Title
Logical Interrogations of Theory and Evidence

Author
Dale, Reid

Publication Date
2022

Peer reviewed|Thesis/dissertation

Logical Interrogations of Theory and Evidence

by

Reid Dale

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Logic and the Methodology of Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Thomas Scanlon, Co-chair
Professor Wesley Holliday, Co-chair
Professor Leo Harrington

Spring 2022

Logical Interrogations of Theory and Evidence

Abstract

Logical Interrogations of Theory and Evidence

by

Reid Dale

Doctor of Philosophy in Logic and the Methodology of Science

University of California, Berkeley

Professor Thomas Scanlon, Co-chair

Professor Wesley Holliday, Co-chair

In the 1941 edition of *Introduction to Logic and to the Methodology of the Deductive Sciences* [45] Alfred Tarski laments that

> [T]he methodology of empirical sciences constitutes an important domain of scientific research. The knowledge of logic is of course valuable in the study of this methodology, as it is in the case of any other discipline. *It must be admitted, however, that up to the present, logical concepts and methods have not found any specific or fertile applications in this domain.* [45, p. xii] [emphasis mine]

This dissertation aims to partially realize Tarski's project of applying mathematical logic to questions beyond the scope of pure mathematics through an investigation of the relationship between a theory and the evidence that (partially) confirms or refutes it. It is constituted by three projects: *On Falsification*, *On Rational Jurisprudence*, and *On the Sufficiency of First-Order Logic*.

The first major chapter, *On Falsification*, is concerned with the question of when a theory is *refutable* with certainty on the basis of sequence of primitive observations. Beginning with the simple definition of falsifiability as the ability to be refuted by *some* finite collection of observations, I assess the literature on falsification and its descendants along the lines of its *static* and *dynamic* components. The static case is

broadly concerned with the question of how much of a theory can be subjected to falsifying experiments. In much of the literature, this question is tied up with whether the theory in question is axiomatizable by a collection of universal first-order sentences. I argue that this is too narrow a conception of falsification by demonstrating that a natural class of theories of distinct model-theoretic interest—so-called NIP theories—are themselves highly falsifiable. The dynamic case, by contrast, is concerned with the question of how falsifiable a single proposition is in the short and long run. Formal Learning theorists such as Schulte and Juhl [35] have argued that long-run falsifiability is characterized by the topological notion of nowhere density in a suitable topological space. I argue that the short-run falsifiability of a hypothesis is in turn characterized by the VC finiteness of the hypothesis. Crucially, VC finite hypotheses correspond precisely to definable sets in NIP structures, pointing to a robust interplay between the static and dynamic cases of falsification. Finally, I end the chapter by giving rigorous foundations for Mayo's [29] conception of severe testing by way of a combinatorial, non-probabilistic notion of *surprise*. VC finite hypotheses again appear as the hypotheses with guaranteed short-run surprise bounds. Therefore, NIP theories and VC finite hypotheses capture the notion of short-run falsifiability.

The next chapter, *On Rational Jurisprudence*, is concerned with the epistemic question of *confirming* a hypothesis—the guilt of a defendant—by way of testimony heard by a juror over the course of an American-style criminal trial. In it, I attempt to settle a dispute between two strands of the legal community over the issue of whether the methods of Bayesian rationality are incompatible with jurisprudential principles such as the Presumption of Innocence. To this end, I prove a representation theorem that shows that so long as a juror would *not* convict the defendant having heard no testimony (the Presumption of Innocence) but would convict upon hearing *some* collection of testimony (Willingness to Convict), then this juror's disposition to convict the defendant is representable as the disposition of a Bayesian threshold juror in Posner's sense. This result indicates that relevant notion of a Bayesian threshold juror is insufficiently specified to render this debate a substantive one.

Finally, *On the Sufficiency of First-Order Logic* is concerned with the limits of sound inference. The starting point of this chapter is a reflection on a principle that Barwise [3] terms the *First-Order Thesis*, namely that "logic is first-order logic, so that anything that cannot be defined in first-order logic is outside the domain of logic." Barwise was chiefly concerned with the relative inexpressiveness of First-Order Logic. Despite this, I argue that First-Order Logic—while not the most expressive abstract logic—is sufficient to represent and carry out any inference a finitistic agent might

carry out. The main mathematical result here is the Σ_1 completeness of the consequence relation $\models_{FO}$ of First-Order logic as a Turing functional. A consequence of Σ_1 completeness is that *any* notion of logical consequence which is machine verifiable given an oracle naming the theory Γ is in fact able to be internalized into a First-Order proof system. While the translation function will generally not preserve semantics on the nose—after all, First-Order Logic is not particularly expressive—the inferential structure of such a system is captured by a first-order system by way of a computable translation function. I then consider a recent argument of Warren's [50] that agents like us in nearby possible worlds can implement the ω rule of inference, a sound but non-recursive pattern of inference. While I do not refute his premise directly, I do argue that there is a clear finitistic interpretation Warren's purported example ω rule, saving the plausibility of the position that agents like ourselves are only capable of performing finitary inferences. I conclude the chapter by reflecting on *Hilbert's Thesis*, a position constituted by two related theses:

1. Hilbert's Expressibility Thesis (HET): All mathematical (extra-logical) assumptions may be expressed in first-order logic, and

2. Hilbert's Provability Thesis (HPT): The informal notion of *provable* is made precise by the formal notion of *provable in first-order logic*.

Kripke [26] has argued that

$$\text{(HET) + (HPT)} \implies \text{Church's Thesis}$$

on the basis of Gödel's Completeness Theorem. The results of this chapter indicate a partial converse. The Σ_1 completeness of $\models_{FO}$ shows that

$$\text{Church's Thesis + In-Principle Machine Verifiability of Proofs} \implies \text{(HPT)}.$$

First-Order Logic is neither the *only* nor *most expressive* logic with Σ_1 complete entailment relation, but the above argument shows that First-Order Logic is sufficiently expressive to simulate any machine-verifiable inferential system.

To Mike Tarling, who fostered my love of mathematics;
my parents, Douglas and Dawn;
Taylor, Pepper, and Paprika.

Contents

Acknowledgments

Throughout my time at Berkeley I have had the great fortune of being able to work very closely with my advisors Tom Scanlon and Wes Holliday. I am grateful for their enthusiasm in supervising a dissertation expressly setting out to fulfill Tarski's program to apply contemporary results mathematical logic to the foundations and methodology of science. I could not have asked for a better pair of advisors, whose complementary areas of expertise have played such a large part in shaping this project.

I am grateful to so many colleagues who have listened to me drone on about my research over the course of many years; these discussions have played a crucial role in my research process. By chapter, I wish to thank Alex Kruckman, Sridhar Ramesh, Oran Gannot, Silvain Rideau-Kikuchi, and Kentarô Yamamoto for discussions regarding falsification, Taylor Davies-Mahaffey for her insights regarding American criminal procedure and deep knowledge of case law and evidentiary standards, and James Walsh and Kentarô Yamamoto for robust discussions regarding the peculiar properties of First-Order Logic.

On a personal level, I have been so fortunate to have had enduring support from so many of the people in my life. In particular, my parents Douglas and Dawn have *always* been there to take my calls and visit Berkeley when I needed it. My father has always shown interest in my research and has himself become quite interested in falsification as it relates to economic theory. My siblings Kyle, Casey, and Kitt have been similarly gracious. Gayle has taken on the role of my adoptive grandmother; her friendship, humor, and counsel has been appreciated.

Taylor has been particularly patient in the waning hours of my PhD program, likely due to a sense of reciprocity for when she was studying for the California Bar Exam. Our pets Pepper and Paprika have been far less patient.

Bob Dumas, who introduced me to mathematical logic at the University of Washington and himself an alumnus of Berkeley Math, has continued to be a mentor in research and entrepeneurship.

Kentarô, James, Alex, Sridhar, Silvain, and Oran have become some of my closest friends. Well before I came to Berkeley, I was one of "just six boys." The others—Jack, Ryan, Connor, Nick, and Dan—as well the honorary seventh member, Ben, have remained steadfast in their friendship for well over half of my life.

I am grateful for the financial support I received throughout my PhD program, including an NSF Graduate Research Fellowship (grants DGE 1106400 and DGE 1752814), a Logic Group Summer Research Grant, and a Berkeley Connect Fellowship.

Chapter 1

Introduction

In the 1941 edition of *Introduction to Logic and to the Methodology of the Deductive Sciences* [45] Alfred Tarski wrote that

> [T]he methodology of empirical sciences constitutes an important domain of scientific research. The knowledge of logic is of course valuable in the study of this methodology, as it is in the case of any other discipline. *It must be admitted, however, that up to the present, logical concepts and methods have not found any specific or fertile applications in this domain.* [45, p. xii] [emphasis mine]

Nearly eighteen years later, in the preface to the 1959 conference proceedings *The Axiomatic Method* [19], Henkin, Suppes, and Tarski lamented that logicians had yet to conclusively demonstrate the utility of mathematical logic in the methodology of the physical sciences, writing that

> it is possible to maintain that the status of axiomatic investigations in physics is not yet past the preliminary stage of philosophical discussion expressing doubt as to its purpose and usefulness. [19, p. VIII]

Tarski and Henkin—two of the founding members of the Group in Logic and the Methodology of Science at UC Berkeley—recognized the ripe potential that logic had for informing methodology beyond the realm of mathematics.

What were the reasons Tarski and others felt that mathematical logic should play a more active role in empirical sciences? In Tarski's 1944 classic paper, *The Semantic Conception of Truth*, he explains that "the study of scientific language constitutes an essential part of the methodological discussion of a science" and that "semantics may have some bearing on any science whatsoever" [44, p. 690]. While Tarski

does not claim to give a full account of the applications logic to the methodology of empirical sciences, he does suggest that logical methods can yield constraints on the *acceptability* of a scientific theory. For example, he defends a "postulate which can be reasonably imposed on acceptable empirical theories and which involves the notion of truth" that

> as soon as we succeed in showing that an empirical theory contains (or implies) false sentences, it cannot be any longer considered acceptable. [44, p. 691]

For Tarski, a proof of logical inconsistency or a proof that the scientific theory contradicts empirical observation was sufficient to conclude the inadequacy of a scientific theory. In other words, a necessary condition for the adequacy of an empirical science is that it is not refuted *with certainty* by either empirical evidence—observations—or the logical evidence in the form of logical inconsistency. Already in 1944 Tarski recognized the *logical* link between theory and evidence.

This dissertation continues in this Tarskian tradition and is expressly aimed at demonstrating the utility of logic—drawing primarily from the methods of model theory and recursion theory—in studying the relationship between theories and the evidence that (partially) confirms or refutes them.

The three projects contained in this dissertation investigate both facets of evidence in the context of scientific, legal, and general inferential systems. Chapter 2, *On Falsification*, is concerned with questions of how amenable a scientific theory is to refutation on the basis of observable data. Chapter 3, *On Rational Jurisprudence*, is concerned with a confirmation problem, namely the guilt of a defendant in an American criminal trial on the basis of the indirect evidence afforded by witness testimony. Finally, Chapter 4, *On the Sufficiency of First-Order Logic* considers the ability of First-Order Logic to simulate *any* machine-verifiable system of inference, shedding light on the limits of the inferential faculties of finitistic agents.

My goal across all of these projects is to demonstrate how, rather than being an insular and esoteric field, contemporary mathematical logic can yield insights into the modes of evidential reasoning.

1.1 On Falsification

In *On Falsification*, I investigate the landscape of refinements of a very modest notion of falsifiability:

Definition 1.1. Let $\mathbb{K}$ be a class of $\mathcal{L}$ structures. We denote the universal theory of $\mathbb{K}$ as

$$\forall_1(\mathbb{K}) = \{\varphi \mid \varphi \text{ is a } \forall_1 \text{ first-order } \mathcal{L}\text{-sentence and } \mathbb{K} \vDash \varphi\}.$$

We say that $\mathbb{K}$ is *falsifiable* provided that the class of models of $\forall_1(\mathbb{K})$ is nontrivial, i.e.

$$\mathrm{Mod}(\forall_1(\mathbb{K})) \neq Str(\mathcal{L}). \qquad \Diamond$$

According to this definition, a class of $\mathcal{L}$ structures $\mathbb{K}$ semantically entailing a *single* nontrivial universal is sufficient to conclude its falsifiability; after all, such a universal $\forall x\varphi(x) \in \forall_1(\mathbb{K}) \setminus \forall_1(\varnothing)$ specifies a Boolean configuration of atomic sentences, i.e. observations, incompatible with the class $\mathbb{K}$. In other words, if $\mathcal{M} \vDash \exists x\neg\varphi(x)$, then $\mathcal{M} \notin \mathbb{K}$. [1]

Various strengthenings of this notion of falsifiability have been defined and studied, and can by and large be split into two classes:

1. The **Static Case**, wherein we ask "just how falsifiable is the class $\mathbb{K}$?," and
2. The **Dynamic Case**, wherein we ask "how much observation is required to falsify a hypothesis, and how quickly can we expect to falsify it?"

The static models of falsifiability typically concern themselves with questions of how close a theory is to being universally axiomatizable; after all, the more universal sentences a theory implies, in principle the more falsfiable the theory becomes.

Simon and Groen [38], in their work on Ramsification and the Second-Order definability of theories, isolate a notion on *pseudoelementary* classes $\mathbb{K}$ they call FITness which they claim isolates the ideal scientific theories: On their account, a pseudoelementary class $\mathbb{K}$ is FIT if and only if it is an ideal scientific theory. They show that for pseudoelementary $\mathbb{K}$, FITness implies universal axiomatizability.[2] Generalizing their definition to arbitrary classes of $\mathcal{L}$-structures $\mathbb{K}$ closed under isomorphism, I show that for finite languages $\mathbb{K}$ being FIT entails that $\mathbb{K}$ is elementary and, in fact, universally axiomatizable. This result substantially generalizes their result over finite languages.

I then turn my attention to an argument given by Chambers et al. [7] that argues that being universally axiomatizable is not sufficient grounds to call a theory falsifiable. Instead, they identify the falsifiable sentences with a class of universal sentences they call UNCAF (a universal negation of a conjunction of atomic formulas).

[1] Of course, one may define a notion of falsifiability relative to a class $\mathbb{K}'$ by replacing the set of first-order universal validities with the set $\forall_1(\mathbb{K}')$ in the definition above.

[2] The proof presented in their paper is incorrect. However, this error is fixed in this disseration.

In turn, I argue that their argument implicitly assumes that the underlying predicates $P \in \mathcal{L}$ exhibit mere Σ_1 behavior, and thus that their argument reaches too far in its conclusions.

As a final foray into the static case of falsification, I consider how falsification intersects with the dividing lines of classification theory. It is not too difficult to show that under very mild restrictions on the language $\mathcal{L}$, NIP theories entail a great deal of nontrivial $\forall_1$ sentences and are highly falsifiable. Of note, in NIP theories each formula φ is equipped with a notion of dimension known as the VC-dimension of a class, which in a sense measures the effective falsifiability of membership in the class of hypotheses it defines.

On the other hand, recent work of Kruckman and Ramsey [27] and, independently, Jeřábek [21], yield examples of NSOP_1 and simple theories which are *un*falsifiable. While there are many NSOP_1 theories which are falsifiable, in a sense NIP is individuated among the dividing lines in model theory as a class of highly-falsifiable theories.

Moving on to the *dynamic* case of falsification, we turn to the account of falsification given by Formal Learning Theorists Juhl and Schulte [35]. For them, a hypothesis $\mathcal{H}$ is identified with a set of possible worlds. They consider the case where $\mathcal{H} \subset 2^\omega$. For them, such a hypothesis is *always falsifiable* provided that regardless of the results of some finite collection of observations, the hypothesis still has the potential to be falsified by some further collection of observational data. This, they show, is equivalent to the topological notion of the *nowhere density* of $\mathcal{H}$ inside 2^ω equipped with the product topology. This account gives a good account of *long-run* falsifiability, but fails to give a satisfactory account of *short-run* falsification as there are no bounds on *how long* it might take an agent to witness a crucial experiment. I define a *sample* along a set X as follows:

Definition 1.2. A *sample* of X is an injective function $f : \omega \to X$. A sample f is *full* provided f is bijective. $\Diamond$

Now, relative to a sample f we define a notion of the *surprise* of a hypothesis along the sample as follows:

Definition 1.3. Let X be a set, $f : \omega \to X$ a sample of X, and $\mathcal{H} \subset 2^X$ a hypothesis. For $Y \subseteq X$ let $\mathcal{H}\restriction_Y = \{h\restriction_Y \mid h \in \mathcal{H}\}$.

The *surprise* of $\mathcal{H}$ is the function

$$S(\mathcal{H}, f, n) = 1 - \frac{|\mathcal{H}\restriction_{f([n])}|}{2^{|f([n])|}}. \qquad \Diamond$$

The surprise of $\mathcal{H}$ along the enumeration f is the relative proportion of the states of the world incompatible with $\mathcal{H}$. If $\mathcal{H}$ is highly surprising, then it is compatible with only a small number of observations along the sample.

Very closely related to the NIP theories discussed in the static case of falsification, the VC finite hypothesis classes are characterized by the ability to obtain uniform bounds on surprise independent of sample.

Finally, we turn our attention to the work of Mayo [29][30], who advocates for a strengthening of Null Hypothesis Statistical Testing as the foundation of statistical testing called *severe testing*. Mayo and other error statisticians ask the question

> When do data x provide good evidence for / a good test of hypothesis H?

The error statistician will invoke some form of a **Severity Principle** to answer this question:

> (Weak Severity Principle) Data x **does not** provide good evidence for H if x is the result of a **test procedure** T with very low probability of uncovering the falsity of H [31, p. 21].

A converse is given by:

> (Full Severity Principle) Data x provides good evidence for H to the extent that test T has been **severely passed** by H [31, p. 21].

However, the definition of severe testing via probabilistic notions is elusive. To round out the discussion of falsification, I show that the notion of surprise I defined in the context of always falsifiability is well-suited to give an account of a *well-defined* combinatorial analogue of severe testing I term *severe surprise*.

Definition 1.4. Let $f : \omega \to X$ be a sample, $\mathcal{H} \subset 2^X$ a hypothesis, $\mathcal{H}^c = 2^X \setminus \mathcal{H}$, $n \in \omega$, and $\epsilon > 0$. We say that $(\mathcal{H}, f, n)$ is *severely surprising* at level ϵ provided the observed data $x \in \mathcal{H}\restriction_{f([n])}$,

$$S(\mathcal{H}, f, n) > 1 - \epsilon$$

and

$$S(\mathcal{H}, f, n) > S(\mathcal{H}^c, f, n). \qquad \Diamond$$

Crucially, surprise is at its core a non-probabilistic notion. Key to this is the observation that surprise is subadditive:

Proposition 1.1. For all $\mathcal{H}, f, n$,

$$S(\mathcal{H}, f, n) + S(\mathcal{H}^c, f, n) \leq 1 \qquad \blacklozenge$$

and

Proposition 1.2. Let $\mathcal{H} \subset 2^\omega$ be dense and codense. Then for all f and n,

$$S(\mathcal{H}, f, n) = S(\mathcal{H}^c, f, n) = 0. \qquad \blacklozenge$$

Under this definition of severe surprise, one can show that VC finite classes are in fact severely surprising if true, uniformly in the size of the sample.

The core message of this chapter is that VC finiteness is in some sense the core concept in falsification for finitistic agents, being robust under arbitrary samples and uniquely endowed with felicitous finite-sample bounds.

1.2 On Rational Jurisprudence

In the next chapter, I turn my attention to a dispute within the legal community regarding the role of probability—in particular, Bayesian rationality—within the criminal justice system. Tribe [46] famously argued against so-called "trial by mathematics," claiming that the proliferation of statistical methods in the testimony of expert witnesses ran afoul of the Presumption of Innocence. This sentiment was echoed by the Supreme Court of Connecticut when, in *State v. Skipper* [42], they ruled that Bayesian testimony was inadmissible in course on account of perceived conflicts with the Presumption of Innocence.

Running counter to this trend, Judge Richard Posner argued that an ideal finder of fact would themselves be a Bayesian, and found no conflict between Bayesian principles and the Presumption of Innocence. For Posner, so long as a juror was Bayes rational and assigned prior probability of guilt $\mathbb{P}(E_G) = \frac{1}{2}$, the juror was fit to serve as a juror.

To resolve this dispute regarding the probabilistic interpretation of the Presumption of Innocence, I introduce a model of a juror's disposition to convict on the basis of witnessing some collection T of testimony as a function f with range $\{C, A\}$ such that

$$f(T) = C$$

if the juror would convict upon hearing all and only the testimony contained in T and would acquit (i.e. $f(T) = A$) otherwise. Under two very mild assumptions, namely that

1. $f(\varnothing) = A$ (Presumption of Innocence), and
2. there exists a transcript T such that $f(T) = C$. (Willingness to Convict),

then there is a Bayesian juror that rationalizes this disposition.

This result cuts at the heart of both competing positions we discussed at the outset, showing that

- nothing is *gained* by mandating that a juror be representable by a Bayesian threshold juror: there is always a prior accommodating their disposition, and
- nothing is *lost* by mandating that a juror be representable by a Bayesian threshold juror: given any disposition, it is indistinguishable from the disposition of a Bayesian threshold juror.

This result indicates that relevant notion of a Bayesian threshold juror is insufficiently specified to render this debate a substantive one.

1.3 On The Sufficiency of First-Order Logic

In the final chapter of the dissertation I address a question that has puzzled me since first learned about the metatheory of First-Order Logic: is there good epistemic justification for restricting our attention to First-Order Logic? I was certainly not the first person to ask that; a version of this phenomenon was described by Barwise as the acceptance of what he calls the *First-Order Thesis*, which asserts that

> logic is first-order logic, so that anything that cannot be defined in first-order logic is outside the domain of logic. [4, pp. 5–6]

Barwise was chiefly concerned with the relative inexpressiveness of First-Order Logic. Despite this, I argue that First-Order Logic—while not the most expressive abstract logic—is sufficient to represent and carry out any inference a finitistic agent might carry out. The main mathematical result here is the Σ_1 completeness of the consequence relation $\models_{FO}$ of First-Order logic as a Turing functional. A consequence of Σ_1 completeness is that *any* notion of logical consequence which is machine verifiable given an oracle naming the theory Γ is in fact able to be internalized into a First-Order proof system. While the translation function will generally not preserve semantics on the nose—after all, First-Order Logic is not particularly expressive—the inferential structure of such a system is able to be witnessed as a first-order system by way of a computable translation. I then consider a recent argument of Warren's that it is a strong metaphysical possibility that agents like us in nearby possible worlds

can implement the ω rule of inference, a sound but highly non-recursive pattern of inference. While I don't refute his premise directly, I do argue that there is a clear finitistic interpretation of the example of the ω rule that Warren has in mind, saving the plausibility of a position that agents like ourselves only perform finitary inferences. I close out the chapter by reflecting on *Hilbert's Thesis*, a position constituted by two related theses:

1. Hilbert's Expressibility Thesis (HET): All mathematical (extra-logical) assumptions may be expressed in first-order logic, and

2. Hilbert's Provability Thesis (HPT): The informal notion of *provable* is made precise by the formal notion of *provable in first-order logic*.

Kripke [26] has argued that

$$\text{(HET) + (HPT)} \implies \text{Church's Thesis}$$

on the basis of Gödel's Completeness Theorem. The results of this chapter indicate a partial converse. The Σ_1 completeness of $\vdash_{FO}$ shows that

$$\text{Church's Thesis + In-Principle Machine Verifiability of Proofs} \implies \text{(HPT)}.$$

First-Order Logic is neither the *only* nor *most expressive* logic with Σ_1 complete entailment relation, but the above argument shows that First-Order Logic is sufficiently expressive to simulate any machine-verifiable inferential system.

Chapter 2

On Falsification

2.1 Introduction

Popper's [32] solution to the demarcation problem says that the distinguishing feature of a scientific theory—construed as an empirical hypothesis—is its *falsifiability*. Various accounts of falsification have emerged over the years; in this chapter, I aim to provide a model-theoretic account of falsification that will aid us in understanding both long-run and short-run properties of various falsificationist strategies.

Broadly speaking, falsification centers on the following question: given a class $\mathbb{K}$ of possible worlds, is there some finite collection of observations about our world W that would allow us to infer $W \notin \mathbb{K}$? If the answer is "yes," the class $\mathbb{K}$ is said to be falsifiable.

Throughout, we suppose that $\mathcal{L}$ is a signature[1] and $\mathbb{K} \subseteq \mathrm{Str}(\mathcal{L})$ a class of $\mathcal{L}$-structures. Epistemically, $\mathcal{L}$ plays the role of the collection of observable relations, functions, and constants that relate objects in the world. We suppose that the world W is itself an $\mathcal{L}$-structure.

An *observable formula* $\varphi(x_1, \ldots, x_n)$ corresponds to a finite Boolean combination of atomic $\mathcal{L}$-formulas. We say that $\varphi(x_1, \ldots, x_n)$ is $\mathbb{K}$-forbidden just in case no $W \in \mathbb{K}$ realizes φ. Model theoretically, this is the same as saying

$$\mathbb{K} \vDash \neg(\exists x_1, \ldots x_n)\varphi(x_1, \ldots, x_n).$$

Of course, this is equivalent to

$$\mathbb{K} \vDash (\forall x_1, \ldots x_n)\neg\varphi(x_1, \ldots, x_n),$$

motivating the following definition:

[1] Not necessarily finite or relational.

Definition 2.1. Let $\mathbb{K}$ be a class of $\mathcal{L}$-structures. We denote the universal theory of $\mathbb{K}$ as

$$\forall_1(\mathbb{K}) = \{\varphi \,|\, \varphi \text{ is a } \forall_1 \text{ first-order } \mathcal{L}\text{-sentence and } \mathbb{K} \vDash \varphi\}.$$

We say that $\mathbb{K}$ is *falsifiable* provided that the class of models of $\forall_1(\mathbb{K})$ is nontrivial, i.e.,

$$\mathrm{Mod}(\forall_1(\mathbb{K})) \neq Str(\mathcal{L}).$$

Let $\mathbb{K} \subseteq \mathbb{K}'$ be two classes of $\mathcal{L}$ structures. We say that $\mathbb{K}$ is falsifiable relative to $\mathbb{K}'$ provided the inclusion

$$\forall_1(\mathbb{K}) \supset \forall_1(\mathbb{K}')$$

is proper. ◊

One may think of $\mathrm{Str}(\mathcal{L})$ as the class of possible worlds relative to a signature $\mathcal{L}$; indeed, it is the largest such collection of possible worlds. However, if one wants to study falsifiability relative to a class of $\mathcal{L}$-structures containing analytic truths beyond the logical validities, one may turn to the relative definition of falsifiability. In the vast majority of this chapter, we will concern ourselves with falsification simpliciter.

An immediate corollary of this definition is that if $\mathbb{K}$ is falsifiable, then any class *stronger* than $\mathbb{K}$ is falsifiable:

Proposition 2.1. Let $\mathbb{K}$ be a falsifiable class. If $\mathbb{K}' \subseteq \mathbb{K}$, then $\mathbb{K}'$ is falsifiable. ⧫

Proof. If $\mathbb{K}' \subseteq \mathbb{K}$ then $\forall_1(\mathbb{K}) \subseteq \forall_1(\mathbb{K}')$. Since $\forall_1(\mathbb{K})$ contains nontrivial universal sentences, so does $\forall_1(\mathbb{K}')$, so $\mathbb{K}'$ is falsifiable. □

In particular, per this definition a class of structures $\mathbb{K}$ need not even be first-order axiomatizable in order to be falsifiable. Per this definition, all that is required of a class of structures to be falsifiable is that there is *some* observable formula φ whose realization is incompatible with the class $\mathbb{K}$.

While this base notion of falsifiability is rather simple in its expression, refinements of the notion of falsification contain a great deal of mathematical complexity. In the remainder of this chapter, we will investigate refinements of falsifiability along two key axes:

1. The **Static Case**, wherein we ask "just how falsifiable is the class $\mathbb{K}$?" relative to the above definition of falsifiability, and

2. The **Dynamic Case**, wherein we ask "how much observation is required to falsify a hypothesis, and how quickly can we expect to falsify it?"

In discussing the static case, we will primarily discuss variants surrounding the *first-order, universal axiomatization* of a class. After all, if a class $\mathbb{K}$ is specified by a universal set of axioms, this means that each one of the theory's axioms expresses the class $\mathbb{K}$ being incompatible with some observation. Others, such as Simon and Groen [38] and Chambers et al. [7], propose even more stringent constraints on $\mathbb{K}$ than its universal axiomatizability to deem them falsifiable.

For the dynamic case, we concern ourselves with how *much* data is necessary to falsify a hypothesis and also *how much data we can afford to omit* and still be guaranteed to falsify a given hypothesis. The work of Juhl and Schulte [35] refines the notion of falsifiability to that of *always falsifiability*, a condition that requires a certain abundance of data that could refute the hypothesis. They go on to show that the notion of always falsifiability is equivalent to the notion of nowhere density in a certain topological space. However, the mere nowhere density of a hypothesis $\mathcal{H}$ does not guarantee that after some number $n = n(\mathcal{H})$ of observations one will witness a bit of data that will aid in falsifying $\mathcal{H}$. We identify a class of hypotheses—those of finite VC dimension—as precisely those hypotheses which yield nontrivial surprise once some fixed number $n = n(\mathcal{H})$ of samples are observed.

As a case study in falsification, we then turn to the framework known as *severe testing*. Developed by Mayo [30], severe testing is a neo-frequentist, neo-Popperian account of what makes a hypothesis *testable*. Motivated in part by the replication crisis in science, Mayo rejects the standard framework of Null Hypothesis Statistical Testing, arguing instead that her notion of severe testing is a necessary and sufficient condition for a hypothesis $\mathcal{H}$ to not be rejected / provisionally accepted in the face of data x. I argue that her basic definition of severe testing is underspecified, but nevertheless argue that the core notion of severe testing can be formalized with the aid of surprise. Crucially, surprise is non-probabilistic: surprise is intimately related to a family of *semi*-probability measure on the space of hypotheses in 2^ω; that is, a function $\mu : 2^{2^\omega} \to [0, 1]$ such that

$$\mu(\mathcal{H}) + \mu(\mathcal{H}^c) \geq \mu(\mathcal{H} \cup \mathcal{H}^c),$$

where $\mathcal{H}^c = 2^\omega \setminus \mathcal{H}$ is the complement of $\mathcal{H}$, with inequality between the two terms possible. As in the discussion of falsification *qua* nowhere density, VC finite hypotheses emerge as the hypotheses with felicitous small-sample properties.

2.2 Falsifiability and Unfalsifiability in Mechanics and Economics

As a warm-up to our investigation of falsification, we investigate falsificational phenomena in physics and economics by way of an analysis of Newtonian Mechanics and the theory of choice.

Newtonian Mechanics

We begin by showing that, in a strong sense, the *framework* of Newtonian Mechanics is unfalsifiable relative to the class of kinematic motions.

For the definition of Newtonian system I follow the formalism given by Arnold [2].

Definition 2.2. [2, p. 8] An n-particle *motion* is a smooth function $x : \mathbb{R} \to \mathbb{R}^{3n}$ such that the graphs of the trajectories of each particle are non-intersecting.

A Newtonian system of n particles is a motion $x : \mathbb{R} \to \mathbb{R}^{3n}$ such that there exists a vector field

$$F : \mathbb{R}^{3n} \times \mathbb{R}^{3n} \times \mathbb{R} \to \mathbb{R}^{3n}$$

such that

$$x''(t) = F(x(t), x'(t), t)$$

for all $t \in \mathbb{R}$. ◊

Let $\mathbb{K}$ be the class of Newtonian systems. Note here that since we do not require $\mathbb{K}$ to be an elementary class in order to be falsifiable, we do not have to exhibit a first-order axiomatization of $\mathbb{K}$.

By an n-particle *kinematic datum* e I mean an equality

$$(x, x', x'')(t_0) = v$$

or inequality

$$(x, x', x'')(t_0) \neq v$$

where $v \in \mathbb{R}^{9n}$ and $t_0 \in \mathbb{R}$. Intuitively, a kinematic datum is a specification of the numerical values of the vector $(x(t), x'(t), x''(t), t) \in \mathbb{R}^{9n} \times \mathbb{R}$. For a set E of kinematic data, let $\pi_t(E)$ be the set of times occurring as values in E. For an element $e \in E$, let t_e be the value of the time coordinate of e.

We say that a set E of kinematic data is *motional* provided all sentences in E are satisfied by some motion.

Proposition 2.2. Let E be a finite motional set of kinematic data. Then there is an n-particle Newtonian system x such that for all $e \in E$, x satisfies all the conditions set out by E.

Thus, the class of Newtonian systems of n particles is unfalsifiable relative to the class of n-particle motions. ♦

Proof. First, we note that if E is motional we may replace all inequalities of E with equalities to prove the claim.[2] Let Y_i denote the trajectory of the i^{th} particle.

Thus, e is equivalent to a system of equations of the form

$$(Y_i, Y_i', Y_i'')(t_j) = v_j = (p_{ij}(t_j), v_{ij}(t_j), a_{ij}(t_j))$$

where p_{ij}, v_{ij}, and a_{ij} represent the position, velocity, and trajectory of the i^{th} particle at time t_j.

Since the data E is motional, there exist n smooth functions $Y_i : \mathbb{R} \to \mathbb{R}^3$ such that the positions of the i^{th} particle Y_i satisfy

$$Y_i(t_j) = p_{ij}.$$

We now show that we may alter this trajectory to ensure that $Y_i'(t_j) = v_{ij}$ and $Y_i''(t_j) = a_{ij}$ for each i, j. The following argument ensures that we can locally alter the Y_i's without intersecting the graphs of the Y_i.

Let I be a closed interval of finite length containing the open interval $[\min((t_j)), \max((t_j))]$. Since the Y_i are all continuous, there exists a compact box[3] $R \subseteq \mathbb{R}^4$ such that a neighborhood of each graph Γ_i of each Y_i restricted to I is contained in R.

Since R is a compact metrizable space and the graphs Γ_i are closed and disjoint, there exists an $r \in \mathbb{R}$ such that the *tubular neighborhoods* of the graphs Γ_i given by

$$U_i(r) = \{x \in \mathbb{R}^4 \mid d(x, \Gamma_i) < r\}$$

are disjoint.

Now, by the existence and uniqueness of ODEs there locally exists a unique solution to the initial value problem

$$(x_i, x_i', x_i'')(t_j) = (p_{ij}, v_{ij}, a_{ij}).$$

[2]Some care must be taken to ensure that the set E remains motional in this case. So long as we replace $x_i(t) \neq p$ with some $x_i(t) = p'$ where p' does not appear in the remaining conditions in E, the set E' will remain motional. Since E is finite and $\mathbb{R}$ is infinite, it is always possible to replace finitely many inequalities with finitely many equalities and remain motional.

[3]By a box I mean a product of intervals.

For each particle i, by taking a small enough interval $I_{i,j}$ around t_j we let $Z_{i,j} : I_{i,j} \to \mathbb{R}^3$ be the solution to this ODE and have the graph of $Z_{i,j}$ contained in the neighborhood $U_i(r)$ constructed above. Perhaps by shrinking the interval on which $Z_{i,j}$ solves the ODE, we may smoothly extend $Z_{i,j}$ to all of $\mathbb{R}$ in a manner such that $Z_{i,j}(t) = 0$ for all $t \notin I_{i,j}$.

Now, by the existence of smooth bump functions, there exists a smooth bump function

$$b_{i,j}(t) : \mathbb{R} \to [0, 1]$$

such that $b_{i,j}(t) = 0$ on some open interval containing t_j and $b_{i,j}(t) = 1$ for all $t > max(I_{i,j})$ and $t < min(I_{i,j})$. We define

$$\tilde{Y}_i(t) = \left(\sum_j b_{i,j}(t)\right) Y_i(t) + \left(1 - \sum_j b_{i,j}(t)\right) \sum_j Z_{i,j}(t)).$$

Then $\tilde{Y}_i(t)$ satisfies the required differential equations.

Finally, we must argue that there exists a force function $F : \mathbb{R}^{3n} \times \mathbb{R}^{3n} \times \mathbb{R} \to \mathbb{R}^{3n}$ such that

$$Y_i''(t) = F(Y_i(t), Y_i'(t), t)$$

for all i. Intuitively, we would like to define

$$F(x, x', t) = \sum_i Y_i''(t) \text{ if } x(t) = Y_i(t)$$

but this is not a continuous function. However, by the construction of U_r, shrinking r as necessary, we can ensure that this map is continuous by defining

$$F(x, x', t) = \sum_i Y_i''(t) \text{ if } (x(t), t) \in U_i(r). \qquad \square$$

This theorem indicates that no matter how many finite points of data we collect regarding the *kinematics* of the system, there will always be some Newtonian theory which accommodates that data. This is the sense in which the framework of Newtonian mechanics fails to be falsifiable.

However, there are natural strengthenings of the class $\mathbb{K}$ of Newtonian motions which are falsifiable.

For example, suppose that our hypothesis is that a particle P is a free particle with motion x with zero initial acceleration relative to a fixed observer's frame of reference. This implies that for all t the resultant force $F(x, x', t)$ is identically zero. Thus, the motion x must follow a straight line.

This hypothesis is highly falsifiable. Since a line in $\mathbb{R}^3$ is determined by two points, the theory of the free particle entails that if $x(t_1)$ and $x(t_2)$ are the positions of the first two observations of the particle, all subsequent observations of the particle must be a member of the line $L(x(t_1), x(t_2)) \in \mathbb{R}^3$. Thus, for every $n > 2$, each subsequent observation carries with it the chance of refuting the claim that the particle is free. This is an instance of the notions of *always falsifiability* and *VC finiteness*, which we will discuss in our treatment of the dynamic case of falsification in section 2.3.

Theory of Choice

We now turn our attention from the falsification of physical theories to the falsification of economic theories of choice. To keep things simple, we model *preference* as a binary relation $\prec$ on a set of choices C, where $x \prec y$ is interpreted as "y is strictly preferred to x." The data $(C, \prec)$ is called a *preference structure*.

A frequently assumed necessary condition for a preference to be considered *rational* is that the preference relation is acyclic; namely, that there is no chain $x_1 \prec x_2 \cdots x_n \prec x_1$ and $x_1 \not\prec x_1$. Let $\mathbb{K}$ be the class of acyclic preference structures.

The class $\mathbb{K}$ is falsifiable: if one observes a configuration

$$\left(\bigwedge_{1 \leq i < n} c_i \prec c_{i+1} \right) \wedge c_n \prec c_1$$

then one can conclude that the underlying choice structure $(C, \prec) \notin \mathbb{K}$. In fact, $\mathbb{K}$ is universally axiomatizable, axiomatized by the collection

$$A_n = (\forall x_1, \ldots, x_n) \neg \left(\left(\bigwedge_{1 \leq i < n} x_i \prec x_{i+1} \right) \wedge x_n \prec x_1 \right)$$

for all $n \in \omega$.

On the other hand, there are common rationality assumptions which are *not* falsifiable.

For example, consider the axioms invoked to prove the representability of a preference structure $(C, \prec)$ by a utility function $u : X \to \mathbb{R}$, that is,

$$x \preceq y \leftrightarrow u(x) \leq u(y).$$

Gilboa [18, p. 51] gives the axioms

1. **Completeness**: $(\forall x, y)(x \preceq y \vee y \preceq x)$,

2. **Transitivity**: $(\forall x, y, z)\,([x \prec y \wedge y \prec z] \rightarrow x \prec z)$, and

3. **Separability**:$(\exists Z \subseteq X)(\forall x, y)(\exists z \in Z)((|Z| \leq \aleph_0) \wedge (x \prec y \rightarrow (x \preceq z \preceq y))$.

Completeness and Transitivity are both $\forall_1$ sentences in the language $\mathcal{L} = \{\prec\}$, but Separability is naturally expressed as a second-order sentence.

Gilboa makes a very interesting argument for the admissibility of the Separability axiom: its unfalsifiability "suggests that [Separability] has no empirical content and therefore does not restrict our theory... Rather, the axiom is a price we have to pay if we want to use a certain mathematical model" [18, p. 52].

The following theorem makes this argument precise.

Theorem 2.1. Let $\mathbb{K} = \mathrm{Mod}(T)$ be the class of models of an $\mathcal{L} = \{\prec\}$-theory T. Let $\mathbb{K}_{sep}$ be the class $\mathcal{M} \in \mathbb{K}$ satsifying Separability. Then $\mathbb{K}_{sep}$ is unfalsifiable relative to $\mathbb{K}$. ⧫

Proof. Let T be a first-order theory and φ a $\forall_1$ sentence such that $\varphi \notin \forall_1(\mathbb{K})$. We wish to show that $\mathbb{K}_{sep} \nvdash \varphi$.

Since $\varphi \notin \forall_1(\mathbb{K})$ there exists some $\mathcal{M} \in \mathbb{K}$ such that $\mathcal{M} \vDash \neg\varphi$. Since φ is a universal sentence, $\neg\varphi$ is existential. Let $\overline{m} \in \mathcal{M}^k$ witness $\neg\varphi$. By the Löwenheim-Skolem theorem there exists a countable elementary substructure $\mathcal{M}' \preceq \mathcal{M}$ containing $\overline{m}$ of size $\aleph_0$. Since $\mathcal{M}' \in \mathbb{K}$ and is of size $\aleph_0$, $\mathcal{M}' \in \mathbb{K}_{sep}$. Therefore $\mathbb{K}_{sep} \nvdash \varphi$, so $\mathbb{K}_{sep}$ is unfalsifiable relative to $\mathbb{K}$. □

Thus, not only is the Separability axiom unfalsifiable relative to the class of *all* $\mathcal{L}$-structures, it is unfalsifiable relative to any first-order axiomatizable theory of preference. In this way the Separability axiom is empirically harmless: we may freely adjoin the Separability axiom to any first-order theory T of choice structures without inadvertently strengthening the observable consequences of T.

2.3 Falsifiability and the Randomness of the Universe

As an application of the results of the previous section, we argue that for many ways of making precise the assertion that "the world is, at a fundamental level, random," the assertion is unfalsifiable.

Defining what it means to be "random," however, poses a great difficulty. To this end, I consider two different formalizations of randomness as it pertains to structures:

1. The evolution of the universe is *generic* in a suitable sense, and

2. The evolution of the universe is generated by a stochastic process.

For a suitable formulation of each of the above cases we will see unfalsifiability arise. To make sense of these two notions we define the notion of a time-indexed structure.

Definition 2.3. Let $\mathcal{L}$ be a relational language. Then the *time-indexed language* $\mathcal{L}_\tau$ is given by

$$\mathcal{L}_\tau = \{R_i(x_1, \ldots, x_m, t) \mid R_i(x_1, \ldots, x_m) \in \mathcal{L}\} \cup \{O(x), \tau(x), <\}.$$

A *time-indexed structure* is an $\mathcal{L}_\tau$ structure satisfying the theory T_τ given by the axioms:

1. Objects and Times are different sorts, i.e.,

$$(\forall x)(O(x) \vee \tau(x) \wedge \neg(O(x) \wedge \tau(x)),$$

2. $(\forall \overline{x}, t)(R(\overline{x}, t) \to (\bigwedge O(x_i) \wedge \tau(t)))$, and

3. the relation $<$ is a linear order on τ. $\Diamond$

A time-indexed $\mathcal{L}_\tau$-structure can be thought of simply as a time-indexed family of $\mathcal{L}$-structures. We use such examples all the time:

Remark 2.1. Let $\mathcal{L} = \{H(x)\}$ be the language consisting of a single unary predicate. We can regard an ω-sequence of coin flips as an $\mathcal{L}_\tau$ structure in a straightforward manner. The domain of the structure $\mathcal{M} = \omega \cup \{c\}$, where c is an object corresponding to the coin. We interpret $O(\mathcal{M}) = \{c\}$, $\tau(\mathcal{M}) = \omega$, and $\mathcal{M} \vDash H(c, n)$ just in case the n^{th} coin flip of c returns heads. ⧫

The Richness of the Universe

We first consider the notion of the randomness of the universe as specified by the notion of a Fraïssé limit.

Definition 2.4. [20, pp. 321-322] A countable class $\mathbb{K}$ of finite $\mathcal{L}$-structures is a Fraïssé class provided $\mathbb{K}$ satisfies the following properties:

1. Hereditary Property (HP): If $\mathcal{M} \in \mathbb{K}$ and $\mathcal{N} \subseteq \mathcal{M}$ is finitely generated then $\mathcal{N} \in \mathbb{K}$,

2. Joint Embedding Property (JEP): If $\mathcal{M}, \mathcal{N} \in \mathbb{K}$ then there exists $\mathcal{Q} \in \mathbb{K}$ such that $\mathcal{M}$ and $\mathcal{N}$ both embed in $\mathcal{Q}$, and

3. Amalgamation Property (AP): If $\mathcal{M}, \mathcal{N}, \mathcal{Q}$ are $\mathcal{L}$-structures and $f_{\mathcal{N}} : \mathcal{M} \to \mathcal{N}$, $f_{\mathcal{Q}} : \mathcal{M} \to \mathcal{Q}$ are embeddings there exists a structure $\mathcal{S}$ and embeddings $g_{\mathcal{N}} : \mathcal{N} \to \mathcal{S}$ and $g_{\mathcal{Q}} : \mathcal{Q} \to \mathcal{S}$ such that

$$g_{\mathcal{N}} \circ f_{\mathcal{N}} = g_{\mathcal{Q}} \circ f_{\mathcal{Q}}. \qquad \lozenge$$

Remark 2.2. The class of finite linear orders is a Fraïssé class. ⧫

Remark 2.3. Let $\mathcal{L}$ be a finite relational language. The class of finite $\mathcal{L}$-structures forms a Fraïssé class. ⧫

For a Fraïssé class $\mathbb{K}$, there is a unique, highly homogeneous, countable structure $\mathbb{K}_{lim}$ into which all and only the members of $\mathbb{K}$ embeds, called the Fraïssé limit.

Definition 2.5. Let $\mathcal{M}$ be an $\mathcal{L}$-structure. The *age* of $\mathcal{M}$, age($\mathcal{M}$), is the class of all finitely-generated $\mathcal{L}$-structures embeddable in $\mathcal{M}$. $\lozenge$

Theorem 2.2. [20, Theorem 7.1.2] Let $\mathbb{K}$ be a Fraïssé class of $\mathcal{L}$ structures. Then there is an $\mathcal{L}$ structure $\mathbb{K}_{lim}$, unique up to isomorphism, such that

1. age($\mathbb{K}_{lim}$) = $\mathbb{K}$,

2. $|\mathbb{K}_{lim}| \leq \aleph_0$, and

3. every isomorphism between finitely generated substructures $\mathcal{M}_1, \mathcal{M}_2 \subseteq \mathbb{K}_{lim}$ extends to an automorphism of $\mathbb{K}_{lim}$. ⧫

Thus, a Fraïssé limit is extremely rich, able to accommodate any finite number of observations. Moreover, when a Fraïssé limit exists for a class $\mathbb{K}$, the first-order theory $\mathbb{K}_{lim}$ is $\forall_1$-conservative over $\mathbb{K}$.

Proposition 2.3. Let $\mathbb{K}$ be a Fraïssé class of $\mathcal{L}$-structures where $\mathcal{L}$ is a finite relational language. If φ is a $\forall_1$ $\mathcal{L}$-sentence, then

$$\mathbb{K}_{lim} \vDash \varphi \leftrightarrow \mathbb{K} \vDash \varphi. \qquad ⧫$$

Proof. If $\mathbb{K}_{lim} \vDash \varphi$, then since every $\mathcal{M} \in \mathbb{K}$ embeds into $\mathbb{K}_{lim}$ and φ is $\forall_1$, $\mathbb{K} \vDash \varphi$.

Conversely, suppose that $\mathbb{K}_{lim} \nvDash \varphi$. Then $\neg\varphi$ is existential, so there is some witness $\overline{m} \subset \mathbb{K}_{lim}$ to the falsity of φ. Since $\mathbb{K}$ = age($\mathbb{K}_{lim}$) and $\langle m \rangle$ is a finitely-generated substructure of $\mathbb{K}_{lim}$, $\mathcal{N} = \langle m \rangle \in \mathbb{K}$ and $\mathcal{N} \vDash \neg\varphi$. Thus $\mathbb{K} \nvDash \varphi$. □

Thus, no new universal sentences are entailed by the Fraïssé limit of $\mathbb{K}$.

We show that the class of finite time-indexed structures forms a Fraïssé class, which we shall see yields unfalsifiabilty of the generic theory relative to the class of time-indexed structures.

Theorem 2.3. Let T_τ be the theory of time-indexed structures. Then

1. the class of finite models of T_τ is a Fraïssé class, and
2. the theory $T_{\tau,lim}$ of the Fraïssé limit of the class is the model companion of the theory T_τ. ⧫

Proof. We need to show that class of finite models of T is a Fraïssé class.

First, since the class $\mathrm{Mod}(T_\tau)$ is universally axiomatizable, its finite models satisfy (HP).

By the axioms of T_τ each model $\mathcal{M} \in \mathrm{Mod}_{fin}(T_\tau)$ can be expressed as

$$\mathcal{M} = ((W_i, t_i))_{i<m}$$

where each W_i is an $\mathcal{L}$-structure (recall $\mathcal{L}_\tau$ was obtained from $\mathcal{L}$) and each t_i is a time.

A necessary and sufficient condition for a map $f : \mathcal{M} \to \mathcal{N}$ with $\mathcal{M}, \mathcal{N} \in \mathrm{Mod}(T)$ is that f restricted to τ is an order embedding and that for each time t_i, $f(W_i) \subset W_{f(t_i)}$ is an embedding of $\mathcal{L}$-structures. From this decomposition of embeddings it is clear that the joint extension property and amalgamation property holds as the class of finite $\mathcal{L}$-structures and the class of finite linear orders are both Fraïssé classes: to jointly embed two finite models of T_τ structures, first jointly embed their temporal component and then jointly embed their $\mathcal{L}$-structures at each time in the intersection of the embedding. Likewise, one may amalgamate by first amalgamating the temporal component and then amalgamating the $\mathcal{L}$-structures over each time.

Thus the theory of the Fraïssé limit $T_{\tau,lim}$ exists and model complete by [20, Theorem 7.4.2]. It remains to show that $T_{\tau,lim}$ is a model companion of T_τ.

Clearly every model of $T_{\tau,lim}$ is a model of T_τ, so it suffices to show that every model of T embeds into a model of $T_{\tau,lim}$. Suppose $\mathcal{M} \vDash T_\tau$. Then since T_τ is $\forall_1$ axiomatizable, all finitely generated substructures of $\mathcal{M}$ are models of T_τ. Moreover, $\mathcal{M}$ embeds into an ultraproduct of its finite substructures since the language is relational, and in turn each finite substructure embeds into the Fraïssé limit of the class. Thus $\mathcal{M}$ embeds into an ultrapower of the Fraïssé limit of the class and hence, since T_{lim} is elementary, a model of T_{lim}. □

As a corollary, we have the following.

Corollary 2.1. The theory $T_{\tau,lim}$ is $\forall_1$-conservative over T; thus, $T_{\tau,lim}$ is relatively unfalsifiable over T_τ. ⧫

Therefore, for this sense of genericity, "the Universe is a generic time-indexed $\mathcal{L}$-structure" is not falsifiable relative to the theory of time-indexed structures.

The Stochasticity of the Universe

We now turn to *probabilistically generated* models of the evolution of the universe.

Definition 2.6. The *discrete time-index language* $\mathcal{L}_\tau^d$ is

$$\mathcal{L}_\tau^d = \{R_i(x_1, \ldots, x_m, t) \mid R_i(x_1, \ldots, x_m) \in \mathcal{L}\} \cup \{O(x), \tau(x), <, S(x)\}. \qquad \Diamond$$

We work with a distinguished class of $\mathcal{L}_\tau^d$ structures $\mathcal{M}$, namely those such that

1. $(\tau(\mathcal{M}), <, S)$ is the structure of $(\omega, <, S)$,
2. $O(\mathcal{M})$ is $[n]$ for some $n \in \omega$,
3. the world $(W, 0)$ is drawn from a probability distribution μ on the state space $\Sigma = Str_{\mathcal{L}}([n])$, and
4. the world $(W, t+1)$ is obtained from (W, t) by way of a time-homogeneous memoryless Markov process, i.e., there exists a stochastic matrix ρ on the state space Σ in ω such that $\mathbb{P}((W, t+1)|(W', t)) = \rho(W, W')$.

Remark 2.4. This construction generalizes the construction of an ω sequence of IID coin flips. In this case, $O(\mathcal{M}) = \{c\}$ is a single object, and each time t is associated to an $\mathcal{L} = \{H(x)\}$-structure where $H(x)$ is a unary predicate meaning "x flipped heads." The basic $\mathcal{L}_\tau^d$ predicate $H(x, t)$ means "x flipped heads at time t."

Let μ be any measure on $\{H, T\}$, i.e., an assignment of $p_H, p_T \in [0, 1]$ such that $p_H + p_T = 1$. The stochastic transition matrix ρ is given explicitly by

$$\rho = \begin{pmatrix} p_H & p_T \\ p_T & p_H \end{pmatrix}.$$

Such a stochastic process generates an $\mathcal{L}_\tau^d$ structure on domain $\omega \cup \{c\}$. ⧫

Now, let $\mathcal{C}$ be a set of pairs (μ, ρ) where μ is a probability distribution on Σ and ρ is a stochastic matrix on Σ.[4] The choice of μ and ρ induce a unique probability measure $\mathbb{P}_{\mu,\rho}$ on Σ^ω. The existential $\mathcal{C}$-theory $T_\mathcal{C}$ in $\mathcal{L}^d_\tau$ is given by:

$$(\exists \overline{x}, \overline{t})\varphi(\overline{x}, \overline{t}) \in T_\mathcal{C} \leftrightarrow (\forall \mu, \rho \in \mathcal{C}) \left(\mathbb{P}_{\mu,\rho}(\varphi(\overline{x}, \overline{t}) \text{ is eventually realized}) = 1\right).$$

Let $\mathcal{C}_+$ be the class of pairs (μ, ρ) of initial distributions μ on Σ with $\mu(W) > 0$ for each $W \in \Sigma$ and stochastic matrices ρ with rows and columns indexed by Σ such that $\rho(W, W') > 0$ for all $W, W' \in \Sigma$.

Let $\varphi_\mathcal{M}(\overline{x}, t)$ be the sentence saying that at time t the $\mathcal{L}$-structure $\mathcal{N}(t)$ is isomorphic to $\mathcal{M}$. To show that every $\mathcal{L}^d_\tau$-satisfiable $\exists_1$ formula in the language $\mathcal{L}^d_\tau$ is a member of $T_\mathcal{C}$ it suffices to show that every formula of the form

$$\left(\bigwedge_i \varphi_{\mathcal{M}_i}(\overline{x}_i, s_i)\right) \wedge \bigwedge_i (s_i < s_{i+1})$$

where

1. $W_i \in \Sigma$,
2. s_i is a term in the language of the successor function $\{S(x)\}$, and
3. the formula $\bigwedge\limits_i s_i < s_{i+1}$ is realizable in $\langle \omega, <, S(x)\rangle$

is realized with probability one for each $\mathbb{P} \in \mathcal{C}_+$.

We demonstrate this by studying an auxiliary Markov process on Σ^m, where m is the number of terms s_i occurring in the formula φ.

Let $(n_1, \ldots, n_m)$ realize the formula $\bigwedge\limits_i s_i < s_{i+1}$. Note that for all $k \in \omega$, $(n_1 + k, \ldots, n_m + k)$ also realizes $\bigwedge\limits_i s_i < s_{i+1}$.

The stochastic process

$$((W_1, n_1), \ldots, (W_m, n_m)) \to ((W'_1, n_1 + 1), \ldots, (W'_m, n_m + 1))$$

is inferred from the data (μ, ρ). The pair (μ, ρ) induce atime-homogeneous Markov Chain on Σ^m as follows: for each $(W_1, \ldots, W_m)$ assign initial probability

[4]Recall that a *stochastic matrix* ρ is a matrix such that the sum of the entries over each row and each column is 1. A stochastic matrix ρ is irreducible provided that for all $\sigma, \sigma' \in \Sigma$ there exists an $n \in \omega$ such that the $\rho^n(\sigma, \sigma') > 0$. In other words, each state is reachable from every other state after some finite number of steps with positive probability.

$\mu^*((W_1, \ldots, W_m))$ according to the probability that $((W_1, n_1), \ldots, (W_m, n_m))$ is realized given (μ, ρ) in the first n_m transitions. Note that by assumptions on μ and ρ,

$$\mu^*((W_1, \ldots, W_m)) > 0$$

for all $W_1, \ldots, W_m$. Likewise, define a stochastic matrix ρ^* on Σ^m by setting

$$\rho^*((W_1, \ldots, W_m), (W_1', \ldots, W_m')) = \prod_i \rho(W_i, W_i').$$

By assumption on ρ, $\rho^*(\sigma, \sigma') > 0$ for all $\sigma, \sigma' \in \Sigma^m$. Thus the data (μ^*, ρ^*) are themselves a time-homogeneous Markov process such that $\mu^*(\sigma) > 0$ and $\rho^*(\sigma, \sigma') > 0$ for all $\sigma \in \Sigma^m$.

In particular, ρ^* is an *irreducible* stochastic matrix and so by standard results in Markov theory [11, Theorem 6.4.4, 6.5.6] there exists a unique stationary distribution η_{ρ^*} on Σ^m capturing the asympototic probability that state $\sigma \in \Sigma^m$ is observed on the k trial, and since expected return times are finite for every irreducible Markov chain, $\eta_{\rho^*}(\sigma) > 0$. Thus, in the long run, with probability one relative to (μ, ρ) the sentence

$$\left(\bigwedge_i \varphi_{\mathcal{M}_i}(\overline{x}_i, s_i) \right) \wedge \bigwedge_i (s_i < s_{i+1})$$

is realized.

In other words, on such a model of the evolution of the universe, every consistent configuration of atomic sentences is realized with probability one according to this process. Hence, this theory is unfalsifiable.

The theories $T_{\mathcal{C}}$ naturally occur as a formal model of the universe as a theormodynamic fluctuation. The idea that the universe is merely a fluctuation has been discarded by many prominent physicists such as Feynman and Carroll; it is worth investigating how these arguments dovetail with the present discussion of their falsifiability.

Feynman argues that we can refute this hypothesis, writing:

> Thus one possible explanation of the high degree of order in the present-day world is that it is just a question of luck. Perhaps our universe happened to have had a fluctuation of some kind in the past, in which things got somewhat separated, and now they are running back together again...
>
> [F]rom the hypothesis that the world is a fluctuation, all of the predictions are that if we look at a part of the world we have never seen before, we

> will find it mixed up, and not like the piece we just looked at. If our order were due to a fluctuation, we would not expect order anywhere but where we have just noticed it... is Every day they turn their telescopes to other stars, and the new stars are doing the same thing as the other stars. We therefore conclude that the universe is not a fluctuation. [17, Lecture 46-5]

On this account, from the fact that we observe order—the aggregate of all of our observations of the universe—we can conclude that the universe is not a fluctuation.

At first glance this argument appears to be an argument from falsification:

1	The Fluctuation Hypothesis entails that the universe is disordered.
2	We observe order in the universe.
3	The Fluctuation Hypothesis is false.

After all, it appears to be framed as a *reductio ad absurdum*, but the inference is more subtle than that. If by the fluctuation hypothesis we understand it to mean that the universe is generated probabilistically in the manner described above, then observing order of arbitrarily large complexity is in fact a deductive consequence of the theory $T_{\mathcal{C}}$.

The tension here comes from a quirk of the probabilistic framework and its relation to first-order logic; while the probability of a *specific* observer witnessing a given highly-ordered conjunction of atomic and negations of atomics formulas will be quite low, nevertheless the theory *predicts* that all such observations will be witnessed. In other words, two notions of *prediction* are at play: in one sense, the theory *entails* that with probability 1 the state that is observed will happen, all the while entailing that the observer in question witnesses a sequence of *low probability states*. Carroll [6] refers to this latter property of the fluctation theory as rendering observers "cognitively unstable" in the sense that the theory in question actively thwarts inductive reasoning as understood by Bayesian confirmation theory.

What Feynman has in mind, most likely, is an anthropic principle of the kind that says we should only affirm/consider theories T which themselves make it highly probably that *our own* inductive reasoning is highly conducive to truth.

Much ink has been spilled over anthropic principles in connection with the hypothesis that the universe is in some manner random [5], but the results of this section indicate that such theories suffer the defect of unfalsifiability. While being unfalsifiable does not refute the *truth* of the hypothesis, it does show that the hypothesis is not amenable to being refuted by way of finitary modes of data acquisition.

2.4 The Static Case: How Much of a Theory is Falsifiable?

The static models of falsifiability typically concern themselves with questions of how close a theory is to being universally axiomatizable; after all, the more universal sentences a theory implies, in principle the more falsfiable the theory becomes.

Simon and Groen [38], in their work on Ramsification and the Second-Order definability of theories, isolate a notion on *pseudoelementary* classes $\mathbb{K}$ they call FITness which they claim isolates the ideal scientific theories: On their account, a pseudoelementary class $\mathbb{K}$ is FIT if and only if it is a scientific theory. They show that for pseudoelementary $\mathbb{K}$, being FIT implies its universal axiomatizability.[5] Generalizing their definition to arbitrary classes of $\mathcal{L}$-structures $\mathbb{K}$ closed under isomorphism, I show that for finite languages $\mathbb{K}$ being FIT entails that $\mathbb{K}$ is elementary and, in fact, universally axiomatizable. This result substantially generalizes their result over finite languages.

I then turn my attention to an argument given by Chambers et al. [7] that argues that being universally axiomatizable is not sufficient grounds to call a theory falsifiable. Instead, they identify the falsifiable sentences with a class of universal sentences they call UNCAF (a universal negation of a conjunction of atomic formulas). In turn, I argue that their argument implicitly assumes that the underlying predicates $P \in \mathcal{L}$ exhibit mere Σ_1 behavior and thus that their argument reaches too far in its conclusions.

As a final foray into the static case of falsification, I consider how falsification intersects with the dividing lines of classification theory. It is not too difficult to show that under very mild restrictions on the language $\mathcal{L}$, NIP theories entail a great deal of nontrivial $\forall_1$ sentences and are highly falsifiable. Of note, in NIP theories each formula φ is equipped with a notion of dimension known as the VC-dimension of a class, which in a sense measures the effective falsifiability of membership in the class of hypotheses it defines.

On the other hand, recent work of Kruckman and Ramsey [27] and, independently, Jeřábek [21], yield examples of NSOP_1 and simple theories which are *un*falsifiable. While there are many NSOP_1 theories which are falsifiable, in a sense NIP is individuated among the dividing lines in model theory as a class of highly-falsifiable theories.

[5] The proof presented in their paper is incorrect. However, this error is fixed in this disseration.

FITness: The Finite Signature Case

We begin our investigation of the static case of falsification by exploring the notion of FITness—the *finite* and *irrevocable*—testability of a theory. Simon and Groen [38] argue that, at least when $\mathbb{K}$ is a pseudoelementary class, $\mathbb{K}$ being FIT is necessary and sufficient for $\mathbb{K}$ to be a scientific theory. They purport to show that if $\mathbb{K}$ is FIT and pseudoelementary, then $\mathbb{K}$ is $\forall_1$ axiomatizable. In this section I show that so long as the signature $\mathcal{L}$ is finite, the requirement that $\mathbb{K}$ is pseudoelementary is unneccessary; all that is needed is that $\mathbb{K}$ is closed under $\mathcal{L}$-isomorphism.

Definition 2.7. Let $\mathcal{L}$ be a language and $\mathbb{K}$ a class of $\mathcal{L}$-structures. $\mathbb{K}$ is said to be *FIT* provided that

i $\mathbb{K}$ is *finitely testable*, i.e., $\mathbb{K}$ is nontrivial:

$$\mathbb{K} \neq \mathrm{Str}(\mathcal{L})$$

and for every $\mathcal{M} \in \mathrm{Str}(\mathcal{L})$,

$$(\forall \mathcal{N} \in \mathrm{Str}(\mathcal{L})[(|\mathcal{N}| < \aleph_0 \wedge \mathcal{N} \subseteq_{\mathcal{L}} \mathcal{M}) \rightarrow \mathcal{N} \in \mathbb{K}]) \rightarrow \mathcal{M} \in \mathbb{K},$$

ii and $\mathbb{K}$ is *irrevocably testable*, i.e., for every $\mathcal{M} \in \mathrm{Str}(\mathcal{L})$

$$\mathcal{M} \in \mathbb{K} \rightarrow (\forall \mathcal{N} \in \mathrm{Str}(\mathcal{L})[(|\mathcal{N}| < \aleph_0 \wedge \mathcal{N} \subseteq_{\mathcal{L}} \mathcal{M}) \rightarrow \mathcal{N} \in \mathbb{K}]). \qquad \Diamond$$

In the case of a finite relational language $\mathcal{L}$, any FIT class $\mathbb{K}$ is universally axiomatizable. This substantially weakens the assumption on $\mathbb{K}$ given in the original paper of Simon and Groen at the cost of working within a more limited class of languages.

Theorem 2.4. Let $\mathbb{K}$ be a FIT class of a structures over a finite relational language $\mathcal{L}$ closed under isomorphism. Then $\mathbb{K}$ is universally axiomatizable. ⧫

Proof. We begin by giving a first-order axiomatization of $\mathbb{K}$. For each finite $\mathcal{N} \in \mathbb{K}$, let $\varphi_{\mathcal{N}}$ be the formula in $|\mathcal{N}|$ many free variables given by $\bigwedge_{\varphi \in diag(\mathcal{N})} \varphi$. This formula expresses the isomorphism type of $\mathcal{N}$ relative to the fixed enumeration $x_1, \ldots, x_n$: if $\mathcal{N} \simeq_{\mathcal{L}(\overline{x})} \mathcal{N}'$ then $\varphi_{\mathcal{N}}$ is equivalent to $\varphi_{\mathcal{N}'}$. Let $\mathbb{K}[n] = \{\mathcal{N} \in \mathbb{K} \,|\, |\mathcal{N}| \leq n\} / \simeq_{\mathcal{L}(\overline{x})}$. Since the language $\mathcal{L}(\overline{x})$ is finite and only includes relations and constant symbols, for each $n \in \omega$ there are only finitely $\mathcal{L}(\overline{x})$-isomorphism classes in $\mathbb{K}$ of size $\leq n$, so $\mathbb{K}[n]$ is finite. Let ψ_n be the sentence

$$\psi_n = \forall x_1, \ldots \forall x_n \left(\bigwedge_{i \neq j} x_i \neq x_j \rightarrow \bigvee_{[\mathcal{M}] \in \mathbb{K}[n]} \varphi_{\mathcal{M}} \right).$$

By construction, each ψ_n is a universal sentence, as the disjunction $\left(\bigvee_{[\mathcal{M}]\in\mathbb{K}[n]} \psi_{\mathcal{M}}\right)$ is a disjunction of finitely many Boolean combinations of atomic formulas.

Let $T_{\mathbb{K}} = \{\psi_n\}_{n\in\omega}$. I claim that $\mathbb{K} = \mathrm{Mod}(T_{\mathbb{K}})$. To see this, suppose that $M \vDash T_{\mathbb{K}}$. Because $\mathbb{K}$ is finitely testable it suffices to show that every finite substructure of $\mathcal{M}$ is a member of $\mathbb{K}$. Let $\mathcal{N}$ be a substructure of $\mathcal{M}$ of size n. Since ψ_n is universal, $N \vDash \psi_n$ and so $\mathcal{N} \vDash \psi_{\mathcal{N}'}$ for some $\mathcal{N}'$ isomorphic to a member of $\mathbb{K}$. Since $\mathbb{K}$ is closed under isomorphism, $\mathcal{N} \in \mathbb{K}$. Thus $\mathcal{M} \in \mathbb{K}$.

Conversely, suppose that $\mathcal{M} \in \mathbb{K}$. To show that $\mathcal{M} \vDash T_{\mathbb{K}}$, it suffices to show that $\mathcal{M} \vDash \psi_n$ for each n. Let $(m_1, \ldots, m_n) \in \mathcal{M}^n$ be a variable assignment. The set $\mathcal{N} = \{m_1, \ldots, m_n\}$ is a set of size $\leq n$ and is a substructure of M. By the irrevocable testability of $\mathbb{K}$, $N \in \mathbb{K}$. Thus, $\mathcal{N} \vDash \varphi_{[\mathcal{N}]}$. Thus $\mathcal{M} \vDash \psi_n$. □

Moreover, a similar argument works to show that a FIT class closed under isomorphism over an arbitrary finite language is universally axiomatizable.

Theorem 2.5. Let $\mathbb{K}$ be a FIT class of structures over a finite language $\mathcal{L}$ closed under isomorphism. Then $\mathbb{K}$ is universally axiomatizable. ⧫

Proof. Same as the above, but by defining the axiom scheme ψ_n as follows. For a function symbol f, we denote the arity of f by $ar(f)$.

Let $\chi_n(x_1, \ldots, x_n)$ be the formula given by

$$\chi_n = \left(\bigwedge_{f\in\mathcal{L}} \left[\bigwedge_{I\subset[n]^{ar(f)}} \bigvee_{0<j<n} f(\overline{x}_I) = x_j \right] \wedge \bigwedge_{c\in\mathcal{L}} \bigvee_{0<i\leq n} x_i = c \right),$$

where $\bigwedge_{f\in\mathcal{L}}$ and $\bigwedge_{c\in\mathcal{L}}$ are understood to be $\top$ in case $\mathcal{L}$ contains no function or constant symbols respectively.

This formula expresses that the set $x_1, \ldots, x_n$ is an $\mathcal{L}$-structure of size $\leq n$, as it expresses that $x_1, \ldots, x_n$ is closed under all function symbols $f \in \mathcal{L}$ and contains all constants $c \in \mathcal{L}$. Since $\mathcal{L}$ is finite, this is a quantifier-free first-order formula.

As above, let $T_{\mathbb{K}}$ be axiomatized by

$$\psi_n = \forall x_1, \ldots, x_n \left(\chi_n \to \bigvee_{[\mathcal{M}]\in\mathbb{K}[n]} \varphi_{\mathcal{M}} \right).$$

A nearly identical argument as before suffices to show that $\mathbb{K}$ is axiomatized by $T_{\mathbb{K}}$. Let $T_{\mathbb{K}} = \{\psi_n\}_{n\in\omega}$. I claim that $\mathbb{K} = \mathrm{Mod}(T_{\mathbb{K}})$. To see this, suppose that $M \vDash T_{\mathbb{K}}$. Because $\mathbb{K}$ is finitely testable it suffices to show that every finite substructure of $\mathcal{M}$

is a member of $\mathbb{K}$. Let $\mathcal{N}$ be a substructure of $\mathcal{M}$ of size n. Since ψ_n is universal, $\mathcal{N} \vDash \psi_n$. Since $\mathcal{N}$ is an $\mathcal{L}$-structure of size at most n, $\mathcal{N} \vDash \chi_n$, so $\mathcal{N} \vDash \psi_{\mathcal{N}'}$ for some $\mathcal{N}'$ isomorphic to a member of $\mathbb{K}$. Since $\mathbb{K}$ is closed under isomorphism, $\mathcal{N} \in \mathbb{K}$. Thus $\mathcal{M} \in \mathbb{K}$.

Conversely, suppose that $\mathcal{M} \in \mathbb{K}$. To show that $\mathcal{M} \vDash T_{\mathbb{K}}$, it suffices to show that $\mathcal{M} \vDash \psi_n$ for each n. Let $(m_1, \ldots, m_n) \in \mathcal{M}^n$ be a variable assignment. The set $\mathcal{N} = \{m_1, \ldots, m_n\}$ is a set of size $\leq n$. If $\mathcal{N}$ is an $\mathcal{L}$-structure, then $\mathcal{N}$ is a substructure of $\mathcal{M} \in \mathbb{K}$ so $\mathcal{N} \vDash \varphi_{[\mathcal{N}]}$ and hence

$$\mathcal{M} \vDash \chi_n(m_1, \ldots, m_n) \to \bigvee_{[\mathcal{M}] \in \mathbb{K}[n]} \varphi_{\mathcal{M}}(m_1, \ldots, m_n).$$

If $\mathcal{N}$ is not a substructure, then the variable assignment satisfies $\mathcal{M} \vDash \neg\chi_n(m_1, \ldots, m_n)$ and so

$$\mathcal{M} \vDash \chi_n(m_1, \ldots, m_n) \to \bigvee_{[\mathcal{M}] \in \mathbb{K}[n]} \varphi_{\mathcal{M}}.$$

Thus $\mathcal{M} \vDash \psi_n$ for all n, so $\mathcal{M} \vDash T_{\mathbb{K}}$. □

Proposition 2.4. Suppose that T is a universally axiomatizable class over a finite signature $\mathcal{L}$.

1. If $\mathcal{L}$ is relational, then $\mathrm{Mod}(T)$ is a FIT class.
2. There exist finitely axiomatizable T which are not FIT. ⧫

Proof. Suppose that $\mathcal{L}$ is relational. We need to show that for all $\mathcal{L}$-structures $\mathcal{M}$, $\mathcal{M} \vDash T$ if and only if $\mathcal{N} \vDash T$ for all finite $\mathcal{N} \subset \mathcal{M}$.

Suppose that $\mathcal{M} \vDash T$. Then since T is universally axiomatizable, $\mathcal{N} \vDash T$ for all substructures $\mathcal{N} \subset \mathcal{M}$. On the other hand, suppose $\mathcal{M} \nvDash T$. Then there exists a $\forall_1$ sentence $\varphi \in T$ such that $\mathcal{M} \vDash \neg\varphi$. Since φ is $\forall_1$, $\neg\varphi$ is $\exists_1$, there exists a witness $\overline{m}$ to $\mathcal{M} \vDash \neg\varphi$. The finitely generated substructure $\mathcal{M}_0 = \langle m \rangle \subset \mathcal{M}$ also satisfies $\mathcal{M}_0 \vDash \neg\varphi$. Since $\mathcal{L}$ is relational, $\mathcal{M}_0$ is a finite structure, so $\mathcal{M} \nvDash T$ ensures that there is a finite $\mathcal{M}_0 \subset \mathcal{M}$ with $\mathcal{M}_0 \nvDash T$. Thus the models of T form a FIT class.

On the other hand, let $\mathcal{L} = \{f(x), g(x), c\}$ and let T be the theory given by the single universal axiom: $(\forall x)\ g(x) \neq x$. We now exhibit an example of an $\mathcal{L}$-structure $\mathcal{M}$ such that $\mathcal{M} \nvDash T$ but $\mathcal{N} \vDash T$ for all finite substructures $\mathcal{N} \subset \mathcal{M}$. Let $\mathcal{M}$ have domain ω and interpret $f(x) = S(x)$ the successor function, $g(x) = x$ the identity function, and $c = 0$. Note that $\mathcal{M}$ has no finite substructures, so vacuously $\mathcal{N} \vDash T$ for all finite structures $\mathcal{N} \subset \mathcal{M}$. However, $\mathcal{M} \nvDash T$, so $\mathcal{M}$ is not FIT. □

FITness: The Arbitrary Language Case

The setting in which Simon and Gröen work involves a distinction between *observational* and *theoretical* scientific terms. Let $\mathcal{L} = \mathcal{L}_o \cup \mathcal{L}_t$ be a language partitioned into the *observational* language $\mathcal{L}_o$ and *theoretical language* $\mathcal{L}_t$. Let Σ be an $\mathcal{L}$-theory. There is, of course, the class of models of the theory:

$$\mathrm{Mod}(\Sigma) = \{\mathcal{M} \mid \mathcal{M} \vDash \Sigma\} \subset \mathrm{Str}(\mathcal{L}).$$

By definition, this class is *elementary*, meaning that it is first-order axiomatizable. However, the class $\mathrm{Mod}(\Sigma)$ is *not* the appropriate class of structures to look at, for if there is a true *o/t* distinction then the scientist only has epistemic access to the *observable* structure. Instead, Sneed [40] isolates the fundamental relation between scientific $\mathcal{L}$-theory Σ and some $\mathcal{L}_o$-structure $\mathcal{N}$ of observations is that of *application*: say that Σ *applies to* $\mathcal{N}$ just in case the $\mathcal{L}_o$-structure $\mathcal{N}$ can be expanded[6] to a full $\mathcal{L}$-structure $\widetilde{N}$ such that $\widetilde{N} \vDash \Sigma$. The *pseudoelementary* class of such structures is given by:

$$\mathrm{Mod}^*(\Sigma) = \{\mathcal{M}|_{\mathcal{L}_o} \mid \mathcal{M} \vDash \Sigma\} \subset \mathrm{Str}(\mathcal{L}_o).$$

In the case of pseudoelementary classes $\mathbb{K} = \mathrm{Mod}^*(\Sigma)$, we are able to drop the hypothesis that $\mathcal{L}$ is a finite language to conclude that an $\mathcal{L}_o$-FIT $\mathbb{K}$ is $\forall_1$-axiomatizable. This is the original result of Simon and Groen [38].

Proposition 2.5. Let Σ be an $\mathcal{L}_o$-*FIT* theory. Then the class $\mathrm{Mod}^*(\Sigma)$ is an elementary class and admits a universal axiomatization.[7] ⧫

Proof. We recall a theorem of model theory [20, Theorem 6.6.7]:

> *Let $\mathcal{L}$ be a first-order language and $\mathbb{K}$ be a pseudo-elementary class of $\mathcal{L}$-structures. Suppose that $\mathbb{K}$ is closed under taking substructures. Then $\mathbb{K}$ is axiomatized by a set of $\forall_1$ $\mathcal{L}$-sentences.*

Since $\mathrm{Mod}^*(\Sigma)$ is pseudo-elementary, it suffices to show that $\mathrm{Mod}^*(\Sigma)$ is closed under substructures. Let $\mathcal{M} \in \mathrm{Mod}^*(\Sigma)$ and let $\mathcal{N} \subset_{\mathcal{L}_o} \mathcal{M}$ be a substructure. To show that $\mathcal{N} \in \mathrm{Mod}^*(\Sigma)$, the *finite testability* implied by $\mathcal{L}_o$-*FIT*-ness tells us that we need only check that for every finite substructure $\mathcal{N}_k \subset_{\mathcal{L}_o} \mathcal{N}$ satisfies $\mathcal{N}_k \in \mathrm{Mod}^*(\Sigma)$. Since every such $\mathcal{N}_k$ is an $\mathcal{L}_o$-substructure of $\mathcal{M}$, the *irrevocability* of $\mathcal{L}_o$-*FIT*-ness ensures that $\mathcal{N}_k \in \mathrm{Mod}^*(\Sigma)$. □

[6] An *expansion* of an $\mathcal{L}_o$ structure $\mathcal{N}$ to an $\mathcal{L}$-structure $\widetilde{N}$ is an $\mathcal{L}$-structure where the domain of $\widetilde{\mathcal{N}}$ is $\mathcal{N}$ and all of the symbols in $\mathcal{L}_o$ are interpreted as is $\mathcal{N}$.

[7] In [38] Simon claims that this result follows from the Łoś-Tarski theorem. However, the Łoś-Tarski theorem applies to elementary classes, whereas the class in question—$\mathrm{Mod}^*(\Sigma)$—is a pseudoelementary class. Thus a new proof is needed.

That is, *FIT*-ness implies that the pseudo-elementary class of $\mathcal{L}_o$-structures expandable to models of Σ is not only *elementary*, but is in fact axiomatizable by *universal* axioms.

A partial converse can be given for the case of relational observational languages $\mathcal{L}_o$:

Proposition 2.6. Suppose $\mathrm{Mod}^*(\Sigma)$ is a universally axiomatized class of $\mathcal{L}_o$-structures, axiomatized by T^*_Σ, such that

i Σ has nontrivial observational consequences, i.e.,

$$\mathrm{Mod}^*(\Sigma) \neq \mathrm{Str}(\mathcal{L}_o),$$

and

ii $\mathcal{L}_o$ is a relational language.

Then Σ is $\mathcal{L}_o$-*FIT*. ♦

Proof. To show that Σ is $\mathcal{L}_o$-*FIT* we must show both *finite testability* and *irrevocable testability*.

Finite testability: Since $\mathrm{Mod}^*(\Sigma) \neq \mathrm{Str}(\mathcal{L}_o)$, it follows that that $\mathrm{Mod}(\Sigma) \neq \mathrm{Str}(\mathcal{L})$, for otherwise *all* $\mathcal{L}_o$ structures would be reducts of models of Σ.

We now need to show that if, for all $\mathcal{M}_k \subset_{\mathcal{L}_o} \mathcal{M}$ finite, $\mathcal{M}_k \in \mathrm{Mod}^*(\Sigma)$, then $\mathcal{M} \in \mathrm{Mod}^*(\Sigma)$. It is a known result [28, Exercise 2.5.20] that any structure $\mathcal{M}$ is $\mathcal{L}_o$-embeddable[8] into an ultraproduct of its finitely-generated substructures. Since $\mathcal{L}_o$ is relational, the finitely-generated substructures are precisely the *finite* substructures. Thus, there is an $\mathcal{L}_o$-embedding $\iota : \mathcal{M} \hookrightarrow \prod_{\mathcal{U}} \mathcal{M}_k$ for $\mathcal{U}$ any nonprincipal ultrafilter over the collection of finite substructures of $\mathcal{M}$. Now, as $\mathcal{M}_k \vDash T^*_\Sigma$ for all finite $\mathcal{M}_k \subset_{\mathcal{L}_o} \mathcal{M}$, $\prod_{\mathcal{U}} \mathcal{M}_k \vDash T^*_\Sigma$. Since T^*_Σ is universally axiomatizable and $\mathcal{M} \subset_{\mathcal{L}_o} \prod_{\mathcal{U}} \mathcal{M}_k$, $\mathcal{M} \vDash T^*_\Sigma$. But this means that $\mathcal{M} \in \mathrm{Mod}^*(\Sigma)$. So Σ is finitely testable.

Irrevocable testability: We need to show that if $\mathcal{M} \in \mathrm{Mod}^*(\Sigma)$ then for all $\mathcal{M}_k \subset_{\mathcal{L}_o} \mathcal{M}$ finite, $\mathcal{M}_k \in \mathrm{Mod}^*(\Sigma)$. Since T^*_Σ is universally axiomatizable, *any* $\mathcal{L}_o$-substructure of $\mathcal{M}$ is a model of T^*_Σ. In particular, each finite $\mathcal{M}_k \subset_{\mathcal{L}_o} \mathcal{M}$ is a model of T^*_Σ and is therefore a member of $\mathrm{Mod}^*(\Sigma)$, as desired.

Hence Σ is $\mathcal{L}_o$-*FIT*. □

[8] Not necessarily elementarily.

The two properties defining *FIT*-ness warrant scrutiny in virtue of their strong implications. We may view the finitely testable hypothesis as a *local compactness* principle: in the stated form it says that if every finite $\mathcal{M}_k \subset \mathcal{M}$ is *consistently expandable to a model of* Σ, so too is $\mathcal{M}$. The irrevocability hypothesis expresses the closure of the class $\mathrm{Mod}^*(\Sigma)$ under (finite) substructures, which together with finite testability implies closure under substructures.

Moreover, when working with a relational language the semantic criterion of *FIT*-ness is equivalent to the universal axiomatizability of the *observable* consequences of the theory. Thus, on the Simon-Groen view, given a universally axiomatizable $\mathcal{L}_o$ theory T *any* $\mathcal{L}_o$-conservative extension of T to an $\mathcal{L}$-theory T' is scientific. For instance, consider adding unary predicate symbols $P_1, \ldots, P_n$ and defining

$$T_m = T \cup \{\forall x \left(\bigvee_{1 \leq i \leq n} P_i(x) \right)\}.$$

the $\mathcal{L}_o$-consequences of T_m are the $\mathcal{L}_o$-consequences of T and so T_m is *FIT* and therefore scientific. By construction, however, the truth of the axioms of T_m are independent from any collection of observational data.

FITness and Finite Generation

The definition of FITness required that membership in a class $\mathbb{K}$ be witnessed by all finite substructures themselves being members of $\mathbb{K}$. However, except in the relational case, a substructure being finitely generated does not imply that that substructure is finite. In this section we consider the analogous notion of FITness obtained by replacing "finite" with "finitely generated" everywhere in the definition of FITness.

Definition 2.8. Let $\mathcal{L}$ be a language and $\mathbb{K}$ a class of $\mathcal{L}$-structures. $\mathbb{K}$ is said to be *fg-FIT* provided that

i $\mathbb{K}$ is *fg testable*, i.e. $\mathbb{K}$ is nontrivial:

$$\mathbb{K} \neq \mathrm{Str}(\mathcal{L})$$

and that for every $\mathcal{M} \in \mathrm{Str}(\mathcal{L})$,

$$(\forall \mathcal{N} \in \mathrm{Str}(\mathcal{L})[(\mathcal{N} \text{ is finitely-generated} \wedge \mathcal{N} \subseteq_{\mathcal{L}} \mathcal{M}) \rightarrow \mathcal{N} \in \mathbb{K}]) \rightarrow \mathcal{M} \in \mathbb{K},$$

ii and $\mathbb{K}$ is *fg-irrevocably testable*, i.e. for every $\mathcal{M} \in \mathrm{Str}(\mathcal{L})$

$$\mathcal{M} \in \mathbb{K} \rightarrow (\forall \mathcal{N} \in \mathrm{Str}(\mathcal{L})[(\mathcal{N} \text{ is finitely-generated} \wedge \mathcal{N} \subseteq_{\mathcal{L}} \mathcal{M}) \rightarrow \mathcal{N} \in \mathbb{K}]). \ \Diamond$$

The FITness and fg-FITness of a class are generally inequivalent.

Proposition 2.7. Let $\mathcal{L} = \{+, -, \times, 0, 1\}$ be the language of rings, and let $\mathbb{K}$ be the class of all rings of positive characteristic, i.e. rings such that there exists an $n \in \omega$ such that $\underbrace{1 + \cdots + 1}_{n \text{ times}} = 0$. Then $\mathbb{K}$ is fg-FIT but not first-order axiomatizable. In particular, $\mathbb{K}$ is not FIT. ⧫

Proof. To show that $\mathbb{K}$ is fg-FIT, it suffices to show that for a ring R, $R \in \mathbb{K}$ just in case every finitely-generated subring of R is in $\mathbb{K}$. Suppose that $R \in \mathbb{K}$. This is witnessed by the quantifier-free formula $\underbrace{1 + \cdots + 1}_{n \text{ times}} = 0$, so any subring $R' \subset R$ is also a member of $\mathbb{K}$. Conversely, if $R \notin \mathbb{K}$ then $\langle 1 \rangle$ is infinite and therefore $\langle 1 \rangle \notin \mathbb{K}$.

To show that $\mathbb{K}$ is not first-order axiomatizable, it suffices to show that $\mathbb{K}$ is not closed under ultraproducts by [8, Theorem 4.1.12]. Note that each finite field F_p is a member of $\mathbb{K}$. Let $\mathcal{U}$ be a nonprincipal ultrafilter on the set of primes. Then $F = \prod_{\mathcal{U}} F_p$ is a field of characteristic zero, thus $F \notin \mathbb{K}$. Hence $\mathbb{K}$ is not first-order axiomatizable. Therefore, by Theorem 2.4, $\mathbb{K}$ is not FIT. □

Moreover, unlike the FIT case, every universally axiomatizable theory is fg-FIT.

Proposition 2.8. Suppose that T is a universally axiomatizable class over an arbitrary signature $\mathcal{L}$. Then T is fg-FIT. ⧫

Proof. We need to show that for all $\mathcal{L}$-structures $\mathcal{M}$, $\mathcal{M} \vDash T$ if and only if $\mathcal{N} \vDash T$ for all finitely-generated $\mathcal{N} \subset \mathcal{M}$.

Suppose that $\mathcal{M} \vDash T$. Then since T is universally axiomatizable, $\mathcal{N} \vDash T$ for all substructures $\mathcal{N} \subset \mathcal{M}$. On the other hand, suppose $\mathcal{M} \nvDash T$. Then there exists an $\forall_1$ sentence $\varphi \in T$ such that $\mathcal{M} \vDash \neg\varphi$. Since φ is $\forall_1$, $\neg\varphi$ is $\exists_1$, there exists a witness $\overline{m} \in \mathcal{M}^k$ to $\mathcal{M} \vDash \neg\varphi$. The finitely generated substructure $\mathcal{M}_0 = \langle m \rangle \subset \mathcal{M}$ also satisfies $\mathcal{M}_0 \vDash \neg\varphi$. Thus $\mathcal{M} \nvDash T$ ensures that there is a finite $\mathcal{M}_0 \subset \mathcal{M}$ with $\mathcal{M}_0 \nvDash T$. Thus the models of T form an fg-FIT class. □

Remarks on Signatures in FITness

In the above discussions regarding FITness, fg-FITness, and universal axiomatizability, it was shown that in the case of a finite relational language, these notions are equivalent without a background assumption on the class $\mathbb{K}$ beyond closure under $\mathcal{L}$-isomorphism. However, these notions began to decouple in the case of languages with constant symbols and function symbols. This behavior is not so surprising; when

converting a function symbol f to a relation symbol by defining R_f by identifying $\forall x \forall y R_f(x, y) \leftrightarrow f(x) = y$, to eliminate the function symbol f from the language completely requires one to include an $\forall_2$ axiom of the form

$$\forall x \exists y R_f(x, y),$$

which in general will *not* be equivalent to a $\forall_1$ sentence. Thus, implicit in the inclusion of function symbols in the language is a $\forall_2$ axiom in a purely relational language.

UNCAF Theories

Motivated by theories of revealed preference in economics, Chambers et al. [7] argue that the empirical content of a theory is captured not by general universal sentences but instead by a special kind of universal sentence they term UNCAF.

Definition 2.9. [7, Definition 4] An UNCAF sentence in a language $\mathcal{L}$ is a universal negation of a conjunction of atomic formulas; that is, a sentence of the form

$$(\forall x_1, \ldots x_n) \neg \left(\bigwedge_{1 \leq i < m} \varphi_i(x_1, \ldots, x_n) \right)$$

where each φ_i is atomic. $\Diamond$

Perhaps surprisingly, they argue that sentences of the form $(\forall x) P(x)$ is *not* falsifiable by virtue of not being UNCAF, while $(\forall x) \neg P(x)$ is.

To argue this point, they write that

> substructures are unsatisfactory as mathematical models for observed data since they correspond to a situation in which the scientist observes the presence or absence of every possible relation among the elements in his data and, therefore, cannot accommodate partial observability.

While I agree with this general point, the conclusion that only UNCAF sentences have empirical content is too strong. For example, let $S(x)$ be the predicate "x is a swan" and $W(x)$ the predicate "x is white." The sentence "all swans are white," when formalized, is equivalent to

$$(\forall x) \neg (S(x) \wedge \neg W(x)),$$

which is not UNCAF owing to the presence of $\neg W(x)$ as a nested subformula. To conclude that this sentence has no falsificational content seems to run counter to the

usual conception of falsification: after all, *if* I were able to produce an example c such that $S(c) \wedge \neg W(c)$, I would immediately be able to infer that $W \notin \mathbb{K}$. However, on their model it is as if, when I go to the local bird sanctuary I am told that I may *only* record instances of white swans. One should not expect to be able to produce a counterexample to "all swans are white" under such constraints!

The way that Chambers et al. circumvent this worry is to note that for each predicate P one may add a new relation symbol $P^{\neg}$ together with the axiom

$$\forall x(\neg P(x) \leftrightarrow P^{\neg}(x)).$$

While this approach does formally work, it is somewhat awkward that this axiom itself is *not* UNCAF, as we see by reducing it to

$$\forall x \neg((\neg P(x) \wedge \neg P^{\neg}(x)) \vee (P(x) \wedge \neg P^{\neg}(x))).$$

Their understanding of falsification *qua* UNCAF-expressibility entangles two separate considerations: first, whether there is *in principle* any falsificational strategy on the basis of some configuration being witnessed by a finite set of data, and *second* whether the model of knowledge acquisition allows one to actually carry out the falsificational strategy. Their account corresponds to a model of knowledge acquisition in which at each stage one gains (at most) one **positive** (relative to $\mathcal{L}$) observation at a time, in a semidecidable fashion.

As an example, suppose that a researcher is observing an agent Ashley and wishes to falsify whether or not her preference relation is complete:

$$\forall x, y\left((x \leq y) \vee (y \leq x)\right).$$

To do so, the observer waits each day d to see whether the agent exhibits some preference relation between a can of Guayakí Enlighten Mint ready-to-drink Yerba Mate and a can of Guayakí Revel Berry that are sitting side-by-side in the office fridge, with no other items in potential consideration.

This experiment, as construed, is doomed to never falsify the experiment. After all, if there is some day d where Ashley surveys the fridge and takes a can of Enlighten Mint but not Revel Berry (resp. Revel Berry but not Enlighten Mint), then $EM \geq RB$ (resp. $RB \geq EM$) and therefore no refutation of the completeness axiom is possible in the context. Likewise, if the day that Ashley takes a can out of the fridge never comes, that *also* does nothing to falsify the completeness axiom.

So, what went wrong? Implicit in their semantics for the experiment is a suppressed existentially-defined quantifier. Let $R_A(x, y, d)$ be the relation that says "on

day d, agent A expressed a weak preference x over y." Then the formula $x \geq y$ in Chambers' terminology would not be $\forall_1$ but instead properly $\forall_2$:

$$\forall x, y \left((\exists t) R_A(x, y, t) \vee (\exists t) R_A(y, x, t) \right).$$

Therefore, the purported example of an unfalsifiable $\forall_1$ sentence is better and more directly modeled as an unfalsifiable $\forall_2$ sentence fully compatible with the standard account of falsification as a universal over an in-principle decidable primitive. What their point indicates is that the standard revealed preference relations in economics are *not* in-principle decidable, but instead are $\exists_1$-definable relative to the empirical relation $R_A(x, y, d)$ via

$$x \geq y := (\exists t) R_A(x, y, d).$$

If we take as epistemically primitive a $\exists_1$-definable relation $R(x, y)$ defined by an $\mathcal{L}$-formula $\exists c \varphi(x, y, c)$ with φ quantifier-free, then their result is clear. A sentence of form $(\forall_1) R(x, y)$ is an $\exists_2$ sentence, while an UNCAF sentence in the language $\mathcal{L}_R = \{R(x, y)\}$,

$$(\forall x_1, \ldots, x_n) \neg \left(\bigwedge_{i,j \in I} R(x_i, x_i) \right),$$

with $I \subset 2^{[n]}$ finite is equivalent to the $\mathcal{L}$-sentence:

$$(\forall x_1, \ldots, x_n) \left(\bigvee_{i,j \in I} \forall c \, (\neg \varphi(x_i, y_i, c)) \right)$$

which is equivalent to a $\forall_1$ sentence in $\mathcal{L}$.

Strength of Theories and their Falsifiability

Contrary to mere falsifiability, FITness, fg-FITness, and UNCAF-axiomatizable are typically not closed upwards under strength.

Proposition 2.9. Let $\mathbb{K}$ be a universally axiomatizable FIT or UNCAF class such that $\mathbb{K}$ has both finite and infinite models. Then the class $\mathbb{K}' \subset \mathbb{K}$ of infinite members of $\mathbb{K}$ is not FIT, fg-FIT, or UNCAF. ⧫

Proof. Let T be a universal axiomatization of $\mathbb{K}$. Then $\mathbb{K}'$ is axiomatized by $T \cup \{\psi_n\}_{n \in \omega}$ where ψ_n is the sentence $(\exists x_1, \ldots x_n) \bigwedge_{1 \leq i \neq j \leq n} x_i \neq x_j$.

Since $\mathbb{K}$ is closed under substructures and admits finite models, $\mathbb{K}'$ necessarily fails to be closed under substructures. Thus $\mathbb{K}'$ is not universally axiomatizable, and in particular is neither FIT nor UNCAF. □

Falsification and NIP

In the preceding sections, we have considered a sequence of refinements to the basic notion of falsifiability. We have seen, under mild conditions on the signature $\mathcal{L}$ and classes $\mathbb{K}$ the web of implications

$$\mathbb{K} \text{ FIT} \longrightarrow \mathbb{K} \ \forall_1 - \text{Axiomatizable} \longrightarrow \boxed{\mathbb{K} \text{ Falsifiable}}$$
$$\uparrow$$
$$\mathbb{K} \text{ has nontrivial UNCAF theory}$$

However, there are a great deal of hypotheses which do not readily fall into this framework at first glance.

For example, we are often interested in testing whether or not a (basic) relation $R(x_1, \ldots, x_n) \in \mathcal{L}$ is equivalent to some other (basic) relation $S(x_1, \ldots, x_n) \in \mathcal{L}$. This is easy to handle directly in our account of falsification; after all,

$$(\forall x_1, \ldots, x_n)(R(x_1, \ldots, x_n) \leftrightarrow S(x_1, \ldots, x_n))$$

is a $\forall_1$ sentence in $\mathcal{L}$ by assumption.

What if, instead, we are probing a more complicated question, such as whether or not $R(x_1, x_2)$ is a line in $\mathbb{R}^n$? In the language of rings augmented by an additional relation symbol

$$\mathcal{L} = \{+, \times, , 0, 1, <, R(x, y)\}$$

this is most easily expressed by the $\exists_2$ formula

$$T_L = (\exists a, b)(\forall x, y)(R(x, y) \leftrightarrow L(x, y; a, b, c))$$

where $L(x, y; a, b)$ is the sentence $ay + bx + c = 0$. Despite being $\exists_2$, this sentence has a great deal of falsificational content owing to the structure of the parametric family $L(x, y; a, b, c)$. From Euclidean geometry that between any two distinct points there exists a unique line. Letting $R^*(a, b, c, d)$ be the sentence

$$R^*(a, b, c, d) = ([(a \neq c) \vee (b \neq d)] \wedge R(a, b) \wedge R(c, d)).$$

Then

$$(\forall x, y)(\forall a, b, c, d)\left(R^*(a, b, c, d) \rightarrow \left(L\left(x, y, \frac{d-b}{c-a}, b - a\frac{d-b}{c-a}\right) \rightarrow R(x, y)\right)\right)$$

which is a nontrivial $\forall_1$ sentence. Thus, while T_L is $\exists_2$ it has nontrivial $\forall_1$ consequences.

These properties of lines is an example of the *VC finiteness* of the class. For the remainder of the section, we assume that $\mathbb{K}$ is an *elementary* class, axiomatized by some first-order set of sentences T.

Definition 2.10. [39, pg. 7-8] Let $\varphi(\overline{x};\overline{y})$ be a first-order formula in disjoint sets of free variables $\overline{x},\overline{y}$. With respect to this partition we say that φ is a *partitioned* formula.

Let $\mathcal{M} \in \mathbb{K}$. We say that $\varphi(\overline{x};\overline{y})$ $\mathcal{M}$-shatters a set $X \subset \mathcal{M}^{|\overline{x}|}$ just in case there is a set $Y \subset \mathcal{M}^{|\overline{y}|}$ such that for every subset $X' \subset X$ there exists $y' \in Y$ such that

$$\mathcal{M} \vDash \varphi(x;y') \leftrightarrow x \in X'$$

for all $x \in X$.

A partitioned formula is *NIP* provided for every $\mathcal{M} \in \mathbb{K}$, no infinite set is $\mathcal{M}$-shattered by φ.

The formula φ has Vapnik-Chervonenkis (VC) dimension, $VC(\varphi) \leq n$ just in case for all $\mathcal{M} \in \mathbb{K}$, no set of size n is $\mathcal{M}$-shattered. If φ has finite VC dimension then φ is said to be *VC finite*.

A theory T is NIP just in case every formula φ is NIP in the class $\mathbb{K} = \mathrm{Mod}(T)$. ♢

For elementary classes $\mathbb{K}$, a formula being NIP is related to its VC finiteness:

Proposition 2.10. Let $\mathbb{K}$ be an elementary class and $\varphi(\overline{x};\overline{y})$ a partitioned first-order formula. If φ is NIP, then φ has finite VC dimension. ⧫

Proof. This is an elementary consequence of the compactness theorem of First-Order Logic [39, Remark 2.3]. □

A first-order formula $\varphi(\overline{x};\overline{y})$ having VC dimension $\leq n$ is first-order expressible by a sentence $VC_n(\varphi)$; moreover, if φ is quantifier-free then the proposition $VC_n(\varphi)$ is a $\forall_1$ sentence.

Proposition 2.11. A formula $\varphi(x;y)$ having VC dimension $\leq n$ is first-order expressible in any language containing φ by a sentence $VC_n(\varphi)$. Moreover, if φ is $\exists_m/\forall_m$, then $VC_n(\varphi)$ is at most $\forall_{m+1}$.

In particular, if φ is quantifier-free then $VC_n(\varphi)$ is a $\forall$ sentence. ⧫

Proof. First, the proposition "$x_1, \ldots, x_n$ is shattered by $\{y_J\}_{J\subseteq[n]}$ in φ" is a Boolean combination of instances in φ:

$$\text{Shatter}_\varphi\left((x_i)_{1\leq i\leq n}, (y_J)_{J\subseteq[n]}\right) \leftrightarrow \bigwedge_{J\subseteq[n]} \bigwedge_{1\leq i\leq n} \Box_{i,J}\varphi(x_i, y_J)$$

where $\Box_{i,J}\varphi$ is $\neg\varphi$ if $i \notin J$ and φ if $i \in J$.

The proposition $VC_n(\varphi)$ is expressed by the following first-order sentence:

$$(\forall(x_i)_{1\leq i\leq n})\forall(y_J)_{J\subseteq[n]}\left(\left(\bigwedge_{1\leq i\neq j\leq n} x_i \neq x_j\right) \rightarrow \neg\,\text{Shatter}_\varphi\left((x_i)_{1\leq i\leq n}, (y_J)_{J\subseteq[n]}\right)\right),$$

as desired. □

Proposition 2.12. Let T be a complete NIP theory in a language $\mathcal{L}$ containing an m-ary relation symbol $R(\overline{x}, \overline{y})$ for some $n > 1$. Then T implies a nontrivial universal sentence. ⧫

Proof. Since T is complete NIP, for some T entails the $\forall_1$ sentence $VC_n(R(\overline{x}; \overline{y}))$ for some $n \in \omega$. Since R non-unary, $VC_n(R(\overline{x}; \overline{y}))$ is not a first-order validity, since the bipartite graph G_n on a disjoint set of vertices $[n] \cup 2^{[n]}$ given by $R(i, X) \leftrightarrow i \in X$ satisfies

$$G_n \vDash \neg VC_n(R(\overline{x}; \overline{y})).$$ □

In fact, since $VC_n(R(x; y)) \rightarrow VC_m(R(x; y))$ for all $m > n$, for a VC finite relation we get a nested chain of $\forall_1$ sentences. As we will see in our account of the dynamic case of falsification, this simple observation has very strong consequences in terms of understanding small-sample falsificational problems.

To explain the restriction about the language, we note that there are NIP unary theories entailing no nontrivial $\forall_1$ sentence. Recall [28, Definition 4.2.17][9] that a theory T is *κ-stable* for a cardinal κ if for every model $\mathcal{M} \vDash T$, $n \in \omega$, and $A \subseteq \mathcal{M}$ of size κ, the space of n-types with parameters in A has size κ

$$|S_n(A)| = \kappa.$$

A theory is stable provided it is κ-stable for some infinite κ. It is well known that stability implies NIP [37, Theorem 4.7].

[9] In the next section we will work with an alternative, equivalent definition of stability better-suited for our purposes.

Proposition 2.13. There exists a stable theory T in a unary language which entails no nontrivial universal sentence. ⧫

Proof. Let T be the theory in the language $\mathcal{L} = \{P(x)\}$ axiomatized by

$$\varphi_n = (\exists x_1, \ldots, x_n) \left(\bigwedge_{i \neq j < n} x_i \neq x_j \wedge \bigwedge_{i<n} P(x_i) \right)$$

and

$$\psi_n = (\exists x_1, \ldots, x_n) \left(\bigwedge_{i \neq j < n} x_i \neq x_j \wedge \bigwedge_{i<n} \neg P(x_i) \right).$$

This theory is clearly $\aleph_0$-categorical: any countable model $\mathcal{M}$ can be partitioned by

$$\mathcal{M} = P(\mathcal{M}) \cup \neg P(\mathcal{M})$$

with each definable set $P(\mathcal{M}), \neg P(\mathcal{M})$ countably infinite. If $\mathcal{M}, \mathcal{N} \vDash T$, then any pair of bijections

$$f_P : P(\mathcal{M}) \to P(\mathcal{N})$$

and

$$f_{\neg P} : \neg P(\mathcal{M}) \to \neg P(\mathcal{N})$$

induce an $\mathcal{L}$-isomorphism

$$f_P \cup f_{\neg P} : \mathcal{M} \to \mathcal{N}.$$

Moreover, T has no finite models, so by Vaught's test [28, Theorem 2.2.6] T is complete. Clearly, $\forall_1(T)$ contains only first-order validities.

This theory is ω-stable. Let A be a set of size $\leq \aleph_0$. The types over A are determined by specifying which coordinates x_i are equal to an element of A and, for those $x_i \notin A$, whether or not $P(x_i)$ or $\neg P(x_i)$. Thus, there are at most $(|A|+2)^n \leq \aleph_0$ types over A. □

Therefore, again under mild conditions on the language

$$\mathbb{K} \text{ NIP} \to \mathbb{K} \text{ Falsifiable}.$$

We observe that VC finiteness is not equivalent to universal axiomatizability.

Proposition 2.14. There exists an NIP T such that T is not universally axiomatizable. There exists a universally axiomatizable T such that T is not NIP. ⧫

Proof. Let DLO be the theory of dense linear orders in the language $\mathcal{L} = \{<\}$. Then T is not universally axiomatizable as all of its models are infinite and the language is relational. Concretely, we know $\mathbb{Q} \vDash DLO$ but no finite subset $X \subset \mathbb{Q}$ is a model of DLO. Since $\mathcal{L}$ is relational, X is a substructure.

On the other hand, let T be the (incomplete) theory of acyclic directed graphs in the language $R(x, y)$. T is universally axiomatizable by the collection φ_n of sentences defined by

$$\varphi_n = (\forall x_1, \ldots, x_n) \neg \left(R(x_n, x_1) \wedge \bigwedge_{1 \leq i < n} R(x_i, x_{i+1}) \right).$$

This class is not NIP as for each n the bipartite digraph G_n on a disjoint set of vertices $[n] \cup 2^{[n]}$ given by $R(i, X) \leftrightarrow i \in X$ satisfies

$$G_n \vDash \neg VC_n(R(x; y))$$

and is a model of T. □

Useful Examples of NIP Theories and VC finite Classes

The preceding section describes the relationship between NIP theories, VC finite classes, and falsification, but further argument is required to demonstrate that these phenomena actually appear in the kinds of hypotheses we seek to falsify.

To this end, the following powerful theorem proves the VC finiteness of a very wide class of geometrically-definable hypotheses.

We assume that the reader is familiar with the notion of an analytic function.

Definition 2.11. [10] Let $\mathcal{A}_{[-1,1]}$ be the set of functions $f : [-1, 1] \to \mathbb{R}$ which extends to an analytic function on an open neighborhood $U \supset [-1, 1]$. Let exp : $\mathbb{R} \to \mathbb{R}$ be the exponential map $\exp(x) = e^x$.

The restricted analytic exponential real field $\mathcal{R}$ is the structure

$$\mathcal{R} = (\mathbb{R}, +, -, 0, <, \exp, (f_j)_{f_j \in \mathcal{A}_{[-1,1]}})$$

The theory $\mathbb{R}_{an,exp}$ is the theory of the structure $\mathcal{R}$. ◊

In this structure, parametric families of equalities and inequalities between analytic functions on compact rectangular domains are definable. The following result shows that such families have finite VC-dimension:

Theorem 2.6. $\mathbb{R}_{an,exp}$ is NIP. Consequently, every first-order definable set in the theory of $\mathbb{R}_{an,exp}$ has finite VC dimension. ⧫

Proof. That $\mathbb{R}_{an,exp}$ is o-minimal is a classical theorem of van den Dries and Miller [10], together with the result that every o-minimal theory T is NIP, which can be found in Simon's book [39, Theorem A.6]. □

This theorem is extremely useful because it implies that not only is any parametric family of algebraic equations over a real field VC finite, but in fact any parametric family of semi-analytic inequalities is VC finite. This is of the utmost importance for examples stemming from physics, as it implies that even definable classes where one includes a model of measurement error can be VC finite.

As an illustration, we consider the example of the family of *fat lines*. By a fat line I mean a set $\tilde{L} \subseteq \mathbb{R}^2$ such that $(x, y) \in \tilde{L}(x, y; a, b, c, r)$ just in case (x, y) is most distance r from the line $ax + by + c = 0$.

Proposition 2.15. The class $\tilde{L}(x, y; a, b, c, r)$ of fat lines in $\mathbb{R}^2$ has finite VC dimension. ⧫

Proof. By Theorem 2.6, any set definable in the theory of $\mathbb{R}_{an,exp}$ is VC finite. Thus it suffices to give an explicit definition of this family. This is easily done: the formula $\varphi(x, y; a, b, r)$ given by

$$|ax + by + c|^2 < r(a^2 + b^2)$$

is a formula in the language of $\mathbb{R}_{an,exp}$ and defines the family of fat lines. □

This suggests an explanation for why many physical theories are so readily falsifiable: many of the predictions of physical theories can be cast in terms of determining membership in a real semianalytic set which expresses being some bounded error away from an analytic or algebraic set.

Falsification and Model-Theoretic Dividing Lines

It turns out that other classification-theoretic conditions on $\mathbb{K}$ fail to guarantee falsifiability.

Recall from classification theory the following definitions:

Definition 2.12. Let $\varphi(\overline{x}; \overline{y})$ be a partitioned formula. We say that

1. [37, Theorem 2.2.3(2)] φ is *stable* if there is no set $(c_n)_{n\in\omega}$ such that for every $k < \omega$,

$$\{\varphi(x, c_0), \varphi(x, c_1), \cdots, \varphi(x, c_{k-1}), \neg\varphi(x, c_k), \neg\varphi(x, c_{k+1}), \cdots\}_{n\in\omega}$$

is consistent.

2. [12, Definition 2.2] φ is $NSOP_1$ if there is no set of tuples $\{c_\sigma \,|\, \sigma \in 2^{<\omega}\}$ such that

   a) (Branch consistency) for every $\tau \in 2^\omega$ the set

   $$\{\varphi(x; c_{\tau|_m}) \,|\, m \in \omega\}$$

   is consistent, and

   b) (Lateral inconsistency) The set

   $$\{\varphi(x, c_{\sigma \cap \langle 1 \rangle}, \varphi(x, c_\gamma)\}$$

   is inconsistent for all $\gamma \supseteq \sigma \cap \langle 0 \rangle$.

A theory T is stable (resp. NSOP_1) provided every formula in T is stable (resp. NSOP_1). ◊

Jeřábek [21], and, independently, Kruckman and Ramsey [27] showed that

Theorem 2.7. Let $\mathcal{L}$ be a language. Then the model companion of the empty theory $T^\varnothing_\mathcal{L}$ exists, is complete, and is

1. stable if $\mathcal{L}$ is unary,

2. unstable NSOP_1 for any non-unary $\mathcal{L}$.

Moreover, $\forall_1(T^\varnothing_\mathcal{L})$ contains only validities, so $\mathrm{Mod}(T^\varnothing_\mathcal{L})$ is not a falsifiable class. ⧫

Proof. We prove only the "moreover" clause. The proof of the rest can be found in [21, Theorem B.1].

Suppose that φ is a $\forall_1$ sentence in $\mathcal{L}$ that is not a validity. We wish to show that $T \nvDash \varphi$. Without loss of generality we assume that

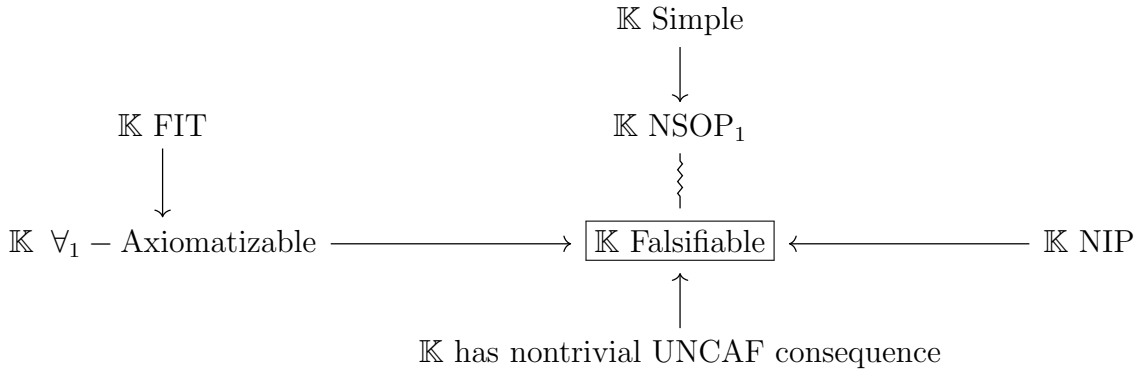
$$\varphi = (\forall x_1, \ldots x_n)\psi(x_1, \ldots, x_n)$$

with $\psi(x_1, \ldots, x_n)$ quantifier-free. Since φ is not a first-order validity, $\neg\varphi$ is satisfiable and is a $\exists_1$ formula. Let $\mathcal{M}$ be an $\mathcal{L}$ structure such that $\mathcal{M} \vDash \neg\varphi$. Since $T^\varnothing_\mathcal{L}$ is the theory of existentially closed $\mathcal{L}$-structures, $\mathcal{M}$ embeds into a model $\mathcal{M}' \vDash T$. By construction, $\mathcal{M}' \vDash \neg\varphi$, so $\varphi \notin T^\varnothing_\mathcal{L}$. □

To be sure, there exist falsifiable NSOP_1 classes. The theory of the random graph, for instance, contains the universal theory of graphs, which in particular includes the non-logical validity

$$\forall x \forall y (R(x, y) \leftrightarrow R(y, x)).$$

In short, NIP classes of structures yield examples of falsifiable structures incomparable to the notions we have thus far discussed. Thus, our picture of the relationship between falsifiable classes of falsifiability now looks like (again assuming mild assumptions on $\mathbb{K}$ and $\mathcal{L}$):



2.5 The Dynamic Case: Falsification and the Accumulation of Evidence

Moving on to the *dynamic* case of falsification, we turn to the account of falsification given by Formal Learning Theorists Juhl and Schulte [35]. For them, a hypothesis $\mathcal{H}$ is identified with a set of possible worlds. They consider the case where $\mathcal{H} \subset 2^\omega$. For them, such a hypothesis is *always falsifiable* provided that regardless of the results of some finite collection of observations, the hypothesis still has the potential to be falsified by some further collection of observational data. This, they show, is equivalent to the topological notion of the *nowhere density* of $\mathcal{H}$ inside 2^ω equipped with the product topology. This account gives a good account of *long-run* falsifiability, but fails to give a satisfactory account of *short-run* falsification as there are no bounds on *how long* it might take an agent to witness a crucial experiment. I define a *sample* along a set X is given by:

Definition 2.13. A *sample* of X is an injective function $f : \omega \to X$. A sample f is *full* provided f is bijective. ◊

Now, relative to a sample f we define a notion of the *surprise* of a hypothesis along the sample as follows:

Definition 2.14. Let X be a set, $f : \omega \to X$ a sample of X, and $\mathcal{H} \subset 2^X$ a hypothesis. The *surprise* of $\mathcal{H}$ is the function

$$S(\mathcal{H}, f, n) = 1 - \frac{|\mathcal{H}\restriction_{f([n])}|}{2^{|f([n])|}}. \qquad \Diamond$$

The surprise of $\mathcal{H}$ along the enumeration f is the relative proportion of the states of the world incompatible with $\mathcal{H}$. If $\mathcal{H}$ is highly surprising, then it is compatible with only a small number of observations along the sample.

Very closely related to the NIP theories discussed in the static case of falsification, the VC finite hypothesis classes are characterized by the ability to obtain uniform bounds on surprise independent of sample.

Finally, we turn our attention to the work of Mayo [29][30], who advocates for a strengthening of Null Hypothesis Statistical Testing as the foundation of statistical testing called *severe testing*. Mayo and other error statisticians ask the question

> When do data x provide good evidence for / a good test of hypothesis H?

The error statistician will invoke some form of a **Severity Principle** to answer this question:

> (Weak Severity Principle) Data x **does not** provide good evidence for H if x is the result of a **test procedure** T with very low probability of uncovering the falsity of H[31, p. 21].

A converse is given by:

> (Full Severity Principle) Data x provides good evidence for H to the extent that test T has been **severely passed** by H[31, p. 21].

However, the definition of severe testing via probabilistic notions is elusive. To round out the discussion of falsification, I show that the notion of surprise I defined in the context of always falsifiability is well-suited to give an account of a *well-defined* combinatorial analogue of severe testing I term *severe surprise*.

Definition 2.15. Let $f : \omega \to X$ be a sample, $\mathcal{H} \subset 2^X$ a hypothesis, $n \in \omega$, and $\epsilon > 0$. We say that $(\mathcal{H}, f, n)$ is *severely surprising* at level ϵ provided the observed data $x \in \mathcal{H}\restriction_{f([n])}$,

$$S(\mathcal{H}, f, n) > 1 - \epsilon$$

and

$$S(\mathcal{H}, f, n) > S(\mathcal{H}^c, f, n). \qquad \lozenge$$

Crucially, surprise is at its core a non-probabilistic notion. Key to this is the observation that surprise is subadditive:

Proposition 2.16. For all $\mathcal{H}, f, n$,

$$S(\mathcal{H}, f, n) + S(\mathcal{H}^c, f, n) \leq 1 \qquad \blacklozenge$$

and, in fact, any dense-codense pair $\mathcal{H}, \mathcal{H}^c \subset 2^X$ surprise 0 along any subsample. This is due to the simple fact that $\mathcal{H}_1 \cap \mathcal{H}_2 = \varnothing$ does not entail that the restriction of the hypotheses to a finite set X_0, $\mathcal{H}_1|_{X_0}$ and $\mathcal{H}_2|_{X_0}$, are disjoint.

Under this definition of severe surprise, one can show that VC finite classes are in fact severely surprising if true, uniformly in the size of the sample.

Thus, VC finiteness emerges as *the* core notion of dynamic case of falsification, being robust under arbitrary samples and uniquely endowed with felicitous finite-sample bounds.

The Formal Learning Theoretic Picture of Falsification

The learning-theoretic analysis of falsification given by Schulte and Juhl identifies *always falsifiability* of a hypothesis $\mathcal{H} \subset 2^X$ with the notion of *nowhere density* in the usual product topology on 2^X.[35, p. 10]

The framework of Formal Learning Theory, especially as developed by Kelly [25], provides a formal model of learning through observation. Schulte and Juhl [35] leverage this framework to give a topological characterization of Popperian falsification. On this model, an agent is idealized as being fed a countably infinite number of observations encoded by natural numbers and determining at each step n whether or not some property P holds of observation x_n: if $P(x_n)$, output 1 and if $\neg P(x_n)$, output 0. Such a sequence is called a *data stream*. Mathematically speaking, a data stream can be thought of simply as an element of Cantor space 2^ω. An *empirical hypothesis* is simply a set of data streams, and thus a subset of 2^ω.

Popper's solution to the demarcation problem says that the distinguishing feature of a scientific theory–construed as an empirical hypothesis–is its *falsifiability*. As Schulte and Juhl explain, a weak form of falsification is that under ideal circumstances the hypothesis can be conclusively ruled out on the basis of some observed relation or relations termed a *crucial experiment* of the theory. However, if the crucial experiment does not rule out the theory, it may be the case that there are no other crucial experiments to run.

A more robust notion of falsifiability—what they term *always falsifiability*—demands a preponderance of crucial experiments: given any finite collection of observations there exists a crucial experiment which may conceivably be run in the future. This notion expresses the idea that the scientific theory might never be *confirmed* at any finite time, since there are always potential observational paths refuting it in the future. Schulte and Juhl demonstrate that the always falsifiability of a hypothesis $\mathcal{H} \subset 2^\omega$ is equivalent to that hypothesis' *nowhere density* in the usual topology on 2^ω.

Within formal learning theory, a *world* is an element of 2^ω and a hypothesis $\mathcal{H}$ is identified with the subset $\mathcal{H} \subset 2^\omega$ of worlds in which $\mathcal{H}$ holds.

Given a hypothesis $\mathcal{H} \subset 2^\omega$, we construct a two-sorted structure $\mathcal{M}_\mathcal{H}$ in the language $\mathcal{L} = \{R(x,y), O(x), W(y)\}$ as follows:

1. The domain of $\mathcal{M}_\mathcal{H}$ is the disjoint union $\omega \cup \mathcal{H}$,
2. The predicate O consists of all of the *observations*: $\mathcal{M}_\mathcal{H} \vDash O(x)$ just in case $x \in \omega$
3. The predicate W consists of all of the *worlds*: $\mathcal{M}_\mathcal{H} \vDash W(y)$ just in case $y \in \mathcal{H}$.
4. $\mathcal{M}_\mathcal{H} \vDash R(x,y)$ just in case $\mathcal{M}_\mathcal{H} \vDash O(x) \wedge W(y)$,$x \in y$, and $y \in \mathcal{H}$.

In other words, the structure on $\mathcal{M}_\mathcal{H}$ is the bipartite graph on ω and $\mathcal{H}$ with each world $w \in \mathcal{H}$ encoding itself:

$$R(\mathcal{M}_H, w) = w.$$

It is worth noting that we can encode a lot of information into this framework. For instance, consider the structure $(\mathbb{Q}, <)$ with $<$ as the usual order on $\mathbb{Q}$. Then, enumerating $\mathbb{Q}^2$ as ω, we may regard the partitioned formula

$$\varphi(x; y_1, y_2) = x > y_1 \wedge x < y_2$$

as a hypothesis $\mathcal{H}_\varphi \subset 2^\omega$. We identify an $h \in \mathcal{H}$ with any one of its codes; that is, $h = \varphi(x; h_1, h_2)$ for some $h_1, h_2 \in \mathbb{Q}$. We then have a definable interpretation of the bipartite graph

$$\mathbb{Q} \cup \mathcal{H}_\varphi(\mathbb{Q})$$

with

$$R(x,h) \leftrightarrow x \in h \leftrightarrow (x \in \mathbb{Q} \wedge (\varphi(x; h_1, h_2)).$$

To make this fully model theoretic, the induced structure we study would not be restricted to only countable models; instead we would concern ourselves with

large, sufficiently saturated $\mathcal{M} \in \mathbb{K}$ and look at the induced bipartite structure with domain

$$\mathcal{M} \cup \mathcal{H}_\varphi.$$

We see that this is even necessary to capture all intervals in $\mathbb{Q}$ with real endpoints; restricting only to parameters $h_1, h_2 \in \mathbb{Q}$ we have only countable many elements in $\mathcal{H}_\varphi(\mathbb{Q})$, but $\mathcal{H}_\varphi(\mathbb{R}) \cap \mathbb{Q} \subset 2^{\mathbb{Q}}$ is strictly larger.

The above construction gives us a way to convert learning questions in model theory with the setup of Formal Learning Theory.

To maintain consistency with the Formal Learning Theory literature, we will work primarily in the standard setting of hypotheses $\mathcal{H} \subset 2^\omega$, knowing that we may choose to encode mathematical structures into this framework as needed.

Popper Dimension, VC dimension, and the Topology of Falsification

The learning-theoretic analysis of falsification given by Schulte and Juhl identifies *always falsifiability* of a hypothesis $\mathcal{H} \subset 2^X$ with the notion of *nowhere density* in the usual product topology on 2^X.[35, p. 10]

There is an equivalent description of the notion of always falsifiability in terms of the fundamental machine-learning theoretic notion of shattering.

Definition 2.16. Let $\mathcal{H} \subset 2^X$ be a hypothesis and $X_0 \subset X$. Then $\mathcal{H}$ is said to shatter X_0 provided that the restriction of $\mathcal{H}$ to X_0,

$$\mathcal{H}\restriction_{X_0} = \{h\restriction_{X_0} \mid h \in \mathcal{H}\},$$

satisfies

$$\mathcal{H}\restriction_{X_0} = 2^{X_0}. \qquad \Diamond$$

One may give an equivalent definition of always falsifiability in terms of shattering, following the account of [36].

Definition 2.17. Let $\mathcal{H} \subset 2^X$ be a hypothesis. Let $f : X_0 \to 2$ be a function defined on a finite subset $X_0 \subset X$. Given such a function, let $\mathcal{H}_f = \{h \in \mathcal{H} \mid h \supset f\}$ be the set of functions in $\mathcal{H}$ extending f.

The Popper dimension δ_P of $\mathcal{H}$ relative to f is the size of the smallest subset of $X \setminus \mathrm{dom}(f)$ *not* shattered by $\mathcal{H}$. More precisely:

$$\delta_P(\mathcal{H}, f) = \min \{|Y| \mid Y \subset (X \setminus \mathrm{dom}(f)) \text{ is not shattered by } \mathcal{H}_f\}.$$

We say that $\mathcal{H}$ is *hereditarily Popper finite* provided $\delta_P(\mathcal{H}, f)$ is finite for all $f : X_0 \to \{0,1\}$ with finite domain. ♢

Proposition 2.17. $\mathcal{H}$ is always falsifiable if and only if $\mathcal{H}$ is hereditarily Popper finite. ⧫

Proof. Suppose that $\mathcal{H}$ is hereditarily Popper finite. To show that $\mathcal{H}$ is nowhere dense, first suppose that $U \subset 2^X$ is a basic open set, say $U = U_s$ for some string s. We need to show that $\mathcal{H} \cap U$ is not dense in U. It suffices to show that there exists a nonempty basic open $V \subset U$ such that $\mathcal{H} \cap V = \varnothing$. Since $\mathcal{H}$ is hereditarily Popper finite, there is a finite n such that $\delta_P(\mathcal{H}, s) = n < \infty$. Then there is a string $t \supset s$ of length $|s| + n + 1$ such that $t \notin \mathcal{H}_s$. Thus $\mathcal{H} \cap U_t = \varnothing$. Since U was arbitrary basic open, $\mathcal{H}$ is nowhere dense.

Conversely, if $\mathcal{H}$ is not hereditarily Popper finite then there exists a finite subset $X_0 \subset X$ and $f : X_0 \to \{0,1\}$ such that all finite subsets $Y_0 \subset X \setminus X_0$ are shattered by $\mathcal{H}_f$. This precisely says that the nonempty basic open set U_f is such that $\mathcal{H} \cap U_f$ is dense. Thus $\mathcal{H}$ is not nowhere dense. □

A stronger condition than hereditary Popper finiteness—the context of Vapnik's PAC learnability—is that of VC finite classes.

Definition 2.18. Let $\mathcal{H} \subset 2^X$ be a hypothesis. The VC dimension of $\mathcal{H}$ is the maximal size of a set shattered by $\mathcal{H}$:

$$\delta_{VC}(\mathcal{H}) = \max\{|Y| \mid Y \subset X \text{ is shattered by } \mathcal{H}\}. \quad ♢$$

Proposition 2.18. 1. If $\mathcal{H}$ is VC finite then $\mathcal{H}$ is hereditarily Popper finite.

2. There exist hereditarily Popper finite $\mathcal{H}$ which are not VC finite.

⧫

Proof. 1. This is essentially [36, Lemma 6.1]. By definition, for all finite $x \subset X$

$$\delta_P(\mathcal{H}, x) \leq VC(\mathcal{H}) + 1,$$

so that if $\mathcal{H}$ is VC finite then $\mathcal{H}$ is hereditarily Popper finite. □

2. Let

$$\mathcal{H} = \{f \in 2^\omega \mid (\forall n \in \omega \text{ even}) f(n) = 0\}.$$

This set is hereditarily Popper finite as $\mathcal{H}$ shatters *no* set containing an even number n, but is VC infinite as $\mathcal{H}$ shatters the collection of odd integers.

Thus the VC finiteness of a hypothesis constitutes a stronger notion of falsifiability than that of always falsifiability. It turns out that the added constraints of VC finiteness are precisely what are needed to yield sample-independent bounds on the *prevalence* of crucial experiments.

Surprise and Observational Studies

Over countable data streams, one can define a probability-independent notion of the *surprise* of a hypothesis $\mathcal{H} \subset 2^X$.

Definition 2.19. Let X be countable. A *sample* of X is an injective function $f : \omega \to X$. A sample f is *full* provided f is bijective. ◊

Definition 2.20. Let X be countable, $f : \omega \to X$ a sample of X, and $\mathcal{H} \subset 2^X$ a hypothesis. The *surprise* of $\mathcal{H}$ is the function

$$S(\mathcal{H}, f, n) = 1 - \frac{|\mathcal{H}\restriction_{f([n])}|}{2^{|f([n])|}}.$$ ◊

The surprise of $\mathcal{H}$ along the enumeration f is the relative proportion of the states of the world incompatible with $\mathcal{H}$. Surprise is a quantitative, probability-independent measure of falsifiability:

Proposition 2.19. Suppose that X is countable, $f : \omega \to X$ is a sample, and $\mathcal{H} \subset 2^X$ is a hypothesis. Then there is a crucial experiment of $\mathcal{H}$ along f at stage n if and only if $S(\mathcal{H}, f, n) > 0$. ⧫

Proof. By definition, a crucial experiment occurs just in case $|\mathcal{H}\restriction_{f([n])}| < 2^{|f([n])|}$, which is equivalent to saying that

$$S(\mathcal{H}, f, n) > 0.$$ □

In the case that $\mathcal{H}$ is VC finite, the Sauer-Shelah lemma allows us to give uniform, enumeration-independent bounds on the surprise of $\mathcal{H}$. The Sauer-Shelah lemma shows that the *growth function* of a hypothesis class is polynomial once the sample size exceeds the VC dimension of the class:

Lemma 2.1. Let $\mathcal{H}$ be a hypothesis class of VC dimension d. Then for all m, the growth function

$$\tau_{\mathcal{H}}(m) = \max\{|\mathcal{H}\restriction_Y| \mid Y \subset X \text{ and } |Y| = m\}$$

satisfies the inequality

$$\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}.$$

In particular, if $m > d+1$ then

$$\tau_{\mathcal{H}}(m) \leq \tau_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d. \quad \blacklozenge$$

Proposition 2.20. Let X be countable and $\mathcal{H} \subset 2^X$ a VC finite hypothesis. Then for every $\epsilon > 0$ there is an $m > 0$ such that for all enumerations $f : X \to \omega$

$$S(\mathcal{H}, f, m) \geq 1 - \epsilon.$$

Moreover, for all enumerations f

$$\lim_{m\to\infty} S(\mathcal{H}, f, m) = 1. \quad \blacklozenge$$

Proof. By the Sauer-Shelah lemma (Lemma 2.1), we have for $m > d+1$ the inequality

$$S(\mathcal{H}, f, m) \leq \tau_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d$$

so that for all enumerations $f : X \to \omega$,

$$S(\mathcal{H}, f, m) \geq 1 - \frac{(em)^d}{2^m d^d}.$$

As $m \to \infty$, $\frac{(em)^d}{2^m e^d} \to 0$ and so

$$\lim_{m\to\infty} S(\mathcal{H}, f, m) = 1. \quad \square$$

On the other hand, in the case of a VC infinite class there exist samples on X with surprise 0 for unbounded time:

Theorem 2.8. Let X be countable and $\mathcal{H} \subset 2^X$ a VC infinite hypothesis. Then for every m there exists a sample $f_m : \omega \to X$ such that for all $k < m$

$$S(\mathcal{H}, f_m, k) = 0 \quad \blacklozenge$$

Proof. Since $\mathcal{H}$ is VC infinite there exists a set $X_m \subset X$ of size m which is shattered. Let f_m be any enumeration of X enumerating X_m first. Then for all $k < m$

$$S(\mathcal{H}, f, k) = 1 - \frac{|\mathcal{H}\restriction_{f([k])}|}{2^k} = 1 - \frac{2^k}{2^k} = 0. \quad \square$$

This result has the following epistemic interpretation: for an agent undertaking observational inquiry, VC infinite classes may take unboundedly long to yield nontrivial surprise.

Falsifiability and Control Studies

In the previous section we saw how an agent in an impoverished epistemic state—only being able to conduct purely observational studies without any way to alter the data stream—is guaranteed short-run falsifiability of a hypothesis $\mathcal{H}$ just in case the hypothesis is VC finite. In this section we characterize the falsifiability of a hypothesis $\mathcal{H}$ in terms of the existence of certain *selectors*—to be thought of as an agent's sequential choice of objects amongst those in X—witnessing crucial experiments.

For example, consider a simplified account of a particle collision experiment wherein at each time t the scientist observes the collision of two elementary particles and the output is recorded. If the hypothesis H in question is a hypothesis concerning the result of a collision between bosons, then the scientist may have to wait an unboundedly long time witnessing irrelevant experiments (e.g. proton-proton collisions). To make the hypothesis *efficiently* falsifiable requires some form of control over the sampling procedure. To this end we define the notion of a selector.

Definition 2.21. A selector $s : \omega \to X$ is an injective sample. ◊

The always falsifiability of a hypothesis is a necessary and sufficient condition for the existence of efficiently falsifying the hypothesis with a selector.

Proposition 2.21. If $\mathcal{H}$ is always falsifiable provided then there exists a selector $s : \omega \to X$ such that for each m, a crucial experiment will be performed by sampling $s([0, m + k])$ where $k = \delta_P(\mathcal{H}, s[m])$.

Moreover, if for every string $x \in 2^{X_0}$ for $X_0 \subset X$ finite there is a selector s such that $s([m]) = x$ and from $m + 1$ onward satisfies that a crucial experiment will be performed by sampling $s([0, m + k])$ where $k = \delta_P(\mathcal{H}, s[m])$. ⧫

Proof. We construct $s : \omega \to X$ in stages:

- (Stage $n = 0$) Suppose $\delta_P(\mathcal{H}, \varnothing) = k < \infty$ Let $s(0), \ldots, s(k - 1)$ enumerate any set $Y \subset X$ witnessing $\delta_P(\mathcal{H}, \varnothing) = k$.
- (Stage $n = m + 1$) Suppose $\delta_P(\mathcal{H}, s([m])) = k < \infty$ and suppose that s is defined on range $[\ell]$. Then define $s(\ell + 1), \ldots, s(\ell + k)$ so as to enumerate any set $Y \subset X$ witnessing $\delta_P(\mathcal{H}, s[m]) = k$.

By construction, s is injective and defined on all of ω, so s is a selector.

The converse is immediate from the definition of hereditary Popper-finiteness. □

We interpret this result as saying that a hypothesis class is always falsifiable just in case an agent able to select data along a selector s as in Proposition 2.21 can falsify $\mathcal{H}$ with sample bounds given by the Popper dimensions $\delta_P(\mathcal{H}, s([n]))$.

VC Finiteness and Inferring Always Falsifiability on Subsamples

In the setup considered above, we looked only at the always falsifiability of a hypothesis over a fixed, countable sample set X. However, in typical scientific inference we typically wish to probe a hypothesis $\mathcal{H}$ with the aid of some (possibly incomplete) sample of the world. In fact, we study hypotheses knowing full well that our sampling capabilities are bounded: we cannot directly perform tests in the ancient past or beyond the observable universe.

This would be no issue if the inference

1	$\mathcal{H}$ is always falsifiable.
2	$\mathcal{H}\restriction_Y$ is always falsifiable.

were true for all subsamples $Y \subset X$. However, this inference is invalid. Recall from topology that the interior of a set X, $\text{int}(X)$, is the union of all open subsets $U \subseteq X$, and the closure of the set X, $\overline{X}$, is the intersection of all closed subsets $C \supseteq X$. Nowhere density of a set X is typically defined by

$$\text{int}(\overline{\mathcal{H}}) = \varnothing.$$

In the case of *finite* subsamples, recall that

Proposition 2.22. The only nowhere dense subset of 2^n is $\varnothing$. ⧫

Proof. The product topology on 2^n is the discrete topology, so the interior and closure operators on subsets $\mathcal{H} \subset 2^n$ are equal to the identity operator. Thus, $int(\overline{\mathcal{H}}) = \varnothing$ just in case $\mathcal{H} = \varnothing$. □

From this observation it follows that

Proposition 2.23. Suppose that $\mathcal{H} \subset 2^\omega$ is nonempty. Then for all $Y \subset X$ finite nonempty, $\mathcal{H}\restriction_Y$ is not nowhere dense. ⧫

Proof. Suppose that $\mathcal{H}$ is nonempty. Then $\mathcal{H}\restriction_Y$ is nonempty, so cannot be nowhere dense by Proposition 2.22. □

This result should not surprise us; after all, if there are only finitely many observations to be made then there will not always exist a further crucial test to perform as required by always falsifiability.

Less trivially, we can exhibit the existence of nowhere dense hypotheses $\mathcal{H}$ such that the inference

$$\begin{array}{l|l} 1 & \mathcal{H} \text{ is always falsifiable.} \\ \hline 2 & \mathcal{H}\restriction_Y \text{ is always falsifiable.} \end{array}$$

fails on a continuum-sized ideal of samples $Y \subset 2^\omega$.

Proposition 2.24. Let $X \subset \omega$ be infinite-coinfinite. Then there exists a nowhere dense $\mathcal{H}_X \subset 2^\omega$ such that for all $Y \subseteq X$, $\mathcal{H}_X\restriction_Y = 2^Y$ and therefore not nowhere dense. ⧫

Proof. Define $\mathcal{H}_X$ as the set of all functions $f : \omega \to \{0, 1\}$ such that $f(x) = 0$ if $x \notin X$.

Because X is coinfinite, the set $\mathcal{H}_X$ is nowhere dense: every $x \notin X$ yields a crucial experiment. Moreover, $\mathcal{H}_X\restriction_X = 2^X$ by definition, and likewise for any $Y \subseteq X$ we have that $\mathcal{H}_X\restriction_Y = 2^Y$. □

While nowhere dense over the full sample set ω, the hypotheses $\mathcal{H}_X$ fail to be always falsifiable on any $Y \subseteq X$.

On the other hand, the stronger notion of VC finiteness is preserved under the implication above:

Proposition 2.25. For all $Y \subseteq X$ the inference rule

$$\begin{array}{l|l} 1 & \mathcal{H} \text{ is VC finite.} \\ \hline 2 & \mathcal{H}\restriction_Y \text{ is VC finite.} \end{array}$$

is valid. ⧫

Proof. Observe that if $Y_0 \subset Y$ is shattered by $\mathcal{H}\restriction_Y$ then Y_0 is itself shattered by $\mathcal{H}$. Thus

$$VC(\mathcal{H}\restriction_Y) \leq VC(\mathcal{H})$$

so $\mathcal{H} \cap 2^Y$ is VC finite. □

VC finiteness does not, however, characterize those hypothesis classes which are hereditarily nowhere dense.

Proposition 2.26. Partition $\omega = \bigcup_{n\in\omega} X_n$ where each $|X_n| = n$. Let $\mathcal{H} = \bigcup_{n\in\omega} 2^{X_n}$. Then

1. for every infinite $Y \subset \omega$, $\mathcal{H}\restriction_Y$ is nowhere dense in 2^Y, and
2. $\mathcal{H}$ is VC infinite. ⧫

Proof. By construction, $\mathcal{H}$ shatters only finite sets, so is hereditarily Popper finite.

$\mathcal{H}$ is VC infinite since, by construction, it shatters arbitrarily large sets. □

Despite this, there is a precise sense in which one can say that if $\mathcal{H} \subset 2^\omega$ is a hypothesis of infinite VC dimension then the structure $\mathcal{M}_\mathcal{H}$ is observationally indistinguishable from a structure $\mathcal{N}$ such that that an infinite set is shattered by the relation R.

Proposition 2.27. Let $\mathcal{H} \subset 2^\omega$ be VC infinite. Then there exists an $\mathcal{N}$ elementarily equivalent to $\mathcal{M}_\mathcal{H}$ such that the interpretation of $\mathcal{H}$ in $\mathcal{N}$, $\mathcal{H}^*$, shatters an infinite set. ⧫

Proof. Let $\mathcal{N}$ be a sufficiently saturated nonprincipal ultrapower of $\mathcal{M}_\mathcal{H}$. Then

$$\mathcal{N} = \omega^* \cup \mathcal{H}^*$$

defines the structure of a hypothesis set on $2^{\omega*}$. $\mathcal{H}^*$ is regarded as a subset of 2^{ω^*} by way of the interpretation of the relation R. That is, we may regard

$$\mathcal{N} \vDash R(n^*, h^*) \leftrightarrow n^* \in h^*$$

as the definition of an embedding $\mathcal{H}^* \subset 2^{\omega^*}$.

By saturation, $\mathcal{N}$ shatters an infinite set as $\mathcal{H}$ shatters arbitrarily large finite sets. □

We note here that the $\mathcal{N}$ as constructed in the above proposition is *elementarily equivalent* to $\mathcal{M}_\mathcal{H}$. This suffices to conclude that, in a strong sense, no finitistic agent will ever be able to discern between $\mathcal{M}_\mathcal{H}$ and $\mathcal{N}$. This is due to the equivalence between elementary equivalence and finitary back-and-forth equivalence.

Definition 2.22. [16, Definition XI.1.1] Let $\mathcal{M}$ and $\mathcal{N}$ be $\mathcal{L}$-structures. A partial function $f : \mathcal{M} \dashrightarrow \mathcal{N}$ with domain $\mathrm{dom}(f) \subseteq \mathcal{M}$ and range $\mathrm{rng}(f) \subseteq \mathcal{N}$ is a *partial isomorphism* provided

1. f is injective,

2. f preserves all relations, function symbols, and constants in $\mathcal{L}$. ◊

Finitary back-and-forth equivalence is a property about being able to extend arbitrary partial isomorphisms with finite domain:

Definition 2.23. [16, Definition XI.1.3] Two $\mathcal{L}$-structures $\mathcal{M}$ and $\mathcal{N}$ are finitarily back-and-forth equivalence provided there is a sequence $(I_n)_{n\in\omega}$ such that

- Every I_n is a nonempty set of partial isomorphisms from $\mathcal{M}$ to $\mathcal{N}$,
- (Forth) For every $f \in I_{n+1}$ and $a \in \mathcal{M}$ there is a $g \in I_n$ with $g \supseteq f$ and $a \in \mathrm{dom}(g)$
- (Back) For every $f \in I_{n+1}$ and $b \in \mathcal{N}$ there is a $g \in I_n$ with $g \supseteq f$ and $a \in \mathrm{rng}(g)$. ◊

The definition of finitary back-and-forth equivalence has an immediate epistemic interpretation. Two structures being back-and-forth equivalent means that any finite quantifier-free relation in $\mathcal{M}$ can be witnessed in $\mathcal{N}$ and vice versa. Thus, no finite amount of observation of quantifier-free formulas can discern between $\mathcal{M}$ and $\mathcal{N}$.[10] Fraissé's theorem relates finitary back-and-forth equivalence with elementary equivalence:

Theorem 2.9. [16, Theorem XI.2.1] Let $\mathcal{L}$ be a finite language. Two $\mathcal{L}$-structures $\mathcal{M}$ and $\mathcal{N}$ are finitely back-and-forth equivalent if and only if they are elementarily equivalent ⧫

Thus, even if $\mathcal{H}$ *happens* to be nowhere dense, VC infinite yet does not shatter an infinite set, in a strong sense $\mathcal{H}$ is observationally indistinguishable from one in which which *does* shatter an infinite set.

This result illustrates an effect of the underlying framework of Formal Learning Theory: it works assuming an agent knows the *extensional specification* of the space of observations (ω) and hypotheses ($\mathcal{H}$) on the nose. However, bounded agents may only grasp the domain of observations and hypotheses intensionally, and thus know the hypothesis and sample domain only up to back-and-forth equivalence.

[10] The astute reading will note that the definition of back-and-forth equivalence requires partial isomorphism between *finitely-generated substructures,* which in the case of a language with function symbols may be infinite. One may remedy this by nothing that any theory T in a language $\mathcal{L}$ containing constant and function symbols is biïnterpretable with a theory T' in a purely relational language, where the finitely generated structures in a relational language are precisely the finite structures.

Viewed in this light, the VC finite classes emerge as precisely the class of nowhere dense hypotheses invariant under observable indistinguishability by finitistically bounded agents.

VC Finiteness is Not a Topological Notion

In this section we argue that the topological and descriptive set theoretic tools relied upon in formal learning theory are too coarse to adequately study the short-run properties of the hypotheses of the sort encountered in machine learning.

A unifying theme of theoretical machine learning is identifying combinatorial notions of dimension on hypotheses such that "finite dimensional iff learnable" is true. These notions of dimensions standardly have the structure of a *nontrivial set-theoretic ideal* on 2^X in the case that X is an infinite set. Two examples of such dimensions are *VC dimension*, characterizing the PAC learnable hypotheses, and *Littlestone dimension*, characterizing the online-learnable hypotheses.

Following the analysis of [9] we investigate the topologies arising from such ideals and conclude that the natural topologies fail to satisfy the standard metrization requirements of Formal Learning Theory. Instead, combinatorial measures of hypotheses are better equipped to handle such questions.

To illustrate this general point, we see that the class of VC finite hypotheses cannot be realized as the nowhere dense sets in a Hausdorff topological space. The arguments here are drawn from the analysis of topological properties of set-theoretic ideals given by Cieselski and Jasinski in [9].

Proposition 2.28. The set of VC finite families on an infinite set X,

$$I_{VC}(X) = \{Y \mid Y \subset X \text{ and } VC(Y) < \infty\}$$

forms a proper ideal in 2^{2^X}. ⧫

Proof. First, $2^X \notin I_{VC}(X)$ since, by definition, 2^X shatters an infinite set. Moreover, it is clear from the definitions that if $Y \in I_{VC}$ and $Z \subset Y$ then $Z \in I_{VC}$ since every set shattered by Z is shattered by Y.

Finally, the Sauer-Shelah lemma implies that if $Y, Z \in I_{VC}$ then $Y \cup Z \in I_{VC}$ since the growth function of the union is polynomial. □

The fact that I_{VC} has the structure of a proper ideal on 2^X means that we may construct a topology in which the VC finite sets are precisely the closed sets. However, this topology is non-Hausdorff.

Proposition 2.29. The collection

$$\tau(2^X) = \{2^X \setminus \mathcal{H} \,|\, \mathcal{H} \in I_{VC}(X)\} \cup \{\varnothing\}$$

of subsets of 2^X forms a non-Hausdorff topology on 2^X. ⧫

Proof. Since $I_{VC}(X)$ has the structure of a set-theoretic ideal, the collection

$$F_{VC}(X) = \{2^X \setminus \mathcal{H} \,|\, \mathcal{H} \in I_{VC}(X)\}$$

is a nontrivial filter on X. Thus, the collection

$$\tau(2^X) = F_{VC}(X) \cup \{\varnothing\}$$

is closed under arbitrary union, finite intersection, and contains $\varnothing$ and 2^X.

This topology is non-Hausdorff: for any $U, V \in \tau(2^X)$,

$$(U \cap V = \varnothing) \to (U = \varnothing \vee V = \varnothing)$$

since $\varnothing \notin F_{VC}(X)$ and $F_{VC}(X)$ is closed under finite intersection. □

Moreover, no Polish space can make all VC finite sets closed.

Proposition 2.30. Let X be countably infinite. Then there are $2^{2^{\aleph_0}}$ VC finite subsets of X. In particular, no Polish topology renders all VC finite sets closed. ⧫

Proof. Since I_{VC} is closed downward, it suffices to show that there is an uncountable VC finite subset of 2^X.

Identifying X with , we may identify the family of intervals $\mathcal{H} = \{(r, \infty) \,|\, r \in \mathbb{R}\}$ with a VC finite subset of 2^X. This family has size $2^{\aleph_0}$, so $|I_{VC}| = 2^{2^{\aleph_0}}$.

Since Polish spaces have at most $2^{\aleph_0}$ many closed sets, no Polish space renders all VC finite subsets closed. □

Finally, we identify a mild condition on topologies guaranteeing that the ideal of nowhere dense sets does not coincide with the ideal of VC finite sets.

Theorem 2.10. Let X be infinite. There is no topology τ on 2^X such that

1. There is a countable disjoint collection of nonempty open sets U_n such that each U_n shatters a set of size $\geq n$,

2. Every VC finite $\mathcal{H}$ is nowhere dense, and

3. $I_{VC}(X) = I_{nd}(X)$. ♦

Proof. We follow the proof strategy outlined in [9, Thm 3.4] by showing that if τ were a topology on X making all VC finite sets nowhere dense then there is a VC infinite nowhere dense subset.

By hypothesis 1, there exists a countably infinite set of disjoint open subsets U_n shattering a set of size $\geq n$.

Let $\mathcal{H}_n \subset U_n$ be a finite hypothesis class shattering a set of size n. Then the hypothesis $\mathcal{H} = \bigcup_{n \in \omega} \mathcal{H}_n$ shatters arbitrarily large subsets by construction, and is nowhere dense as each $\mathcal{H}_n$ is finite and concentrated on a single open set U_n. □

In particular, the standard results of descriptive set theory—requiring that the topology in question be Polish and hence Hausdorff—do not apply to *any* topology rendering the learnable sets nowhere dense.

2.6 Rigorous Foundations for Severe Testing

Null Hypothesis Statistical Testing (NHST) is a ubiquitous method of statistical inference. As Wasserman [51, Chapter 10] describes it, the basic data of a Null Hypothesis Statistical Testing consists of

1. A space Θ of probability distributions on sample space Ω,
2. A partition $\Theta = \mathcal{H}_0 \cup \mathcal{H}_1$
3. A random variable $T : \Omega \to \mathbb{R}$ called the test statistic,
4. A critical value $c \in \mathbb{R}$.

In the setup of a two-sided test, it is assumed that the null hypothesis $\mathcal{H}_0$ is a *single distribution*, i.e. $\mathcal{H}_0 = \{\theta_0\}$, and therefore unambiguously determines a probability measure $\mathbb{P}_{\mathcal{H}_0}$ that we may use to infer probability statements about the test statistic T. To *reject* hypothesis $\mathcal{H}_0$, a statistical version of modus tollens is invoked, by replacing $\mathcal{H}_0 \to (T(x) \leq c)$ with $\mathbb{P}_{\mathcal{H}_0}(T(x) > c) < \epsilon$:

1	$\mathbb{P}_{\mathcal{H}_0}(T(x) > c) < \epsilon$
2	$T(x) > c$
3	$Rej(\mathcal{H}_0)$

.

Mayo and other error statisticians instead advocate for a modern recasting of NHST—severe testing—as the appropriate framework guiding the use of statistical methods. Central to the error statistician is the question:

> When do data x provide good evidence for/a good test of hypothesis H?

The error statistician will invoke some form of a **Severity Principle** to answer this question:

> (Weak Severity Principle) Data x **does not** provide good evidence for H if x is the result of a **test procedure** T with very low probability of uncovering the falsity of H [31, p. 21].

A converse is given by:

> (Full Severity Principle) Data x provides good evidence for H to the extent that test T has been **severely passed** by H [31, p. 21].

The error statistician naturally asks *which* hypotheses are amenable to error-theoretic analysis. This question is of utmost importance as the Full Severity Principle suggests the following account of scientific content: the hypotheses H that have scientific content are precisely those which are severely testable. But what is the definition of severe testing?

The notion of severe testing as described by Mayo is defined as follows:

Definition 2.24. A hypothesis H passes a severe test relative to experiment E with data x if (and only if):

i x agrees with or "fits" H (for a suitable notion of fit), and

ii experiment E would (with very high probability) have produced a result that fits H less well than x does, if H were false or incorrect. [29, p. 99] ◊

We turn now to discussing prongs (i) and (ii) in the above definition.

Regarding (i), Mayo writes that "fit" should at the very least be

$$\mathbb{P}(x; H) > \mathbb{P}(x; \neg H)$$

arguing that

> any measure of evidential relationship, degree of confirmation, probability, etc., can be regarded as supplying a fit measure. Severity can then be assessed by computing the error probability required in (ii). [29, p. 124]

If the notation $\mathbb{P}(x; H)$ is unfamiliar, that is for good reason: Mayo explains that

> I am using ";" in writing $\mathbb{P}(x; H)$—in contrast to the notation typically used for a conditional probability, $\mathbb{P}(x|H)$—in order to emphasize that severity does *not* use a conditional probability which, strictly speaking, requires that the prior probabilities $\mathbb{P}(H_i)$ be well-defined for an exhaustive set of hypotheses. [29, p. 102]

This is on the face of it a key departure from the framework of NHST, which requires us to *only* work with probability sentences involving $\mathbb{P}_{\mathcal{H}_0}$. A serious difficulty for this account of severe testing is that no general construction of $\mathbb{P}(x; \neg\mathcal{H})$ is given.

For example, let $\mathcal{H}_r \subset 2^\omega$ be the a statement such as "the long run relative frequency of heads in a countable sequence coin flips is equal to r." The complement $\mathcal{H}_r^c \subset 2^\omega$ can be decomposed as the disjoint union

$$\mathcal{H}_r^c = \mathcal{H}_\uparrow \cup \bigcup_{s \neq r} \mathcal{H}_s$$

where $\mathcal{H}_\uparrow$ is the set of all infinite binary strings with non-convergent limiting relative frequency as well as $\mathcal{H}_{r'}$ for all $r' \neq r$. Moreover, $\mathcal{H}_\uparrow$ and $\mathcal{H}_s$ are dense and codense in 2^ω, so the complement $\mathcal{H}_r^c$ has rich topological structure. There is no clear way to construct a probability measure that amalgamates all $\mathcal{H}^c$ into a probability measure $\mathbb{P}_H(x; \mathcal{H}^c)$ if one does not avail oneself to an aggregation function such as a Bayesian prior.

Even restricting only to the probability distributions on $\{0, 1\}$, which we identify with the interval $[0, 1]$, non-Bayesian methods of aggregating families of probability distributions—such as the Maximum Likelihood Estimator—are generally not well-defined. Recall the definition of the Maximum Likelihood Estimator [51, Definition 9.7]

Definition 2.25. Let Θ be a family of distributions over Ω and $X_1, \ldots, X_n : \omega \to \mathbb{R}$ be an IID set of random variables. The likelihood function is given by

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

where the $f(X_i; \theta)$ are the probability density functions of the random variables X_i with respect to distribution θ. $\Diamond$

The maximum likelihood estimator is typically defined as *the* value $MLE(\theta, n) \in \Theta$ maximizing $\mathcal{L}_n(\theta)$. However, this definition is misleading, as $MLE(\theta, n)$ may fail to exist or to be unique.

First, the Maximum Likelihood Estimator may fail to be unique. Let $\Theta = \{0, 1\}$ be the space of distributions asserting that all flips of a coin are heads or tails. Confronted with observations $\overline{x} = (H, T)$, the likelihood functions have values

$$\mathcal{L}_2(0) = 0 = \mathcal{L}_2(1)$$

and so *both* 0 and 1 maximize the likelihood function relative to θ.

Second, the Maximum Likelihood Estimator may fail to exist within Θ. Let $\mathcal{H}_0 = \{\frac{1}{2}\}$ and $\mathcal{H}_1 = \mathcal{H}_0^c = [0, \frac{1}{2}) \cup (\frac{1}{2}, 1]$. Suppose that $\overline{x} = (H, T)$. A routine calculation [51, Example 9.10]

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} p_H^{X_i}(1 - p_H)^{1-X_i} = p_H^S \times (1 - p_H)^{n-S}$$

where S is the number of heads in the sequence of coin flips. Taking the derivative of $\mathcal{L}_n(\theta)$ and setting it to zero we find that $\theta = \frac{1}{2}$ is the unique maximum likelihood estimator on $[0, 1]$ with likelihood $\frac{1}{4}$. While $\frac{1}{2} \notin \mathcal{H}_1$, this does not on its own show that the Maximum Likelihood Estimator does not exist in $\mathcal{H}_1$. In fact, *any* paritition of a compact connected space $\Theta = \mathcal{H}_1 \cup \mathcal{H}_2$ into nonempty subfamilies of distributions will suffer this defect since the existence of a $\theta \in \mathcal{H}_i$ maximizing likelihood is guaranteed only if $\mathcal{H}_i$ is closed. However, it is not hard to see that the image $\mathcal{L}_n(\mathcal{H}_1) = [0, \frac{1}{4})$. In other words, the likelihood function on $\mathcal{H}_1$ is arbitrarily close to $\frac{1}{4}$ but never obtains that value. So, no maximum likelihood estimator exists on $\mathcal{H}_1$. Thus, the standard frequentist method of aggregating probability distributions in light of data is not even generally well-defined, and cannot serve as a definition of $\mathbb{P}(x; \neg\mathcal{H})$.

Regarding condition (ii) in the definition of severe testing, Mayo requires the satisfaction of the following conditional: if $\mathcal{H}$ is false, then E would have produced a result that fits $\mathcal{H}$ less well than x does with high probability. In this conditional, we assume that $\mathcal{H}$ is false and tasked with computing *some* probability given $\neg\mathcal{H}$ and the specification of the experiment E. This poses a serious problem for her account of severe testing; she gives no general theory of semantics for the probabilistic statements comprising the definition of severe, as no method for determining how to construct a probability distribution $\mathbb{P}_{E,\neg\mathcal{H}}$ from experiment E is given. With no way to determine what a "good" probability distribution is, it is difficult to make sense of this.

For instance, consider the case of the hypothesis $\mathcal{H}_0$ expressing "all flips of a coin are heads." Let E_n be the experiment given by flipping the coin some very large number n of times. Then, given such an experiment E,

Proposition 2.31. Let $\epsilon > 0$. There exists a probability distribution $\mathbb{P}_n$ on $\{H, T\}$ such that the probability p_H of flipping at least one tails T on n IID flips is $< \epsilon$. ⧫

Proof. Let $\mathbb{P}$ be a probability distribution on $\{H, T\}$. Let p_H be the probability of heads. The probability p of flipping at least one tails T on n i.i.d. flips is

$$\mathbb{P}(\text{At least one Tails}) = 1 - \mathbb{P}(\text{All Heads}) = 1 - p_H^n.$$

Let $\epsilon > 0$. Then

$$\mathbb{P}(\text{At least one Tails}) = 1 - p_H^n < \epsilon$$

is equivalent to saying

$$p_H > (1 - \epsilon)^{\frac{1}{n}}.$$ □

The upshot is that the specification of the experimental setup itself does not determine *a priori* the relevant probability distribution is. Moreover, this hypothesis is as falsifiable as it can be: for each coin flip, $\mathcal{H}$ is compatible with only a single outcome. Yet, on a strict reading of Mayo's definition, $\mathcal{H}$ cannot be severely tested.

Rather, it seems to me that the combinatorics of the hypothesis—not any notion of probability—are what make $\mathcal{H}$ severely testable. The critical element of Mayo's definition of severe testing is that the data x be compatible with $\mathcal{H}$ and—simultaneously—highly incompatible with $\neg\mathcal{H}$. That is, for a hypothesis $\mathcal{H}$ to be severely tested by an experiment with n observations x, we would require:

1. $x \in \mathcal{H}\restriction_{[n]}$, and
2. $\mathcal{H}\restriction_{[n]} \ll \mathcal{H}^c\restriction_{[n]}$.

These requirements are precisely captured by the previously-defined notion of *surprise*: if $\mathcal{H}\restriction_{[n]} \ll \mathcal{H}^c\restriction_{[n]}$ then $\frac{\mathcal{H}\restriction_{[n]}}{2^n} < \frac{\mathcal{H}\restriction_{[n]}}{\mathcal{H}^c\restriction_{[n]}} \approx 0$. This motivates the following definition:

Definition 2.26. Let $f : \omega \to X$ be a sample, $\mathcal{H} \subset 2^X$ a hypothesis, $n \in \omega$, and $\epsilon > 0$. We say that $(\mathcal{H}, f, n)$ is *severely surprising* at level ϵ provided the observed data $x \in \mathcal{H}\restriction_{f([n])}$,

$$S(\mathcal{H}, f, n) > 1 - \epsilon$$

and

$$S(\mathcal{H}, f, n) > S(\mathcal{H}^c, f, n).$$ ◊

Taking a step back, it is worth relating this definition back to probability theory. Crucially, the definition of surprise is non-probabilistic in the sense that there probabilistic *frequency semi-measure* lurking in the background:

Definition 2.27. Let $f : \omega \to X$ be a sample and $n \in \omega$. For every $\mathcal{H} \in 2^\omega$ $\mu_{f,n} : 2^\omega \to [0,1]$ as follows:

$$\mu_{f,n}(\mathcal{H}) = \frac{|\mathcal{H}\restriction_{f([n])}|}{2^n}. \qquad \lozenge$$

Proposition 2.32. Let $f : \omega \to X$ be a sample and $n \in \omega$. Then $\mu = \mu_{f,n}$ is a bounded sub-additive function on 2^{2^ω}. More precisely, for $\mathcal{H}_1, \mathcal{H}_2 \in 2^\omega$ we have

$$\mu(\mathcal{H}_1 \cup \mathcal{H}_2) \leq \mu(\mathcal{H}_1) + \mu(\mathcal{H}_2).$$

Moreover, a necessary and sufficient condition for the inequality above to be strict for $\mathcal{H}_1, \mathcal{H}_2 \in 2^\omega$ is that

$$\mathcal{H}_1\restriction_{f([n])} \cap \mathcal{H}_2\restriction_{f([n])} \neq \varnothing.$$

Finally, every dense-codense $\mathcal{D} \subset 2^\omega$ satisfies $\mu(\mathcal{D}) = \mu(\mathcal{D}^c) = 1$ so that

$$\mu(\mathcal{D} \cup \mathcal{D}^c) = 1 < 2 = \mu(\mathcal{D}) + \mu(\mathcal{D}^c). \qquad \blacklozenge$$

Proof. Expanding terms, we find that

$$\begin{aligned}
\mu(\mathcal{H}_1) + \mu(\mathcal{H}_2) &= \frac{|\mathcal{H}_1\restriction_{f([n])}| + |\mathcal{H}_2\restriction_{f([n])}|}{2^n} \\
&= \frac{|(\mathcal{H}_1\restriction_{f([n])} \setminus \mathcal{H}_2\restriction_{f([n])})| + 2|(\mathcal{H}_{1\restriction f([n])} \cap \mathcal{H}_2\restriction_{f([n])})| + |(\mathcal{H}_2|_{f([n])} \setminus \mathcal{H}_1\restriction_{f([n])})|}{2^n} \\
&\geq \frac{|(\mathcal{H}_1\restriction_{f([n])} \setminus \mathcal{H}_2\restriction_{f([n])})| + |(\mathcal{H}_1\restriction_{f([n])} \cap \mathcal{H}_2\restriction_{f([n])})| + |(\mathcal{H}_2\restriction_{f([n])} \setminus \mathcal{H}_1\restriction_{f([n])})|}{2^n} \\
&= \mu(\mathcal{H}_1 \cup \mathcal{H}_2)
\end{aligned}$$

By the above chain of equalities, we see that

$$\mu(\mathcal{H}_1) + \mu(\mathcal{H}_2) - \mu(\mathcal{H}_1 \cup \mathcal{H}_2) = \frac{|(\mathcal{H}_1\restriction_{f([n])} \cap \mathcal{H}_2\restriction_{f([n])})|}{2^n}.$$

Now, while $\mathcal{H}_1 \cap \mathcal{H}_2 = \varnothing$, that does not imply that $(\mathcal{H}_1\restriction_{f([n])} \cap \mathcal{H}_2\restriction_{f([n])}) = \varnothing$.

Finally, suppose that $\mathcal{D}$ is dense-codense. Then $\mathcal{D}|_{f([n])} = 2^{f([n])} = \mathcal{D}^c|_{f([n])}$, so

$$\mu(\mathcal{D}) = \mu(\mathcal{D}) = \frac{|2^{f([n])}|}{|2^{f([n])}|} = 1. \qquad \square$$

The relationship between $\mu_{f,n}$ and the surprise function $S(-, f, n)$ is that

$$S(\mathcal{H}, f, n) = 1 - \mu_{f,n}(\mathcal{H}).$$

Therefore

Proposition 2.33. For all $\mathcal{H}, f, n$,

$$S(\mathcal{H}, f, n) + S(\mathcal{H}^c, f, n) \leq 1$$ ♦

Proof. Immediate from the sub-additivity of $\mu_{f,n}$ □

Remark 2.1. While the surprise function $S(\mathcal{H}, f, n)$ has a direct epistemic interpretation, the above the *co-surprise function* has structure reminiscent of a probability measure. Let $S^{\text{co}}(\mathcal{H}, f, n) = S(\mathcal{H}^c, f, n)$. Then

1. $S^{\text{co}}(\varnothing, f, n) = S(2^\omega, f, n) = 0$,
2. $S^{\text{co}}(2^\omega, f, n) = S(\varnothing, f, n) = 1$,
3. $S^{\text{co}}(\mathcal{H}, f, n) + S^{\text{co}}(\mathcal{H}^c, f, n) = S(\mathcal{H}^c, f, n) + S(\mathcal{H}, f, n) \leq 1$, and
4. if $\mathcal{H}_1 \subset \mathcal{H}_2$ then $S^{\text{co}}(\mathcal{H}_1, f, n) \leq S^{\text{co}}(\mathcal{H}_2, f, n)$ as $\mathcal{H}_1^c \supseteq \mathcal{H}_2^c$.

Co-surprise fails to be finitely additive, as any dense-codense subset $\mathcal{D}$ of 2^ω has $S^{\text{co}}(\mathcal{D}, f, n) = 0 = S^{\text{co}}(\mathcal{D}^c, f, n)$ as

$$\mathcal{D}\restriction_{f([n])} = 2^{f([n])} = \mathcal{D}^c\restriction_{f([n])},$$

so is not a measure.

In fact, this function is very far from being a measure in the sense that the maximal Boolean subalgebra of 2^{2^X} on which $S^{co}(-, f, n)$ is a finitely additive probability measure is rather small. For each $s \in 2^{f([n])}$, let

$$\mathcal{J}_s = \{h \in 2^X \mid h \supseteq s\}.$$

It is clear that

$$\mathcal{H}\restriction_{f([n])} \cap \mathcal{H}^c\restriction_{f([n])} = \varnothing$$

just in case for some $S \subseteq f([n])$

$$\mathcal{H} = \bigcup_{s \in S} \mathcal{J}_s,$$

for if $\mathcal{J}_s \cap \mathcal{H} \neq \varnothing$ and $\mathcal{J}_s \cap \mathcal{H}^c \neq \varnothing$ then

$$\mathcal{H}\restriction_{f([n])} \cap \mathcal{H}^c\restriction_{f([n])} \supseteq \{s\} \neq \varnothing.$$

From this observation it is we may conclude that the Boolean algebra

$$\mathcal{A}_{f,n} = \{\mathcal{J}_s \mid s \in 2^{f([n])}\} \subseteq 2^{2^X}$$

is *the* maximal Boolean algebra on which S^{co} is a probability measure.

This Boolean algebra is naturally isomorphic to $2^{f([n])}$ generated by the assignment $\mathcal{J}_s \mapsto s$. Under this identification, $S^{co}(-, f, n)$ coincides with the uniform measure on $2^{f([n])}$:

$$S^{co}(\mathcal{J}_s, f, n) = \frac{1}{2^n}.$$

Remark 2.2. There is a notion of *conditional* co-surprise analogous to conditional probability. Since S^{co} is monotonic, if $\mathcal{H}, \mathcal{J} \in 2^{2^X}$ then

$$S^{co}_{\mathcal{J}}(\mathcal{H}) = \frac{S^{co}(\mathcal{H} \cap \mathcal{J})}{S^{co}(\mathcal{J})}$$

is a well-defined function on 2^{2^X} with range $[0, 1]$ whenever $S^{co}(\mathcal{J}) \neq 0$. ⧫

Thus, while 2^{2^X} has large cardinality, $S^{co}(-, f, n)$ is only a probability measure on a subalgebra of size 2^n, identifiable with the uniform measure on 2^n. ⧫

Proposition 2.33 implies that

Proposition 2.34. Suppose that $(\mathcal{H}, f, n)$ is such that $S(\mathcal{H}, f, n) \geq 1 - \epsilon$ for $0 < \epsilon < \frac{1}{2}$. Then $(\mathcal{H}, f, n)$ is severely surprising at level $\epsilon < \frac{1}{2}$.

In particular, if $\epsilon < \frac{1}{2}$ at most one of $\mathcal{H}, \mathcal{H}^c$ is severely surprising at level ϵ. ⧫

Proof. Immediate from the above inequality on surprise. □

On the other hand, it is possible for *neither* $\mathcal{H}$ nor $\mathcal{H}^c$ to be severely surprising along sample f by observation n.

Proposition 2.35. Let $\mathcal{H} \subset 2^\omega$ be dense-codense. Then for all f and n,

$$S(\mathcal{H}, f, n) = S(\mathcal{H}^c, f, n) = 0.$$ ⧫

Proof. We show the argument for $\mathcal{H}$ assuming density; the case of $\mathcal{H}^c$ follows by codensity of $\mathcal{H}$. Since $\mathcal{H}$ is dense, $\mathcal{H}\restriction_{f([n])} = 2^{f([n])}$, so

$$S(\mathcal{H}, f, n) = 1 - \frac{|2^{f([n])}|}{|2^{f([n])}|} = 0.$$ □

Nevertheless, VC finite classes of hypotheses provide us with a great wealth of severely surprising hypotheses:

Proposition 2.36. Let $\mathcal{H} \subset 2^X$ be VC finite. Then for every $\frac{1}{2} < \epsilon < 1$ there exists an $n = n(\epsilon)$ such that for every injective sample $f : \omega \to X$, if $x \in 2^{f([n])} \cap \mathcal{H}$ then $\mathcal{H}$ is severely surprising at level ϵ. ⧫

Proof. By Proposition 2.20, there exists an $n = n(\epsilon)$ such that $S(\mathcal{H}, f, m) > 1 - \epsilon$ for all injective $f : \omega \to X$ and for all $m > n$.

Thus, it remains to show that $S(\mathcal{H}, f, m) > S(\mathcal{H}^c, f, m)$ for all $m > n$. There are two ways to see this. First, and most directly, by Proposition 2.33 $S(\mathcal{H}, f, m) > 1-\epsilon$ implies that $S(\mathcal{H}^c, f, m) < \epsilon$.

We can obtain better bounds, however, by the VC finiteness of $\mathcal{H}$. Since $\mathcal{H}$ is VC finite, without loss of generality we may assume n is taken to be sufficiently large so that $|\mathcal{H}\restriction_{f([m])}| = p(m)$ for some polynomial, with $|\mathcal{H}^c\restriction_{f([m])}| \geq 2^m - p(m) \approx 2^m$. Thus not only is, $S(\mathcal{H}^c, f, m) < S(\mathcal{H}, f, m)$ for all $m > n$, we have in fact that

$$\begin{aligned}
\frac{S(\mathcal{H}^c, f, m)}{S(\mathcal{H}, f, m)} &= \frac{1 - \frac{2^m - \tau_{\mathcal{H}}(m) + |\mathcal{H}\restriction_{f([n])} \cap |\mathcal{H}^c\restriction_{f([n])}|}{2^m}}{1 - \frac{\tau_{\mathcal{H}}(m)}{2^m}} \\
&= \frac{\tau_{\mathcal{H}}(m) - |\mathcal{H}\restriction_{f([n])} \cap \mathcal{H}^c\restriction_{f([n])}|}{2^m - \tau_{\mathcal{H}}(m)} \\
&= O\left(\frac{\tau_{\mathcal{H}}(m)}{2^m - \tau_{\mathcal{H}}(m)}\right) \\
&= O\left(\frac{m^{deg(\tau_{\mathcal{H}})}}{2^m}\right)
\end{aligned}$$

noting that since

$$|\mathcal{H}\restriction_{f([n])} \cap \mathcal{H}^c\restriction_{f([n])}| \leq |\mathcal{H}\restriction_{f([n])}| = \tau_{\mathcal{H}}(m)$$

we have

$$\tau_{\mathcal{H}}(m) - |\mathcal{H}\restriction_{f([n])} \cap \mathcal{H}^c\restriction_{f([n])}| \leq \tau_{\mathcal{H}}(m). \qquad \square$$

So, if $\mathcal{H}$ is VC finite then not only is $\mathcal{H}$ severely surprising to level ϵ if the data x is compatible with $\mathcal{H}$, but the ratio between $S(\mathcal{H}, f, m)$ and $S(\mathcal{H}^c, f, m)$ shrinks at a rate of $O\left(\frac{m^{deg(\tau_{\mathcal{H}})}}{2^m}\right)$.

For an explicit example, the hypothesis of $\mathcal{H}$ that "all coin flips are heads" is VC finite. If $\mathcal{H}$ is true, $\mathcal{H}$ will be severely surprising to level ϵ so long as the number of flips n satisfies

$$n > \log(\epsilon^{-1}),$$

and since $|\mathcal{H}^c\restriction_{f([n])}| = 2^n$ we have that

$$S(\mathcal{H}^c, f, n) = 0$$

for all f and n.

As in our discussion of the Formal Learning Theoretic account of falsification, VC finite classes emerge as a distinguished class of highly-testable hypotheses.

2.7 Conclusion : Shattering as the Fundamental Concept of Falsification

The stated goal of this chapter was to take stock of the various modes of falsification that have been studied since Popper's initial definition of falsifiability. Over the course of this examination, the central importance of the notion of *shattering* became clear:

1. failing to shatter a set X is the same as saying that $\mathcal{H}$ has forbidden configuration over X,
2. the class of NIP structures has uniquely strong falsificational content as compared to the NSOP_1 dividing line in classification theory.
3. the nowhere density of a hypothesis can be defined in terms of shattering,
4. The VC finite classes are precisely the nowhere dense classes closed under elementary equivalence, and
5. The VC finite classes are uniquely suited for severe testability, as viewed through the lens of severe surprise.

It is no doubt that Vapnik himself—one of the originators of VC dimension—would not be surprised by the primacy of the notion of shattering. While he phrased his results primarily in terms of the equivalent notion of uniform two-sided convergence of the ERM method of learning, he gives a probabilistic analogue to the above characterizations of VC finite hypotheses as *the* class of effectively falsifiable hypotheses, writing that

> if for some some [hypothesis $\mathcal{H}$] conditions of uniform convergence do not hold, the situation of nonfalsifiability will arise.[48, page 49]

It is my hope that this chapter has bridged the gap between the Vapnikian account of probabilistic falsification as the study of VC finite classes with the combinatorial, logical, and topological accounts of falsification we have heretofore discussed.

Chapter 3

On Rational Jurisprudence

3.1 Introduction

In *State v. Skipper* [42], the Supreme Court of Connecticut ruled that Bayesianism directly conflicts with the presumption of innocence, stating:

> Because Bayes' Theorem requires the assumption of a prior probability of paternity, i.e., guilt, its use is inconsistent with the presumption of innocence in a criminal case such as this.... If we assume that the presumption of innocence standard would require the prior probability of guilt to be zero, the probability of paternity in criminal cases would always be zero.... In other words, Bayes' Theorem can only work if the presumption of innocence disappears from considerations. [42, at 623]

Thus, the court argues, the jury cannot simultaneously hold the presumption of innocence and update their credences according to Bayes' rule without trivializing the enterprise of criminal trial by licensing *only* verdicts of "not guilty." Committed to the presumption of innocence, the court therefore rejects the use of Bayesian inference by the jury in a criminal setting. This caution is shared by many legal scholars, most notably Tribe [46], who argues that many forms of probabilistic reasoning in the trial setting—including Bayesian inference—violate the presumption of innocence.

On the other hand, the economic analysis of law undertaken by Judge Richard Posner [33] suggests that a trier of fact *ought* to be Bayesian. Posner models the ideal juror as an instance of what I term a *Bayesian threshold juror*: an agent who updates her credence of guilt according to Bayes' rule and moves to convict just in case, at the conclusion of the trial, the credence she assigns to guilt is above some threshold value θ sufficiently close to 1. To ameliorate the worries of those

skeptical of the Bayesian paradigm's compliance with jurisprudential norms such as the presumption of innocence, Posner proposes a simple solution: require of the juror that her credence of guilt at the outset of the trial is exactly 50%.

In the following sections I will argue that such restrictions on an ideal Bayesian juror fail to meaningfully constrain a juror's disposition to render a conviction or acquittal. My argument relies on a formalization of the notion of "Bayesian juror," wherein a juror's disposition to convict based on an observed sequence of testimonies is modeled as a function

$$f : \mathcal{T}_S \to \{C, A\}$$

where $\mathcal{T}_S$ is a set of collections of testimony that can be presented in a court of law, C represents "conviction," and A represents "acquittal." The disposition function f is interpreted as $f(T) = C$ just in case, having heard all and only testimony $T = \{t_1, \ldots, t_n\}$ over the course of the trial, the juror would vote to convict. That the juror's disposition be a function of the collection of testimony heard over the course of the trial is essential to modeling American criminal trials, as this is taken to be a constitutional right of the defendant, as described in *Turner v. Louisiana*:

> The requirement that a jury's verdict 'must be based upon the evidence developed at the trial' goes to the fundamental integrity of all that is embraced in the constitutional concept of trial by jury...
>
> In the constitutional sense, trial by jury in a criminal case necessarily implies at the very least that the 'evidence developed' against a defendant shall come from the witness stand in a public courtroom where there is full judicial protection of the defendant's right of confrontation, of cross-examination, and of counsel.[47]

By testimony I refer not to the content of an agent's testimony but to information of the form "Alice testified that S, to which the defendant's attorney objected on grounds X, Y, and Z." Events such as this can effect a change in a juror's belief that S, but the juror does not herself witness S. The fundamental assumption I make is

> **Testimonial Consistency Axiom (TCA):** Let $\mathcal{T}_S$ be a collection of possible testimonial events. For all subsets $T \in \mathcal{T}_S$, both the guilt and the innocence of the defendant are consistent with T.

For example, while the *semantic content* of a witness' testimony may be inconsistent with either the guilt or innocence of the defendant, the witness' act of testifying to that effect is consistent with both guilt and innocence. After all, such a witness may be mistaken or lying. While there are some apparent violations of this assumption

(e.g., a witness testifying that the defendant had murdered him), it generally holds true in the actual trial context.

At minimum, the **Presumption of Innocence** (PoI) places the following constraint on a juror's disposition to convict at the outset of the trial: absent any testimony, the juror must not convict. In the notation of dispositions,

$$f(\varnothing) = A.$$

Moreover, it is assumed that *some* set of testimony would compel a juror to convict; in other words, there exists a set $T \in \mathcal{T}_S$ of testimony such that

$$f(T) = C.$$

Call this the **Willingness to Convict** (WtC).

Using this formalism, I critically evaluate the interactions between Bayesianism and contemporary American legal theory. To this end, I focus on two key questions:

1. Are Posner's Bayesian threshold jurors rational agents in that they maximize the expectation of some utility function?

2. Does Posner's model of Bayesian threshold jurors materially constrain a juror's disposition beyond the (PoI) and (WtC)?

Using recent work of Easwaran [15], I argue that the answer to Question 1 is "Yes" by exhibiting a utility function that a Bayesian threshold juror optimizes. However, by using the (TCA) I argue that the answer to Question 2 is "No"; while Bayesian threshold jurors are rational qua the standard decision-theoretic account of rationality, *all* dispositions satisfying (PoI) and (WtC) can be realized as the disposition of some Bayesian threshold Juror.

This result calls into question the utility of modeling an ideal juror's inferential structure as a Bayesian threshold juror, as the disposition of *any* juror satisfying (PoI) and (WtC) is rationalizable. For instance, consider the disposition that renders a conviction so long as at least two witnesses testify, regardless of the content of the testimony. This disposition satisfies the (PoI) and the (WtC), and by the representation theorem is represented by a Bayesian threshold juror. Such a person would be ill-suited to be a juror, yet the Posnerian account cannot rule them out.

A natural response to this result would be to add further constraints to rule out such contrived dispositions. In the final section of this paper multiple proposals in this vein are analyzed. As we will see, each of these proposals avail themselves to serious objections on both epistemic and jurisprudential grounds.

Ultimately, existing proposals to give probabilistic foundations to normative legal reasoning fail to do so, and the truth of principles such as (TCA) cast serious doubt that any such approach can succeed.

3.2 A Formal Model of a Juror's Reasoning

In this section I present the formal model, partially sketched in Section 1, of a juror's reasoning in a criminal trial. The atoms in this model are individual testimonies. I use the word "testimony" here very broadly: it can include any sensory information that a juror perceives during the course of the trial that may inform her rendering of a verdict, including the statements by witnesses, legal counsel, and judges made during the trial as well as qualitative information such as the demeanor and body language of the witness.

Juror Dispositions

A transcript T of a trial is a collection of testimonies, to be understood as the contents of a trial as perceived by the juror. Given a collection S of possible testimonies, the set $\mathcal{T}_S = \mathcal{P}(S)$ is the set of transcripts over S.

Recall from Section 1 that a juror's disposition to convict (hereafter "disposition") is simply a function

$$f : \mathcal{T}_S \to \{C, A\}$$

where $f(T) = C$ (respectively A) is read as "the juror is disposed to convict (respectively acquit) on the basis of transcript T."

So far, this model does not include any consideration of the material guilt or innocence of the defendant. To remedy this, we define a set of possible worlds relative to a set S of possible testimonies by setting

$$W_S = \mathcal{T}_S \times \{G, I\},$$

where G is shorthand for "materially guilty" and I is shorthand for "materially innocent" of the charges alleged by the prosecution. In other words, a world is a pair $w = (T, x)$ consisting of a transcript T and a value $x \in \{G, I\}$ corresponding to the material guilt or innocence of the defendant. It is critical to note that we are licensed in including both a world (T, G) and (T, I) by the axiom (TCA), which ensures that both material guilt and innocence are consistent with transcript T.

On W_S we define two distinguished classes of events. First, for a transcript T define

$$E_T = \{(T, G), (T, I)\},$$

the two-world set consisting of a transcript T and the two possible values for guilt, G and I. Then E_T is the event of receiving transcript T from the trial. Second, define

$$E_G = \{w \mid w = (T, G) \text{ for some } T \in \mathcal{T}_S\},$$

the collection of worlds where the defendant is materially guilty. Thus, E_G is the event of the defendant being materially guilty.

3.3 Bayesian Analysis in the Law

In this section some of the key literature surrounding the interaction between Bayesian epistemology and legal epistemology is reviewed from a formal perspective.

Throughout, I make two crucial assumptions. First, I assume that all evidence is presented as testimony. At first blush this may seem to be an unrealistic assumption, but in court all physical evidence is accompanied by some testimony authenticating or otherwise speaking to its relevance and veracity. For example, merely exhibiting a firearm during the course of a murder trial bears little relevance to the case at hand *unless* someone testifies to salient facts concerning, for instances, its ownership, fingerprints found on the firearm, and the matching of the firearm to bullets recovered at the scene of the crime.

Second, I assume that the guilt of a defendant is *materially* independent from the act of any witness testifying. This is of course not to say that testimony lends no inductive weight to the case at hand, but rather that such a testimonial act never *necessitates* guilt or innocence.[1]

This is not to say that there are no logical inferences to be made regarding the *probabilities* of guilt and the veracity of the testimonies, only that there is no direct deductive relation between them.

Fundamentals of Bayesian Inference

To set the stage, we review the basics of the Bayesian account of rationality. Let S be a collection of sentences containing "true" $\top$, "false" $\bot$, and closed under Boolean operations; namely, conjunction $\wedge$, disjunction $\vee$, and negation $\neg$. For the purposes of this paper we will assume that S is finite.

We assume that the semantics of the sentences in S is understood extensionally; in other words, the sentence S is identified with the collection of worlds w in some ambient universe of possible worlds W such that s holds true in w. Under this semantics we may think of a subset $T \subset S$ of sentences as its corresponding *event*

$$E_T = \{w \in W \,|\, T \text{ is true in } w\}.$$

[1]This might not hold if, for instance, the defendant is charged with murder of person A and then A testifies during the trial. Then the act of A testifying contradicts the guilt of the defendant.

The Bayesian model of rationality supposes that each agent A is equipped at the outset with a *prior* $\mathbb{P}$, which is a certain kind of mathematical object that encodes the degree of belief, or *credence*, they afford each sentence $s \in S$. The structure of this prior $\mathbb{P}$ is that of a probability measure on S:

Definition 3.1. A probability measure on a set S of sentences is a function

$$\mathbb{P} : \mathcal{P}(S) \to [0, 1]$$

such that

1. $\mathbb{P}(E_\perp) = 0$,
2. $\mathbb{P}(E_\top) = 1$, and
3. If $E_T \cap E_{T'} = \varnothing$, then

$$\mathbb{P}(E_T) + \mathbb{P}(E_{T'}) = \mathbb{P}(E_T \cup E_{T'}). \qquad \Diamond$$

The Bayesian model of rationality requires that as evidence accumulates, the agent A updates her degrees of belief on the basis of taking conditional probabilities:

Definition 3.2. Let E be an event and $\mathbb{P}$ a prior. The *posterior distribution* of $\mathbb{P}$ given E is the conditional probability measure defined on $\mathcal{P}(S)$ given by

$$\mathbb{P}_E(T) = \frac{\mathbb{P}(E \cap T)}{\mathbb{P}(E)}. \qquad \Diamond$$

For further details regarding the epistemic interpretation of conditional probabilities and the general theory of Bayesian Rationality, the reader is directed to the excellent survey by Earman [13].

State v. Skipper and the Court's Error

Recall from the introduction the argument presented by the Supreme Court of Connecticut against the use of Bayesian reasoning in the setting of a criminal trial:

> Because Bayes' Theorem requires the assumption of a prior probability of paternity, i.e., guilt, its use is inconsistent with the presumption of innocence in a criminal case such as this.... If we assume that the presumption of innocence standard would require the prior probability of guilt to be zero, the probability of paternity in criminal cases would always be zero.... In other words, Bayes' Theorem can only work if the presumption of innocence disappears from considerations. [42, at 623]

While I argue that this argument is incorrect, the precise *way* in which it is incorrect motivates the definition of a juror's disposition. The court's argument seems to be:

1. (Presumption of Innocence) The defendant is to be presumed innocent until proven guilty.
2. (Principle of Bayesian Inference) A Bayesian juror must update their beliefs according to Bayes' rule when presented with evidence during the course of the trial.
3. If the jury finds, after hearing a set of testimony T, the conditional probability of guilt given the testimony $\mathbb{P}(E_G \,|E_T) = 0$ then the jury must acquit.
4. The presumption of innocence implies that any prior adopted by the jury must satisfy
$$\mathbb{P}(E_G) = 0.$$
5. (Conditioning) Conditioning by *any* set of testimony E_T will yield a posterior probability of 0:
$$\mathbb{P}(E_G \mid E_T) = 0$$
6. (Conclusion) For any criminal case, the jury must acquit.

That Premise 4 is a misunderstanding of Bayesian epistemology is discussed in [1], but the precise nature of this error is very illustrative of the difference between *credence* and *decision* that our framework of juror dispositions and Bayesian threshold jurors distinguishes between. On the court's view, in order for a Bayesian agent to presume innocence would mean that the Bayesian could not entertain the mere *possibility* of guilt. This is duplicitous: criminal trials are predicated on countenancing the possibility of *both* guilt and innocence at the outset. As we will see, Judge Richard Posner *also* finds Premise 4 faulty, instead arguing that the Presumption of Innocence requires that $\mathbb{P}(E_G) = \frac{1}{2}$. The remainder of this section will be spent analyzing his account of Bayesian threshold jurors.

Posner's Even-Odds Proposal

Among other things, Judge Richard Posner is in part known for his economic approach to the law. In particular, his analysis of factfinding centers on the use of Bayes' Theorem[33, p. 1486]. Posner does "make clear at the outset that I do *not* propose that juries or judges be instructed in the elements of Bayesian theory... The

most influential model of rational decision making under conditions of ineradicable uncertainty... it can be of great help, as we shall see, in evaluating the rationality of rules of evidence."[33, p. 3] Nevertheless, Posner models jurors as agents whose credences form a probability measure, which are updated in light of new evidence stemming from the testimony offered during the course of a trial. Moreover, in this model Posner interprets the burden of persuasion and the burden of proof beyond a reasonable doubt probabilistically:

> In the typical civil trial... it is enough to justify a verdict for the plaintiff that the probability that his claim is meritorious exceeds, however slightly, the probability that it is not...
>
> Type I errors are more serious than Type II errors in criminal cases therefore are weighted more heavily in the former by the imposition of a heavy burden of persuasion on the prosecution... Judges when asked to express proof beyond a reasonable doubt as a probability of guilt generally pick a number between .75 and 0.95.[33, pp. 34–36]

Therefore, in the context of a criminal trial an ideal rational juror is modeled as:

Definition 3.3. A juror j assessing the guilt G of some defendant on the basis of testimony T a collection of possible testimonial events is *Bayesian* in case: is Bayesian just in case j

i The juror has a prior probability measure $\mathbb{P} : \mathcal{P}(W) \to [0, 1]$,

ii (Conditionalization) Assigns probability $\mathbb{P}(E_G \,|E_T)$ to guilt when the juror has heard all and only the testimonies T, and

iii (θ-Verdict Rule) There is a fixed θ with $0.5 < \theta < 1$ such that the juror j renders a conviction if and only if $\mathbb{P}(E_G \,|\, E_T) \geq \theta$. $\Diamond$

It is important to note here that rendering a verdict of convict on the basis of exceeding a threshold θ requires some work to justify in the framework of classical decision theory since an explicit utility function is not presented. Somewhat recent results of Easwaran [15, p. 828] provide such a utility function. The idea is to give a reward R to an agent just in case that agent correctly believes that a proposition P is true, and to give a penalty $-W$ if the agent incorrectly believes that P is false. It is required that the agent believes at least one of P or $\neg P$, but not both. In our current context we may interpret "belief" as "votes to convict" and "disbelief" as "votes to acquit."

Definition 3.4. A *doxastic state* on a set S of sentences closed under Boolean operations is a function

$$d : S \to \{0, 1\}$$

such that $d(s) = 1$ implies $d(\neg s) = 0$.

Let s be a proposition and $R, W > 0$ be real numbers. The (R, W)-weight of s is defined as

$$\eta_{R,W}(s) = \begin{cases} R & \text{if } s \text{ is true,} \\ -W & \text{if } s \text{ is false.} \end{cases}$$

The *score* of a doxastic state is given by

$$\sigma_{R,W}(f) = \sum_{s \in S} d(s)\eta_{R,W}(s). \qquad \Diamond$$

A doxastic state encodes the a binary belief function: if $d(s) = 1$ then the agent believes s, and if $d(s) = 0$ then the agent does *not* believe s. The score of the doxastic state encodes the correctness of the agent's doxastic state.

Easwaran shows that

Theorem 3.1. [15, p. 828] For a given probability function $\mathbb{P}$, a doxastic state maximizes expected score iff it believes all propositions s such that $\mathbb{P}(s) > \frac{W}{R+W}$ and believes no propositions s such that $\mathbb{P}(s) < \frac{W}{R+W}$. Both believing and not believing are compatible with maximizing expected score if $\mathbb{P}(s) = \frac{W}{R+W}$. ⧫

In our setting, a Bayesian juror votes to convict the defendant when the posterior probability of guilt exactly equal to the threshold. The above theorem says that either choice maximizes expected score. Thus, for a threshold $0 < \theta < 1$, a Posnerian juror maximizes expected $(1 - \theta, \theta)$-score, and so Posnerian jurors are representable as a Bayesian agent with respect to *some* utility function.

This choice of utility function, however, avails itself to criticism. One way to interpret the score function above in the judicial context is to identify with W the average net social cost of wrongful conviction and R the average net social benefit of a correct conviction. While simple in its expression, making decisions according to such a score function runs into some difficult challenges.

For instance, there is little reason to think that the overall values of W and R would be the same across different crimes. The variation amongst crimes for the values of W and R would therefore adjust the value of the probability threshold θ. Therefore, if a juror is to render her verdict on the basis *only* of the posterior probability of guilt, expected *score* might be optimized but expected net social benefit will not.

One might object to this picture by saying that while it is true that the precise values of W and R vary from crime to crime, it is the judge that sentences the defendant and judges have a great deal of discretion in determining the sentence. However, for many offenses mandatory minimum sentencing renders it impossible for the judge to appropriately calibrate the punishment of the defendant once convicted.

Beyond the Bayesian model of a juror's reasoning described above, Posner also proposes the following probabilistic interpretation of the Presumption of Innocence:

> Ideally we want the trier of fact to work from prior odds of 1 to 1 that the plaintiff or prosecutor has a meritorious case... Although bias is clearest when the judge or jury not only has a prior belief about the proper outcome of the case but also holds the belief unshakably–that is, refuses to update it on the basis of evidence–it is not a complete response to a charge of bias that the judge or juror has an "open mind" in the sense of being willing to adjust his probability estimate in the light of the evidence presented at the trial. Any rational person will do that... His prior odds, if he is a Bayesian, will still have an influence on his posterior odds and hence... on his decision. [33, p. 1514]

Posner's solution does guarantee a form of the Presumption of Innocence: provided that the threshold θ is chosen to be above 50%, no Bayesian juror conforming to the constraints that Posner outlines would convict absent any evidence. However, that is all it guarantees. Since testimony is presumed logically independent from the material facts at hand, it is perfectly consistent to ensure that no matter what testimony is afforded the juror will convict as soon as testimony of any kind is given. More formally:

Proposition 3.1. Let $0 < \theta < 1$ and that T satisfies (TCA). Then there exists a prior probability $\mathbb{P}$ such that

$$\mathbb{P}(E_G) = \frac{1}{2}$$

but for any nonempty collection T of testimony

$$\mathbb{P}(E_G \mid E_T) \geq \theta. \qquad \blacklozenge$$

Proof. By the it suffices to show that we can ensure that

$$\mathbb{P}(E_G \mid E_T) \geq \theta,$$

but the conditional extension lemma (Lemma 3.1) ensures this. □

In other words, Posner's proposal–constraining only the *priors* of the jurors–is only sufficient to guarantee that juror acting in accordance with Posner's rule will not convict at the outset of the trial, and moreover is compatible with *guaranteed* conviction as soon as the first testimony is offered. By this result, constraining the prior probability of guilt to yield 1 to 1 odds only ensures that a juror's disposition cannot be to convict at the outset of the trial.

3.4 Rationalizing Juror's Dispositions

Having defined a formal model of a juror's reasoning in the preceding section, we are now in a position to evaluate whether Posner's view places any constraints on a threshold Bayesian juror's disposition beyond the Presumption of Innocence and the Willingness to Convict.

Let $\theta \in (0,1)$. We call a juror disposition $f : \mathcal{T}_S \to \{C, A\}$ *θ-rationalizable* provided that there exists a prior $\mathbb{P}_f$ on $\mathcal{P}(W)$[2] such that

$$f(T) = C \leftrightarrow \mathbb{P}_f(E_G|E_T) \geq \theta.$$

In other words, a disposition f is θ-rationalizable just in case the verdicts reached by f on all transcripts T can be realized as an instance of a threshold juror determining that the probability of the defendant's guilt meets or exceeds θ at a trial specified by transcript T.

The aim of this section is to prove a representation theorem that states that *all* juror dispositions satisfying (PoI) and (WtC) are the dispositions of *some* threshold Bayesian juror.

Theorem 3.2. Suppose that S is a finite collection of testimonies. Let

$$f : \mathcal{T}_S \to \{C, A\}$$

be a juror's disposition such that

1. $f(\varnothing) = A$ (PoI)
2. there exists a transcript T such that $f(T) = C$. (WtC)

Suppose that $\frac{1}{2} < \theta < 1$. Then there exists a prior $\mathbb{P}_f$ on W_S θ-rationalizing f such that

$$\mathbb{P}_f(E_G) = \frac{1}{2}.$$

[2]Where W is constructed as in Section 2

Moreover, $\mathbb{P}_f$ can be taken to be open-door in the sense that $\mathbb{P}_f(E_G \mid E_T) \notin \{0,1\}$ for any transcript T. ⧫

This representation theorem shows that constraints set forth by Posner to analyze the efficiency of the trial system place no meaningful constraint whatsoever on the dispositions of the finders of fact in question beyond their nontriviality.

Proof. Suppose that $f : \mathcal{T}_S \to \{C, A\}$ is a disposition satisfying the hypotheses of the theorem statement and that $\frac{1}{2} < \theta < 1$. The support of C (resp. A) is the set $f^{-1}(C) = \{T \in \mathcal{T}_C \mid f(T) = C\}$ (resp. $f^{-1}(A)$). By definition, $f^{-1}(C)$ and $f^{-1}(A)$ partition $\mathcal{T}_C$, and by the assumption of the theorem they are nonempty.

We define $\mathbb{P}_f$ as a weighted combination of two measures defined in terms of the supports of C and A. Let $n_C = |f^{-1}(C)|$ and $n_A = |f^{-1}(A)|$.

Set

$$\mathbb{P}_{f,C}(\{(T,x)\}) = \begin{cases} \theta n_C^{-1} & \text{if } T \in f^{-1}(C) \text{ and } x = G, \\ (1-\theta) n_C^{-1} & \text{if } T \in f^{-1}(C) \text{ and } x = I, \\ 0 & \text{if } T \notin f^{-1}(C). \end{cases}$$

and

$$\mathbb{P}_{f,A}(\{(T,x)\}) = \begin{cases} (1-\theta) n_A^{-1} & \text{if } T \in f^{-1}(A) \text{ and } x = G, \\ \theta n_A^{-1} & \text{if } T \in f^{-1}(A) \text{ and } x = I, \\ 0 & \text{if } T \notin f^{-1}(A). \end{cases}.$$

Both $\mathbb{P}_{f,C}$ and $\mathbb{P}_{f,A}$ induce probability measures on W_S. Since $\mathbb{P}_{f,C}(E_G) = \frac{\theta n_C}{n_C} = \theta > \frac{1}{2}$ and $\mathbb{P}_{f,A}(E_G) = \frac{(1-\theta)n_A}{n_A} = 1 - \theta < \frac{1}{2}$ there exists some $\alpha \in (0,1)$ such that the measure

$$\mathbb{P}_f = \alpha \mathbb{P}_{f,C} + (1-\alpha)\mathbb{P}_{f,A}$$

has $\mathbb{P}_f(E_G) = \frac{1}{2}$.

Moreover,

$$\mathbb{P}_f(E_G | E_T) = \frac{\mathbb{P}_f(E_G \cap E_T)}{\mathbb{P}_f(E_T)} = \frac{\mathbb{P}_f(T,G)}{\mathbb{P}_f((T,G),(T,I))} = \begin{cases} \theta & \text{if } T \in f^{-1}(C) \\ (1-\theta) & \text{if } T \in f^{-1}(A) \end{cases}$$

ensuring that $\mathbb{P}_f$ θ-rationalizes f and is open-door. □

3.5 Further Constraints on Juror's Dispositions?

Having seen that Posner's proposal puts no constraints on a juror's disposition beyond the Presumption of Innocence and the Willingness to Convict, one may attempt to rescue this account by placing further constraints on the definition of a Bayesian

threshold juror to pare down the class realizable juror dispositions. The final section of this paper is intended to survey obstacles that possible additional constraints on Bayesian accounts of rational jurors face. Broadly speaking, the additional constraints fall into three separate categories:

1. The *uniformity* of the prior with respect to a well-chosen sample space,
2. The *objectivity* of the prior, and
3. The rate of convergence to a verdict of "Guilty" of a prior.

I argue that, ultimately, none of these constraints suffice to save the Bayesian account of rational jurors.

Constraint One: Mandating Uniform Priors

The first potential save we will consider is to restrict the class of Bayesian threshold jurors to include only uniform priors with respect to an appropriate sample space. At first glance, a uniform prior of guilt on some (appropriately large) collection X of persons is appealing. For sake of simplicity, suppose that only one person x is guilty of the crime in question; that is, $E_G = \{x\}$. Then

$$\mathbb{P}(E_G) = \frac{1}{|X|}.$$

So long as X has at least two members and the threshold $\theta > \frac{1}{2}$, a juror with this prior will not convict at the outset of the trial. While this is a perfectly well-defined prior, one runs into trouble with how a juror is to update her credences on the basis of witness testimony. After all, when the axiom TCA holds, the occurrence of testimony T is consistent with the guilt or innocence of *each* $y \in X$. This implies that the set of possible worlds in which testimony T occurs is not expressible as a subset of X.

Faced with this obstruction, an advocate for assigning a uniform prior probability of $\frac{1}{|X|}$ to guilt must either supply us with a sufficiently rich sample space that the ideal juror ought to update her credences according to or argue that constraining the numerical probabilities to be proportional to the size of the set of potential perpetrators is sufficient to constrain the ideal juror. The latter option is untenable, as the proof of Proposition 3.1 shows: merely constraining the numerical prior probability of guilt to be low poses no constraint on how rapidly the juror may converge to an assessment of guilt.

As an illustration, suppose that the sample space X consists of the residents of Manhattan. A witness testifies that the perpetrator of the crime has brown eyes. Call this testimonial event T. If the witness is taken to be certainly correct, the event E_T is

$$E_T = \{x \in X \mid x \text{ has brown eyes}\} \subseteq X,$$

and the updated probability of the guilt of the defendant d is:

$$\mathbb{P}(E_G \mid E_T) = \begin{cases} 0 & d \notin E_T \\ \frac{1}{|E_T|} & d \in E_T \end{cases}$$

However, if it is acknowledged that the witness might be mistaken, the testimony does not conclusively rule out any member of X; in other words,

$$E_T = X.$$

Therefore the sample space X has insufficient expressive power to facilitate nontrivial probabilistic reasoning.

Constraint Two: Requiring Objectivity of a Juror's Prior and the Principal Principle

A second constraint one may impose on a Bayesian threshold juror is that the juror's prior be in some sense *objective*. A convenient way to formalize this is bay way of Lewis' Principal Principle, which can be expressed as follows:

> **Principal Principle:** Assume we have a number x, proposition A, time t, rational agent whose evidence is entirely about times up to and including t, and a proposition E that (a) is about times up to and including t and (b) entails that the chance of A at t is x. In any such case, the agent's credence in A given E is x. [52]

The difficulty with this approach is that, in the context of a trial, a Bayesian Threshold juror will not witness an event E of the sort referenced in the statement of the Principal Principle.

An excellent example of this exact issue playing out in case law can be found in *State v. Spann* [43]. During the course of the trial an expert witness testified, on the basis of Bayesian analysis, that from a prior probability of 50% of paternity the defendant's blood test rendered a 96.55% probability of paternity upon posterior updating. The court recounts the cross-examination of an expert witness by defense's counsel:

> On cross-examination defense counsel brought out the fact that the probability of paternity percentage was based on that fifty-fifty assumption. The expert described it as a "neutral" assumption... [h]er characterization of the evidence was that its "purely objective" nature was "one of the beauties of the test"; that it "makes no assumption other than everything is equal"; and that "the jury simply has objective information" ... Counsel noted that even if it were conclusively proven that defendant had been out of the country at the time when conception could have occurred, this expert still would have concluded that the probability the defendant was the father was 96.55%. Counsel's observation was correct; the expert's opinion had no relation whatsoever to the the facts of the case. [43, p. 590]

Defense's observation was astute. There was no *mathematical* error in the expert testimony, being a straightforward application of Bayes' Theorem. The expert's reliance on the "fifty-fifty" assumption underlies an even deeper issue: the expert testimony imposes undue constraints on the *structure* of the juror's priors *beyond* the simple constraints of "guilty" vs. "not guilty." Proper application of Bayesian updating requires knowledge of the *full* structure of the juror's prior, incorporating not only their prior assessment of guilt but also the other pieces of testimony they had heard, their background assumptions regarding the veracity of expert testimony, their understanding of general causal laws, and the like.

We reconstruct the probabilistic analysis we see in *State v. Spann*. Suppressed in the testimony is an underlying *sample space*. The basic data of the sample space is a tuple:

$$(b_X, b_Y, p(X, Y))$$

where b_X, b_Y represent the blood types of a pre-selected pair of people, X and Y, each a member of the following: $\{A+, A-, AB+, AB-, B+, B-, O+, O-\}$ and p represents the paternal relationship between the two; either "True" if the first person is a parent of the second, and "False" otherwise. This yields a sample space of size

$$8 \times 8 \times 2 = 128.$$

The event "X is a parent of Y" has size 64 in the sample space—one half the size of the total—and the expert advises us to adopt the prior that

$$\mathbb{P}(\text{“}X \text{ is a parent of } Y\text{”}) = \frac{1}{2}.$$

This probability model cannot update on an event of the form "X was in a different country from Y's mother for the 5 years before and after Y's birth" as it is not coextensive with any subset sample space: the sample space is far too coarse. Thus the

expert testimony carries with it a suppressed underlying model of the possible states of the world, which with good reason are insufficient for the purpose of updating in a Bayesian manner.

Presumably, an expert advising the jury on the probability p of the defendant's guilt on the basis of evidence such as DNA matching intends for their testimony would move the juror's credence of guilt to be p, absent any other testimony. A juror's updating of the prior in this way would, however, generally not be an instance of the Principal Principle. After all, the agent conditionalizes on testimonial data of the form "the expert testified that the probability of the defendant's guilt is p," which by the TCA axiom is logically independent from the guilt of the defendant.

The obstruction to applying the Principal Principle in this case is the missing premise that *if* the expert testifies that the probability of the defendant's guilt is p *then* the probability of the defendant's guilt is p, a premise which would refute the TCA.

Constraint Three: Objective Relevance Standards and Objective Likelihood Ratios

Another potential constraint one might consider to further constrain a Bayesian threshold juror is to require that the juror's conditional updating is compatible with some notion of an objective likelihood ratio.

Posner advances such an argument, but as we will see there is an awkward tension in Posner's analysis between objective and subjective probabilities. One might get the impression from Posner's description that there are objective probabilities governing the computation of likelihood ratios, as in the following excerpt:

> Suppose that the new piece of evidence is testimony by bystander Z that he saw X shoot Y. Suppose further that the prior odds $\Omega(H)$ are 1 to 2 that X shot Y, while the probability that Z would testify that he saw X shoot Y if X did shoot Y is .8 and the probability that he would testify that he saw X shoot Y if X did not shoot Y is .1, so that the likelihood ratio is 8. The posterior odds that X shot Y will therefore be 4 to 1...
>
> [A]ltering posterior odds may not have much or even any social value even if the likelihood ratio of the new evidence is high, as in our shooting example, where it was 8. The value of the evidence will depend on the prior odds and on the decision rule. Suppose that the prior odds (as a consequence of the previously presented evidence) that X shot Y are not 1 to 2 but 1 to 10 and that for X to be held liable for the shooting the

> trier of fact must consider the odds that he did it to be at least 1.01 to 1 (the preponderance standard). Then the new evidence, since it would lift the posterior odds above the threshold (multiplying the prior odds by a likelihood ratio of 8 yields posterior odds of only 1 to 1.25), would have no value. [33, pp. 1486–7]

It is important to note that in general one cannot posit an agent-independent likelihood ratio of a piece of evidence E: if an agent's prior probability of hypothesis H is $\mathbb{P}(H) = \theta$, then any agent with prior probability distribution $\mathbb{P}$ is assured that the likelihood ratio is at most $\frac{1}{\theta}$. In other words, the likelihood ratio afforded to evidence is *inseparable* from the structure of the agent's prior probability *measure* $\mathbb{P}$, not just its numerical values. Therefore an attempt to save Posner's account on the basis of something like objective likelihood ratios is doomed to fail: any prescribed value $L(E, H)$ will result in inconsistent assessments of probability for many agents.

The strange hybrid of subjective and objective probabilities appears again in Posner's account of the relevance standard of the Federal Rules of Evidence (FRE 401), where Posner interprets it within an economic framework: "In Bayesian terms, evidence is relevant if its likelihood ratio is different from one and irrelevant if it is one." [33, p. 1522] This Bayesian gloss fails to emphasize that the assessment of whether or not a piece of evidence is relevant in the sense that its likelihood ratio is different from 1 depends on the structure of the factfinder's prior probability measure.

To see this, note that having a likelihood ratio $L_{\mathbb{P}}(E, H)$ equal to 1 is equivalent to evidence E being probabilistically independent from H. The measure extension lemma—Lemma 3.1—entails that if a piece of evidence E is logically independent from all preceding evidence, then there exist probability distributions $\mathbb{P}$ in which E is probabilistically independent of the rest of the evidence, i.e. *irrelevant*, and probability distributions in which the evidence is probabilistically dependent, i.e. *relevant*. On Posner's account, FRE 401 is at best underspecified, and at worst fangless: all testimonial evidence is both *potentially* relevant and *potentially* irrelevant. Worse yet, even if one constrains the class of relevant testimony, the effect on the posterior probability is unconstrained.

Constraint Four: Restricting Convergence Rates

The final constraint we consider in this section is that one might attempt to save the Bayesian account by requiring that a juror not be too quick in reaching a verdict by limiting the degree to which any given piece of testimony can affect a juror's beliefs. For instance, one might demand that the *ratio* between prior and posterior belief in

guilt is bounded by some fixed amount for all testimonies T, e.g. by requiring that

$$\frac{2}{3} \leq \frac{\mathbb{P}(E_G|E_{t_1} \cap \cdots \cap E_{t_m+1})}{\mathbb{P}(E_G|E_{t_1} \cap \cdots \cap E_{t_m})} \leq \frac{3}{2}.$$

For a juror who convicts only when $\mathbb{P}(E_G|E_{t_1} \cap \cdots \cap E_{t_m} \cap E_{t_m+1}) > \theta > \frac{3}{4}$ and for whom $\mathbb{P}(E_G) = \frac{1}{2}$, such a rule would ensure that, at least conceivably, *one* piece of testimony at any time would be insufficient to render a verdict of "Guilty" since $\mathbb{P}(E_G|E_{t_1}) < \frac{3}{4} < 1$.

That said, for any rule of the form

$$\frac{\mathbb{P}(E_G|E_{t_1} \cap \cdots \cap E_{t_m+1})}{\mathbb{P}(E_G|E_{t_1} \cap \cdots \cap E_{t_m})} \leq 1 + \gamma$$

where $\gamma > 0$, verdict threshold $\theta > 0$, and prior probability of guilt $\mathbb{P}(E_G) = \frac{1}{2}$ there exists jurors who will convict just so long as there are

$$m > \frac{log(2\theta)}{\log(1+\gamma)}$$

pieces of testimony, a disposition Posner would surely want no juror to have.

Similarly, it is easily conceivable that we would *want* a juror to be able to convict having witnessed a single piece of testimony. For example, consider a defendant who—defying their pretrial pleading of "Not Guilty"—has a sudden change of heart on the stand and confesses to the crime at hand, a juror can hardly be faulted in deciding to convict on that basis alone.

Outlook on Further Constraints

The candidate constraints considered in this section range from constraints regarding the uniformity, objectivity, and the rate of convergence of a Bayesian threshold juror's prior to conviction. Each of these candidates faced severe challenges, and either were mathematically ill-defined, were not operationalizable in the trial setting, or constrained the class of rational dispositions too much. I do not claim that this list of candidate constraints is exhaustive; however, the prospects of a formal solution to the problems with the threshold Bayesian account laid out by the representation theorem appear bleak.

3.6 Conclusion

In the Introduction to this chapter, we saw a deep divide between two factions. On one side we have the anti-Bayesian current, with Tribe and the judges of the Supreme

Court of Connecticut the vanguard members advocating for the *inadmissibility* of Bayesian and other probabilistic forms of reasoning from the criminal trial system on the basis of a perceived conflict with the Presumption of Innocence. The opposing, Bayes-rationalist side of this dispute, exemplified by Judge Posner, claim that to the contrary that rationality *requires* these forms of reasoning, lest the criminal justice system fall victim to a strain of irrationality.

The analysis of this paper suggests a mundane resolution to this dispute: there is neither harm in *nor* necessity to demand a juror be Bayes rational; so long as a juror's disposition satisfies the Presumption of Innocence and the Willingness to Convict, that juror's disposition is indistinguishable from a Bayesian threshold juror's regardless of the underlying causal source of her dispositions. This result strikes at the heart of both the strongly anti-Bayesian and pro-Bayesian accounts: if you demand that Bayesian inference be banned in all its forms, there is no way to discern this on the basis of an agent's dispositions. Likewise, for the pro-Bayesian account, the representation theorem demonstrates that *nothing is gained by demanding that an agent be Bayes rational*. Thus, Posner's notion of a Bayesian juror is insufficiently specified to render this debate a substantive one.

3.7 Appendix: Lemmata from Probability Theory.

A very general extension theorem [34, p. 70] goes as follows:

Theorem 3.3. Let $\mathcal{C}$ be a Boolean algebra of subsets of a set Ω and let $\mu : \mathcal{C} \to [0,1]$ be a finitely additive, positive bounded measure. Suppose that $A \in \mathbb{P}(\Omega) \setminus \mathcal{C}$. Write

$$\mu_l(A) = \sup\{\mu(B) \,|\, B \in \mathcal{C} \wedge B \subset A\}$$

and

$$\mu_u(A) = \inf\{\mu(B) \,|\, B \in \mathcal{C} \wedge B \supset A\}.$$

Then for any $d \in [\mu_l(A), \mu_u(A)]$[3] there exists a finitely additive, positive bounded measure

$$\tilde{\mu} : \mathcal{C}\langle A\rangle \to [0,1]$$

such that

$$\tilde{\mu}(A) = d \qquad \blacklozenge$$

This lemma gives necessary and sufficient conditions to extend *finitely additive* measures to larger Boolean algebras.

More generally, we have a great deal of control over extending measures to ensure certain conditional probability inequalities hold:

Definition 3.5. Let $\mathcal{C}$ be a Boolean algebra on X. We say that B is logically independent from $\mathcal{C}$ provided that for all $A \in \mathcal{C}$,

$$A \neq X, \varnothing \to (A \cap B \neq \varnothing \wedge A^c \cap B \neq \varnothing) \qquad \lozenge$$

In other words, B intersects every nontrivial Boolean combination of elements of $\mathcal{C}$ nontrivially. When construing the set X as a set of possible worlds, this is the same as proposition B being logically independent of any set of propositions in $\mathcal{C}$.

Lemma 3.1. Suppose that $A \in \mathcal{C}$, $\mathbb{P}(A) \notin \{0,1\}$, $\mathbb{P}$ a probability charge on $\mathcal{C}$ and B is logically independent from $\mathcal{C}$. Then for all $\theta \in [0,1]$ there is an extension $\widetilde{\mathbb{P}}$ of $\mathbb{P}$ to $\mathcal{C}\langle B\rangle$ such that

$$\widetilde{\mathbb{P}}(A|B) = \theta. \qquad \blacklozenge$$

[3]Note that $\mu_l(A) \leq \mu_u(A)$ so there is always at least one extension to the measure μ.

Proof. We apply the extension theorem twice: write

$$\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c)$$

and for any $C \in \mathcal{C}$ let $\mathbb{P}_C(D) = \mathbb{P}(D \cap C)$. This is an unnormalized measure on the algebra $\mathcal{C}_C = \{C \cap D \,|\, D \in \mathcal{C}\}$. When $\mathbb{P}(C) \notin \{0, 1\}$ the measures $\mathbb{P}_C$ and $\mathbb{P}_{C^c}$ are both positive bounded charges on $\mathcal{C}_C$ and $\mathcal{C}_{C^c}$. Now as B is logically independent from $\mathcal{C}$ the set $B \cap A$ (resp. $B \cap A^c$) is logically independent of $\mathcal{C}_A$ (resp. $\mathcal{C}_{A^c}$). By the charge extension theorem we may extend the charge $\mathbb{P}_A$ (resp. $\mathbb{P}_{A^c}$) to a charge $\widetilde{\mathbb{P}_A}$ on $\mathcal{C}_A \langle B \cap A \rangle$ (resp. $\widetilde{\mathbb{P}_{A^c}}$ on $\mathcal{C}_A \langle B \cap A^c \rangle$) assigning to it any value $\rho_0 \in [0, \mathbb{P}(A)]$ (resp. $\rho_1 \in [0, 1 - \mathbb{P}(A)]$). Since the algebras $\mathcal{C}_A \langle B \cap A \rangle$ and $\mathcal{C}_A \langle B \cap A^c \rangle$ are disjoint, we may define a probability charge $\widetilde{\mathbb{P}}$ on $\mathcal{C} \langle B \rangle$ via the formula

$$\widetilde{\mathbb{P}} = \widetilde{\mathbb{P}_A} + \widetilde{\mathbb{P}_{A^c}}.$$

Let $\theta \in [0, 1]$. Then we wish to express θ as

$$\theta = \widetilde{\mathbb{P}}(A|B) = \frac{\widetilde{\mathbb{P}}(A \cap B)}{\widetilde{\mathbb{P}}(A \cap B) + \widetilde{\mathbb{P}}(A^c \cap B)} = \frac{\rho_0}{\rho_0 + \rho_1}$$

This is a continuous function in (ρ_0, ρ_1) taking on the values of 0 ($\rho_0 = 0$ and $\rho_1 = \mathbb{P}(A^c)$) and 1 ($\rho_0 = \mathbb{P}(A)$ and $\rho_1 = 0$) and so by taking a suitable line inside $[0, \mathbb{P}(A)] \times [0, \mathbb{P}(A^c)]$ one can apply the Intermediate Value Theorem to conclude that there are ρ_0, ρ_1 rendering

$$\mathbb{P}(A|B) = \theta$$

true. □

This proposition tells us that when adjoining a logically independent event to our algebra $\mathcal{C}$, $\mathbb{P}$ may be consistently extended in such a way as to render the conditional probability $\widetilde{\mathbb{P}}(A|B)$ *any* value whatsoever. This indicates that merely placing constraints on the *numerical* values of prior probability of an event A places *no* constraint on the posterior probability when updating by a logically independent event except in the degenerate cases where $\mathbb{P}(A) \in \{0, 1\}$.

3.8 Appendix: Remarks on the Utility Functions Rationalized by Threshold Bayesian Jurors

Easwaran's score function reflects a coarse grain of accuracy. In this section we generalize his result by considering the case of epistemic scores determined by the four possible outcomes of a juror's verdict: the conviction of a guilty defendant, the conviction of an innocent defendant, the acquittal of a guilty defendant, and the acquittal of an innocent defendant. Let

$$S = \{GC, NGC, GA, NGA\}$$

be the state space corresponding to these four possible outcomes. A collection $\{(s, \alpha_s)\}_{s \in S}$ with each $\alpha_s \in \mathbb{R}$ determines a utility function:

$$U(s) = \sum_{t \in S} \chi_t(s)\alpha_s$$

where

$$\chi_t(s) = \begin{cases} 1 & \text{if } s = t, \\ 0 & \text{otherwise.} \end{cases}$$

Relative to a probability measure $\mathbb{P}$ on $\{G, NG\}$, the expected utility of each verdict is given by:

$$E(U; C) = \mathbb{P}(E_G)\alpha_{GC} + (1 - \mathbb{P}(E_G))\alpha_{NGC},$$

and

$$E(U; A) = \mathbb{P}(E_G)\alpha_{GA} + (1 - \mathbb{P}(E_G))\alpha_{NGA}.$$

To maximize expected utility, an agent will convict just in case

$$E(U; C) \geq E(U; A).$$

Simple algebraic manipulation shows that this occurs precisely when

$$\mathbb{P}(E_G) \geq \frac{\alpha_{NGA} - \alpha_{NGC}}{\alpha_{GC} - \alpha_{NGC} - \alpha_{GA} + \alpha_{NGA}}.$$

This computation entails that being a threshold juror is rationalizability according to more flexible epistemic utility functions.

Chapter 4

On the Sufficiency of First-Order Logic

4.1 Introduction

Barwise [4] defines the *first-order thesis* as the "widespread" view that "logic is first-order logic, so that anything that cannot be defined in first-order logic is outside the domain of logic." He goes on to write that

> [a]s logicians we do our subject a disservice by convincing others that logic is first-order logic and then convincing them that almost none of the concepts of modern mathematics can really be captured in first-order logic. Paging through any modern mathematics book, one comes across concept after concept that cannot be expressed in first-order logic.... First-order logic is just an artificial language constructed to help investigate logic, much as the telescope is a tool constructed to help study heavenly bodies. From the perspective of the mathematician in the street, the first-order thesis is like the claim that astronomy is the study of the telescope. [4, pp. 5–6]

After all, the semantics of first-order logic is coarse: it is famously unable to discriminate between countable and uncountable structures. Barwise's concern seems chiefly with the *semantic* deficiencies of first-order logic.

On the other hand, there is a sense in which first-order logic is sufficient for the kinds of deductive reasoning that finitistic agents are able to perform: any recursively enumerable deductive system—very generally construed—can be realized as an instance of a first-order deductive system.

Thus, even while there are strictly more expressive logics than first-order logic with a recursively enumerable consequence relation, first-order logic is nevertheless sufficient to translate and mechanize its reasoning.

In the next section, we evaluate the argument recently put forward by Warren [50] claiming that it is conceivable that agents much like ourselves in certain spacetimes known as Malament-Hogarth spacetimes may be able to perform infinitary, non-recursive deductive arguments. For him, "obviously the cleanest argument for infinite inference would be to point to an example. But it would need to be an example that couldn't plausibly be redescribed in finite terms" [50, p. 397] . While I don't directly argue against the metaphysical possibility of ω inference, I argue that the inference Warren has in mind can be readily redescribed as an instance of modus ponens relative to a validity in the theory of Malament-Hogarth spacetimes together with an empirical observation. Those who default to the position that ω inference is *not* metaphysically possible may live to fight another day.

Finally, we discuss how the Σ_1 completeness of first-order entailment dovetails with the so-called *Hilbert Thesis*, which consists of two sub-theses:

1. Hilbert's Expressibility Thesis (HET): All mathematical (extra-logical) assumptions may be expressed in first-order logic, and

2. Hilbert's Provability Thesis (HPT): The informal notion of *provable* is made precise by the formal notion of *provable in first-order logic*.

Kripke [26] argued that

$$\text{(HET) + (HPT)} \implies \text{Church's Thesis.}$$

The above arguments suggest a partial converse to Kripke's conclusion; the Σ_1 completeness of $\vdash_{FO}$ suggests that

$$\text{Church's Thesis + In-Principle Machine Verifiability of Proofs} \implies \text{(HPT).}$$

4.2 Sufficiency of First-Order Logic for Recursive Proof

In this section we are focused solely on the **inferential structure** of a collection of propositions, regardless of their underlying semantics. To this end, we model a system of inference as a closure operator on a set of words over some recursive set S of sentences. Without loss of generality we identify S with ω.

We show that the class of *recursive* closure operators can in fact be mechanized within first-order logic in a precise manner.

An *inferential closure operator* over S is a function

$$\nabla : \mathcal{P}(S) \to \mathcal{P}(S)$$

satisfying

1. (Monotonicity) $\Gamma \subseteq \nabla(\Gamma)$
2. (Homomorphism) $\Gamma_1 \subseteq \Gamma_2$ entails $\nabla(\Gamma_1) \subset \nabla(\Gamma_2)$
3. (Idempotence) $\nabla^2(\Gamma) = \nabla(\Gamma)$

and perhaps more requirements. However, we will not make use of these properties: all that will matter for our purposes is the *Turing functionality* of this operator.

Definition 4.1. A *recursive deductive closure operator* is an inferential closure operator such that there exists a Σ_1 formula ψ_∇ in the language of arithmetic augmented by a single unary predicate $P(x)$ such that for all $\Gamma \subseteq S$ and φ, the model (ω, Γ) where the interpretation of predicate P is Γ satisfies

$$\varphi \in \nabla(\Gamma) \leftrightarrow (\omega, \Gamma) \vDash \psi_\nabla(\varphi). \qquad \Diamond$$

We think of ∇ as determining a notion of entailment:

$$\varphi \in \nabla(\Gamma) \leftrightarrow \Gamma \vdash_\nabla \varphi.$$

In other words, a recursive deductive closure operator is one for which membership in the deductive closure is uniformly recursively enumerable, uniform over the base theory $\Gamma \subseteq S$. Examples of such closure operators are ubiquitous in mathematical logic.

Of course, First-Order Logic, Intuitionistic Logic, Classical Propositional Logic, all have recursive deductive closure operators, witnessed by enumerating the consequences of an oracle-fed Γ by way of enumerating the proofs relative to any of their proof systems.

However, there exist strictly stronger logics still satisfying this property.

Definition 4.2. Let $L(Q_1)$ be First-Order Logic augmented with the quantifier Q_1 with the following semantics: $\mathcal{M} \vDash (Q_1 x_1, \ldots, x_n)\varphi(x_1, \ldots, x_n)$ just in case there are $\geq \aleph_1$ many $m \in \mathcal{M}^n$ such that $\mathcal{M} \vDash \varphi(m)$. $\Diamond$

This logic is not compact for arbitrary sets of sentences. However,

Theorem 4.1. [4, Corollary III.1.2.2 and III.1.2.3] The logic $L(Q_1)$ is countably compact, i.e., if $|\Gamma| \leq \aleph_0$ and each finite $\Gamma_0 \subset \Gamma$ is satisfiable, then Γ is satisfiable.

Moreover, for each recursive signature $\mathcal{L}$, the $L(Q_1)$ entailment relation on $L(Q_1)$ $\mathcal{L}$-sentences is a recursive deductive closure operator. ⧫

Keisler [24] showed that the logic $L(Q_1)$ has an effective proof system such that for countable sets of $L(Q_1)$ sentences Γ, we have $\Gamma \vDash \varphi$ just in case $\Gamma \vdash_1 \varphi$.

The proof system for $L(Q_1)$ consists of modus ponens together with the following axioms [23, Definition IV.3.1.1]:

1. All universal closures of first-order validities in the language $L(Q_1)$,
2. $\neg Qx(x = y \vee x = z)$,
3. $\forall x(\varphi \to \psi) \to (Qx\varphi \to Qx\psi)$,
4. $(Qy\exists x\varphi) \to (\exists xQy\varphi \vee Qx\exists y\varphi)$, and
5. $Qx\varphi(x) \leftrightarrow Qy\varphi(y)$ where $\varphi(x, \dots)$ is a formula in $L(Q_1)$ in which y does not occur, and $\varphi(y, \dots)$ is obtained by subsituting each instance of y by an instance of x.

We write $\Gamma \vdash_1 \varphi$ to mean that φ is provable from Γ by way of the above proof system. We denote the axioms 2-5 in the above by $\mathcal{A}_1$. Note that $\mathcal{A}_1$ is a recursive set of axioms.

A set Γ of $L(Q_1)$ sentences is *consistent* just in case the above proof system does not prove a contradiction from Γ. The completeness theorem for $L(Q_1)$ asserts the following.

Theorem 4.2. [24, p. 13] Let Γ be a set of $L(Q_1)$ sentences in a countable signature. Then Γ is satisfiable if and only if Γ is consistent. ⧫

Consequently:

Corollary 4.1. Let Γ be a countable set of $L(Q_1)$ sentences and φ an $L(Q_1)$ sentence. Then $\Gamma \vDash \varphi$ entails $\Gamma \vdash_1 \varphi$.

Moreover, given an oracle naming Γ, the consequence relation $\vDash$ is recursively enumerable in this oracle. ⧫

Proof. Suppose $\Gamma \vDash \varphi$ but $\Gamma \nvdash_1 \varphi$. Then $\Gamma \cup \{\neg\varphi\}$ is countable and consistent but unsatisfiable, contradicting the completeness theorem of $L(Q_1)$.

The "moreover" clause follows because the set of proofs in $L(Q_1)$ from Γ is recursively enumerable in Γ. □

There is a strong sense in which the inferential structure of $L(Q_1)$ can be simulated by first-order logic. Given a formula $Qy\varphi(y, \overline{x})$ we may recursively map it to the first-order formula $\varphi_{Qy}(\overline{x})$ generated by the assignment $QyR(\overline{(}x), y)$ to a new predicate symbol $R_{Qy}(\overline{x})$. This translation function is clearly well-defined; let φ^* denote the translation of an $L(Q_1)$-formula φ into the first-order formula described above. By replacing all instances of φ with φ^* in the axioms $\mathcal{A}_1$ of the proof system for $L(Q_1)$ with $\varphi_{Qy}(\overline{x})$ we see that the axioms are all first-order sentences and

$$\Gamma \vdash_1 \varphi \leftrightarrow (\Gamma^* \cup \mathcal{A}_1^*) \vdash_{FO} \varphi^*.$$

If Γ is recursive/recursively enumerable relative to some oracle, so is $(\Gamma^* \cup \mathcal{A}_1^*)$.

Of course, the formula φ^* will generally admit different *models* than φ since $QyP(y)$ will have only uncountable models if satisfiable, while P_{Qy} will have countable models by the Löwenheim-Skolem theorem.

Thus, while first-order logic is strictly less expressive than $L(Q_1)$, first-order logic is nevertheless able to internalize the consequence relation on $L(Q_1)$ for all countable sets of $L(Q_1)$ sentences. Therefore, a "deductive" version of Lindström's characterization of First-Order Logic is not possible:

Corollary 4.2. There are logics strictly stronger than First-Order Logic with recursive deductive closure operators. ⧫

However, we might still hold out hope that First-Order Logic is *sufficiently strong* to internalize the consequence relation of all recursive deductive closure operators in a manner similar to its representation of the proof system of $L(Q_1)$. A trivial generalization of a well-known theorem in recursion theory shows that this is indeed the case.

Let $\mathcal{L}$ be a recursive language consisting of

– countably many constant symbols c,

– countably many function symbols f for each arity $n \in \omega$, and

– countably many relation symbols R for each arity $n \in \omega$.

Then:

Proposition 4.1. The set V of valid first-order sentences in the recursive signature $\mathcal{L}$ is Σ_1-complete. Moreover, the first-order deductive closure operator Turing-reduces every other Σ_1 Turing functional. ⧫

Proof. By the completeness theorem of first-order logic, every valid first-order sentence is provable, so the set of valid first-order sentences in a recursive signature is recursively enumerable, i.e., Σ_1. Moreover, the proof of the undecidability of first-order logic in [16, Thm 4.1] goes by proving that the set V solves the halting problem, which is Σ_1-complete [41, Thm II.4.2].

The "moreover" clause follows because one may encode the halting of an oracle machine into first-order logic in the exact same way as in the above proof. □

From this we may conclude that First-Order Logic is sufficient to simulate the deductive structure of *any* recursive deductive closure operator. Independently, Walsh [49] has given a similar argument.

Corollary 4.3. Every recursive deductive closure operator ∇ is Turing-reducible to the First-Order deductive closure operator. ⧫

More directly, this result says that if $\vdash$ is some Turing functional capturing the notion of proof in some logical system, then there are computable translation function $\varphi \mapsto \varphi^*$ and $\Gamma \mapsto \widehat{\Gamma}$ such that

$$\Gamma \vdash \varphi \Leftrightarrow \widehat{\Gamma} \vdash_{FO} \varphi^*.$$

However, as we will see in the following section, Warren [50] argues that there is—at least in principle—plausible physical theories in which an agent can invoke a properly infinitary rule of inference, the ω rule of inference. The ω-rule is not a Σ_1 rule of inference as it allows for the infinite use of an oracle, whereas Σ_1 rules allow only finite use. If Warren's account is correct, my argument for the sufficiency of first-order logic fails.

4.3 On a Purported Instance of the ω Rule

A central premise of the sufficiency argument for first-order logic given above is that first-order logic is able to simulate any *uniform, recursively enumerable* deductive closure operator. Recently, Warren [50] has argued that an agent may have the ability to invoke the ω rule of inference in our own reasoning. In the paper, he illustrates his point by way of an example from the literature on the physical possibility of

performing supertasks—specifically, instances of the ω rule of inference—such as deciding on the basis of some physical process whether a given $\forall_1$ sentence of arithmetic is true or false in finite time.

Recall [8, p. 81] the notion of an ω-model:

Definition 4.3. An ω-model is an $\mathcal{L}$-structure $\mathcal{M}$ such that $\mathrm{dom}(\mathcal{M}) = \omega$.

The ω-consequence relation $\Gamma \vDash_\omega \varphi$, where Γ is a set of first-order $\mathcal{L}$-sentences and φ a first-order $\mathcal{L}$-sentence is given by

$$\Gamma \vDash_\omega \varphi \leftrightarrow (\forall \mathcal{M} \text{ an } \omega \text{ model})(\mathcal{M} \vDash \Gamma \rightarrow \mathcal{M} \vDash \varphi). \qquad \lozenge$$

The ω rule of inference is the following infinitary deductive pattern:

$$\begin{array}{l|l} 1 & \psi(0), \psi(1), \ldots, \psi(n), \ldots \\ \hline 2 & \forall x \psi(x) \end{array}$$

The ω rule is *sound* on ω-models; after all, it precisely expresses the truth condition for the quantifier $\forall x$ on an ω-model. However, the above rule is by construction not even expressible as a finite string. Nevertheless, we can inductively define this closure operator to get a well-defined entailment relation between sets of sentences. We construct $\vdash_\omega$ by way of induction:

1. If $\gamma \in \Gamma$ then $\Gamma \vdash_\omega \gamma$.
2. If $\Gamma \vdash_\omega \varphi_1, \ldots, \varphi_n$ and $\varphi_1, \ldots, \varphi_n \vdash_{FO} \psi$ then $\Gamma \vdash_\omega \psi$.
3. If for all $n \in \omega$ $\Gamma \vdash_\omega \psi(n)$ then $\Gamma \vdash_\omega \forall x \psi(x)$.

The ω rule is non-recursive in the sense that from a recursive base theory Γ the ω rule entails *all* $\forall_1$-truths in $\langle \omega, +, \times, 0, S(x) \rangle$:

Proposition 4.2. Let PA be Peano Arithmetic. Then for every true $\forall_1$ sentence φ in arithmetic, $PA \vdash_\omega \varphi$.

Since the $\forall_1$ theory of arithmetic is not recursive, $\vdash_\omega$ is a non-recursive provability relation. ⧫

Proof. Since $\omega^n \simeq \omega$ definably in the language $\mathcal{L}$, φ is provably equivalent to a sentence of the form $\forall x \psi(x)$ with $\psi(x)$ a quantifier-free sentence in the language of rings. By the definition of $\forall$,

$$\omega \vDash \forall x \psi(x) \Leftrightarrow \bigwedge_{n \in \omega} \omega \vDash \psi(n) \Leftrightarrow \{\psi(n)\}_{n \in \omega} \vdash_\omega \forall x \psi(x) \Leftrightarrow PA \vdash_\omega \forall x \psi(x).$$

Therefore the ω rule entails $\forall x\psi(x)$ if and only if $\omega \vDash \forall x\psi(x)$.

However, the universal theory of arithmetic is co-r.e. complete and hence not recursive. Therefore, the ω proof system $\vdash_\omega$ is not a recursive relation. □

Using the fact that Peano Arithmetic proves that every formula ψ in n free variables is equivalent to a formula in m free variables for all $0 < m < n$, a simple induction argument shows that all of True Arithmetic (TA) is provable from PA given the ω-rule:

Corollary 4.4. $PA \vdash_\omega TA$. ♦

Thus, being able to perform an ω inference is a *highly* uncomputable process. Warren's purported example of an instance of the ω rule is partially mechanized with the aid of a hypothetical physical process. A classic paper of Earman and Norton [14] demonstrates the existence of models of General Relativity in which the truth of arbitrary $\forall_1$ sentences of arithmetic can be determined by way of physical experiment. In broad strokes, this experimental setup is accomplished by exploiting the fact that in General Relativity there can exist two observers O_1 and O_2 such that O_2's "past light cone contains the entire world-line of" O_1 [14, p. 23]. Suppose that we wish to determine whether or not the arithmetic sentence $\forall x\varphi(x)$ is true, with $\varphi(x)$ quantifier-free. Earman and Norton set out to determine this by having a terrestrial scientist send an idealized computer into the vicinity of a spacetime singularity. Over the infinite time horizon of this idealized computer, it will determine whether each instance $\varphi(n)$ is true or not. If the computer finds a counterexample to $\forall x\varphi(x)$, it sends a physical signal back to the scientist and then halts. Of course, relative to the computer's frame of reference infinite time has elapsed, but if set up properly the entire worldline of the computer is observable to the terrestrial observer in finite time. So, relative to the terrestrial observer's frame of reference there is a time t_0 such that if no signal indicating that a counterexample was received by time t_0, then $\forall x\varphi(x)$ is true.

One may in this case be tempted to argue that an instance of the ω-rule is at play: after all, the experimental setup involves observer O_1 checking each instance $\varphi(0), \varphi(1), \ldots$ in order to verify $\forall x\varphi(x)$. This is the position that Warren adopts, writing in the case of testing Goldbach's conjecture that

> when the computation [determining the truth of $\forall xGB(x)$] fails to halt, we first accept each of $GB(0), GB(1), GB(2), GB(3), \ldots$. We do this without any explicit proofs of these claims, using instead the evidence of the computation. Then, with the computational interval complete, we conclude $\forall xGB(x)$. *I think that the best and most natural description of*

> *this reasoning is that we would, on the basis of the computation, accept each of the infinitely many premises, and then infer the conclusion—Goldbach's Conjecture—from these infinitely many premises with omega reasoning.* [50, p. 17] [emphasis mine]

Warren defends his position by arguing that the pattern of reasoning in the Malament-Hogarth setting is *not* best understood as a form of induction. The induction argument for Goldbach's Conjecture would go

1	$GB(0)$
2	$\forall x(GB(x) \rightarrow GB(Sx))$
3	$\forall x GB(x)$

Warren rightly points out that

> there is no sense in which, at any point, the Goldbach computation itself checks or establishes the premise "$\forall x(GB(x) \rightarrow GB(Sx))$."... It is possible that, even if we use mathematical induction to infer the conjecture, omega reasoning is still used in securing the induction premise. On the basis of the computation we accept all of the infinitely many conditionals — $(GB(0) \rightarrow GB(1))$, $(GB(1) \rightarrow GB(2))$, $(GB(2) \rightarrow GB(3))$, ... — and then, on this basis, accept "$\forall x(GB(x) \rightarrow GB(Sx))$". So this way of using induction doesn't avoid the omega rule. [50, p. 17].

This argument does, I think, dispense with the competing account by induction. However, induction is not the only relevant pattern of reasoning. Neither inferential pattern that Warren considers here invoke any premises with *physically observable content*. That is, they model the reasoning of the agent as *only* involving mathematical premises such as $GB(n)$. Rather, the most straightforward account of the observer's inference to $\forall x GB(x)$ can be viewed as a simple application of Modus Ponens. For each $\forall_1$ sentence $\varphi = \forall x \widehat{\varphi}(x)$ of arithmetic, let $\blacksquare\varphi$ be the statement "experiment E_φ was performed and no signal indicating a counterexample to φ was observed by time t_0." For a true $\forall_1$ sentence of arithmetic φ, the observer reasons to φ by way of simple Modus Ponens:

1	$\blacksquare\varphi \rightarrow \varphi$	Validity in Malament-Hogarth Spacetimes
2	$\blacksquare\varphi$	Experimental Observation
3	φ	Modus Ponens, 1, 2

Examining premises (1) and (2) in detail, we find that no properly infinite reasoning takes place. For Premise (1), one may argue $\blacksquare\varphi \to \varphi$ by contraposition and the properties of universal and existential quantifiers. First, if the experiment does not take place then $\neg\blacksquare\varphi$ holds, so $\neg\varphi \to \neg\blacksquare\varphi$. Therefore we reduce to the case where the experiment takes place. In this case, if $\neg\varphi$ holds, then there is some $n \in \omega$ such that $\neg\widehat{\varphi}(n)$ holds. By the specification of the experimental setup, there exists some time $t(n) < t_0$ where a signal indicating that φ has been refuted is received by the observer. This is precisely the truth condition for $\neg\blacksquare\varphi$ given that the experiment takes place. Therefore, φ is a logical consequence of $\blacksquare\varphi$ in the Malament-Hogarth setting. This is not, properly speaking, a *reflection principle*. Instead, it expresses the soundness of our experimental apparatus. Therefore the inferential pattern given above shows that there is a simple first-order argument to φ in the Malament-Hogarth setting.

Unlike Premise (1), $\blacksquare\varphi$ is *not* a validity in Malament-Hogarth Spacetimes. Rather, $\blacksquare\varphi$ is a contingent premise. The truth conditions for $\blacksquare\varphi$ are purely physical: $\blacksquare\varphi$ can be verified or refuted simply by determining whether or not some signal was received by a set, known time t_0.

Of course, the truth of $\blacksquare\varphi$ is partially grounded in the truth of the instances $\widehat{\varphi}(n)$, but it is *also* partially grounded in the observation of a specific physical state. In fact, it is instead a red herring that the observer in the above scenario observes each $\widehat{\varphi}(n)$ *prior* to inferring φ. The inference to φ by the observer does not require the observer to accept *a single* instance $\widehat{\varphi}(n)$ prior to t_0.

To see why, let us slightly modify the experimental setup. Rather than a signal being sent to the observer just in case a counterexample to φ is found, let the signal be sent first to a receiver which will in turn send a signal at a predetermined time $t_1 > t_0$. The receiver sends our observer the signal "True" if no counterexample to φ was found by time t_0 and "False" otherwise. In this case, the observer

1. receives no signal regarding any particular instance $\widehat{\varphi}(n)$ and

2. infers φ correctly on the basis of the truth of some observable physical phenomena.

Therefore, the particular instances $\widehat{\varphi}(n)$ are inadmissible to the observer's pattern of inference; the receiver censors the signals until all of the information comes in. Nevertheless, the observer is able to infer φ on the basis of the directly observable "black box" sentence $\blacksquare\varphi$.

In Warren's initial setup, one might be tempted to believe that the inferential picture is described by the narrative: "The observer observes $\widehat{\varphi}(n)$ for each n by time t_0, and on this basis infers φ." The above example shows that one can instead

consistently believe that the narrative in the initial example can be redescribed: "The observer observes confirmation of $\varphi(n)$ for each n by time t_0; separately, the observer concludes on the basis of some physical fact that φ at time t_0." Both the inference and the conclusion of the experiment occur at time t_0, but that does not entail that each of the infinite premises is used in the agent's inference.

I readily admit that this does not constitute a knock-down argument against the metaphysical possibility of some realization of the ω rule—I would not be surprised if there were no such knock-down argument. However, Warren's purported example itself does not necessitate an instance of ω reasoning: the inference Warren has in mind can can be readily redescribed as an instance of modus ponens relative to a validity in the theory of Malament-Hogarth spacetimes together with a contingent empirical observation.

4.4 Reflections on Hilbert's Thesis

The *Hilbert Thesis*, first defined by Barwise [3, p. 41], is the hypothesis that

1. Hilbert's Expressibility Thesis (HET): All mathematical (extra-logical) assumptions may be expressed in first-order logic, and

2. Hilbert's Provability Thesis (HPT): The informal notion of *provable* is made precise by the formal notion of *provable in first-order logic*.

Hilbert's Provability Thesis (HPT)—explicitly affirming the sufficiency of first-order deduction for provability—is reminiscent of Church's Thesis that the informal notion of computability is made precise by the notion of Turing computability. Kahle argues that Church's Thesis is disanalagous to Hilbert's thesis in the following regard:

> One could try to put in parallel the different first-order axiom systems with the different functions calculated by different Turing machines, such that the different non-logical axioms would correlate to the different states and transition tables of a Turing machine. This parallel is insofar[sic] defective, as there exists a *universal Turing machine* which can encode the different machines in just one, while—due to Gödel—such a unified first-order axiom system cannot exist. [22, Section 4]

In other words, Gödel's incompleteness theorem shows that—unlike the situation of Church's thesis—no single first-order *theory* is sufficient to capture all mathematical inferences, while there *is* a universal Turing machine. Kahle notes that there is well-known way to salvage HPT by requiring only that that all proofs be in-principle

formalizable in *some* theory T—as opposed to a fixed theory T—as Gödel's Incompleteness Theorem applies only to fixed T.

By contrast, Kripke [26] argued that

$$\text{(HET)} + \text{(HPT)} \implies \text{Church's Thesis}$$

on the basis of Gödel's completeness theorem. The argument runs as follows:

> Suppose one has any valid argument whose steps can be stated in a first-order language. It is an immediate consequence of the Gödel completeness theorem for first-order logic with identity that the premises of the argument can be formalized in any conventional formal system of first-order logic. Granted that the proof relation of such a system is recursive (computable), it immediately follows in the special case where one is computing a function (say, in the language of arithmetic) that the function must be recursive (Turing computable). [26, p. 81]

The Σ_1 completeness of the first-order entailment relation $\vdash_{FO}$ (Corollary 4.3) entails

$$\text{Church's Thesis} + \text{In-Principle Machine Verifiability of Proofs} \implies \text{(HPT)}.$$

This is because the Σ_1 completeness of $\vdash_{FO}$ implies that any deductive system with in-principle machine-verifiable inferences can be simulated within the first-order proof system at the cost of potentially weakening the expressive strength of the underlying logic. As we saw in the example of $L(Q_1)$, the translation from an $L(Q_1)$ sentence φ to a first-order sentence φ^* will generally result in a sentence with different truth conditions and hence different models.

At first glance this appears to be a defect of the first-orderization process. However, this process actually isolates *sufficient first-order conditions* to carry out the argument: despite the truth-conditions of the original sentences φ being non-first-order, a first-order approximation of them suffices to witness the validity of the argument. First-Order Logic is, of course, not the only such logic: any logic with Σ_1 complete entailment relation will also be able to simulate any recursively enumerable deduction relation. But First-Order Logic suffices.

Bibliography

[1] Ronald J Allen et al. "Probability and Proof in State v. Skipper: an Internet Exchange". In: *Jurimetrics* 35 (1994), p. 277.

[2] Vladimir Igorevich Arnol'd. *Mathematical Methods of Classical Mechanics*. Vol. 60. Springer Science & Business Media, 2013.

[3] Jon Barwise. "An Introduction to First-Order Logic". In: *Handbook of Mathematical Logic*. Elsevier, 1977, pp. 5–46.

[4] Jon Barwise et al. "Chapter I: Model-Theoretic Logics: Background and Aims". In: *Model-Theoretic Logics*. Springer-Verlag, 1985, pp. 3–23.

[5] Nick Bostrom. *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Routledge, 2013.

[6] Sean M Carroll. "Why Boltzmann Brains are Bad". In: *Current Controversies in Philosophy of Science*. Routledge, 2020, pp. 7–20.

[7] Christopher P Chambers, Federico Echenique, and Eran Shmaya. "The Axiomatic Structure of Empirical Content". In: *American Economic Review* 104.8 (2014), pp. 2303–19.

[8] Chen Chung Chang and H Jerome Keisler. *Model Theory*. Elsevier, 1990.

[9] Krzysztof Ciesielski and Jakub Jasinski. "Topologies Making a Given Ideal Nowhere Dense or Meager". In: *Topology and its Applications* 63.3 (1995), pp. 277–298.

[10] Lou van den Dries and Chris Miller. "On the Real Exponential Field with Restricted Analytic Functions". In: *Israel Journal of Mathematics* 85.1-3 (1994), pp. 19–56.

[11] Rick Durrett. *Probability: Theory and Examples*. Vol. 49. Cambridge university press, 2010.

[12] Mirna Džamonja and Saharon Shelah. "On $\triangleleft^* -$ *maximality*". In: *Annals of Pure and Applied Logic* 125.1 (2004), pp. 119–158. ISSN: 0168-0072. DOI: https://doi.org/10.1016/j.apal.2003.11.001. URL: https://www.sciencedirect.com/science/article/pii/S0168007203001015.

[13] John Earman. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. MIT Press, 1992.

[14] John Earman and John D Norton. "Forever is a Day: Supertasks in Pitowsky and Malament-Hogarth spacetimes". In: *Philosophy of Science* 60.1 (1993), pp. 22–42.

[15] Kenny Easwaran. "Dr. Truthlove or: How I Learned to Stop Worrying and Love Bayesian Probabilities". In: *Noûs* 50.4 (2016), pp. 816–853.

[16] H-D Ebbinghaus, Jörg Flum, and Wolfgang Thomas. *Mathematical Logic*. Springer Science & Business Media, 2013.

[17] Richard P Feynman, Robert B Leighton, and Matthew Sands. *The Feynman Lectures on Physics, Vol. I: The New Millennium Edition: Mainly Mechanics, Radiation, and Heat*. Vol. 1. Basic Books, 2011.

[18] Itzhak Gilboa. *Theory of Decision Under Uncertainty*. Vol. 45. Cambridge University Press, 2009.

[19] Leon Henkin, Patrick Suppes, and Alfred Tarski. *The Axiomatic Method. with Special Reference to Geometry and Physics*. North-Holland Amsterdam, 1959.

[20] Wilfrid Hodges. *Model Theory*. Vol. 42. Cambridge University Press, 1993.

[21] Emil Jeřábek. "Recursive Functions and Existentially Closed Structures". In: *Journal of Mathematical Logic* 20.01 (2020).

[22] Reinhard Kahle. "Is There a "Hilbert Thesis"?" In: *Studia Logica* 107.1 (2019), pp. 145–165.

[23] Matt Kaufmann. "Chapter IV: The Quantifier" There Exist Uncountably Many" and Some of Its Relatives". In: *Model-Theoretic Logics*. Springer-Verlag, 1985, pp. 123–176.

[24] H Jerome Keisler. "Logic with the Quantifier "There Exist Uncountably Many"". In: *Annals of Mathematical Logic* 1.1 (1970), pp. 1–93.

[25] Kevin T Kelly. *The Logic of Reliable Inquiry*. Oxford University Press, 1996.

[26] Saul A Kripke. "The Church-Turing 'Thesis' as a Special Corollary of Gödel's Completeness Theorem". In: *Computability: Turing, Gödel, Church, and Beyond* (2013), pp. 77–104.

[27] Alex Kruckman and Nicholas Ramsey. "Generic Expansion and Skolemization in NSOP1 Theories". In: *Annals of Pure and Applied Logic* 169.8 (2018), pp. 755–774.

[28] David Marker. *Model Theory: an Introduction*. Vol. 217. Springer Science & Business Media, 2006.

[29] Deborah G Mayo. "Evidence as Passing Severe Tests: Highly Probable Versus Highly Probed Hypotheses". In: *Scientific Evidence: Philosophical Theories and Applications*. Ed. by Peter Achinstein. JHU Press, 2005. Chap. 6, pp. 95–127.

[30] Deborah G Mayo. *Statistical Inference as Severe Testing*. Cambridge University Press, 2018.

[31] Deborah G Mayo and Aris Spanos. "Introduction and Background". In: *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science* (2010), p. 28.

[32] Karl Popper. *The Logic of Scientific Discovery*. Routledge, 2005.

[33] Richard A Posner. "An Economic Approach to the Law of Evidence". In: *Stan. L. Rev.* 51 (1998), p. 1477.

[34] KPS Bhaskara Rao and M Bhaskara Rao. *Theory of Charges: a Study of Finitely Additive Measures*. Academic Press, 1983.

[35] Oliver Schulte and Cory Juhl. "Epistemology, Reliable Inquiry and Topology". In: (1996).

[36] Yevgeny Seldin and Bernhard Schölkopf. "On the Relations and Differences Between Popper Dimension, Exclusion Dimension and VC-Dimension". In: *Empirical Inference*. Springer, 2013, pp. 53–57.

[37] Saharon Shelah. *Classification Theory: and the Number of Non-Isomorphic Models*. Elsevier, 1990.

[38] Herbert A Simon. "FIT, Finite, and Universal Axiomatization of Theories". In: *Philosophy of Science* 46.2 (1979), pp. 295–301.

[39] Pierre Simon. *A Guide to NIP theories*. Cambridge University Press, 2015.

[40] Joseph D Sneed. *The logical structure of mathematical physics: Second Edition, Revised*. D. Reidel Publishing Company, 1979.

[41] Robert I Soare. *Recursively Enumerable Sets and Degrees: A study of Computable Functions and Computably Generated Sets*. Springer Science & Business Media, 1999.

[42] *State v. Skipper*. 1994.

[43] *State v. Spann*. 1993.

[44] Alfred Tarski. "The Semantic Conception of Truth and the Foundations of Semantics (1944)". In: *Alfred Tarski: Collected Papers* 2 (1986), pp. 661–699.

[45] Alfred Tarski and Jan Tarski. *Introduction to Logic and to the Methodology of the Deductive Sciences*. 24. Oxford University Press, 1994.

[46] Laurence H Tribe. "Trial by Mathematics: Precision and Ritual in the Legal Process". In: *Harv. L. Rev.* 84 (1970), p. 1329.

[47] *Turner v. Louisiana*. 1965.

[48] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer science & business media, 2013.

[49] James Walsh. "The Modeling in Model-Theoretic Skepticism". In: *Preprint* (2021).

[50] Jared Warren. "Infinite Reasoning". In: *Philosophy and Phenomenological Research* 103.2 (2021), pp. 385–407.

[51] Larry Wasserman. *All of Statistics*. Springer Science & Business Media, 2004.

[52] Brian Weatherson. "David Lewis". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University, 2016.