# The Topology of Surprise

Based on joint work with Nick Bezhanishivili & David Fernadez-Duque.

# I. The Surprise Examination Paradox

The Student knows for sure that the date of the exam has been fixed in one of the five (working) days of next week. But he doesn't know in which day.

But then the Teacher announces that the **exam's date will be a surprise**: even in the evening before the exam, the student will still not be sure that the exam is tomorrow.

It is assumed that Teacher **cannot lie**: she *always tells the truth*, and moreover the *Student infallibly knows this*!).

NOTE: There are weaker versions of the puzzle in which Teacher's veracity is in doubt. But such a weakening provides an "easy" way to avoid the paradox.

So in this talk I will focus on the above "strong" version.

# Paradoxical Argumentation

Intuitively, one can prove (by backward induction, starting with Friday) that (given the assumption that he knows the Teacher told the truth) *the exam cannot take place in any day of the week*.

But this is a CONTRADICTION: he also knows that there will definitely be an exam next week! So, using this argument, the student comes to “know” that the announcement is false: the exam CANNOT be a surprise, IN CONTRADICTION with the assumption that the Teacher told the truth!

NOTE: In the ‘weakened’ versions, Student has a way out: concluding that the Teacher lied, he dismisses her announcement, and he confidently waits to somehow ‘know’ in advance when the exam comes. (After all, he proved the exam won't be a surprise!)

BUT THEN..., whenever the exam will come (say, on Wednesday) it WILL indeed be a surprise!

# Paradoxical Argumentation step-by-step

The Student can **prove that, if Teacher didn't lie, then the exam will NOT be on Friday**:
since if it was on Friday, he'd know it on Thursday evening (since there will be no other mornings left by Friday), and so there'd be no surprise (contrary to Teacher's statement).

Now there are *only 4 possible days left* (Monday-Thursday).

**Can the student repeat the above argument to exclude Thursday as well**? IF so, then he can *keep iterating the argument* to **eliminate all days**, thus reaching a *contradiction*!

*Is he entitled to do this*?

Well..., it **all depends** on how we interpret Teacher's statement!

# Two Interpretations of Teacher's Statement

Teacher's statement can be interpreted in *two different ways*: a *non-self-referential* way, and *self-referential* one.

# 1. Non-self-referential interpretation

Some authors (e.g. J. Gerbrandy) think that what Teacher meant was:

"*You wouldn't know in advance the day of the exam,* **without any help from me** *(i.e. if you are* **not** *using even the information that I am just announcing now)*".

In this case, the elimination process **stops here**: the Student *can only exclude Friday*. He **cannot reuse** Teacher's statement after this, since *nothing guarantees that Teacher's statement will stay true* **after** *it was announced*.

This is similar to so-called *Moore sentences*: "You don't know it, but you are dirty on your forehead". This sentence may be *true before* it is announced, but *not after* (-since after hearing it, you know you're dirty).

# 2. Self-referential interpretation

But this is NOT what Teacher says: she does NOT refer to some kind of counterfactual future.

She says: “the exam WILL be a surprise”, in the **actual future** (as it will unfold after she says this!)

Hence, most authors adopt a second, self-referential interpretation of Teacher’s announcement. According to this view, what the Teacher means is:

“*You will not know in advance the exam day, period (i.e. even after hearing* **this** *announcement)*!”

This version seems to lead to paradox! (E.g. Gerbrandy thinks this version is a genuine, Liar-like paradox.)

# II. Knowing the Actual World

**Can I ever get to know the actual state of the world**?

Well, it *depends*!

In fact, it depends on three things:

1. **the actual world**: some worlds are knowable, others are not;
2. **what can I observe**: the *topology* of observable evidence;
3. **my background information** (known *a priori*, or obtained from any other source than observations, e.g. *communication*): the **current space** (or subspace, or partition cell).

# Knowing the Answer to a Question

You may think that knowability of the "actual world" sounds like a very metaphysical and unapplied problem.

But in an epistemic model, the "possible worlds" are just abstractions: *the most refined descriptions of the world that are relevant for the given purpose*.

Typically, such a description simply consists of (all) **the answer(s) to the relevant question(s)** (i.e. the questions that are relevant in a given context or situation): e.g. "*when will the exam be*?". The "actual" world consists of the *true* answer(s).

So what we are studying here is just the **possibility of learning the true answer to some question**, if *given enough observable evidence* (and maybe also given some background information).

*"Knowability of the World" = Possibility of Answering the Relevant Question(s).*

# The Epistemic Interpretation of Topology

Given a topological space $(X, \mathcal{T})$, the *points* of the space $X$ represent the **(epistemically) possible worlds** (or "possible states"): all *possible descriptions of the actual world*, that are compatible with the agent's background information.

A proposition is '**known**' if it is *true in all possible worlds*.

The **open sets** in $U \in \mathcal{T}$ represent the agent's **potential evidence** about the world: at world $x$, a neighborhood of $x$ (-set $U \in \mathcal{T}$ with $x \in U$) is a *piece of (true) evidence that could/will in principle be 'observed' by the agent (or somehow become available to her) in the future*.

We assume that the agent **will only observe true evidence**: a neighborhood of the point representing the actual world.

*Conversely*, we assume that *every such true observable evidence the agent* **can** *in principle be observed* by the agent.

# Why a Topology?

But why would the family of “evidence” form a topology?

Directly observable properties represent *direct, or “basic”*, *evidence*.

**Derived Evidence**:
**Finite intersections** of basic evidences $B = U_1 \cap \ldots, \cap U_n$ are also (indirect) evidence (based on cumulating observations).

But an arbitrary (possibly infinite) **union** $A = \bigcup_i U_i$ of such evidences is also a (derived) evidence: w

whenever $A$ is true (in some world $x \in A$), then $A$ is entailed by some evidence $U_k \subseteq A$ with $x \in U_k$.

So *open sets* represent properties of the world that are **inherently knowable**, or “*verifiable*”, by observable evidence:

the propositions that *can be known* through observations *whenever they are true*.

# Epistemic updates: moving to a subspace

Given the space $(X, \mathcal{T})$, the act of *learning some (true) piece of information* $A \subseteq X$ can be modeled as an **update**:
the worlds $x \notin A$ are no longer epistemically possible, so they are deleted from the model.

Such an update shrinks the space of possibilities: from the original space $(X, \mathcal{T})$, we move to the **subspace** $(A, \mathcal{T}_A)$, wherr

$$\mathcal{T}_A := \{U \cap A : U \in \mathcal{T}\}$$

is the **subspace topology** (on the subset $A$).

NOTE 1: If $A$ is open ($A \in \mathcal{T}$), then we can interpret the update with $A$ as an **observation** (of some observable evidence). But in general, an agent can also learn information that is **NOT observable by the agent** ($A \notin \mathcal{T}$): for instance, $A$ can be communicated to her by another agent.

NOTE 2: This is learning of "hard facts": information that is guaranteed to be true with absolute certainty.

# Interior as Knowability Operator

Read $x \in \mathrm{Int}(A)$ as "$A$ is **knowable**" in world $x$: *there exists some "true" observable evidence for* $A$ (i.e. some $U \in \mathcal{T}$ with $x \in U$ and $U \subseteq A$).

This is a notion of *knowability through observations*, or more generally *evidence-based* knowability:

$x \in \mathrm{Int}(A)$ holds iff there exists some true evidence ($U \in \mathcal{T}$ with $x \in U$) s.t. $A$ **will be known after observing** $U$: i.e., known in the subspace $(U, \mathcal{T}_U)$.

# Conditional Knowability

If I am given *more background information*, more things may become known or knowable!

Suppose that someone tells me that the real world belongs to a set $A$ (=subset of the space $X$).

**Given this piece of information $A$, $B$ is knowable in world $x$** iff

$$x \in \mathrm{Int}_A(B), \qquad \text{where}$$

$$\mathrm{Int}_A(B) = \{x \in X | \exists U \in \tau(x \in A \cap U \subseteq B)\} = A \cap \mathrm{Int}((X - A) \cup B)$$

is the interior in the *subspace topology* $(A, \tau_A)$*, on* $A$:

*after learning* $A$ *in world* $x$*,* $B$ *will become knowable.*

# Back to Our Question: can one know the world?

When is a **world** $x$ **knowable** (*given some background info* $A$)?

In other words, *when is the proposition* $\{x\}$ *knowable at world* $x$, *given* $A$?

This means that

$$x \in Int_A\{x\},$$

i.e. **there exists some evidence (=open set)** $U$ **s.t.** $\{x\} = A \cap U$.

Equivalently: $\{x\}$ *is open in the subspace topology on* $A$.

Equivalently: $x$ **is isolated in (the subspace topology on)** $A$.

Dually: the world is **unknowable given** $A$ if $x$ is a **limit point of** $A$.

# The Epistemic Interpretation of Cantor Derivative

The **(Cantor) derivative of** $A$ is *the set of its limit points*:

$$d(A) = \{x \in X | \forall U \in \mathbf{T}(x \in U \to (U - \{x\}) \cap A \neq \emptyset)\}.$$

Using our above interpretation, 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** $A$".

# III. Self-reference: Perfect Sets

A **perfect set** is a set $B$ with $d(B) = B$.

A set $B \subseteq A$ is **perfect in** $A$ if it is perfect in the subspace topology on $A$; i.e. if $d_A(B) = B$, where

$d_A(B) := A \cap d(B)$ is the derivative in the subspace $(A, \tau_A)$.

The **perfect core** of $A$ is the *largest perfect (in $A$) subset of $A$*.

The perfect core can be calculated by a classical procedure: the **iterative Cantor-Bendixson process**.

# Cantor-Bendixson process and Cantor-Bendixson rank

For any set $A \subseteq X$, define a transfinite sequence

$$d^0(A) = A,$$

$$d^{\alpha+1}(A) = A \cap d(d^\alpha(A))$$

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

Then this a descending sequence

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

which must reach a **fixed point**; i.e. there exists an ordinal $\alpha$ s.t.

$$d^{\alpha+1}(A) = d^\alpha(A).$$

The smallest such ordinal is the **(Cantor-Bendixson) rank of** $A$.

If $\alpha = rank(A)$ is the Cantor-Bendixson rank of a set $A$, then **the above fixed point** $d^\alpha(A)$ **is the perfect core of** $A$.

# Notation for Perfect Core

We denote the perfect core of a set $A$ by $d^\infty(A)$.

Formally, the largest fixed point definition can be encoded in a mu-calculus definition:

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

Here, $\nu P$ denotes the "*largest fixed point*" construction, so this says that $d^\infty(A)$ is the largest fixed point of the set-operator (acting on subsets of $X$) given by

$$P \;\mapsto\; A \cap d(P).$$

# 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 $d^{\infty}(A)$ can thus be understood as the *self-referential version of Cantor's derivative*: in English, it can be rendered as the statement

> "$A$ **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.

# IV. The logic of derivative and perfect core

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

**Semantics** uses **topo-models** $M = (X, \tau, \| \bullet \|)$: a topological space $(X, \tau)$ with a valuation function (mapping every atomic sentence $p$ into a subset $\|p\| \subseteq X$).

**Recursive definition** of $\|\varphi\|_M \subseteq X$: the usual clauses for Booleans; knowledge is the global universal modality, $\Diamond$ is the derivative modality, $\Diamond^\infty$ is the perfect core modality, and $\langle\varphi\rangle\psi$ is the update modality (that evaluates $\psi$ in the subspace $\|\varphi\|$):

$$\|K\varphi\|_M = X \text{ iff } \|\varphi\|_M = X, \quad \text{and } \|K\varphi\|_M = X \text{ otherwise;}$$

$$\|\Diamond\varphi\|_M = d(\|\varphi\|_M); \quad \| \odot \varphi\|_M = d^\infty(\|\varphi\|_M);$$

$$\|\langle\varphi\rangle\psi\|_M = \|\psi\|_{M|\varphi}, \quad \text{where}$$

$M|\varphi$ is the relativization (restriction) of $M$ to the subspace $\|\varphi\|_M$.

# Abbreviations

$$\Box\varphi \;:=\; \neg\Diamond\neg\varphi,$$

$$\mathcal{K}\varphi \;:=\; \varphi \wedge \Box\varphi.$$

$\mathcal{K}$ is "**knowabilty**", and matches the *interior operator*:

$$\|\mathcal{K}\varphi\| = Int(\|\varphi\|).$$

$$\widehat{K}\varphi \;:=\; \neg K\neg\varphi$$

is **epistemic possibility**, modeled by the *global existential modality*.

# Sound and Complete Axiomatization

- Axioms and Rules of Propositional Logic
- Necessitation Rule and Distribution Axiom for $K$, $\Box$ and $[\varphi]$.
- Positive and negative introspection for knowledge:

$$K\varphi \Rightarrow KK\varphi \qquad \neg K\varphi \Rightarrow K\neg K\varphi$$

- Positive Introspection of Knowability: $\mathcal{K}\varphi \ \Rightarrow \ \mathcal{K}\mathcal{K}\varphi$
- 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:

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

$$\langle\varphi\rangle\neg\theta \ \Leftrightarrow \ (\varphi \wedge \neg\langle\varphi\rangle\theta), \qquad \langle\varphi\rangle\widehat{K}\theta \ \Leftrightarrow \ (\varphi \wedge \widehat{K}\langle\varphi\rangle\theta)$$

$$\langle\varphi\rangle\Diamond\theta \ \Leftrightarrow \ (\varphi \wedge \Diamond\langle\varphi\rangle\theta), \qquad \langle\varphi\rangle\odot\theta \ \Leftrightarrow \ \odot\langle\varphi\rangle\theta$$

# Completeness

**Proposition.** (Baltag, Bezhanishvili and Fernandez-Duque)

*The above axiomatic system (for the logic of Cantor derivative and perfect core) is sound and complete wrt the class of all topological models.*

*The logic of Cantor derivative and perfect core has the Finite Model Property and hence it is decidable.*

The **proof** is *non-trivial*.

Filtrations don't seem to work with the perfect core modality in the general case.

Goldblatt and Hodkinson use a vey complicated filtration to deal with "tangled derivative" (a version of the perfect core) in the case when the space satisfies the separation axiom $TD$.
Their method does not seem to work in general.

# Detour: Relational Models

Like the one of Goldblatt and Hodkinson, our proof goes via detour through *relational models*, corresponding a special case of topological spaces: **Alexandroff spaces**.
They is the same as the *standard Kripke semantics* for $\Diamond$ on **relational models** $(X, to)$, with $\to$ *irreflexive and weakly transitive*.

**Weak Transitivity**: if $w \to s \to v$, then either $w = v$ or $w \to v$.

$w \models \odot\varphi$ iff $\exists$ an infinite chain of (not necessarily distinct) worlds

$w = w_0 \to w_1 \to w_2 \to \ldots \to w_n \to \ldots,$ with $w_n \models \varphi$ for all $n$.

In particular, finite topo-models are relational models in this sense.

Goldblatt and Hodkinson only deal with the simpler case of (Alexandroff) topologies that are **TD** : assume $\to$ *transitive*.

# Proof Sketch

We first use reduction axioms to eliminate update modalities.

Then we start from the canonical model $\Omega$ (comprising all maximally consistent theories accessible from some fixed theory): note though that the Truth Lemma fails for our logic $\mathcal{L}$ in the canonical model.

Next, for any given finite set of formulas $\Sigma$, we select a special submodel of the canonical model $\Omega^{\Sigma}$ (called the $\Sigma$-**final model**), consists of $\Sigma$-**final theories**: the ones whose cluster is locally definable by some formula in $\Sigma$.

Showing that the Truth Lemma does hold in $\Omega^{\Sigma}$ for $\Sigma$-formulas requires some work.

# Proof Sketch Continued

Another key ingredient in our proof is the fact that $\Omega^{\Sigma}$ 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 $\Omega^{\Sigma}$ below some fixed **finite** stage of the Cantor-Bendixson process.

Moreover, **this observation can be generalized to the full topological mu-calculus** based on Cantor derivative: *all the required fixed points are reached below some fixed finite stage (that depends only on the set* $\Sigma$.

This can be used to prove **completeness and decidability for the full topological mu-calculus**.

# V. A Topological Analysis of the Surprise Exam

Recall that, when modelling a given problem, our space contains as points or "possible worlds" **all and only the relevant possibilities**: the possible answers to the relevant questions.

Since the only initial question is about the day of the exam, there are *5 such points* $x_1, x_2, x_3, x_4, x_5$, corresponding to the 5 working days of the next week:

e.g. in world $x_1$ the exam will be on Monday, in world $x_2$ the exam will be on Tuesday, etc.

But **what should be the topology**? What is the "potential evidence"?

# The topology of advance (negative) observations

Unhelped by Teacher, Student gathers evidence only by going to class every morning, and seeing if there is an exam or not.

Given the story, we need a topology for which the interior operator corresponds to "knowing **in advance**" (i.e., before the exam takes place).

This means we must consider as evidence **only advance observations** (i.e. only evidence that can be gathered *before* the exam): but these are **negative observations** (=observing that the exam has not taken place today, or in any past day).

So, for our purposes, only the absence of an exam is "observable" before the exam. The exam itself is "not observable" (since then the exam's date will be **trivially knowable**, as it indeed it is *after* the exam starts, but... that is too late).

# The Topo-Model

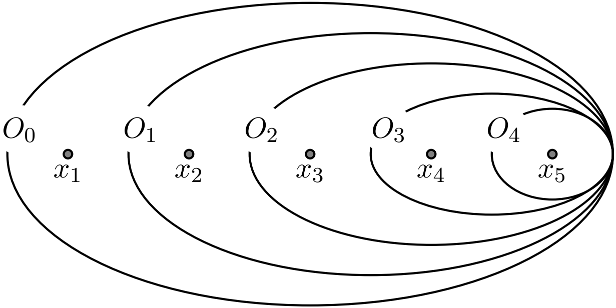


Figure: The initial space $X = O_0$.

Here, $O_0$ corresponds to the trivial observation (before Monday) that the exam will be in one of the 5 days of next week; $O_1$ corresponds to the negative observation after Monday morning: that the exam was not on Monday (hence it will be in one of the remaining four days); etc.

# Passing of a day without exam

When Monday morning passes with no exam, our student observes $O_1 = X - \{x_1\}$ ("no Monday exam"), which thus becomes actual, hard evidence (rather than a potential observation). This **update** $< \neg x_1 >$ corresponds to moving to the subspace $O_1$:

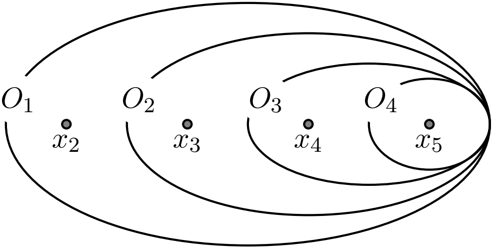


Figure: The subspace $O_1$, obtained after Monday morning.

# Second day passes

If the exam is not on Tuesday, then after Tuesday morning the space is updated with $\neg x_2$, shrinking the space to $O_2$:

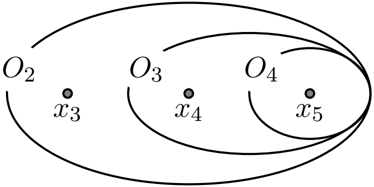


Figure: The subspace $O_2$, obtained after Tuesday morning.

# Next Two Days

The next two updates (in case next mornings pass without exam) yield:

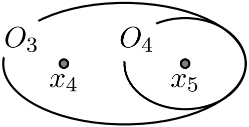


Figure: The space $O_3$, after Wednesday morning.

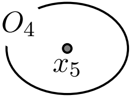


Figure: The space $O_4 = \{x_5\}$ after Thursday morning.

# Surprise: the non-self-referential version

If we adopt Gerbrandy's **non-self-referential interpretation** of the "surprise" announcement, as saying that

> *"(If I haven't said what I am saying now) you'd never know the exam date in advance (without any help from me)"*,

then the Teacher's announcement is formalized as:

$$\textsc{surprise} \;:=\; \Diamond\top,$$

where $\Diamond$ is the derivative modality wrt the evidential topology (and $\top := p \vee \neg p$ is any tautology).

This induces an update $< \Diamond\top >$, which shrinks the initial space $X$ to its Cantor derivative $d(X)$.

# Announcing "surprise": the first derivative

In the initial space $X$ (i.e. before the passing of any day), the Teacher's announcements moves us to the subspace $d(X)$:

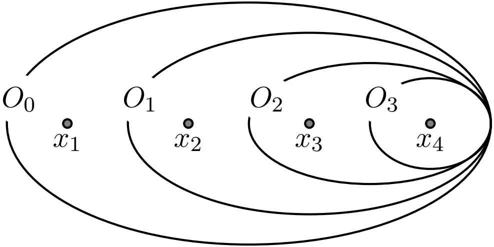


Figure: The subspace $d(X)$, obtained after Teacher's announcement.

# Conclusion

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.

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. 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!

And (according to Gerbrandy) she did NOT mean to claim that the sentence will still be true after it is announced!

# Conclusion Continued

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.

But (according to the above interpretation) **this does not in any way contradict the truthfulness of Teacher's announcement**:

the announcement *was true* at the moment when it was announced.

In this analysis the apparent 'paradox' only points to the *existence of sentences that change their truth value after being announced*.

But this is a well-known (and by now well-understood, and non-paradoxical) fact:

e.g. Moore sentences: "*You don't know that you failed the exam, but you did*".

# Criticism of the non-self-referential interpretation

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 **only counterfactually**: if no such announcement was made).

But then in what sense can one still claim that the Teacher was truthful in her announcement about "will" happen?

For most people, "it will be a surprise" means that it *will* be so (in the actual future), not that it would have been so in some other possible future.

# Surprise: self-referential version

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. even after hearing* **this** *very announcement)".*

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

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

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 $\mu$-calculus notation, we can write the statement as

$$\textsc{surprise}^{\infty} \ := \ \nu P. \langle P \rangle \Diamond \top.$$

# Equivalent Formulation: the perfect core

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 $\textsc{surprise}^{\infty}$ 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 $\textsc{surprise}^{\infty}$.

# Calculating the Perfect Core

If a Teacher who is known never to lie made this announcement $\textsc{surprise}^{\infty} = \odot\top$, that would induce an update that shrinks the original space $X$ to its perfect core $X^{\infty}$.

We can calculate this perfect core in the Surprise Exam topo-model, by the Cantor-Bendixson process, iterating the taking of the derivative, until reach a perfect set.

We can visualize this as a **repeated iteration of the non-self-referential (version of the) announcement**:

Teacher announces $\Diamond\top$ again, to stress that the Student still cannot know the date (even after the first announcement); and then announces it again, etc.

# Second announcement: second derivative

The result of the second announcement of $\Diamond\top$ is the subspace given by the second derivative $d^2(X) = d(d(X))$:

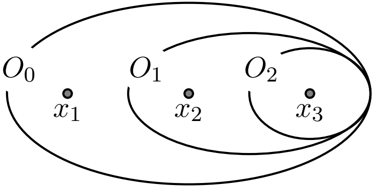


Figure: The subspace $d^2(X)$, after Teacher's 2nd announcement.

while after the next iterations of negative announcements we obtain:

# Next Iterations

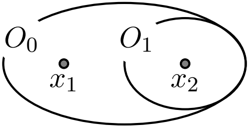


Figure: The subspace $d^3(X)$, after the 3rd negative announcement.

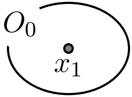


Figure: The space $d^4(X) = \{x_1\}$ after the 4th announcement.

*Next negative answer* yields $d^5(X) = d(\{x_1\}) = \emptyset$: the **fixed point**!

# Paradox?

Since the perfect core is empty, we seemed to have reached **a contradiction**!

Moreover, note that each step of the Cantor-Bendixson (visualised above as a iterated update with non-self-referential statement $\Diamond\top$) corresponds to a step in the intuitive process of elimination (by backward induction) that the underlies Student's reasoning in the Surprise story!

So... is Student's reasoning correct after all?!

And.. is the self-referential version a genuine paradox?!

# Contradiction Resolved

The version in terms of iterated non-self-referential announcements gives away the solution:

if Teacher keeps saying “You will not know”, she **cannot keep telling the truth**: at some point (depending on the actual exam day) her statement is a lie!

The contradiction actually shows that a **non-lying Teacher CANNOT answer “No” 5 times**.

Indeed, depending on the actual day of the exam, Teacher will have to eventually answer “Yes”.

For instance, if the Teacher plans to have the exam on Wednesday, then she will answer “No” only twice.

The third time the question is asked, a truthful Teacher must have answered:

> *“Well, YES, given all the information I already provided you, you should be able to know the exam’s day by the evening before the exam.”*

# Solution to the Paradox

CONCLUSION: Since we assumed Teacher always tells the truth, she **must eventually stop saying "You will not know"**!

This conclusion has consequences for the self-referential version, which compresses all the 5 successive SURPRISE announcements into one equivalent, self-referential announcement: $\text{SURPRISE}^{\infty}$.

An infallibly truth-telling Teacher simply **CANNOT** make the self-referential announcement $\text{SURPRISE}^{\infty}$ involved in the story: since if she did, the announcement would be false (and hence the Teacher would have told a lie).

**But this doesn't mean that the statement $\text{SURPRISE}^{\infty}$ in itself is false for sure, when it is not announced**: all we (and the Student) can prove is that the sentence WOULD be false IF announced by the Teacher.

If *unannounced*, the statement $\text{SURPRISE}^{\infty}$ *may be true or false* (depending on the day of the exam).

# The Source of Apparent Paradoxicality

The apparent paradox was due to the fact that **this fixed point is empty**.

'Paradoxes' such as Surprise Exam occur when the **only perfect subset of $A$ is** $\emptyset$.

**'Non-paradoxical cases' (e.g. our real-number story) occur in the case of non-empty perfect sets**.

A model/space $X$ is said to be **'paradoxical'** iff $d^{\infty}(X) = \emptyset$ iff $\Box^{\infty}\bot$ holds (at any state) in $X$.

# $TD$ implies paradoxicality

All finite $TD$ spaces (satisfying the separation axiom $TD$) are paradoxical.

As a consequence, the logic of $TD$ spaces does NOT have finite model property (since $\Box^{\infty}\bot$ is valid on finite $TD$ spaces, but not on all $TD$ spaces).

Also as a consequence, to represent non-paradoxicality in a finite topo-model, we need a non-$TD$ space.

# Example of non-paradoxical space

The Teacher marks a point $x$ on the real line $R$. He announces the students that the point is in the set

$$A = \{0\} \cup \{\frac{1}{n} : n \in N, n \geq 2\} \cup \{\frac{1}{n} + \frac{1}{n^m} : n, m \in N, n, m \geq 2\} \cup [1, 2].$$

(We assume Teacher cannot lie, so this is hard information.)

QUESTION: If the Student is allowed to do measurements (of any arbitrary precision $> 0$) of the position of the point, *can she know the exact position*?

ANSWER: **NO, if the point is a limit point of $A$**; i.e. if
$x \in d(A) = = \{0\} \cup \{\frac{1}{n} : n \geq 2\} \cup [1, 2]$.
**YES, otherwise**; i.e. if
$x \in A - d(A) = \{\frac{1}{n} + \frac{1}{n^m} : n, m \in N, n, m \geq 2\}$.

# Example Continued

The Teacher (truthfully) announces:

"*Without my help (i.e.* **not** *using the information that I am announcing now), you would never know the exact position of the point* (no matter what measurements you perform)!"

After updating with the announcement, we are in the subspace

$$Lim(A) = d(A) = \{0\} \cup \{\frac{1}{n} : n \geq 2\} \cup [1, 2].$$

QUESTION: Can **now** (after this announcement) the Student know the exact position?

ANSWER: **NO, if $x$ is a limit point of** $d(A)$; i.e. $x \in d(d(A)) =$
$= \{0\} \cup [1, 2]$.
**YES, otherwise**; i.e. if $x = \frac{1}{n}$, for some $n \geq 2$.

# Example Continued

After the announcement, the Teacher (truthfully) announces again: "*Even after hearing my previous announcement, without any* **further** *help (i.e.,* **not** *using the information that I am announcing now), you still would never know the exact position of the point*!"

After updating with the announcement, we are in the subspace

$$d(d(A)) = \{0\} \cup [1, 2].$$

QUESTION: Can **now** (after this new announcement) the Student know the exact position?

ANSWER: **NO, if $x$ is a limit point of $d(d((A))$**, i.e.

$$X \in d(d(d(A))) = [1, 2].$$

**YES, otherwise** (if $x = 0$).

# Example Concluded

Suppose that, once again, the Teacher (truthfully) announces: "*Even after hearing all my previous announcements, without any further help (=***not** *using the information that I am announcing now), you still would never know the exact position of the point*!"

After updating with the announcement, we are in the subspace

$$X \in d(d(d(A))) = [1, 2].$$

**All points are limit points in this set**, so $[1, 2]$ is a "perfect set":

$$d([1, 2]) = [1, 2].$$

QUESTION: Can **now** the Student know the exact position?

# Example Solved

ANSWER: **Definitely NO**!

Moreover, the Student now **knows this for certain** (*that he* **cannot know** the exact position): since it is true regardless of where the point is (as long as in $[1, 2]$).

So the Teacher **can** keep truthfully announcing the same thing again and again, any number of times ("Even after hearing my previous announcement, you'd still never know the exact position). But all these announcements would be **redundant**: the Student already has this information.

We reached a **fixed point**: the set $[1, 2]$ is perfect (i.e. equal to its own derivative).

# Self-Referential Announcement

We could have reached the fixed point faster, if the Teacher made a **self-referential announcement**, right at the start:

after first announcing (as before) that the point is in the set

$$A = \{0\} \cup \{\frac{1}{n} : n \in N, n \geq 2\} \cup \{\frac{1}{n} + \frac{1}{n^m} : n, m \in N, n, m \geq 2\} \cup [1, 2].$$

he (truthfully) announces:
"*You will never know the correct position of the point, period (i.e. even after hearing* **this** *announcement).*"

The effect of this self-referential announcement is to update $A$, by going straight to the largest perfect set included in $A$, i.e. to $[1, 2]$.

Teacher's self-referential announcement was **true**, and it **remains true** (after it was announced). Moreoever, after the announcement the Student **knows all this**:

he now knows that $x \in [1, 2]$, and he knows that he will never know the correct position of $x$, no matter how accurate measurements he performs.

# Conclusions

The self-referential version of the Surprise Exam is NOT a Liar-like paradox.

The sentence $\textsc{surprise}^{\infty}$ has in any case a **definite truth value**, unlike the Liar sentences.

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, and hence **known to be false**).

# Conclusions continued

A Teacher who is known not to lie CANNOT truthfully make the announcement $\textsc{surprise}^{\infty}$ in the situation described in the puzzle (though she CAN do so in other situations, e.g. our numerical example).

But, contrary to 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).

The derivative operator is monotonic, so **the type of self-referentiality involved in the Surprise story is innocuous**!

# “Conclusions concluded”

*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**:

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.

# Multi-Agent Version

**Syntax**:

$$\varphi ::= \quad p \quad | \quad \neg\varphi \quad | \quad \varphi \wedge \varphi \quad | \quad \langle a \rangle \varphi \quad | \quad \Diamond^{\infty} \varphi$$

with $a \in A$ come from a set $A = \{1, \ldots, n\}$ of "agents".
**Semantics**: we interpret this language on **multi-agent topo-models** $M = (X, \tau_1, \ldots, \tau_n, \| \bullet \|)$: $n$ topologies on a set $X$, and a valuation function map.

$$\|\langle a \rangle \varphi\| = d_a(\|\varphi\|)$$

is the Cantor derivative of $\|\varphi\|$ wrt the topology $\tau_a$, and

$$\|\Diamond^{\infty} \varphi\| = \nu P.\|\varphi\| \cap d_1(P) \cap \ldots d_n(P)$$

is the perfect core of $\|\varphi\|$ wrt to the *joint derivative*.
**Open Questions**: Axiomatization? Decidability? FMP?

# Application: The Numbers' Dialogue

Alice and Bob have each a natural number drawn on his/her forehead. Each can see only the other's number, but not his/her own.

*Hard Background Information*: the two numbers are related (one way or the other) by the successor function

$$n \mapsto n+1,$$

i.e. one of them is the immediate successor of the other.

They are asked, repeatedly:

> *"Do you know your own number (on your forehead)?"*

They are supposed to answer *truthfully and simultaneously*, without engaging in any other communication (e.g. no telling to each other the numbers, no waiting for the other's answer before they answer etc).

# The Dialogue Continued

The first time they're asked, they both answer "*I don't know*".
The second time, they both answer "*I don't know*".
But the third time, Alice answers "*Yes, now I know my number*"
(while Bob still says *I don't know*).
QUESTIONS: What are the numbers? And what will Bob say the next time he's asked?

# Topological analysis

This is a $2$-agent topological information frame $(X, \tau_1, \tau_2)$, with

$$X = \{(n_1, n_2) : n_1, n_2 \in N, \text{ with } n_1 = n_2 + 1 \text{ or } n_1 + 1 = n_2\}$$

and with partitional topologies $\tau_1$ for child 1 (Alice), and $\tau_2$ for child 2 (Bob), given by:

$$\tau_1 = \{\emptyset, X\} \cup \{O_1(m) : m \in N\},$$

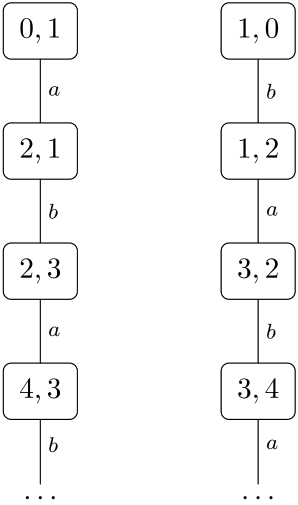
where the observation $O_1(m) = \{(n_1, n_2) \in X : n_2 = m\}$ corresponds to Alice observing that Bob's number is $m$; and similarly

$$\tau_2 = \{\emptyset, X\} \cup \{O_2(m) : m \in N\},$$

where the observation $O_2(m) = \{(n_1, n_2) \in X : n_1 = m\}$ corresponds to Bob observing that Alice's number is $m$.

*Equivalently*: take the standard (S5) Kripke model for this. Basic open sets in Alice's topology $\tau_1$ are the equivalence classes (partition cells) with respect to $a$'s equivalence relation, and the same for $\tau_2$ wrt $b$'s relation.

# Standard (S5 Kripke) Model as a Topology



# Topological analysis continued

A point $x \in X$ is a *common limit point* if it is a limit point in *both* topologies. The *joint derivative* of $A$ is the set of common limit points:

$$d(A) = d_{\tau_1}(A) \cap d_{\tau_2}(A).$$

Each pair of simultaneous answers "*I don't know*" by both agents eliminates the points that are isolated in *either* of the two topologies, thus amounting to taking the *joint derivative* $d(A)$ of the previous subspace $A$. We obtain a sequence of subspaces

$$X = d^0(X) \supseteq d(X) = d^1(X) \supseteq d(d(X)) \supseteq \ldots d^n(X) \supseteq \ldots$$

EXERCISE: *Show that, for every point $x = (n_1, n_2) \in X$, there exists a finite stage $N \geq 1$ s.t. $x \in d^{N-1}(X) - d^N(X)$.* Hence, *the dialogue always finishes in finitely many steps*: at some stage one of the children will know his/her number.

# Paradoxical Version

We can create a paradoxical self-referential version of the story, similar to the Surprise Exam Paradox:
a Teacher (who is guaranteed to tell the truth) announces:

> *"If you don't communicate with each other (or look in the mirror), you will NEVER know the numbers (even after you are hearing this)."*

The two agents will now be forced to **eliminate all numbers**: it is easy to see that

$$d^{\omega}(X) = \bigcap_{n \in N} d^{n}(X) = \emptyset.$$

Once again, the "paradox" is due to the fact that $\emptyset$ is the only perfect set in this topology.

# Parikh's Transfinite Dialogues

Parikh 1992 uses joint derivative to analyze Number Puzzles, and he considers other versions, by **changing the function connecting the two numbers.**

An example is the function $g$, given by:

$$g(n) = 1, \text{ if } n = 2^k \text{ for some } k > 0$$

$$g(n) = n + 2, \text{ if } n \text{ is odd}$$

$$g(n) = n - 2, \text{ otherwise.}$$

This time the dialogue may go **transfinitely**: some points will only become isolated after *a transfinite number of iterations of $d$*!

See Parikh 1992 for details. In fact, it is proved there that: for every computable ordinal $\alpha$ there exists a recursive function $g$, and $g$-related numbers $(n_1, n_2)$ s.t. the above dialogue will stop in exactly $\alpha$ steps.

NOTE: As before, we can also create a paradoxical self-referential version, since $\emptyset$ is still the only perfect set in this topology.

# References

1. R. Parikh (1992), *Finite and Infinite Dialogues*, Logic from Computer Science.
2. G. Bezhanishvili, L. Esakia and D. Gabelaia (2011), *Spectral and $T0$-Spaces in $d$-Semantics*, Lecture Notes in AI.
3. Balbiani & Uridia (2014),*Completeness and Definability of a Modal Logic Interpreted over Iterated Strict Partial Orders*, Advances in Modal Logic 9.
4. Goldblatt & Hodkinson (2017), *Spatial Logic of Tangled Closure Operators and Modal Mu-calculus*, Ann. Pure Appl. Log. 168 (5); (2018) *The Finite Model Property for Logics with Tangle Modality*, Studia Logica 106(1).
5. **Baltag, Bezhanishivili & Fernandez-Duque (2021), The Topological Mu-Calculus, Proceedings of LICS 2021**.
6. **Baltag, Bezhanishivili & Fernandez-Duque (2022), The Topology of Surprise, Proceedings of KR 2022. Awarded the Ray Reiter Best Paper Prize 2022, sponsored by Artificial Intelligence Journal.**