that $T' \longrightarrow T''$ and $\odot\varphi \in T''$. The fact that T' is $\odot\varphi$-final ensures that $T'' \longleftrightarrow T'$, as desired.

(2) $\Rightarrow$ (3): It is obvious that (3) is a weaker statement.

(3) $\Rightarrow$ (1): Assume given Σ-final theories T', T'', with $T \longrightarrow T' \longleftrightarrow T''$ and $\varphi \in T', T''$. Then there exists some $\theta \in \Sigma$ such that T' is θ-final. Apply Lemma B.10 to obtain the desired conclusion.

(3) $\Rightarrow$ (4): Obvious again. For all $n \geq 1$, just take $T_{2n-1} := T'$ and $T_{2n} := T''$.

(4) $\Rightarrow$ (3): Let $T = T_0 \longrightarrow T_1 \longrightarrow \ldots T_n \longrightarrow \ldots$ be an infinite chain of Σ-final theories, such that $\varphi \in T_n$ for all n. By the Finite Depth Lemma B.6, this cannot be a strict chain (-in fact even its initial segment of length $|\Sigma| - 1$ must be non-strict): so there exist indices $n' < m'$ such that $T \longrightarrow^* T_{n'} \longleftrightarrow T_{m'}$. We need to prove now the stronger statement (3). If we have $T = T_{n'}$, then we get $T = T_{n'} \longrightarrow T_{m'} \longrightarrow T_{n'} = T$, so by taking $n := m'$ and $m := n'$, we obtain $T \longrightarrow T_n \longleftrightarrow T_m$, as desired. If however we have $T \neq T_{n'}$, then we get $T \longrightarrow T_{n'} \longleftrightarrow T_{m'}$, so by taking $n := n'$ and $m := m'$, we reach again the desired conclusion. □

We have now all the ingredients to immediately prove a Truth Lemma for the final model (and thus our completeness result). But, for later use in the decidability proof, it is convenient to state a more general form of this Truth Lemma, by abstracting the relevant properties of the final model into a definition: we consider submodels of the final model satisfying closure properties that are (the syntactic counterpart of the existential parts of) the above Final $\Diamond$- and $\odot$-Witness Lemmas.

Definition B.12 (Perfect Submodels). *A submodel of the Σ-final model Ω^Σ is* perfect *if the underlying set $M \subseteq \Omega^\Sigma$ satisfies the following two conditions:*

(1) *for every theory $T \in M$ and every formula $\Diamond\varphi \in T \cap \Sigma$, there exists some φ-final theory $T' \in M$ with $T \longrightarrow T'$;*

(2) *for every theory $T \in M$ and every formula $\odot\varphi \in T \cap \Sigma$, there exist $\odot\varphi$-final theories $T', T'' \in M$ with $T \longrightarrow T' \longleftrightarrow T''$, and*

(3) *for every theory $T \in M$ and every formula $\widehat{K}\varphi \in T \cap \Sigma$, there exists some φ-final theory $T' \in M$ with $T \sim T'$.*

Examples: Lemmas B.8 and B.11 show that *the Σ-final model is a perfect submodel* (of itself). Later, for our decidability proof, we will see examples of *finite* perfect models.

The key result underlying our completeness and decidability proofs is the following.

Lemma B.13 (Truth Lemma). *Let Σ be a finite set of formulas, closed under subformulas, and let $\mathbf{M} = (M, \longrightarrow, \| \cdot \|)$ be a perfect submodel of the Σ-final model Ω^Σ. Then for all formulas $\varphi \in \Sigma$, we have:*

$$\|\varphi\|_{\mathbf{M}} = \{T \in M : \varphi \in T\}$$

Proof. By structural induction on φ. The atomic case and Boolean cases are standard, so we consider only the modal cases.

The case $\varphi := \Diamond\psi$: For one direction, assume that $\Diamond\psi \in T$. By condition (1) in the definition of perfect submodels, there exists some theory $T' \in M$ with $T \longrightarrow T'$ and $\psi \in T'$. By the induction hypothesis, we get $T' \models_{\mathbf{M}} \psi$, and hence $T \models_{\mathbf{M}} \Diamond\psi$, as desired.

For the converse, assume that $T \models_{\mathbf{M}} \Diamond\psi$. By the semantics, there must exist $T' \in M$ with $T \longrightarrow T'$ and $T' \models_{\mathbf{M}} \psi$. By the induction hypothesis, we get $\psi \in T'$, and so we conclude that $\Diamond\psi \in T$ (by the Canonical $\Diamond$-Witness Lemma B.2), as desired.

The case $\varphi := \widehat{K}\psi$: This case is analogous, but using the $\widehat{K}$-Witness Lemma.

The case $\varphi := \odot\psi$: For one direction, assume that $\odot\psi \in T$. By condition (2) in the definition of perfect submodels, there exist theories $T', T'' \in M$ with $T \longrightarrow T' \longleftrightarrow T''$ and $\odot\psi \in T', T''$. From $\odot\psi \in T, T', T''$, we obtain $\psi \in$

T, T', T'' (by the Fixed Point Axiom), and hence (by the induction hypothesis) we have that T, T' and T'' satisfy ψ in the model $\mathbf{M}$. But then the infinite sequence $T \longrightarrow T' \longrightarrow T'' \longrightarrow T' \longrightarrow T'' \longrightarrow \ldots$ shows that $T \models_{\mathbf{M}} \odot\psi$.

For the converse, assume that $T \models_{\mathbf{M}} \odot\psi$. By definition, there must exist an infinite chain $T = T_0 \longrightarrow T_1 \longrightarrow \ldots \longrightarrow T_n \longrightarrow \ldots$, with $T_n \in M$ (hence, T_n is Σ-final) and $T_n \models_{\mathbf{M}} \psi$ for all n. By the induction hypothesis, we get $\psi \in T_n$ for all n. Applying the Final $\odot$-Witness Lemma B.11, we conclude that $\odot\psi \in T$, as desired. □

We can now finish our completeness proof.

Proof of Theorem 3.3 (Weak Completeness): Fix a consistent formula θ, and let Σ be the (finite) set consisting of θ as well as all subformulas of θ. Fix a Σ-final theory $T_0 \in \Omega^\Sigma$ with $\theta \in T_0$ (-such a theory exists by the Lindembaum Lemma combined with the Final Lemma B.7), and consider the canonical model $\mathbf{\Omega} = (\Omega, \longrightarrow, \| \cdot \|)$ for T_0. Since $\theta \in T_0 \in \Omega^\Sigma$ and the Σ-final model Ω^Σ is perfect, we can apply to it the Final Truth Lemma B.13 to conclude that $T_0 \models \theta$ in Ω^Σ. Hence, our axiomatic system is complete for the class of weakly-transitive relational models. By Lemma A.1, we can add irreflexivity, so the system is also complete for the class of irreflexive and weakly-transitive relational models. But, as already mentioned, this class coincides with the class of Alexandroff topo-models, so the system is also complete for topo-models.

For decidability and FMP, we need to do a bit more work.

C Proof of decidability

Given Theorem 3.2, it is enough to prove the decidability of the static logic $\mathcal{L}$.

Definition C.1 (Defects and defect-depth). *If a submodel (determined by a set) $M \subseteq \Omega^\Sigma$ is* not *perfect, then every pair $(T, \Diamond\varphi) \in M \times \Sigma$ providing a counterexample to the clause (1) of the definition of perfect models is called a* $\Diamond$-defect *of M. A* correction *of the defect $(T, \Diamond\varphi)$ is a φ-final theory $T' \in \Omega^\Sigma$ with $T \longrightarrow T'$. Similarly, every counterexample $(T, \odot\varphi) \in M \times \Sigma$ to the clause (2) of the same definition is called a* $\odot$-defect *of M. A a* correction-pair *of the defect $(T, \odot\varphi)$ is a pair (T', T'') of $\odot\varphi$-final theories with $T \longrightarrow T' \longleftrightarrow T''$.*

The defect-depth $\mathrm{ddpt}(M)$ *of the submodel (determined by) M is the maximum depth of the defects of M, defined as*

$$\max\{\mathrm{dpt}(T) : (T, \varphi) \textit{ is a defect of } M \textit{ for some } \varphi \in \Sigma\}.$$

By the Finite Depth Lemma B.6, we have $0 \leq \mathrm{ddpt}(M) \leq |\Sigma| - 1$ *(for all final submodels M).*

Obviously, M is perfect iff it has no defects (of any of the two kinds). To prove FMP, it is clear that it is enough to show the following:

Lemma C.2. *Let Σ be a finite set of formulas closed under subformulas and under the additional clause: if $\odot\psi \in \Sigma$ then $\Diamond\odot\psi \in \Sigma$. For any Σ-final theory T_0, there exists a finite perfect submodel $M \subseteq \Omega^\Sigma$ with $T_0 \in M$.*

Proof. We recursively construct an infinite sequence of finite submodels

$$M_0, M_1, \ldots, M_n, \ldots$$

of the Σ-final model Ω^Σ:

- For each $\widehat{K}\varphi \in T_0 \cap \Sigma$, choose $T_\varphi \in \Omega^\Sigma$ such that $\varphi \in T_\varphi$. Put $M_0 := \{T_0\} \cup \{T_\varphi : \widehat{K}\varphi \in T_0 \cap \Sigma\}$.
- Given M_n, put $M_{n+1} := M_n$ if M_n is perfect. Otherwise, for each defect $(T, \Diamond\varphi)$, choose a correction $T' \in \Omega^\Sigma$; we'll refer to T' as the *designated correction* of that defect. Similarly, for each defect $(T, \odot\varphi)$, choose a correction-pair (T', T''); we'll refer to T' and T'' as the *designated corrections* of that defect. Define Δ_n to be the set of all $S \in \Omega^\Sigma$ which are a designated correction of some defect of M_n. Then, let $M_{n+1} = M_n \cup \Delta_n$.

By induction, it is clear that *all models M_n are finite* (-the induction step uses the fact that a finite model has only finitely many defects, since Σ is finite).

Claim 1. If (S, φ) is a defect of M_{n+1}, then $\varphi \in S \in M_{n+1} \setminus M_n$, and S is a designated correction of some defect (T, ψ) of M_n, with $T \longrightarrow S$.

Proof of Claim 1. Let (S, φ) be a defect of M_{n+1}. By definition, we then have $\varphi \in S \in M_{n+1}$. Suppose that we also have $S \in M_n$. We consider two cases. *Case 1*: assume that (S, φ) is *also a defect of* M_n. Then, by the construction of M_{n+1}, this defect has a designated correction $S' \in M_{n+1}$, or else a correction-pair $(S', S'') \in M_{n+1}$. But then the theory S', or the pair (S', S''), testify that (S, φ) is *not* a defect of M_{n+1}, contradicting our initial premise. *Case 2:* assume that (S, φ) is *not* a defect of M_n. In this case, there must exist witnesses $T' \in M_n$ or $(T', T'') \in M_n \times M_n$ attesting that (S, φ) is not a defect of M_n. But since $M_n \subseteq M_{n+1}$, the same witnesses also attest that (S, φ) is not a defect of M_{n+1}), again contradicting our initial premise.

So we must have $S \in M_{n+1} \setminus M_n$. But then (by the construction of M_{n+1}) S must be a designated correction to some defect (T, ψ) of M_n. If (T, ψ) is a $\Diamond$-defect, then (by the definition of its corrections) we have $T \longrightarrow S$, and we are done. If (T, ψ) is a $\odot$-defect, then S is part of a correction pair (S, S') or (S', S). In the first case (by the definition of correction pairs) we have $T \longrightarrow S$, and we are done; in the second case, we have $T \longrightarrow S' \longleftrightarrow S$, which together with the fact that $T \neq S$ (since $T \in M_n$ and $S \in M_{n+1} \setminus M_n$) gives again $T \longrightarrow S$ (by weak transitivity).

Claim 2. If M_{n+2} is not perfect, then $\mathrm{ddpt}(M_{n+2}) < \mathrm{ddpt}(M_n)$.

Proof of Claim 2. Let (T_{n+2}, φ_{n+2}) be a defect of M_{n+2}. Then, by applying Claim 1 twice, there must exist some defect (T_n, φ_n) of M_n (with $T_n \in M_n$), as well as some defect (T_{n+1}, φ_{n+1}) of M_{n+1}, s.t. $T_{n+2} \in M_{n+2} \setminus M_{n+1}$ is a designated correction of (T_{n+1}, φ_{n+1}) (and hence $T_{n+1} \longrightarrow T_{n+2}$), and similarly $T_{n+1} \in M_{n+1} \setminus M_n$ is a designated correction of (T_n, φ_n) (hence $T_n \longrightarrow T_{n+1}$)). Also, from $T_{n+2} \in M_{n+2} \setminus M_{n+1}$, $T_{n+1} \in M_{n+1} \setminus M_n$ and $T_n \in M_n$, we infer that $T_{n+2} \neq T_{n+1} \neq T_n \neq T_{n+2}$.

By Lemma B.5, from $T_n \longrightarrow T_{n+1} \longrightarrow T_{n+2}$ we obtain that $\mathrm{dpt}(T_n) \geq \mathrm{dpt}(T_{n+1}) \geq \mathrm{dpt}(T_{n+2})$, with one of the inequalities being strict if either $T_{n+1} \not\longrightarrow T_n$ or $T_{n+2} \not\longrightarrow T_{n+1}$, in which case we get $\mathrm{ddpt}(M_{n+2}) < \mathrm{ddpt}(M_n)$, hence $\mathrm{ddpt}(M_{n+2}) < \mathrm{ddpt}(M_n)$, and we are done. So from now on we can assume that $T_n \longleftrightarrow T_{n+1} \longleftrightarrow T_{n+2}$, which together with $T_n \neq T_{n+2}$ gives us $T_n \longleftrightarrow T_{n+2}$ (by weak transitivity).

To prove Claim 2, we look at the shape of the defect (T_{n+2}, φ_{n+2}), considering the two possible cases:

Case 1: (T_{n+2}, φ_{n+2}) *is a $\Diamond$-defect*, say $\varphi_{n+2} = \Diamond\theta \in T_{n+2}$. Since $T_n \longleftrightarrow T_{n+2}$, we have $\Diamond\Diamond\theta \in T_n$. By our axioms, we have either $\theta \in T_n$, or $\Diamond\theta \in T_n$. In the first case, we have $T_{n+2} \longrightarrow T_n \ni \theta$ and $T \in M_n \subseteq M_{n+2}$, which gives a witness in M_{n+2} for clause (1) applied to $\varphi_{n+2} = \Diamond\theta \in T_{n+2}$, contradicting the assumption that (T_{n+2}, φ_{n+2}) is a defect of M_{n+2}. In the second case, if $\Diamond\theta \in T_n \in M_n$, then by our construction there must exist some $S_{n+1} \in M_{n+1}$ with $T_n \longrightarrow S_{n+1} \ni \theta$ (either because $(T_n, \Diamond\theta)$ was *not* a defect of M_n hence such a theory S_{n+1} already existed in M_n, or else because the defect $(T_n, \Diamond\theta)$ has a designated correction in M_{n+1}). But $T_{n+2} \longleftrightarrow T_n \longrightarrow S_{n+1}$ implies that $T_{n+2} \longrightarrow^* S_{n+1}$, which together with the fact that $T_{n+2} \neq S_{n+1}$ (since $S_{n+1} \in M_{n+1}$ while $T_{n+2} \in M_{n+2} \setminus T_{n+1}$) gives us that $T_{n+2} \longrightarrow S_{n+1}$ (by weak transitivity). So we again conclude that $\theta \in S_{n+1} \in M_{n+1} \subseteq M_{n+2}$ is a witness in M_{n+2} for clause (1) applied to $\varphi_{n+2} = \Diamond\theta \in T_{n+2}$, contradicting the assumption that (T_{n+2}, φ_{n+2}) is a defect.

Case 2: (T_{n+2}, φ_{n+2}) *is a $\odot$-defect*, say $\varphi_{n+2} = \odot\theta \in T_{n+2}$. Since $T_n \longleftrightarrow T_{n+2}$, we have $\Diamond\odot\theta \in T_n \in M_n$. Since $\odot\theta \in \Sigma$ (because this is a defect of M_{n+2}), we also have $\Diamond\odot\theta \in \Sigma$ (by the additional closure requirement of our lemma). By construction, there must exist some $S_{n+1} \in M_{n+1}$ with $T_n \longrightarrow S_{n+1} \ni \odot\theta$ (again either because $(T_n, \Diamond\odot\theta)$ was *not* a defect so that such a theory S_{n+1} already existed in M_n, or because the defect $(T_n, \Diamond\odot\theta)$ has a designated correction in M_{n+1}). But then we can repeat this argument on $(S_{n+1}, \odot\theta)$; by construction, there must exist $\odot\theta$-final $S'_{n+2}, S''_{n+2} \in M_{n+2}$, with $S_{n+1} \longrightarrow S'_{n+2} \longleftrightarrow S''_{n+2}$ (again either because such theories already existed in M_{n+1}, or because the defect $(S_{n+1}, \odot\theta)$ has designated corrections in M_{n+2}). From $T_{n+2} \longleftrightarrow T_n \longrightarrow S_{n+1} \longrightarrow S'_{n+2}$, we obtain that $T_{n+2} \longrightarrow^* S'_{n+2} \longleftrightarrow S''_{n+2}$. If $T_{n+2} \longrightarrow S'_{n+2}$, then the pair (S'_{n+2}, S''_{n+2}) gives witnesses in M_{n+2} for clause (2) applied to $\varphi_{n+2} = \odot\theta \in T_{n+2}$, thus contradicting the assumption that (T_{n+2}, φ_{n+2}) is a defect of M_{n+2}. On the other hand, if $T_{n+2} \not\longrightarrow S'_{n+2}$, then by weak transitivity we must have $T_{n+2} = S'_{n+2} \longrightarrow S''_{n+2} \longleftrightarrow S'_{n+2}$, and so the pair (S''_{n+2}, S'_{n+2}) gives again witnesses in M_{n+2} for clause (2) applied to $\varphi_{n+2} = \odot\theta \in T_{n+2}$, again in contradiction with the assumption that (T_{n+2}, φ_{n+2}) is a defect of M_{n+2}.

Given Claim 2, let now $N := 2 \cdot |\Sigma|$, where $|\Sigma|$ is the size of Σ.[17] We claim that M_N *is a perfect submodel*.

To show this, assume that this is not the case. Then of course none of the submodels M_n with $n \leq N$ are perfect. By repeatedly applying Claim 2, we have

$$\mathrm{ddpt}(M_0) > \mathrm{ddpt}(M_2) > \ldots > \mathrm{ddpt}(M_N).$$

This contradicts the fact that $0 \leq \mathrm{ddpt}(M_0) \leq \mathrm{dpt}(M_0) \leq |\Sigma| - 1$ (by the Finite Depth Lemma B.6): the set $\{0, 1, \ldots, |\Sigma| - 1\}$ has cardinal $|\Sigma|$, so it cannot contain $\frac{N}{2} + 1 = |\Sigma| + 1$ distinct natural numbers. □

Proof of FMP and Decidability: Fix a consistent formula θ, and let Σ be a finite set containing θ, and closed under subformulas and under the additional clause in the previous lemma (if $\odot\psi \in \Sigma$ then $\Diamond\odot\psi \in \Sigma$). Fix as before a Σ-final theory $T_0 \in \Omega^\Sigma$ with $\theta \in T$, and let $\mathbf{M}$ be the finite perfect submodel constructed in the above Lemma. Since $\theta \in T_0 \in M$ and $\mathbf{M}$ is perfect, we can apply the Final Truth Lemma B.13 to conclude that $T_0 \models \theta$ in $\mathbf{M}$. Our submodel $\mathbf{M}$ is a finite weakly transitive relational model, but by the technique in the proof of Lemma A.1, we can convert it into an equivalent model, that is finite, irreflexive and weakly-transitive. But this is nothing but a finite topo-model, so we have proved (strong) FMP for the topological semantics. Decidability immediately follows.

To finish the proof of Theorem 3.4, we need to look at the complexity of the decision problem for the static logic.

D PSPACE completeness

We may obtain a PSPACE complexity bound for the static logic from our decidability proof. First note that the validity problem for $\mathcal{L}$ is PSPACE-hard, as it embeds S4, which is PSPACE-complete (Chagrov and Zakharyaschev 1997). So we focus on the upper bound.

We begin with a PSPACE algorithm for satisfiability in the K-free fragment $\mathcal{L}^{\odot}_{\Diamond}$. First, some preliminary definitions. We work with a set of formulas Σ closed under subformulas, single negations and such that if $\odot\varphi \in \Sigma$, then $\Diamond\odot\varphi \in \Sigma$. For a formula φ, the size of such a Σ containing φ is polynomial on the length of φ. A Σ*-type* is a subset $\Phi \subseteq \Sigma$ such that $\psi \wedge \theta \in \Phi$ implies that $\psi \in \Phi$ and $\theta \in \Phi$, $\psi \vee \theta \in \Phi$ implies that $\psi \in \Phi$ or $\theta \in \Phi$, and $\neg\psi \in \Phi$ if and only if $\psi \notin \Phi$. A Σ *cluster-type* is a multiset $\mathcal{C}$ of Σ-types where each type can occur at most twice and such that if $\Phi, \Psi \in \mathcal{C}$ are such that $\Phi \neq \Psi$:

1. if $\psi \in \Phi$ then $\Diamond\psi \in \Psi$,
2. if $\Diamond\psi \in \Phi$ and $\psi \notin \Phi'$ for any $\Phi' \in \mathcal{C}$ with $\Phi' \neq \Phi$, then $\Diamond\psi \in \Psi$,
3. if $\odot\psi \in \Phi$ then $\odot\psi \in \Psi$, and
4. if $\psi, \neg\odot\psi \in \Phi$ then $\Box\neg\odot\psi \in \Phi$.

A *defect* of $\mathcal{C}$ is either any formula $\Diamond\psi \in \Phi \in \mathcal{C}$ such that $\psi \notin \Phi'$ for any $\Phi' \in \mathcal{C}$ with $\Phi' \neq \Phi$, or a formula $\odot\psi \in \Phi \in \mathcal{C}$ such that there is at most one $\Phi' \in \mathcal{C}$ with $\psi \in \Phi'$. Note that Σ cluster-types represent *irreflexive* clusters.

We will define a generalization of satisfiability whose use of space is easier to control. An instance is a sequence

[17] We can lower this bound somewhat, taking instead the size of the set $\{\varphi : \Diamond\varphi \in \Sigma \text{ or } \odot\varphi \in \Sigma\}$.

$(\varphi, b, \Sigma, \mathcal{K}\psi_1, \ldots, \mathcal{K}\psi_m, n)$, where $\varphi, \Box\psi_i \in \Sigma$, $b \in \{1, 2\}$, and $n \in \mathbb{N}$. We want to check if $\varphi \wedge \mathcal{K}\psi_1 \wedge \ldots \wedge \mathcal{K}\psi_m$ is satisfiable in a model of depth n, whose root cluster has b instantces of φ (so that if $b = 2$, the root cluster witnesses $\odot\varphi$). The original satisfiability problem can be reduced to $(\varphi, 1, \Sigma, |\Sigma|)$, where Σ is the least set containing φ with the required closure properties.

We solve an instance $(\varphi, b, \Sigma, \mathcal{K}\psi_1, \ldots, \mathcal{K}\psi_m, n)$ using the following steps:

1. Choose a cluster type $\mathcal{C}$ such that φ occurs at least b times in $\mathcal{C}$, and every $\Psi \in \mathcal{C}$ has $\psi_i, \Box\psi_i \in \Psi$ for all $i = 1, \ldots, m$. Accept if $\mathcal{C}$ has no defects, and reject if no such $\mathcal{C}$ exists.
2. If $n = 0$, reject. Otherwise, let $\psi'_1, \ldots, \psi'_{m'}$ be all formulas such that $\Box\psi'_i \in \bigcup C$.
   (a) For every defect of $\mathcal{C}$ of the form $\Diamond\theta$, solve the instance $(\theta, 1, \Sigma, \mathcal{K}\psi'_1, \ldots, \mathcal{K}\psi'_{m'}, n-1)$.
   (b) For every defect of $\mathcal{C}$ of the form $\odot\theta$, solve the instance $(\theta, 2, \Sigma, \mathcal{K}\psi'_1, \ldots, \mathcal{K}\psi'_{m'}, n-1)$.
3. Reject if any of the above instances rejects, otherwise accept.

It can be verified by induction on n that $(\varphi, b, \Sigma, \mathcal{K}\psi_1, \ldots, \mathcal{K}\psi_m, n)$ has an accepting computation iff $\varphi \wedge \mathcal{K}\psi_1 \wedge \ldots \wedge \mathcal{K}\psi_m$ is satisfiable on a model of depth $\leq n$ where φ occurs at least b times on the root; indeed, we are simply building a model step-by-step. So it remains to check that the algorithm can be implemented in polynomial space.

First observe that we may restrict the size of each cluster $\mathcal{C}$ to have at most $2|\Sigma|$ elements. This is because each formula needs to occur at most twice, and removing additional Σ-types from $\mathcal{C}$ will not create new defects. Since each Σ-type is $O(|\Sigma|)$ in size, we need $O(2|\Sigma| \cdot |\Sigma|) = O(|\Sigma|^2)$ space to store $\mathcal{C}$. With this in mind, we may prove by induction on n that the algorithm requires $O(n|\Sigma|^2)$: each recursive call in Step 2 uses $O\big((n-1)|\Sigma|^2\big)$, and we may reuse the same space so we do not need additional space for the multiple calls. In addition, we need to store $\mathcal{C}$, which takes $O(|\Sigma|^2)$ space. We may also store the list of defects that have been processed ($O(|\Sigma|)$ space, but this can be avoided if we simply treat defects in some pre-established order). Thus we need $O(|\Sigma|^2 + (n-1)|\Sigma|^2) = O(n|\Sigma|^2)$ space, as claimed.

Finally, we extend the algorithm to the full static language $\mathcal{L}$. This is done as follows: first, let Σ_K be the set of formulas ψ such that $K\psi \in \Sigma$. We non-deterministically choose a set $\Pi = \{\psi_1, \ldots, \psi_m\} \subseteq \Sigma_K$ such that $\neg\varphi \notin \Pi$; these will be the formulas of the form $K\psi$ true in our target model. For each $\theta \in \Sigma \setminus \Pi$, we solve the instance $(\theta, 1, \Sigma, \mathcal{K}\psi_1, \ldots, \mathcal{K}\psi_m, |\Sigma|)$, and accept if all such instances are accepted, otherwise reject. This algorithm is correct since we can amalgamate each model of $\theta \wedge \mathcal{K}\psi_1 \wedge \ldots \wedge \mathcal{K}\psi_m$ to obtain a model of $\big(\bigwedge_{\psi \notin \Pi} \widehat{K}\psi\big) \wedge \big(\bigwedge_{\psi \in \Pi} K\psi\big)$; this amalgamated model will be a model of φ. It is a PSPACE algorithm, since we only need to store the set Π and the list of $\theta \in \Sigma \setminus \Pi$ that have to be treated, which takes $O(|\Sigma|)$ space, in addition to the space already required to solve each instance $(\theta, 1, \Sigma, \mathcal{K}\psi_1, \ldots, \mathcal{K}\psi_m, |\Sigma|)$, which as we have seen is polynomial.

This finishes the proof of Theorem 3.4.