Proof Verification Technology and Elementary Physics

Ernest Davis
Dept. of Computer Science
New York University
New York, NY 10012
davise@cs.nyu.edu

July 7, 2018

Abstract

Software technology that can be used to validate the logical correctness of mathematical proofs has attained a high degree of power and sophistication; extremely difficult and complex mathematical theorems have been verified. This paper discusses the prospects of doing something comparable for elementary physics: what it would mean, the challenges that would have to be overcome; and the potential impact, both practical and theoretical.

Memories of Jonathan Borwein

I knew Jon Borwein only briefly and slightly, but my few interactions were extremely memorable.

I first encountered Jon in connection with a collection of essays on the ontology of mathematics that my late father, Philip Davis, and I were putting together. Jeremy Avigad recommended him to me as a contributor, writing that "he has a lot to say about lots of things", which was certainly true. Jon and David Bailey agreed to write a chapter, and contributed a marvelous essay (Bailey and Borwein, 2015), spanning the world of experimental mathematics from computations of the partition function, to reciprocal series for π, to Ising integrals, to protein structure, to chimera states in oscillator arrays. Being a rather fussy editor, I asked for revisions, and then for more revisions, until on the third go-around Jon informed me, politely but firmly, that this was the final version.

Some time later, Jon generously invited me to present a talk at the 2016 meeting of ACMES. I didn't know how I fit in, since I barely do mathematics at all, and certainly don't do experimental mathematics, but Jon was very encouraging, and I ended up giving a talk which was an early version of the paper below.

The highlight of my visit to ACMES was certainly my dinner with Jon, Judi, and friends that evening. Jon, as his friends know much better than I, was in person an ebullient, larger-than-life character and a wonderful raconteur; the conversation wandered from tales of mathematicians to the cleverness of octopi. It was worth going out to London, Ontario just for that evening.

In the months following, I had a couple of pleasant email exchanges with Jon: one about whether mathematicians worked through the proofs of the theorems they use, one about a historical point — a supposed medieval invention of a random number generator. (It proved to be fictitious.) I very much looked forward, then, to further interactions with him. I wish that I had had the chance to know him much better and much longer.

1 Mathematical proof verification software

One of the major accomplishments of late nineteenth and early twentieth century mathematics was the determination that essentially every rigorous mathematical proof can in principle be fully formalized as symbolic logical inference over set theory. To be precise, there are three statements here:

1. Practically[1] every mathematical concept can be defined in set-theoretic terms; and therefore every mathematical proposition can be formulated as a proposition in set theory.
2. Practically every mathematical proposition that has been rigorously proved, when cast into set theory, can be proved from standard axiomatizations of set theory using first-order logic.
3. Proofs in first-order logic can be characterized purely in terms of rules for manipulating strings of symbols; no understanding of the symbols, or mathematical intuition, or anything of the kind, is required.

The central landmark in establishing these facts was Whitehead and Russell's *Principia Mathematica*, though many other mathematicians, logicians, and philosophers both before and after were involved. The validity of a proof expressed in this symbolic form can be checked by a simple computer program that verifies that the sequence of assertions in the proof conforms to a set of rules for manipulating symbols. The verification program need understand nothing about the content of the proof, and the identical verification program will work for proofs in virtually every subfield of mathematics.

The software instantiation of this logical theory has been the development of *mathematical proof verification systems*. Over the past fifty years, software environment such as Isabelle/HOL (Nipkow, Paulson, and Wenzel, 2002), and others have been developed, which allow a user to formulate symbolic encodings of proofs of mathematical theorems, which the software can then check for correctness. Substantial libraries of basic theorems and lemmas to draw on have been created, and some number of advanced, difficult proofs of major theorems have been formally verified, including:

- The prime number theorem, both using the analytical proof based on the zeta-function (Harrison, 2009) and the "elementary" proof due to Selberg and Erdős (Avigad, Donnelly, Gray, & Raff, 2007).
- The Feit-Thompson theorem that every simple group of odd order is cyclic (Gonthier et al. 2013).
- The Kepler optimal packing theorem (Hales et al. 2015).

More or less, it seems safe to claim that;

- Any proof that is standardly taught in undergraduate math courses either already has been verified with this technology or could be a fairly small amount of work.
- Practically any theorem in the mathematical literature that has been proved could be verified with this technology; however, any given theorem might well require very substantial amounts of expert labor. This obviously does not apply to exceptionally complex proofs, such as

[1] I do not know to what extent the experts agree on which, if any, kinds of theorems lie outside generalizations (1) and (2). As far as I know, there is essentially universal agreement that (1) and (2) are valid across most subfields in mathematics. Whether set theory is the *best* foundation for mathematics, or whether it is *important* for mathematics to have foundations at all, are separate questions.

the categorization of finite simple groups, which presumably would require truly impossible amounts of expert labor, or the proof of Mochizuki's ABC theorem, which, as of the time of writing, is not fully understood by anyone other than Mochizuki himself.

The question I wish to explore in this paper is this:

Can a software technology comparable to mathematical proof technology be constructed that would allow the expression and validation of arguments in elementary physics, particularly those that connect theory and observation?

It will be convenient, for purpose of reference, to give this hypothetical project a name; I will dub it PAVEL.

Disclaimer: this paper is exploratory and discursive; it neither presents established results nor constructs a tight argument. Moreover, my own limitations for carrying out this kind of investigation will soon become all too obvious to the reader; I do not know as much philosophy of science as I should for this purpose, and my knowledge of physics is altogether inadequate. The reason that the discussion in this paper is limited to elementary physics is that that's all the physics I know. (I will briefly discuss more advanced physics in section 3.8.1, relying entirely for my information on (Laughlin and Pines, 2000).) However, to paraphrase Donald Rumsfeld, at 61 years old, one largely does analysis with the knowledge and abilities that you have, and not those that one would like to have.

The paper will proceed as follows. Section 2 will further discuss aspects of formal mathematical proof and of proof verification software further, since those are our primary comparanda and starting points. Section 3, which is the bulk of this paper, discusses the PAVEL project: What it would look like, and what it might accomplish. As part of this discussion, we will set up a straw man as a proposed architecture for PAVEL; the process of knocking down that straw man will help clarify what PAVEL should look like. Section 4 present a formalization of a simple word problem of the kind that might be used in PAVEL. Section 5 review the history of related ideas and proposals. Section 6 discusses possible impact of a successful implementation of PAVEL on the philosophy of science. Section 7 will summarize and will discuss directions forward.

2 Formal proof and proof technology in mathematics

To begin with, let us consider the case of mathematics in more depth. We will discuss briefly the value of the logic-based theory of mathematical proof and of proof-verification technology and their limitations; this will be useful as a point of comparison for discussing the potential value and limitations of pursuing these in the context of physics.

Logic-based analysis of mathematical concepts and proofs provides a normative model for rigorous argumentation in mathematics, which is perfectly well-defined, and which applies to practically every proof throughout the discipline. We will note some limits on the significance of this below; however, those limits do not make this finding any less significant or astonishing.

Moreover, logic-based analysis of mathematics led to the development of mathematical logic, a field that is of enormous inherent interest; provides results important for other areas of mathematics, e.g. the unsolvability of Diophantine equations; and is central to computation theory. The practical consequences throughout computer technology are incalculable.

It is certainly important to keep in mind the limits of logical analysis as a characterization of mathematics. It is presumably of little or no value in developing a *cognitive* theory of mathematical

understanding and reasoning; that is, a psychological theory of how professional mathematicians, lay people, children, or animals understand mathematical concepts and arguments (Dehaene, 1997). In *historical* studies, the twentieth-century logical analysis is treacherous to use as a framework; it can lead one to a "Whig history" point of view in which, let us say, Newton's conception of a point at infinity or Euler's conception of a function is viewed as a defective version of our own perfect understanding. Even as regards contemporary mathematics, it has been argued that the logical sense of proof does not encompass all that we mean by proof, and that the formulation of mathematical concepts in set-theoretic terms does not encompass all that we mean by those concepts (P.J. Davis, 1993). The formal viewpoint omits the social role of proofs; proofs are one form of communication among mathematicians. But, again, these limitations do not negate the enormous importance of this kind of analysis.

Moreover, thus far the impact of either the theory or the technology on the daily labors of the mass of professional mathematicians, working in, say, partial differential equations, or homology theory, or ideal theory, has been much less than the notoriety of mathematical logic and of theorems such as those of Gödel's in popular mathematics and among philosophers of mathematics might suggest. Few, if any, undergraduate math majors at American universities require a course in mathematical logic; and more than one well-regarded math department does not offer *any* regular course in mathematical logic. As for proof verification software, most mathematicians are probably only dimly aware that it exists at all.

The impact of the *technology* of proof verification systems has been enormously less than the *theory* of mathematical logic. Still, it has had a significant impact in certain areas, and may well have greater impact in the future. Perhaps its greatest impact to date is as part of a wide range of activities in implementing logical reasoning on computer systems. This body of work in general has had many practical applications, including logic-based programming languages, automated software and hardware verification, knowledge-based artificial intelligence (AI) reasoners and expert systems. Broadly speaking, these kinds of systems lie along a spectrum, with different trade-offs of the expressivity and depth of the representation, on the one hand, versus efficiency of inference, on the other. Mathematical proof verification lies on the extreme end of favoring expressivity at the expense of efficiency; nonetheless, technical developments here have impact on similar project with more directly practical applications.

In particular, proof verification is closely related to logic-based software and hardware verification. Much more work has been invested in software and hardware verification than in mathematical proof verification because of its direct practical significance. The goal of these kinds of verification system can range from limited verification, determining that the software is free from specific kinds of bugs, to complete verification that the program works correctly in all respects. Bug-checking verification is currently a very powerful technology which can be applied to enormous, complex programs such as operating systems, and complex hardware architectures, such as state-of-the-art CPUs.

Complete verification of software correctness is much more difficult. A major obstacle is that it is extremely hard even to state complete specifications for what a complex program should do; the specification statement ends up being almost as long, and much less intelligible, than the program. Therefore, verification of a formal specification works best for functionalities where the logical specification of the desired functionality is much simpler than its implementation, such as mathematically-oriented software. For example, Harrison (2006) carried out the formal verification of library functions that do floating-point computation of trigonometric functions; the verification raised some interesting subtle issues of correctness beyond what is usually considered in numerical analysis.

In the long term, we can hope to see other kinds of impact on mathematical practice:

- Confidence in highly complex proofs can be increased.
- The development of representation might be a step toward content-based search for theorems in the mathematical literature. Currently, it is often easier to reprove a lemma than to find it in the literature.
- Ultimately, this is a step toward a "general AI mathematician"; an AI that carry out all, or many, of the activities of a research mathematician, either by itself or in partnership with a human.

2.1 What hasn't been done for math

A number of limitations of the technology should be noted.

Obviously, we do not have AI programs that can *generate* proofs of a general kind in advanced math or even in college-level math. The technology for symbolic manipulation, in systems like MAPLE and MATLAB has become extraordinarily sophisticated (Bailey & Borwein, 2015) this will suffice for most proofs in high-school and some fraction of proofs in some areas of math. Beyond that, a handful of interesting original proofs have been generated by computers, either using general theorem-proving technology (e.g. the Robbins conjecture (McCune, 1997)) or using programs specifically written for a particular case (e.g. the four-color theorem (Appel & Haken, 1977).) But we are far from having a program that can generate the kinds of proofs required of undergraduate math majors.

We are nowhere near having an AI program that can read the mathematical literature and "understand" it, in the sense of translating it to a formal representation, or even a program that can do most of this with occasional assistance from a "human in the loop". There has been some work on the much more limited task of translating word problems stated in English into a representation and then solving the equations. For instance Kushman et al. (2014) report a program that achieves an overall accuracy of 68.7% on textbook problems that translate into two equations in two unknowns.

A more immediate issue is user-unfriendliness. By all accounts, the learning curve for this technology is extremely challenging and the user interface uninviting. Consequently, when a new theorem is verified it is much more likely that an expert on verification has learned the math involved in the theorem than a mathematician who is an expert in the area of the theorem has learned to use the verification technology. Verifying the Feit-Thompson theorem involved a six-year collaborative effort by a team of fifteen mathematicians[2] (Gonthier, 2013). If the technology were easy to use, then one could imagine the "mathematician in the street" taking the trouble to master in order to check that their proofs are correct; but currently that seems far off.

2.2 Word problems

Another part of math, particularly elementary math, is word problems.

Let us pass over the large problems of natural language processing and of knowledge base construction and focus on the representational problem: How can the content of a word problem plausibly be expressed in a logical representation that describes the real world situation and that suffices for the solution of the problem, when combined with the relevant mathematical theory? The problem formulation should as far as possible be a direct expression of the *meaning* of the natural language formulation of the problem. That is, we want as much as possible of the reasoning needed to find the solution to be made explicit in the proof structure built on the formulation, and as little

[2]This does not, of course, imply that it required ninety man-years of work.

reasoning as possible done implicitly in the process of translating the natural language expression into the formal problem specification.

Tables 1 and 2 illustrate what I have in mind, for one well-known brain teaser.

Some comments about the formalization in tables 1 and 2. The representation uses a sorted, first-order logic with theories of time, dimensioned quantities and vectors, and Euclidean geometry, that I have developed for representing physical theories (Davis, in prep.) The semantics is straightforward, and the intended meaning is hopefully self-evident. There is a partial account in (Davis, Marcus, and Frazier-Logue, 2017). `Typewriter font` is used for object-level symbols; *Italics* are used for sortal symbols. Non-logical symbols have an initial upper-case letters, object-level variables have an initial lower-case letter, and sortal variables use Greek letters. Sorts of symbols are declared in a form modeled on declarations in typed programming languages such as Java. Thus, for example, the declaration

`VectorFrom(x,y:`*Point*`)` $\rightarrow$ *Vector[Distance]*

means that `VectorFrom` is a function symbol, taking two arguments, `x` and `y`, both of which are *Point*s, and returning a value which is a vector of dimension *Distance*.

The problem formulation in tables 1 and 2 combined with suitable basic axioms and definitions of the dimensions involved, time, and Euclidean space will support a proof of the conclusion `ArcLength(Z,T0,TC) = 150 * Mile.`

The complexity of tables 1 and 2 together with the domain axiomatization not shown here, as compared to the simplicity, both of the natural language expression, and of the mathematical forms that a human reasoner might write down or think through in solving this problem, might be taken as a sign that we are seriously on the wrong track here. In particular the gap between the phrase "the bird flies back and forth between the two trains" and the complex axioms 6 and 7 is concerning. Certainly any human being would find it much easier to solve the problem directly from the natural language formulation than to translate the natural language into the formulas in these tables. (I myself spent some hours getting them right, and I have thirty years' practice in writing these kinds of formalisms.) More than that, one might well worry that it would be easier to write a program that could solve these kinds of problem than to write one that could generate these axiomatizations.

There are a number of partial answers, at different levels. First, the gap from "flies back and forth" to axioms 6 and 7 can be bridged by positing an intermediate form,[3] such as `Until(t,`ϕ`,`ψ`)`, meaning "Starting at time `t`, ϕ remains true at least until ψ becomes true." Axiom 6 can then be worded,

$6'$ $\forall_{\texttt{ta}}$ `T0` $<$ `ta` $<$ `TC` $\wedge$ `V(ta,Place(B)) = V(ta,Place(TrA))` $\Longrightarrow$

```
   Until(ta,Place(B) = Place(TrB),
                 Velocity(B) =
                   Vec*(150 * Mile/ Hour, Direction(VectorFrom(Place(B), Place(TrB))))).
```

and axiom $7'$ would be analogous.

Getting to these intermediate representation $6'$ and $7'$ from "flies back and forth", seems considerably more doable, though certainly not a solved problem; and the process of getting from the intermediate forms $6'$ and $7'$ to axioms 6 and 7 can easily be completely specified.

Second, while a shallower semantic analysis might *often* suffice to build a computer program that solves word problems, in the same way that human students sometimes learn to solve math

[3]Technically speaking, the operator `Until` here can be viewed as "syntactic sugar", or as a temporal modal operator, or, if one performs some "representational tinkering" on the arguments, as a first-order predicate.

Problem: Two trains 100 miles apart are speeding toward one another. One is going 75 mph, the other is going 25 mph. A bird flies back and forth between them at 150 mph. How far does the bird travel before the trains collide?

Sorts: *Object, Time, Duration, Point, Distance, Speed, Real*

Sortal Functions:
Fluent[α] — Function from *Time* to sort α.
Vector[α] — If α is a real-valued dimension, then a vector of dimension α.
For example, *Vector[Speed]* is the sort of velocities.
Vector[Real] is the sort of dimensionless vectors.
$\alpha \otimes \beta$ — Infix operator: Dimension α times dimension β.
For example, *Duration* $\otimes$ *Speed* = *Distance*.
$\alpha \oslash \beta$ — Dimension α divided by dimension β.
For example, *Distance* $\oslash$ *Duration* = *Speed*

Constant Symbols:
`TrA`→*Object* – the first train.
`TrB`→*Object* – the second train.
`B`→*Object* – the bird.
`T0`→*Time* – the initial time.
`TC`→*Time* – the time the two trains collide.
`Mile`→*Distance* – a mile
`Hour`→*Duration* – an hour
Standard numerals → *Real*.

Function Symbols:
`Place(x:`*Object*`)` → *Fluent[Point]*. The function tracking the position of object `x` over time.
`Velocity(x:`*Object*`)` → *Fluent[Speed]*. The function tracking the velocity of object `x` over time.
`Magnitude(v:`*Vector[α]*`)` → α. Magnitude of vector `v`. $|\vec{v}|$.
`Direction(v:`*Vector[α]*`)` → *Vector[Real]*. Direction of `v`. $\vec{v}/|\vec{v}|$.
`V(t:`*Time*`, q:`*Fluent[α]*`)` → α. Value of fluent `q` at time `t`.
`VectorFrom(x,y:`*Point*`)` → *Vector[Distance]*. The vector `y` − `x`.
`Vec*(s:`α`, v:`*Vector[β]*`)` → *Vector[*$\alpha \otimes \beta$*]*.
Scalar `s` of dimension α times vector `v` of dimension β.
`x:`α `* y:`$\beta \rightarrow \alpha \otimes \beta$.
Infix operator `x` $*$ `y` where `x` has dimension α and `y` has dimension β.
`x:`α `/ y:`$\beta \rightarrow \alpha \oslash \beta$.
Infix operator `x/y` where `x` has dimension α and `y` has dimension β.

Table 1: Formalization of a word problem: Sorts and Symbols

Problem Statement:

1. `Magnitude(VectorFrom(V(T0,Place(TrA)),V(T0,Place(TrB)))) = 100*Mile`.
   The two trains are initially 100 miles apart.

2. `V(TC,Place(TrA)) = V(TC,Place(TrB))`
   The two trains collide at time `TC`.

3. $\forall_{\texttt{t}}$ `T0` $<$ `t` $<$ `TC` $\implies$
   `V(t,Velocity(TrA)) =`
   `Vec*(25 * Mile/Hour, Direction(VectorFrom(V(t,Place(TrA)),V(t,Place(TrB)))).`

   Between `T0` and the collision, train `TrA` moves at 25 mph toward train `TrB`.

4. $\forall_{\texttt{t}}$ `T0` $<$ `t` $<$ `TC` $\implies$
   `V(t,Velocity(TrB)) =`
   `Vec*(75 * Mile/ Hour, Direction(VectorFrom(V(t,Place(TrB)),V(t,Place(TrA)))).`

   Between `T0` and the collision, train `TrB` moves at 75 mph toward train `TrA`.

5. `V(T0,Place(B)) = V(T0,Place(TrA))`.
   The bird starts at train `TrA`.

6. $\forall_{\texttt{ta,tb}}$ `T0` $<$ `ta` $<$ `TC` $\wedge$ `V(ta,Place(B)) = V(ta,Place(TrA))` $\wedge$ `ta` $<$ `tb` $<$ `TC` $\wedge$
   $[\forall_{\texttt{tx}}$ `ta` $<$ `tx` $\leq$ `tb` $\implies$ `V(tx,Place(B))` $\neq$ `V(tx,Place(TrB))`$]$ $\implies$
   `V(tb,Velocity(B)) =`
   `Vec*(150 * Mile/ Hour,`
   `Direction(VectorFrom(V(tb,Place(B)), V(tb,Place(TrB)))).`
   If the bird is at train `TrA` at time `ta`, and it does not reach train `TrB` any time between `ta` and `tb` inclusive, then at time `tb` it is moving toward `TrB` at 150 mph.

7. $\forall_{\texttt{ta,tb}}$ `T0` $<$ `ta` $<$ `TC` $\wedge$ `V(ta,Place(B)) = V(ta,Place(TrB))` $\wedge$ `ta` $<$ `tb` $<$ `TC` $\wedge$
   $[\forall_{\texttt{tx}}$ `ta` $<$ `tx` $\leq$ `tb` $\implies$ `V(tx,Place(B))` $\neq$ `V(tx,Place(TrA))`$]$ $\implies$
   `V(tb,Velocity(B)) =`
   `Vec*(150 * Mile/ Hour,`
   `Direction(VectorFrom(V(tb,Place(B)), V(tb,Place(TrA)))).`
   If the bird is at train `TrB` at time `ta`, and it does not reach train `TrA` any time between `ta` and `tb` inclusive, then at time `tb` it is moving toward `TrA` at 150 mph.

Evaluate: `ArcLength(T0,TC,Place(Z))`.

Table 2: Formalization of a word problem: Problem Formulation

problems by pattern matching against problems that they have seen before, I would argue that solving these problems *robustly* will require a semantic representation of the depth of tables 1 and 2. For instance, to answer the particular question "How far will the bird fly?", a computer does not actually have to understand what is meant by "back and forth" at all; it suffices to understand that the bird is flying at 150 mph. However, that will not suffice if you change the problem statement or the question:

- How many times is the bird exactly 10 miles from one or the other train?"
- Is there any time at which the distance from the bird to the first train and the distance to the second train are both simultaneously decreasing?
- Suppose that whenever the bird reaches a train, it rests for a minute. How far does it fly in that case?

For any of these, you will need a level of understanding comparable to tables 1 and 2

The objection that people find it easier to solve the problem than to work through the notation of tables 1 and 2, though often raised as a derisive dismissal of logic-based notations, really has no weight at all. Working through any description of how a cognitive task is carried out is almost always more difficult than performing the task. I can guarantee that if somebody builds a system based on machine learning that solves the bird problem, that will also be harder to understand than solving the bird problem.

In general, what is the state of the art in representing math word problems in this way? I don't know of any systematic study; it would be interesting to carry one out. But my guess would be that problems in high school level or freshman college level math – that is, elementary problems in Euclidean geometry and trigonometry, basic algebra, differential and integral calculus through the first three college courses, and combinatorics — would rarely if ever present difficulties.

Probability theory might often be challenging. The Kolmogorov formulation of probability theory suffices for all formal mathematical theorems in probability theory (as far as I know); if you want to prove the central limit theorem, say, or the existence of limiting distribution for a Markov chain, you can state it and prove it within the Kolmogorov formulation. Likewise, if a word problem can be easily cast in terms of a sample space, then it can be represented and solved. For instance, if we wish to answer the question, "What is the probability that a five-card hand is a flush (including straight flush)?", then it is straightforward to axiomatize the combinatorics and prove that $\#\texttt{Hand} = C(52,5)$, $\#\texttt{Flush} = 4 \cdot C(13,5)$ and therefore $\texttt{Prob(Flush|Hand)} = 4 \cdot C(13,5)/C(52,5) = 0.00198$

However, in many cases the derivation is much more problematic. Consider the following well-known puzzle:[4]

A. John has two children and at least one of them is a boy. What is the probability that he has two sons? **Answer:** 1/3.

B. John has two children; the older is a boy. What is the probability that John has two sons? **Answer:** 1/2.

C. John has two children; at least one is a boy born on Tuesday. What is the probability that John has two sons? **Answer:** 13/27.

I'm ignoring here the slight correlation in days of birth due to twins, the even slighter correlation in sex due to identical twins, and the fact that male births are not exactly 50% of all births.

[4]The "Monty Hall" problem is even trickier, and has tripped up professional mathematicians.

(Peter Winkler (email to the author, 12/28/17) has pointed out that almost any real world situation where you know that John has 2 children and one is a boy — for instance, if you are told that he has two children, and then you run into him with one child, who is a boy — conforms to the analysis in (A) or (C) rather than the one in (B). However, he reports running into one real-world exception: A friend of his was pregnant with fraternal twins, and had some kind of genetic test that gives positive results if either fetus has a Y chromosome. In that case, the analysis in (A) held; there was a 1/3 chance that she was bearing two boys.)

If you consider $\texttt{Prob}(\phi|\psi)$ to be a sentential operator then the probabilities to be evaluated are easily expressed:

A. $\texttt{Prob}(\#\{\texttt{x}|\texttt{Child}(\texttt{x},\texttt{John}) \wedge \texttt{Sex}(\texttt{x}) = \texttt{Male}\} = 2 \mid$
$\exists_{\texttt{y},\texttt{z}}\{\texttt{x}|\texttt{Child}(\texttt{x},\texttt{John})\} = \{\texttt{y},\texttt{z}\} \wedge \texttt{y} \neq \texttt{z} \wedge \texttt{Male}(\texttt{y}))$

B. $\texttt{Prob}(\#\{\texttt{x}|\texttt{Child}(\texttt{x},\texttt{John}) \wedge \texttt{Sex}(\texttt{x}) = \texttt{Male}\} = 2 \mid$
$\exists_{\texttt{y},\texttt{z}}\{\texttt{x}|\texttt{Child}(\texttt{x},\texttt{John})\} = \{\texttt{y},\texttt{z}\} \wedge \texttt{y} \neq \texttt{z} \wedge \texttt{Male}(\texttt{y}) \wedge \texttt{Older}(\texttt{y},\texttt{z}))$.

C. $\texttt{Prob}(\#\{\texttt{x}|\texttt{Child}(\texttt{x},\texttt{John}) \wedge \texttt{Sex}(\texttt{x}) = \texttt{Male}\} = 2 \mid$
$\exists_{\texttt{y},\texttt{z}}\{\texttt{x}|\texttt{Child}(\texttt{x},\texttt{John})\} = \{\texttt{y},\texttt{z}\} \wedge \texttt{y} \neq \texttt{z} \wedge \texttt{Male}(\texttt{y}) \wedge \texttt{Born}(\texttt{y},\texttt{Tuesday}))$

But I don't know of any logical formalization which will allow one to go from forms like the above to stochastic models in which the specified probabilities can be calculated.

Furthermore, stochastic models whose complexity seems quite moderate when presented in an applied probability textbook, such as the k-gram model of language production, end up being much more intricate when written out in full in a logical notation. The elegant mathematical formulas used to describe such models in the research literature often turn out, on careful analysis, to be a morass of implicit quantifiers of implicit scope and ambiguous variable symbols, superscripts, and subscripts, meaningful only to someone who reads the accompanying text and understands what is intended.

Mathematically, statistics is largely a subfield of probability, but it seems to gravitate toward that class of probability problems that are particularly difficult to formulate logically. I suspect that many word problems in statistics would be extremely difficult to represent in a reasonable way that supports the statistical inference.

3 Physics

With the example of mathematics in mind, as inspiration and point of comparison, we can now enter on the main topic of the question. Vaguely put, can we carry out this same kind of project for physics? More specifically, can we achieve the following:

- Represent some significant part of the content of physics, including both foundational theories and the experimental and observational results that they rest on, in a formal language?
- Characterize some significant part of reasoning and argumentation in physics, particularly the reasoning that connects foundational theories to "real world" situations, in a formal theory of reasoning?
- Implement the representation and reasoning mechanisms in a technology for argument verification for physics?

3.1 The potential value of this undertaking

If PAVEL can be built, then it seems to me that both the finished product and the work involved in developing the product are likely to have significant payoffs, in a number of different directions.

First, the work involved in PAVEL might shed some light on issues in the philosophy of science. That will be easier to discuss after we have looked at specific issues, so I am deferring it to section 6.

Second, work on PAVEL would be a step toward in developing AI that can do flexible, powerful commonsense physical reasoning. Gary Marcus and I have argued at length elsewhere (Davis and Marcus, 2014, 2016) that approaches to physical reasoning based on simulation, which currently entirely dominate AI physical reasoning, are insufficient for many of the kinds of problems that a general purpose AI will confront, besides being implausible as general cognitive models. It is certainly the case that physicists, in reasoning about physical situations, use a wide variety of reasoning techniques beyond simulation. It seems likely, therefore, that analyzing the kinds of reasoning needed to do physics may open up the space of automated reasoning techniques available to AI reasoners.

Third, it may be possible to integrate the reasoning in PAVEL with program verification technology, and thus to formally verify the validity of programs that control physical devices, in safety-critical ways: airplanes, robots, nuclear reactors and so on. A major accomplishment in program verification, some years ago (Souyris et al. 2009) was the verification of the control software for the Airbus airplane. However, that verification only proved that the *program* won't crash; it didn't prove that the *airplane* won't crash. In work closer to PAVEL, Jeannin et al. (2015) formally verified a hybrid system for the avoidance of aircract collision. Their domain axiomatization is similar in flavor to the axiomatizations we develop in this paper, but are quite specialized to the problem under discussion.

Finally, PAVEL would be a step toward the "super-AI-scientist" fantasized by the many "AI as messiah" enthusiasts; an AI that can achieve an integrated, total, understanding of *all* of science and thus can solve those of our problems that can be solved that way. In fact, it seems to me that solving the issues involved in PAVEL is a *necessary* step; the super-AI-scientist *must* have the kind of general understanding that is encoded in PAVEL.

Paleo (2012) similarly argue in favor of expressing arguments in physics in proof-theoretic terms, arguing that this will clarify existing debates in the philosophy of science and "open new conceptual bridges between the disciplines of Physics and Computer Science."

3.2 The Bayesian formulation

In thinking about PAVEL, I find it helpful to keep in mind the Bayesian approach to scientific hypothesis and data (Jaynes, 2003); (Rosenkrantz, 1977), (Howson & Urbach, 2006) (Strevens, 2005), partly as a framework to make things concrete, partly as a foil to work against.

The basic Bayesian formulation of scientific theorizing is straightforward. There is a space Φ of possible scientific theories; that is, each hypothesis $h \in \Phi$ is a complete theory of physics. There is a space Δ of possible total data collections; that is, each element $D \in \Delta$ is a combined record of all the outcomes of all the experiments and observations ever performed. We are given one particular collection of data $D \in \Delta$. We are looking for the most likely theory given the data; that is, $\text{argmax}_{h\in\Phi} P(h|D)$. So now, as always, we use Bayes' Law:

$$\text{argmax}_{h\in\Phi} P(h|D) = \text{argmax}_{h\in\Phi} P(D|h)P(h)$$

All that's left is to set the priors $P(h)$, to compute the conditional probabilities $P(D|h)$, and

to find the maximum of the expression. Within reason, the exact values of the priors don't matter much anyway, since their contribution is soon swamped by the data. That is, you think of each imaginable theory of physics as a generative stochastic process that outputs data, and thus defines a probability distribution $P(\cdot|h)$ over Δ. You imagine a prior distribution over all such processes. Then you match the observed data to the predicted data.

One thing that's appealing about this is that it completely eliminates the need for scientific induction as a separate mode of reasoning. There is no need to address the difficult question of what it means for data to support a hypothesis; Bayes' law allows you to turn that into the much more straightforward question of whether a hypothesis predicts data.

The hypotheses in Φ must all be mutually exclusive or the method doesn't work. They cannot be theories in the logical sense, organized in a lattice of generality, because the probability of a more general theory is necessarily less than a narrow theory. Given any premise or data, the conditional probability of $\forall_x B(x) \implies A(x)$ is cannot be greater than the conditional probability of $\forall_x B(x) \wedge C(x) \implies A(x)$, because the first sentence implies the second. If, therefore, Φ included more and less general theories, the maximum would never land on the most general theories; those are always the least probable. In Bayesian models, therefore, all the hypotheses are maximally specific. For example, in the Hierarchical Bayesian Models theory (Tenenbaum et al., 2011), all the theories in Φ are generative stochastic models that generate data. The choice therefore, is not between "$\forall_x B(x) \implies A(x)$" and "$\forall_x B(x) \wedge C(x) \implies A(x)$". Rather the choice is between

> H1(p): $\forall_x B(x) \implies A(x)$ and A occurs randomly with probability p among entities that are not B;
>
> vs.
>
> H2(p): $\forall_x B(x) \wedge C(x) \implies A(x)$ and A occurs randomly with probability p among entities that are not both B and C;

Here p is a parameter that will be optimized (viz. set to the measured frequency of A in the two referent sets). Since H1 no longer implies H2, there is now nothing to prevent us from assigning a higher prior probability to H1 than to H2.

As is well known, Bayesian theories are equivalent to minimum description length theories under the information-theoretic correspondence $I(\phi) = -\log_2 P(\phi)$. That is: you choose an optimal encoding for hypotheses based on their prior probabilities, or, conversely, you set the prior probability to be exponential in the length of the theory: $P(h) = 2^{-I(h)}$ where $I(h)$ is the number of bits needed to express h. For each hypothesis h, you choose an optimal encoding for possible data outputs where $I(D|h) = -\log_2 P(D|h)$. So overall you have attained an expression of length $I(D) = I(D|h) + I(h)$. Choosing the most probably value of h given D is then equivalent to using Occam's razor to choose the shortest expression of the data; that is, we find the simplest, most elegant theory that explains the data.

In some ways, this seems enormously appealing, almost inevitable; in other ways it seems completely far-fetched ((Sober, 2002) is a sharp critique.) The idea that there exists a space Φ of fully formed physical theories prior to making any observations and the idea that there is a space Δ of possible data collections that exists independent of the physical theory — all the theory does is to change the conditional probability distribution over Δ — do not correspond to our experience of how science actually progresses. What we see, rather, is a tight mutual dependence between theory and data. On the one hand, in the development of science, the data collected thus far affects, not just the choice of theories, but even the language that the theories are expressed in. On the other hand, the choice of which experiments to carry out or observations to make depends on what is known about the physics. As we will discuss further in section 3.5, an experimental device or design and the interpretation of its behavior as data depend critically on knowledge of physics; if

the physics of world were otherwise, then the experiment would be not merely inconclusive, it would be meaningless or impossible.

A Bayesian might justify the spaces Φ and Δ with the following *Gedanken* experiment. Let us fix the scientific investigator under discussion: perhaps a new born baby (Gopnik, 2012), perhaps a scientific community over millennia. Imagine now the collection Ω of all epistemically possible physical worlds; or, at least, all those consistent with the existence of a baby/scientific community. (This is somewhat similar to Tegmark's (2009) Level IV multiverse.) We insert a clone of the investigator into each possible world. The investigator's task in each world is to find out which world he is in; or at least to get some information about that. We now, from the outside, observe all these investigators in all these worlds. At a certain point, we stop him; we find out what data he have seen and we ask him what physical theory he now believes, or what set of alternative theories he has under consideration. For each world $w \in \Omega$, let D_w be the collection of data that the investigator has compiled in w and let Φ_w be the set of alternative theories that the investigator in w reports. Then $\Delta = \{D_w | w \in \Omega\}$ and $\Phi = \bigcup_{w \in \Omega} \Phi_w$.

The fact that, in different worlds, the investigator will perform different experiments and make different observations is merely the standard scenario in decision theory in which the space of possible actions may depend on prior observations. It slightly complicates Bayesian inference, but does not fundamentally alter it.

The reason that this view seems alien (the Bayesian can continue) is that, due to our own cognitive limitations, we are not used to taking such a large view; we are used to looking at the development of science through a much narrower window. However, fundamentally, behind the scenes, this is what is going on. In fact the ultimate AI scientist will be able to take exactly this view of things; it will take into account *all* of the scientific data D that has been collected and chose the best among *all* possible scientific theories $h \in \Phi$, up to limits of computational power.

The transformation of a theory of physics — that is, a collection of physical laws — into a stochastic model elicits starkly varying reactions from different people. To a Bayesian, this is natural, indeed inevitable; trying to do inference without a distribution is like trying to bake a cake without an oven. To a logicist, burdening an elegant, well-motivated logical theory with an ugly, arbitrary probability distribution is adding an unnecessary excrescence; it is like trying to bake a cake with a blowtorch. As we will discuss in section 3.5, the relation between theory and a scientific theory, in general will carve out a strangely shaped, lower-dimensional manifold[5] in the space Δ of all data collections; and defining a natural distribution over such a manifold is a problematic and ill-defined undertaking. It is hard enough to characterize the sense in which the observations of the tides, for example, can be explained in terms of Newton's law of gravity. The question, "What is the probability distribution over observations of the tides, given Newton's law?" seems a truly strange one.

We will not pursue this argumentation back and forth further here; however, in the course of our discussion, we will refer back to this as a possible frame of reference. An implementation of this approach by Kemp and Tenenbaum (2009) will be discussed in section 5.4.2.

3.3 Straw man: The tee-shirt model of PAVEL

At this point I want to put up a straw man proposal for an approach to building PAVEL; the process of knocking it down will serve as an effective frame for making the points I want to make.

The straw man is this: We express the famous laws of physics in a formal logic. These are the

[5]You may argue that because of noise, the theory does not correspond to a lower-dimensional manifold, it corresponds to a probability distribution centered on the manifold. That hardly helps, because now the probability distribution of the projection onto the manifold depends strongly on largely arbitrary assumptions about the noise.

axioms of our system. Everything else is proved from those axioms. In our Bayesian formulation, this collection of axioms is the hypothesis h.

I call this "the tee-shirt model", because tee-shirts printed with a few elegant equations are popular among the geekier part of the population. Full disclosure: As an undergraduate I owned and wore a Maxwell's equations sweatshirt. Less snarkily, I will also call this approach "the foundational approach" when that is more appropriate.

Now, the tee-shirt model is *exactly* the equivalent of what is done in mathematical proof verification systems. The basic axioms given are the ZFC axioms of set theory (or some other similar foundational set); everything else in math is defined in terms of sets and all proofs can ultimately be traced back to the foundational axioms. At the other extreme, it is hard to imagine that anyone would propose anything like the tee-shirt model for chemistry or biology, let alone for the cognitive or social sciences, with the possible exception of economics. But physics occupies a middle ground here, and it seems as though the tee-shirt model should be more or less attainable. I will argue that, at least in our current state of understanding, the tee-shirt model is nowhere close to right, for quite a number of reasons.

3.4 The equations are more complicated than their tee-shirt version

To begin with a rather minor point: the actual equations of physics are often more complicated than than they appear on tee shirts.[6]

To take a simple example: On the tee-shirts Newton's theory of universal gravitation might well be given in two equations;

$$F = G\frac{m_i m_j}{r^2} \qquad \text{Universal law of gravitation}$$

$$F = m\frac{d^2x}{dt^2} \qquad \text{Newton's 2nd law}$$

But actually, a force is a vector with a direction, and Newton's second law applies to the vector sum of all the forces incident on a particle. Forces and positions are functions of time. We need to exclude forces by a particle on itself. So for point particles, the equations become

$$i \neq j \implies \vec{F}_{i,j}(t) = G\frac{m_i m_j \cdot \hat{\theta}(\vec{x}_j(t) - \vec{x}_i(t))}{|\vec{x}_j(t) - \vec{x}_i(t)|^2}$$

$$m_i \frac{d^2\vec{x}_i(t)}{dt^2} = \sum_{j \neq i} \vec{F}_{i,j}(t)$$

The indices i, j range over particles. We use $\hat{\theta}(\vec{v})$ to mean the direction of vector $\vec{v}$: $\hat{\theta}(\vec{v}) = \vec{v}/|\vec{v}|$.

If we want to have extended objects, then things become still more complicated. We can develop a theory of eternal extended objects constructed from particle by introducing a predicate $c(p_i, p_j)$, meaning "particle p_i is connected to particle p_j." The object is then the set of particles within the transitive closure of the relation c.

[6]The Lagrangian for the Standard Model, given in full in (Gutierrez, 1999), is 36 lines long and has something like 170 terms and 1000 symbols. However, Gutierrez does claim that he has printed tee shirts with the whole thing.

For a rigid object, ignoring contact forces between the objects — that is, allowing objects to freely interpenetrate — we get the following rules:

$$c(p_i, p_j) \implies c(p_j, p_i)$$

$$c(p_i, p_j) \implies |x_j(t) - x_i(t)| = d_{i,j}$$

$$\vec{F}_{i,j}(t) = -\vec{F}_{j,i}(t)$$

$$\neg c(p_i, p_j) \implies \vec{F}_{i,j}(t) = G\frac{m_i m_j \cdot \hat{\theta}(\vec{x}_j(t) - \vec{x}_i(t))}{|\vec{x}_j(t) - \vec{x}_i(t)|^2}$$

$$m_i \frac{d^2\vec{x}_i(t)}{dt^2} = \sum_{j \neq i} \vec{F}_{i,j}(t)$$

The second equation above expresses the rigidity constraints by requiring the distance between connected particles to be constant. The third equation is Newton's third law.

For elastic objects, the second equation above, characterizing the constraint between connected particles, is replaced by Hooke's law:

$$c(p_i, p_j) \implies \vec{F}_{i,j}(t) = k_{i,j}(|\vec{x}_j(t) - \vec{x}_i(t)| - d_{i,j}) \cdot \hat{\theta}(\vec{x}_j(t) - \vec{x}_i(t))$$

The formulation for continuum mechanics is similar, but replaces the force by force density, the relation between connected particles by the corresponding partial differential equations, and the summation by an integral.

These don't have quite the same panache on a tee shirt. This observation does not refute the possibility of using a foundational model to build Pavel, but it does suggest that formulating the the foundational equations correctly may take more care than one might suppose.

3.5 The Grounding of Physics in Observation and Experiment

The most serious objection to the tee-shirt model is it ignores the problem of expressing the connection between the terms in the equations and the ways that these are manifested in the world that a physicist interacts with.

A hypothetical student who merely knows the above equations and has worked through their mathematical consequences can hardly be said to have an adequate understanding of gravity. She additionally needs to understand the consequences of these equations in the observable world; *how* they explain falling objects in the everyday setting; the weight of objects, as perceived and as measured on scales of various designs; the motion of planets in the solar system; the tides; and so on.

None of these observations in itself validates the entire Newtonian theory of universal gravity; each corresponds to part of the theory, with some degree of indirectness. Measuring the time that an object takes to fall various distances gives indirect information about the acceleration, but none about the forces, the masses, or the distance to the center of the earth. Feeling the weight of an

object being held gives fairly direct but very imprecise information about the force of gravity on the object (what you are directly experiencing is the normal force of the object on your hand). Using a spring scale gives indirect information about the weight of the object, in the form of the height of an indicator, mediated by the compression of a spring. Using a balance scale gives information about the weight of the object being weighed as compared to standard weights mediated by the law of the lever. For the observations of the planets, which were the major source for the theory of gravity, the data was a record, over time, of the direction from the earth to the planets, the earth itself, of course, being a platform with a complex movement. It took the combined genius of Kepler and Newton to show how these measurements related to the equations of gravity, and even so, the astronomical observations did not give any information about the absolute distance of the planets. Until the observations became precise enough for the effect of one planet on another to be measurable, they likewise gave no information about the relative masses of the planets. The tides, correctly explained, are an effect of the spatial derivative of the gravitational force of the moon, as reflected, though a complex mechanism, in a twice-daily rising and falling of sea level at every sea-coast location.

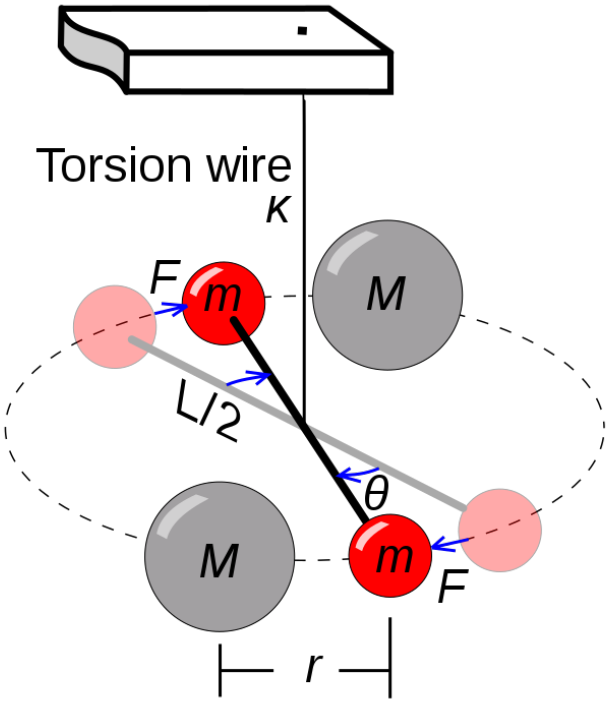
As experiments and theories become more complex, the relation between the observations and the theory generally become more indirect, at least in some respects. Cavendish's experiment (figure 1) to determine the gravitational constant (from his point of view, to determine the mass of the earth), for the first time succeeded in creating a setting in which the masses and the distances could all be directly measured. But the measurement of the minuscule gravitational force created ($1.74 \cdot 10^{-7}$ Newton) is quite indirect: The torsion coefficient for the wire is determined by timing the oscillation period for the small balls twisting back and forth on the wire; the force needed to twist the wire is then calculated from the small angular deflection created.

The deeper the science, the more indirect the experimental evidence. The relation between Schödinger's equation and the experiments that support it are very indirect. You need to know a lot of physics to understand how gravitational wave detectors work or how the Higgs boson was detected.

At the other end of the spectrum, we have been speaking of measuring distances and time as though those were atomic percepts. But measurements rely on measuring devices; measuring small distances requires an accurately calibrated ruler, and measuring durations of time requires a clock. Designing high quality rulers (see for instance Berger, 2010 pp. 116-120) or clocks requires some physics and some engineering. (The foundations of theories of measurement is analyzed in (Suppes, Luce, Kranz & Tversky, 1974).)

Moreover, measurements are taken separately, and the experimenter assumes that they remain [close to] constant from one stage of the experiment to the next. In the Cavendish experiment, you first measures the torsion coefficient using oscillation; and then you assume that that same coefficient is valid when you are measuring the gravitational force between the balls. You first weigh the balls on a scales and then place them in the apparatus. We are thus drawing on a basic commonsense understanding of world in reasoning about the experiment, but we also know that that the commonsense view is insufficient.

Therefore, in Pavel's encoding of the relation of Cavendish's experiment to the law of universal gravitation, the statement of the law of gravity is only a small part of the physics knowledge that you need, and the final actual measurements — the masses of the objects, the length of the rod, the oscillation period, and the displacement of the balls — are only a very small part of the description of the situation. Most of the knowledge of physics — the relevant part of h, in our Bayesian formulation — has to do with the properties of parts of the apparatus: most obviously, that the wire will exert a force against twisting proportional to the angle of twist, but also that the rod remains (reasonably) straight, that the masses of the balls remain (close to) constant in between being weighed and being placed in the apparatus. Almost all of the data — the relevant part of D in the Bayesian formulation



Drawing by Chris Burks. From the Wikipedia article, "Cavendish Experiment".

Figure 1: Cavendish's experiment

— is a description of the design of the apparatus and of the procedure followed. The representation of the procedure must, at least implicitly, characterize all the things you *didn't* do in the course of the experiment: you didn't cut the rod shorter after measuring it or chop a chunk out of the balls after weighing them.

Moreover, a *full* description of the experiment should in principle include a description of the measurement apparatus and how it is used. The oscillation period was 20 minutes; but what kind of clock did you use? The small balls weigh 1.61 pounds; but what kind of scales did you use? Life being finite, the regress here cannot be infinite; and it would seem to bottom out, partly in systems of circular support (e.g. two independent rulers or clocks confirm one another), partly in direct perception (e.g. the ticks on the ruler look equally spaced), partly in some physical assumptions in the reasoning system that are made and not justified (e.g. that the masses do not change between being weighed and being put in the experimental set-up); and, at the individual level, in trust in the scientific community.

This last issue of trust is a major epistemic difference between mathematics and physics. In principle, a mathematician can check the proof of every theorem she is using; in practice, mathematicians do work through the proofs of many of the basic results in their area, and, even in our time, some mathematicians are known for their care in checking the proofs of the theorems they use. (MathOverflow, 2016). By contrast, a physicist must trust both that the suppliers of scientific equipment are not sabotaging her lab by sending her defective instruments and equipment, and that other physicists are accurately reporting their experimental results. Even in principle, a scientist cannot rerun all the experiments that underlie her theory. Some require unique equipment (the Hubble telescope, the CERN accelerator); others, such as astronomical observations, can only be made from particular locations at particular times (or must be made at multiple locations simultaneously). In mathematics, the communal aspect is important (Martin and Pease, 2015); in physics and the other sciences, it is inescapable.[7]

To calculate the mass of the earth, Cavendish additionally needed to know the radius of the earth, which, at least in Cavendish's time, in turn was based on all kinds of *geographic* knowledge — knowing the north-south distance between two cities and comparing the angle of shadows at noon on the same day, and such. (The radius of the earth is also one important starting point for much of the knowledge of astronomical distances.) One doesn't necessarily think of pacing out the distance from Cyrene to Alexandria as a physics experiment, but these measurements certainly have implications for physics, and they are all part of the data D in our Bayesian formulation.

From the standpoint of the foundational approach, all this information consists of rules for translating human-scale realities into boundary conditions. That seems like an strange characterization, but, in the foundational approach, there is nothing else that it can be, as far as I can see. There are the differential equations, which are the foundational dynamics laws, and then there are the boundary conditions, and there is no room for anything else to enter in.

3.6 Is the complexity of grounding different in physics than math?

I have argued that, in our reasoning system, the fundamental laws can only be a small part of the content. One might respond that an analogous situation holds with mathematics. Only a very small part of the mathematical knowledge of a mathematical proof verifier consists of the base ZFC axioms of set theory; most of the content is the definition of more complex mathematical concepts — the real numbers, the Gamma function, the regular dodecahedron, the class of NP-complete

[7]Large levels of trust are needed in any such enterprise. That is why conspiracy theorists, who are willing to distrust any evidence that runs against their theory, are so crazy and so unanswerable; and why any violation of trust — by scientists, by technologists, by the media — is so damaging, not just to the specific instance, but to the entire scientific/technological enterprise.

languages, Lie algebras, and so on — as set-theoretic constructions. Similarly, we could start with the foundational elements of physics, define things like the Cavendish experiment as a construction over the foundational elements, and then prove the behavior of the experiment from the foundational laws.

In principle, this is presumably possible; in fact, as we will discuss in section 3.7, it is an important principle of physics that in principle this is possible. In practice, however, it is so far from being possible as to be not worth discussing. In mathematics the reduction to set theory is reasonably straightforward; any mathematician could work out the set-theoretic definition of the Gamma function and the rest of them, perhaps occasionally looking up some forgotten definition in Wikipedia or MathWorld. By contrast, characterizing the internal structure of the wire in Cavendish's experiment in terms of the atomic structure of its material, and proving that when twisted it exerts a restoring force proportional to the angle (rather than, for example, breaking, deforming, disintegrating, exerting a negative force, or exerting a force that is non-linear in the angle, within the angle range under discussion) are extremely difficult. We will discuss this issue of argumentation further in section 3.8.

In mathematics, one sometimes gets out of these difficulties by positing the properties that you want; define a "Cavendish wire" to be one that, on twisting, exerts a restorative force proportional to the angle of twist, and then define the Cavendish experiment as using a Cavendish wire. But in this context, that doesn't help; we now have to prove that there exist Cavendish wires, and that the wire that was actually used in the experiment is a Cavendish wire.

3.7 Claims to universality

A distinguishing feature of physics, as compared to other disciplines, is that it makes claims to universality of a certain kind. Specifically, physics makes one very general universal claim, which I will get to, but it also makes a number of more limited, but still very broad claims. Let me discuss a few, in increasing order of generality.

Historically, perhaps the first important finding of this kind was Laplace's successful explanation of all the motions of the planets then known in terms of Newton's law, which he published in his five-volume opus *Méchanique Céleste* (1799-1825). (The precession of the perihelion of Mercury, which requires general relativity, was reported by Le Verrier in 1859.)

Second: In chapter 1 of his *Lectures on Physics*, Richard Feynman (1964) wrote,

> If, in some cataclysm, all of scientific knowledge were to be destroyed and only one sentence passed on to the next generation, what statement would contain the most information in the fewest words? I believe that it is the *atomic hypothesis* ... that *all things are made of atoms* — little particles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another.

As further confirmation of the centrality of the atomic hypothesis, we may note that the reality of atoms was a matter of fierce debate in the late nineteenth century and the first two decades of the twentieth, with Mach and others vehemently arguing that they were just a theoretical construct. The establishment of the physical reality of atoms, by Jean Perrin, Einstein, and others, was one of the major accomplishments of the early part of the twentieth century (less well known than relativity or quantum theory, because it was the consolidation of an established doctrine rather than a revolutionary new one).

Though fundamental, atoms are not on the tee-shirt; you will not get rich selling tee-shirts reading "All things are made of atoms". They are also not foundational, in the current view of things;

an atom is the lowest energy state solution to the quantum electrodynamical equation describing a system with k electrons orbiting a nucleus with k protons. Atoms are not universal; there are no atoms in neutron stars. What Feynman's rather vague "all things" means is, presumably, "all matter within the terrestrial setting".

The fact that the atoms are fundamental but not foundational is not, in itself, an argument against grounding our reasoning system in foundational theories. One might say the same of the construction of real numbers from set theory. Real numbers predate infinite sets, certainly historically, almost certainly cognitively; and I know many mathematicians who, faced with Feynman's hypothetical cataclysm, would much prefer that mankind remember the reals rather than remember ZFC.

A third universalizing statement seems to me important, though difficult to state precisely. (This is discussed, in a somewhat more limited form, in (Laughlin and Pines, 2000).) The claim is more or less this: Taking the influx of radiation from outside earth to be an exogenous boundary condition, practically all physical events and physical properties of things that people encounter on earth are consequences of the earth's gravity together with non-relativistic quantum mechanics (Schrödinger's equation) applied to the electromagnetic interactions of atomic nuclei and electrons. There are some number of exceptions — the tides, the occasional meteor, radioactive decay, the things that happen inside sophisticated physics experiments — but those are largely known, and otherwise it is a very reliable rule. That is, if you make some physical observation or encounter a physical phenomenon, whether in meteorology, earth science, biology, chemistry, material science, or whatever, then it is overwhelmingly likely that this is a consequence of these two theories. The presumption is that it would not be necessary to invoke quantum chromodynamics, or the weak force, let alone to posit physical processes or entities previously unknown to physics. Moreover, these theories are mathematically simple: the equation of terrestrial gravity is extremely simple, and the necessary quantum mechanics, "can be written down simply and is completely specified by a handful of known quantities: the charge and mass of the electron, the charges and masses of the atomic nuclei, and Planck's constant" (Laughlin and Pines, 2000).

The final statement is completely universal. The claim is that anything in the universe that happens, happens by virtue of physical changes to physical substances, governed by universal physical law:

> Schematically, physicalism can be thought of as the claim that the physical facts determine all the facts. ... In developing a claim of this sort, we need to do two things: first provide some dependence relation that explicates the thought that one set of facts "determines" another; second, decide what kinds of facts are to count as physical. Physicalist positions have been articulated in terms of a variety of dependence relations, including supervenience (there can be no change without physical change), realization (non-physical properties are s second-order, properties of physical properties), and token identity (everything (concrete) that instantiates a non-physical property also instantiates a physical property, to name but a few. ... [T]he causal level must be "causally closed" with respect to the higher level; there is no "downward causation" from the higher level to the lower level. (Hendry, 1999).

Laplace's finding is easily expressed in a logical system; one simply states that Newtonian gravitation exactly characterizes the motion of the planets. I can see, more or less, how to represent Feynman's atomic hypothesis; one can state that all solids, bodies of liquids, bodies of gas, and unions of these are a set of atoms; or that the mass of all the matter within a given spatial region at a time is equal to the sum of the masses of the atoms. However, I have no idea how to formalize the latter two statements. Indeed, their logical status is not clear to me; I don't know whether they are statements *in* physics or meta-level statements about physics or heuristics for carrying out research

in physics.

However, it does seem that there can be experimental *evidence* for these claims. For example, Rubner's 1894 demonstration that conservation of energy holds within a dog is an important experiment for *physics*, because it demonstrates that the principle holds for living creatures, which is not obvious on the face of it. More generally, the justification for these two claims rest on an enormous body of experimental evidence showing the profound regularities in chemical behavior, material behavior, biochemical behavior, and biological behavior; and the theoretical analysis and experimental evidence that demonstrates, as far as it goes, that chemical and material behavior can be explained in terms of physics, that biochemistry can be explained in terms of chemistry, and that biology can be explained in terms of biochemistry. Conversely, any phenomenon that is puzzling and not explained, such as the reversal of the earth's magnetic field, is necessarily to some extent evidence against the claims. (In a Bayesian theory, if a positive outcome is evidence for a claim, then necessarily a negative outcome is evidence against it.) All of these are, in principle, part of the data D to be considered.

Also, it seems to me, these claims indicate that Occam's razor, as used by physicists, involves something more than just the minimum description length principle. When you make a new experimental finding, then the MDL principle gives you brownie points (so to speak) if you can explain it in terms of known laws of physics, because you can use that to compress the description of the data. That in turn translates back into a increased probability for those laws and hence into predictive power. But I don't see any justification for the MDL principle giving you brownie points as a reward for speculating that the new findings ought to somehow be explicable using known laws of physics.

3.8 Argumentation in physics

From the AI perspective, the difficulties discusses above are mostly problems of *representation*. Even greater are the difficulties of *reasoning* — how one can characterize an argument and implement the validation of arguments in a computer program.

Rigorous mathematical proof consists entirely on deductive reasoning: The conclusion is a logically necessary consequence of the assumptions. In actual mathematical discourse, there are certainly informal arguments, but, as discussed at the start of this article, the great discovery that powers verification technology is that, in the vast body of math that is considered rigorously proved, it is possible to eliminate all informal, "hand-waving" arguments and to fill in all logical gaps.

However, such an undertaking does not seem to be close to possible in physics. Unlike math, it is not possible to ground the reasoning about physical systems on the human scale in deductive inference from the foundational theories; the complexities are simply too large.

3.8.1 Deduction from the absolute foundations

One extreme form of the tee-shirt approach to Pavel is to start from a minimal set of absolutely fundamental concepts and laws, and do everything deductively from there. This idea is demolished in (Laughlin and Pines, 2000); I really cannot do better than to quote from them at a little length, and then I have nothing to add.

> We know that [the Schrödinger equation for electrodynamics] is correct ... But it cannot be solved accurately when the number of particles exceeds about 10. No computer existing, or that will ever exist, can break this barrier because it is a catastrophe of dimension. If the amount of computer memory required to represent the quantum wavefunction of one particle is N, then the amount required to represent the wavefunction of

> k particles is N^k. It is possible to perform approximate calculations for larger systems, and it is through such calculation that we have learned why atoms have the size they do, why chemical bonds have the length and strength they do, why solid matter has the elastic properties it does, why some things are transparent while others reflect or absorb light .... With a little more experimental input for guidance, it is even possible to predict atomic conformations of small molecules, simple chemical reaction rates, structural phase transitions, ferromagnetism, and sometimes even superconducting transition temperatures .... *But the schemes for approximating are not first-principles deductions but are rather art keyed to experiment* [emphasis added] and thus tend to be the least reliable precisely when reliability is most needed, i.e. when experimental information is scarce, the physical behavior has no precedent, and the key questions have not yet been identified. There are many notorious failures of alleged *ab initio* computation methods, including the phase diagram of liquid ^{3}He and the entire phenomenology of high-temperature superconductors .... Predicting protein functionality or the behavior of the human brain from these equations is patently absurd.

This is from 2000; certainly we can now compute much more than we could eighteen years ago, and for all I know, some of the specific examples that Laughlin and Pines mentioned may be outdated.[8] Moreover, these kinds of calculations may be a good fit for quantum computing, when that technology becomes practical. But as far as I can determine, the general point still holds, and will continue to hold for the foreseeable future.

It would certainly be immensely desirable to include in PAVEL the kinds of arguments based on "art keyed to experiment" that connect quantum theory to the many phenomena mentioned by Laughlin and Pines. However, I do not have the knowledge to discuss the logical structure of these arguments or what would be involved in incorporating them into PAVEL.

3.8.2 Argumentation in elementary physics

Let me return to the level of physics that I understand. In arguments that use elementary physics to analyze real-world situations, we can characterize a variety of non-deductive forms of reasoning:

- **The closed world assumption.** It is assumed that everything that will affect the outcome of the experiment has been accounted for.
- **Ignoring irrelevant issues.** A description of Cavendish's experiment need not specify the geographic location where the experiment was performed. (By contrast, the latitude is critical in a description of Foucault's pendulum.)
- **Ignoring small quantities.** In some cases, the value of some small quantity is known, or can be bounded, and it is assumed without proof that, because it is small, its impact on the analysis is small. In other cases, the value of a quantity is not known with any precision, but it is assumed to be small and further assumed to have a small impact on the analysis.
- **Approximation.** "Assume a spherical cow" as the old joke says. Surfaces are taken to be flat, densities are taken to be uniform, resistances are taken to be linear, and so on.

  Certainly approximation, and order-of-magnitude reasoning which is similar, can sometimes be carried out deductively. If an upper bound on the inaccuracy of the approximation is known, it may be possible to answer Boolean questions with certainty or to give an upper bound on

[8] Hendry (1999) similarly argues that molecular structures cannot be calculated from Schrödinger's equation. Rather, given the structure, it is possible to use quantum mechanics to calculate various physical values.

the inaccuracy of numerical calculation. In a probabilistic setting, if an upper bound on the variance is known, then it may be possible to compute a lower bound on the certainty of the answer to a Boolean question or an upper bound on the variance of a numerical answer.

- **Idealization and abstraction.** Almost every analysis of a physical situation idealizes the entities involved and abstracts the relations between them. One reasons about a physical electronic circuit in terms of a circuit diagram. In a mechanics problem, a string is taken to be massless and one-dimensional. Continuum mechanics is an abstraction of the actual particles structure of matter.

  Moreover, a single argument may use multiple idealizations of the same thing. Analyses of chemical reactions, for example, will often combine an molecular model of substances, to describe the reaction, with a continuous model, or multiple continuous models, to describe the fluid mechanics and thermodynamics. An analysis of the tides caused by a planet's moons might well first calculate the moon's orbit approximating the planet as a point mass, and then use the planet's extent and material composition in calculating the tidal effects.

The same physical object and even the same physical situation may have many different possible models, depending on what is the range of behaviors under consideration, the accuracy desired, and the measurements being made. Consider a pendulum on a string. You have the following choices, among others (Beech, 2014).

- The setting can be two-dimensional or three-dimensional. It can even be one-dimensional, if you simply set up the problem in terms of the Lagrangian $\mathcal{L}(\theta) = m(r\dot{\theta})^2/2 + mgr\sin(\theta)$, where θ is the angle from vertical downward.
- The bob can be a point mass, a circle or sphere, or a more complex shape.
- There are many different options for the string:
  - It can be considered like a rod, holding the weight at a fixed distance from the attachment point; or a hard constraint maintaining an upper bound on the distance from the bob to the attachment point; or a soft constraint, exerting an elastic force when stretched beyond a fully extended position. In the Lagrangian formulation mentioned above, the string is completely abstracted away, into the formulation of the energy function.
  - It can be one dimensional or three dimensional.
  - It can be massless or massed.
  - It can bend along its length or twist along its axis or both.
  - It can be immutable, or it can snap, or it can be cut.
- Dissipative forces can include air resistance or friction at the attachment point or both; various kinds of approximations can be used.
- The frame within which the pendulum is set up can be fixed, or it can be attached to a rotating earth. This option would hardly cross one's mind, except that it is critical in Foucault's pendulum.
- Gravity can be a uniform field, or a Newtonian field, or follow general relativity

Different circumstances call for different idealizations. A problem in a freshman course would probably use a two-dimensional setting, a point mass, and a string of fixed length. A problem in an advanced mechanics class might simplify the analysis to a one-dimensional Lagrangian formulation or might complicate it by positing an extended mass or a three-dimensional setting. Cavendish's

experiment requires a three-dimensional setting, an extended bob, (the two weights on the rod) and a cord that twists along its axis. Foucault's pendulum requires a three-dimensional setting, a cord of fixed length, and a frame attached to the rotating earth. Smith, Battaglia, and Vul (2013) describe a psychological experiment in which subjects were ask to predict the trajectory of a pendulum if its cord is cut in mid flight; this requires a fixed length string that can be cut. Reasoning about a yo-yo requires an extended object and a flexible one-dimensional string. Reasoning about a cord swinging freely requires a one-dimensional string with constant density. The pendulum in a grandfather clock is connected to a mechanism that adds energy at every swing. In the Poe story, "The Pit and the Pendulum", the cord is a brass rod, the bob is "a crescent of glittering steel, about a foot in length from horn to horn; the horns upward, and the under edge evidently as keen as that of a razor;" and the frame gradually descends.

It is tempting to propose that one should always use the most detailed possible model. But this is hardly feasible; not only does the complexity of calculations go up rapidly, but, more seriously, so does the kind of information needed. If you approximate a cord in a pendulum as a distance constraint, all you need to know is its length; if you want a detailed model you need to know additionally its radius and its material characteristics. In a given situation, these may be unspecified or hard to determine. (Again, of course, the Bayesians will tell you blithely that, if you don't know them, you should use a probability distribution over the range of values.)

3.9 Reasoning about things that are partially understood

Physical reasoning can be applied to phenomena that are only partially understood, such as plate tectonics, the planetary magnetic fields, the million-degree temperature in the sun's corona, and lightning (Dwyer and Uman, 2013). Feynman (1964) book-ends his volume-long textbook on electromagnetism as follows:

> [End of chapter 1] Let us end this chapter by pointing out that among the many phenomena studied by the Greeks, there were two very strange ones: that if you rubbed a piece of amber, you could lift up little pieces of papyrus, and that there was a strange rock from the island of Magnesia which attracted iron. It is amazing to think that these were the only phenomena known to the It is amazing to think that these were the only phenomena known to the Greeks in which the effects of electricity or magnetism were apparent. [Feynman seems to have forgotten lightning.] (Feynman 1964, end of chapter 1]
>
> [End of chapter 37] We now close our study of electricity and magnetism. In the first chapter we spoke of the great strides that have been made since the early Greek observations of the strange behavior of amber and of lodestone. Yet in all our long and involved discussion, we have never explained why it is that when we rub a piece of amber we get a charge on it nor have we explained why a lodestone is magnetized ... So you see this physics of ours is a lot of fakery — we start out with the phenomena of lodestone and amber, and we end up not understanding either of them very well [end of chapter 37]

Almost sixty years later, these phenomena are certainly *better* understood, but none are perfectly understood; in particular the triboelectric effect, in which rubbing one material with another creates an electric charge "is not very predictable" (Wikipedia, triboelectric effect). Nonetheless, a lot of physical reasoning about these *is* possible, through a combination of fundamental principles, experimental evidence, approximations, and speculative reconstruction of structure and mechanisms. Such explanations typically fail to match observed reality in some respects or fail to distinguish the circumstances where the phenomenon occurs from those where it doesn't. Nonetheless, these expla-

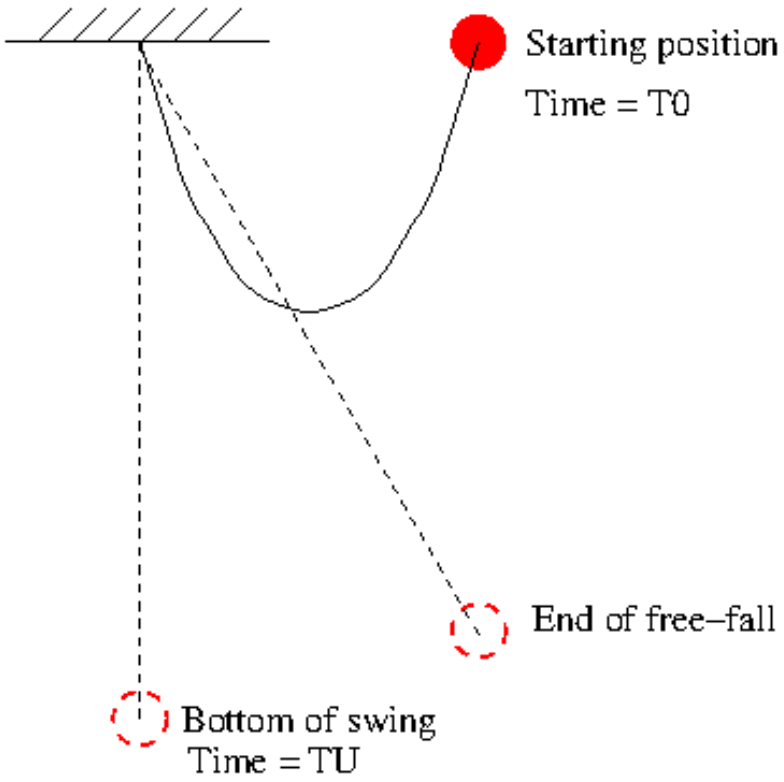


Figure 2: Dropping a pendulum on a string

nations are considered valid as far as they go; no one seriously proposes that these phenomena are evidence of a fundamental physical process that lies outside of the known fundamental theories.

4 An example word problem

To illustrate what would be involved in formalizing simple physics reasoning in PAVEL, we present a formalization of the following simple word problem:

> **Problem:** A 1 kg pendulum bob on a 1 meter inelastic string is dropped from the point 0.5 meters directly to the right of the attachment point. How long will it take to reach a point directly below the attachment point? What force will the string be exerting on the bob at that point? (Figure 2).

Table 3 shows the sorts and the sortal operators. Table 4 shows the language of geometry and kinematics used. Table 5 shows the theory associated with string. Table 6 shows the dynamic laws of physics. Finally table 7 shows the formulation of the problem. What is missing here is the purely mathematical theory (the theory of the reals, vector algebra, and vector calculus); the axioms governing the relation between dimensions; and the purely kinematic theory.

Sorts: *Object, String, PseudoObject, Time, Point, Real, Duration, Distance, Speed, Acceleration, Mass, Force*

Sortal Functions:

Fluent[σ] — Function from *Time* to sort α.
Vector[α] — If α is a real-valued dimension, then a vector of dimension α.
$\alpha \otimes \beta$ — Infix operator: Dimension α times dimension β.
$\alpha \oslash \beta$ — Infix operator: Dimension α divided by dimension β.

Notes: *Acceleration, Force,* and *Momentum* are here scalar dimensions, not vectors.
Here and in the tables below, sortal variables σ and τ range over all sorts; variables α and β range over the additive dimensions (i.e. real-valued dimensions with a natural sense of zero and of addition) *Real, Duration, Distance, Speed, Acceleration, Mass,* and *Force.*

Table 3: Physics Word Problem: Sorts

Constant Symbols
`Meter` → *Distance*.
`Second` → *Duration*.
`X` → *Vector[Real]*. Horizontal dimensionless unit vector.
`Z` → *Vector[Real]*. Vertical dimensionless unit vector.

Function Symbols:
`Dist(pa:`*Point*`, pb:`*Point*`)` → *Distance*.
Distance between *Point*s `pa` and `pb`.
`Magnitude(v:`*Vector[*α*]*`)` → α. The magnitude $|\vec{v}|$.
`PointPlusVec(p:`*Point*`, v:`*Vector[Distance]*`)` → *Point*
The sum $\mathbf{p} + \vec{v}$ of point $\mathbf{p}$ plus vector $\vec{v}$.
`VecFrom(pa:`*Point*`,pb:`*Point*`)` → *Vector[Distance]*.
Vector `pb` − `pa` where `pa` and `pb` are *Point*s.
`VecMinus(u:`*Vector[*α*]*`, v:`*Vector[*α*]*`)` → *Vector[*α*]*. Vector $\vec{v} - \vec{u}$.
`ScalarTimesVec(x:`α`, v:`*Vector[*α*]*`)` → *Vector[*$\alpha \otimes \beta$*]*. The scalar product $x \cdot \vec{v}$.
`DotProd(u:`*Vector[*α*]*`, v:`*Vector[*β*]*`)` → $\alpha \otimes \beta$. Dot product $\vec{u} \cdot \vec{v}$.
`Direction(v:`*Vector[*α*]*`)` → *Vector[Real]*.
Dimensionless direction of $\vec{v}$. $\vec{v}/|\vec{v}|$.
`V(t:`*Time*`, q:`*Fluent[*α*]*`)` → α. Value of fluent `q` at time `t`.
`VelocBefore(p:`*Fluent[Point]*`)` → *Fluent[Vector[Speed]]*.
Derivative of $\vec{x}(t)$, where $\vec{x}$ is a *Point*-valued *Fluent*, evaluated from the left (see text.)
`VelocAfter(p:`*Fluent[Point]*`)` → *Fluent[Vector[Speed]]*.
Derivative of $\vec{x}(t)$, where $\vec{x}$ is a *Point*-valued *Fluent*, evaluated from the right.
`DerivOfVeloc(p:`*Fluent[Vector[Speed]]*`)` → *Fluent[Vector[Acceleration]]*.
First time derivative of $\vec{v}(t)$, where $\vec{v}$ is a velocity, evaluated from the left.
`OPlace(o:`*Object*`)` → *Fluent[Point]*.
The fluent tracking the location of *Object* `o` over time.
`QPlace(q:`*PseudoObject*`)` → *Fluent[Point]*.
The fluent tracking the location of *PseudoObject* `q` over time.

Predicate symbols:
`Zero(x:`α`)` — Scalar x has zero value.
`Positive(x:`α`)` — Scalar x is positive.
`Continuous(p:` *Fluent[Point]*`t :`*Time*`)`.
Point-valued *Fluent* `p(t)` is a continuous function of time in a neighborhood of time `t`
`TwiceDifferentiable(p:`*Fluent[Point]*`, t:`*Time*`)`.
Point-valued *Fluent* `p(t)` is twice differentiable in a neighborhood of time `t`.

Table 4: Physics word problem: Geometric and kinematic primitives

Function symbol:
`Length(s:`*String*`)`. Length of *String* `s`.

Predicate Symbols:
`EndOf(q:`*PseudoObject*`, s:`*String*`)`. *PseudoObject* `q` is an end of *String* `s`.
`Fixed(q:`*PseudoObject*`)`. *PseudoObject* `q` is fixed in position.
`Attached(q:`*PseudoObject*`, o:`*Object*`)`. *Object* `o` is attached to *PseudoObject* `q`.
`Taut(s:` *String:*`, t:`*Time*`)`. *String* `s` is taut at time `t`.
`Yanking(s:`*String*`, t:`*Time:*`.` At time `t`, *String* `s` yanks the objects attached to it. See text.
`Yanked(o:`*Object*`, t:`*Time*`)`. At time `t`, *Object* `o` is yanked by some string it is attached to.

Axioms:
S.1. $\forall_{\texttt{s:}String}\ \exists_{\texttt{qa,qb}}$ `EndOf(qa,s)` $\wedge$ `EndOf(qb,s)` $\wedge$ `qa` $\neq$ `qb` $\wedge$
$[\forall_{\texttt{qc}}$ `EndOf(qc,s)` $\Longrightarrow$ `qc=qa` $\vee$ `qc=qb`$]$.
Every string has exactly two ends.

S.2. $\forall_{\texttt{t:}Time;\texttt{s:}String,\texttt{qa,qb:}PseudoObject}$
`EndOf(qa,s)` $\wedge$ `EndOb(qb,s)` $\wedge$ `qa` $\neq$ `qb` $\Longrightarrow$
`Distance(V(t,QPlace(qa)), V(t,QPlace(qb)))` $\leq$ `Length(s)`.
The distance between the end of a string is at most the length of the string.

S.3 $\forall_{\texttt{s:}String;\texttt{q:}PseudoObject;\texttt{oa,ob:}Object}$ `EndOf(q,s)` $\wedge$ `Attached(oa,q)` $\wedge$ `ob` $\neq$ `oa` $\Longrightarrow$
$\neg$`Attached(ob,q)` $\wedge$ $\neg$`Fixed(q)`.

S.4. $\forall_{\texttt{o,q}}$ `Attached(q,o)` $\Longrightarrow$ $\forall_{\texttt{t}}$ `V(t,OPlace(o)) = V(t,QPlace(q))`.
If *Object* `o` is attached to end `q` of a *String* then `o` and `q` are always in the same place.

S.5. $\forall_{\texttt{q}}$ `Fixed(q)` $\Longrightarrow$ $\forall_{\texttt{ta,tb:}Time}$ `V(ta,QPlace(q)) = V(tb,QPlace(q))`.
A fixed end of a string is always in the same place.

S.6. $\forall_{\texttt{s:}String;\texttt{t:}Time}$ `Taut(s,t)` $\Leftrightarrow$
$[[\forall_{\texttt{q}}$ `EndOf(q,s)` $\Longrightarrow$ `[Fixed(q)` $\vee$ $\exists_{\texttt{o}}$ `Attached(q,o)]]` $\wedge$
`[Distance(V(t,QPlace(qa)), V(t,QPlace(qb))) = Length(s)]]`.
Definition: A string is taut at time `t` if both ends are either fixed or attached to an object and the distance between the ends is equal to its length.

S.7 $\forall_{\texttt{s:}String;\texttt{t:}Time}$ `Yanking(s,t)` $\Leftrightarrow$
`Taut(s,t)` $\wedge$
$\exists_{\texttt{qa,qb}}$ `EndOf(qa,s)` $\wedge$ `EndOf(qb,s)` $\wedge$
`Positive(DotProd(VecMinus(V(t,VelocBefore(Place(qa))),`
`V(t,VelocBefore(Place(qb))))`
`VecFrom(V(t,Place(qa)),V(t,Place(qb)))))`.
Definition: String `s` is yanking at time `t` if it is fully extended at `t`, and if the velocity of the two ends at time `t` are such that it would be overextended if they continued in their motion.

S.8 $\forall_{\texttt{o:}Object;\texttt{t:}Time}$ `Yanked(o,t)` $\Leftrightarrow$
$\exists_{\texttt{s,q}}$ `Attached(o,q)` $\wedge$ `EndOf(q,s)` $\wedge$ `Yanking(s,t)`.
Object `o` is yanked at *Time* `t` if it is attached to some *String* that is yanking.

Table 5: Theory of strings

Constant symbol:
`Kilogram` → *Mass*

Function Symbols:
`MassOf(o:`*Object*`)` → *Mass*. The mass of *Object* `o`.
`GravForceOn(o:`*Object*`)` → *Fluent[Vector[Force]]*. The gravitational force on *Object* `o`.
`ForceOn(oa:`*Object*`, ob:`*Object*`)`. The force executed on `oa` by `ob`.
`TotalForceOn(oa:`*Object*`)` → *Fluent[Vector[Force]]*. The total force executed on `oa`.

Axioms:

P.1. $\forall_{\texttt{o}:Object;\texttt{t}:Time}$ `Continuous(OPlace(o),t).`
Objects move continuously.

P.2. $\forall_{\texttt{o}:Object;\texttt{t}:Time}$ $\neg$`Yanked(o,t)` $\Longrightarrow$
`TwiceDifferentiable(OPlace(o),t)` $\wedge$
`ScalarTimesVec(Mass(o),V(t,DerivOfVeloc(VelocityBefore(OPlace(o))))) =`
`V(t,TotalForceOn(o)).`
Newton's second law, except when there is an impulse from a string.

P.3. $\forall_{\texttt{o}:Object;\texttt{t}:Time}$ `GravForceOn(o,t) =`
`ScalarTimeVec(`-9.8 `* MassOf(o) * Meter / (Second * Second), Z).`
Terrestrial gravitational force.

P.4. $\forall_{\texttt{o}:Object;\texttt{s}:String;\texttt{qa},\texttt{qb}:PseudoObject;\texttt{t}:Time}$
`Attached(o,qa)` $\wedge$ `EndOf(qa,s)` $\wedge$ `EndOf(qb,s)` $\wedge$ `Fixed(qb)` $\wedge$ `Yanking(s,t)` $\Longrightarrow$
`V(t,VelocAfter(OPlace(o))) =`
`VecMinus(V(t,VelocBefore(OPlace(o))),`
`ScalarTimesVec(DotProd(V(t,VelocBefore(OPlace(o))),`
`Direction(VectorFrom(V(t,QPlace(qb)),V(t,QPlace(qa))))),`
`Direction(VectorFrom(V(t,QPlace(qb)),V(t,QPlace(qa))))).`
When an object "collides" with the end of a string and the other end is fixed, then the velocity after the collision is the component of the velocity before the collision in the direction tangent to the taut string.

P.5. $\forall_{\texttt{o}:Object;\texttt{s}:String;\texttt{t}:Time}$ $\neg$`Taut(t,s)` $\Longrightarrow$ `Zero(V(t,ForceOn(s,o))).`
If a string is not taut, it is not exerting any force.

P.6. $\forall_{\texttt{o}:Object;\texttt{s}:String;\texttt{qa},\texttt{qb}:PseudoObject;\texttt{t}:Time}$
`Taut(t,s)` $\wedge$ `EndOf(qa,s)` $\wedge$ `EndOf(qb,s)` $\wedge$ `qa` $\neq$ `qb` $\wedge$ `Attached(o,qa)` $\Longrightarrow$
`[Zero(Magnitude(V(t,ForceOn(s,o)))) ∨`
`Direction(V(t,ForceOn(s,o))) =`
`Direction(VectorFrom(V(t,QPlace(qa)), V(t,QPlace(qb))].`
The force exerted by a taut string on an object attached at one end is parallel to the direction to the other end.

Table 6: Physics word problem: Laws of physics

B→*Object* – The bob.
S→*String* – The string.
QA→*PseudoObject* – The end of the string attached to B
QB→*PseudoObject* – The fixed end of the string.
T0→*Time* – The initial time.
TU→*Time* – The time when the bob is directly under the attachment point.

Axioms:
F.1. EndOf(QA,S) ∧ EndOf(QB,S) ∧ QA ≠ QB.

F.2. Attached(B,QA).

F.3. Fixed(QB).

F.4. Length(S) = Meter.

F.5. MassOf(B) = Kilogram.

F.6. V(T0,OPlace(B)) =
PointPlusVec(V(T0,QPlace(QB)), ScalarTimesVec(0.5 * Meter, X)).

F.7. Direction(VectorFrom(V(TU,QPlace(QB)),V(TU,OPlace(B)))) =
ScalarTimesVec(−1, Z).

F.8. $\forall_{t:Time}$ Direction(VectorFrom(V(t,QPlace(QB)),V(t,OPlace(B)))) =
ScalarTimesVec(−1, Z)) ⟹
t ≥ TU.
TU is the first time when the bob is below the attachment point.

F.9 $\forall_{t:Time}$ V(t,TotalForceOn(B)) = V(t,ForceOn(S,B)) + V(t,GravForceOn(B)).
Closed world assumption: The only forces on the bob are gravity and the string.

Evaluate: (TU−T0). **Evaluate:** V(TU, ForceOn(S,B)).

Table 7: Physics word problem: Problem formulation

The formalization in tables 3-7 should be largely self-explanatory, but a few points require explanation.

A "pseudo-object" (introduced in Davis, 1988) is a geometrical feature that moves around with an object: The center of a spherical object, the surface of an object, the apex or base of a cone, the hole in a donut, and so on. In this case, we mark the two ends of the string as pseudo-objects.

The long-winded and unappealing symbols that we have used for vector and function operators — `PointPlusVec(x,v)` instead of simply $\mathtt{x}+\vec{\mathtt{v}}$, and so on — are there in order to keep our system of sorts simple. Standard mathematical notation, and many math-oriented programming languages such as MATLAB, enormously overload standard symbols such as '+' and '·'. In a practical implementation of PAVEL, this might end up being worthwhile; for this simple example, it seemed better to keep the sorting system simple and burden ourselves with separate symbols.

Since the velocity of the bob is discontinuous at the moment when the end of the string is reached, we define the velocity of an object **o** before time `t` to be the limit of the derivative of its position at time $\mathtt{t}'$ as $\mathtt{t}'\rightarrow\mathtt{t}^-$ and the velocity after `t` analogously. (The definition would be included in the kinematic axioms, not enumerated here.)

In section 3.5 we raised the issue of the assumption that masses and so on remain constant from one stage of an experiment to another. In our formalization here, we have unabashedly cheated on all such concerns by using time-independent symbols for every quantity or relation that does not change over time in this particular problem. For instance `MassOf(o)` is presumed to be a time-invariant property of an object `o`; `Attached(o,q)` is assumed to be a time-invariant relation between object `o` and pseudo-object `q`; and so on.

We use a simple theory of non-elastic, one-dimensional, massless strings, governed by the following rules, enumerated in tables 5 and 6

- A string has two ends (axiom S.1) which cannot be more than a fixed distance apart (the length of the string) (axiom S.2).
- The end of a string may be *attached* to a single point object, or it may be *fixed* in space (presumably actually attached to some fixed frame, but we did not include the frame in our formulation here) (axiom S.3). If it is attached to an object, then the object and the end of the string are always at the same point (axiom S.4). If the end of the string is fixed, then it is always at the same point (axiom S.5).
- A string is *taut* if both ends are either attached or fixed, and if it is fully extended; that is, the distance between the two ends is equal to the length of the string (axiom S.6)
- An inelastic event involving the string, called a *yanking* occurs when the string is taut, and the difference in velocities between the two ends has a positive component in the direction from one end to the other (axiom S.7). Note that if difference is orthogonal to the direction, as in the case when the string is swinging in a circle, that is not considered a yanking.
- If an object is attached to an end of a string undergoing a yanking then it is said to be yanked (axiom S.8)
- A string exerts no force if it is not taut (axiom P.5)
- A string that is taut and not yanking exerts on an attached object a non-negative force in the direction along the string (axiom P.6)
- If one end of a string yanks on an object, and the other end is fixed, then the velocity of the object changes discontinuously. Specifically, its velocity after the event is equal to the

component of its velocity before the event in the direction orthogonal to the direction of the string (axiom P.4 — there may well be some more elegant way to axiomatize this.)

Combining these with the statement (axiom P.2) that, when not yanked, the object obeys Newton's second law, these suffice to determine that, after falling vertically to the length of the string, the bob will swing back and forth on a circle, and the centripetal force that the string exerts on the bob will be exactly what is needed to keep it on that path (the component of gravity in the direction of the string plus the centrifugal force). If the centripetal force were less then that, then the distance between the end of the string would be greater than the length, which is impossible; if it were greater, it would pull the bob within the circle; the string would cease to be taut, and the bob would instantaneously fall back, which is also impossible.

The rule for the changes in velocity if the string is attached to objects at both ends and a yanking event occurs is similar, but more complicated; it is not included here.

The problem formulation requires a closed world assumption (axiom F.9) that the total force on the bob is the sum of gravity and the force from the string. Almost any problem formulation in physical reasoning has to have some kind of closed world assumption, that states that everything that will interfere with the system has been accounted for. In this case, it would be better to have a general rule of physics that the total force on an object is the sum of the forces, and then to have the individual problem statement assert that the only forces on the bob are gravity and the string. However, that would require adding "sets of forces" as a sort and summation over sets as an operation, so we went with this simpler, less general, formulation instead.

In general, there is always a choice to be made about how general to make the formulation of the theory and how much to tailor it to the specifics of the problem at hand. If you have only a single problem in mind, then the decision is essentially stylistic: using a general representation makes the argument that the theory generalizes more plausible, using a more tailored one makes the exposition simpler. The more problems you address, the more is gained by generality, but it remains to some extent a matter of taste. (Tailoring the representation of a general theory to the specifics of one or a few problems violates the "no function in structure" rule of (de Kleer and Brown, 1985). On the other hand, if one is going to choose among idealizations the one that best fits the problem, as I have argued above, then that principle has been given up in any case.)

On the whole word problems in physics are simpler and more idealized than experimental set ups. A reasonable axiomatization of the Cavendish experiment at a comparable level of detail would probably be two or three times longer.

5 Historical context and related work

There is a long history of work more or less along the lines of PAVEL. That history has three primary threads: in physics, in philosophy, and in AI. The physics and philosophy threads both largely begin with Hilbert; the AI thread is largely separate.

Corry (2004) gives a very detailed account of the physics and philosophical work up through the work of Hilbert; I have not found a comprehensive review of the work since Hilbert.

5.1 Before Hilbert

Newton's *Principia* is substantially presented as deductions from axioms, in imitation of Euclid. In modern times Heinrich Hertz's (1894) *Die Prinzipien der Mechanik* was the first attempt to

formulate the laws of mechanics in axiomatic form. It was notable for its exclusion of force as a fundamental concept, and using only time, space, and mass.

5.2 Hilbert's sixth problem and the axiomatization of physics

In Hilbert's famous collection of 23 mathematical problems, proposed at the 1900 International Congress of Mathematicians, number 6 was the axiomatization of physics (Corry, 2004).

> **Mathematical Treatment of the Axioms of Physics.** The investigations on the foundations of geometry suggest the problem: To treat in the same manner, by means of axioms, those physical sciences in which already today mathematics plays an important part; in the first rank are the theory of probabilities and mechanics.

Hilbert further explained:

> As to the axioms of the theory of probabilities, it seems to me desirable that their logical investigation should be accompanied by a rigorous and satisfactory development of the method of mean values in mathematical physics, and in particular in the kinetic theory of gases. ... Boltzmann's work on the principles of mechanics suggests the problem of developing mathematically the limiting processes, there merely indicated, which lead from the atomistic view to the laws of motion of continua.

In general, mathematicians have been unenthusiastic about Hilbert's sixth problems. It is very much an outlier among his 23 problems; whatever it is, it isn't mathematics,[9] and it is not at all clear what would count as a solution. Benjamin Yandell (2002), in his 400-page book on Hilbert's problems, dismissed the sixth problem in a mere four pages.

There seem to be three general projects involved in Hilbert's sixth problem.

First, the axiomatization of probability theory. This was accomplished by Kolmogorov, at least as far as the measure space interpretation goes. As discussed in section 2.2, I am not convinced that the likelihood model, which permits probabilities of individual propositions, is axiomatized to the point that it supports analysis of real-world situations.

Second, the axiomatization of the foundations of physics; these, of course, were radically transformed in the three decades after Hilbert's speech. Hilbert himself devoted substantial research energy to the formulations of quantum theory and general relativity; he and Emmy Noether were in communication with Einstein about general relativity during the years that Einstein was developing the theory.

As best as I can ascertain, the current status is as follows:

- General relativity is completely axiomatized. It would be feasible to formulate the theory as axioms in a proof-verification system and to prove consequences such as the rotation of the perihelion of Mercury, the possibility of black holes, gravitational lenses, gravitational waves, and so on.
- Schrödinger's equation for non-relativistic quantum mechanics is easily axiomatized — it is just a partial differential equation — and its consequences can be proved, up to the limits

[9] Incidentally, the fact that Hilbert included this problem and spent a great deal of time working on it tells strongly against the common idea that Hilbert was a pure formalist, who viewed the meaning of mathematical symbols as unimportant (Corry, 2004).

discussed in section 3.8.1. However, if one adds Born's law, which governs the probabilistic collapse of the wave function following an observation, then the situation becomes much less clear. As far as I can find, most so-called "axiomatizations" of quantum physics that include Born's law (e.g. Cappellaro 2012, chap. 3) are fine as regards the physics, but do not specify what is the probabilistic logic used (if that is necessary) or give a useful characterization of an observation, or state the independence assumptions. It is not clear to me that we are currently in a position to characterize axiomatically experiments whose outcomes depend on Born's law.[10] I do not know how severe a limitation that is; for example, how many, if any, of the explanations of phenomena enumerated in the above quote from Laughlin and Pines would be affected.

Ludwig's (1985, 1987) *An Axiomatic Basis for Quantum Mechanics* develops an axiomatic theory, and, further, presents a metatheory of axiomatizations of physical theory. It includes an extensive, though very abstract, discussion of the relation between the theory and its macroscopic manifestations. Unfortunately, I am not at all in a position to evaluate what is the scope of what he accomplished; apparently the discussion is extremely difficult and relentlessly abstract, even for expert readers (Vogt, 1997).

Boender, Kammüller and Nagarajan (2015) have Coq to verify protocols in quantum communication and quantum cryptography, but this is far from the physics experiments that we are discussing, and though it uses probabilities, it requires only a very limited theory.

- Quantum field theory is in a much less certain state; the axiomatizations that have been proposed, such as the Wightman axioms, have severe limitations. This remains an open problem.

Also, as is well known, finding a satisfactory theory that encompasses both general relativity and quantum theory is unsolved.

Third, the explanation of continuum mechanics in terms of particle mechanics;[11] more generally, the explanation of macroscopic behaviors in terms of foundational theories. This is a more open-ended project, since there are several forms of continuum mechanics, and an open-ended collection of macroscopic behaviors.

One particularly important and difficult problem of this kind has been to complete the derivation of thermodynamics from statistical mechanics begun by Maxwell and Boltzmann. A recent study which develops a substantial formal foundation, is Wallace (to appear).

In general, it seems to me fair to say that what a physicist usually means by "axiomatization" is quite different from a mathematician means, and still more from what a logician means. When a physicist claims to have "axiomatized" a theory, what he/she generally has done is to have enumerated a set of foundational rules for an abstract theory which, generously supplemented by the physicists' own understanding of the concepts involved and by a variety of facts too obvious to be worth mentioning, will support various kinds of informal arguments. ((Ludwig 1985, 1987) is certainly an exception.)

[10] It has been suggested to me that it will be easier to find a logical formulation of the "many-worlds" interpretation of quantum mechanics or, alternatively, the theory of quantum decoherence than the Copenhagen interpretation. That may be so; but I can't find that anyone has produced a logical formulation of either of these interpretations either.

[11] Slemrod (2013) writes, "Historically a canonical interpretation of this '6th problem of Hilbert' has been taken to mean passage from the kinetic Boltzmann equation for a rarefied gas to the continuum Euler equations of compressible gas dynamics as the Knudsen number ϵ approaches zero." I do not know what is the basis for this rather narrow interpretation.

5.3 Philosophy

There is a long philosophical literature on axiomatizing physics, particularly particle dynamics, either in a strictly logical notation or in some other formalism. Some early work include part VII of Russell's (1903) *The Principles of Mathematics*, a precursor to *Principia Mathematica*, entitled "Matter and Motion"; and Hamel (1912, 1921) *Elementare Mechanik* and *Grundbegriffe der Mechanik*. (Hamel was a student of Hilbert's)

In Vienna in the 1920s, a group of philosophers, mathematicians, and physicists called "The Vienna Circle" (Sigmund, 2017) embarked on a formidably ambitious project to investigate the foundations of science, called "logical positivism" or "logical empiricism". Following the models of Whitehead and Russell's (1910) in *Principia Mathematica*, and of Wittgenstein's (1922) *Tractatus*, they attempted to demonstrate that scientific theory could be built up logically from basic observations. They planned to produce a large series of books, the *International Encyclopedia of Unified Science,* which would formalize the foundations of all the sciences — physical, biological, and social. Twenty monographs in the series were published, in two volumes.

The Vienna Circle held regular meetings from 1924 to 1936; at any given time, there were 10 to 20 people involved. The central figures at the start were the physicist Moritz Schlick, who served as chair, the sociologist Otto Neurath, and the mathematicians Otto Hahn and Philipp Frank. In 1926, the Circle were joined by Rudolf Carnap, who became the leading exponent of logical positivism; his book *The Logical Structure of the World* became a Bible of the movement. (Gödel was also a participant in the meetings of the Circle; however, he does not seem to have ever subscribed to the tenets of logical positivism.)

During the 1950s, there seems to have been an explosion of interest in the subject. Most notably, in 1957, Leon Henkin, Patrick Suppes, and Alfred Tarski (1958) organized a ten-day international symposium at Berkeley on *The Axiomatic Method: With Special Reference to Geometry and Physics*; Part II consisted of 13 papers on "Foundations of Physics". (Part I had to do with geometry; part III was miscellaneous.) In particular, the papers by Adams on rigid body and particle mechanics, by Noll on continuum mechanics, by Hermes on axiomatizing mechanics, and by Suppes, by Walker, and by Ueno on relativistic kinematics give sets of precise axioms that could easily be formalized in a logical notation, and used in a proof verifier. These all lie within the foundational paradigm; they are concerned with formulating basic axioms, not with drawing connections to experiment or observation, except in a very general sense.

There are a number of striking gaps. Carnap was not involved, despite being a good friend of Tarski's and nearby at UCLA; nor are there any citations to his work or any of the other logical positivist work. Hilbert's sixth problem is never mentioned as a context or motivation. Despite the fact that the organizers were Henkin, Suppes, and Tarski, none of the papers in the physics section use logical notation or refer to the concepts of mathematical logic; of course, it is not an especially congenial notation for physics theories. (Several of the papers in parts I and III do use logical notation and reference mathematical logic.) Feynman's dictum notwithstanding, atoms are never mentioned, as far as I can tell.

In their preface to the proceedings of the symposium, Henkin, Suppes, and Tarski (1958) expressed some reservations about whether the project of axiomatizing physics was a reasonable one:

> Much foundational work in physics is still of the programmatic sort, and it is possible to maintain that the status of axiomatic investigations in physics is not yet past the preliminary stage of philosophical doubt as to its purpose and usefulness.

An even sharper critique arguing for the unsuitability in physical reasoning, not merely of axiomatic logic, but of any kind of rigorous mathematics, was Schwartz (1960) "The Pernicious

Influence of Mathematics on Science."

A number of important papers along the same lines precede the conference e.g. McKinsey, Sugar, and Suppes (1953). But after the conference, this line of research seems to have gradually petered out.[12] Richard Montague (1974) wrote a paper on deterministic physics, illustrated with an axiomatization of the gravitational theory of a finite collection of particles, written in logical notation. In the last few decades there have been some further sporadic studies of this kind e.g. (Sant'Anna 1999).

In recent years, some philosophers seeking a mathematical framework for science have turned to Bayesianism, discussed earlier in section 3.2.

A fascinating study, not easily characterized in terms of the above categories, is Strevens' (2013) *Tychomancy*. Drawing extensively on the cognitive psychology of probabilistic reasoning, Strevens attempts to justify the probabilistic reasoning underlying Maxwell's amazing derivation of the distribution of velocities among particles in a gas; he includes also a discussion of the reasoning involved in Darwinian evolution.

5.3.1 Is PAVEL a bad reinvention of logical positivism?

In many ways the previous undertaking that most resembles my proposal for PAVEL was logical positivism. Like PAVEL, logical positivism, as applied to physics, attempted to draw a logical line all the way from the theory to the experience of the scientist doing measurements or observations and to characterize the way in which the theory explains the data and the data supports the theory.

That is not the most encouraging of precedents. The general consensus is that logical positivism was thoroughly demolished by Wittgenstein, Popper, Quine, Kuhn, Lakatos, and others, and that it is an entire dead end — a wholly unworkable approach to the analysis of the scientific method. "The fundamental assumptions of the positivist world view ... lie shattered" (Bhaskar, 1979). Is PAVEL trying to revive a long-dead horse?

Obviously, I don't think so. I think that there are reasons for optimism.

First, the general consensus may be overstated. A philosophical programme that makes ambitious claims is apt to get strong rejoinders, but demonstrating that it has limitations and flaws does not establish that it has nothing of value to offer. Moreover, part of the disrepute of logical positivism is that it became associated with the psychological theory of behaviorism; but the philosophy of science in no way depends on that. There are some indications that the pendulum in the philosophical world may be swinging back.

Second, one issue that the logical positivists were never able to resolve to their own satisfaction was the nature of the ultimate data. The bedrock data from which theory is built are supposed to be "protocol statements" expressing "direct perception", but that turns out to be a very slippery notion. We are in a better position to deal with that now. Perception is better understood now than in 1930. If we want, we could use computer vision to start with actual sensor input. Whether or not this would have satisfied Carnap or early Wittgenstein as an epistemically primitive starting point, it is clearly a well-defined and motivated starting point. The analogous question then becomes, at what level do we move from opaque computer vision procedures to representations with semantics, but that is much more of an engineering question.

[12] I am necessarily relying here on the fact that I have failed to find much later work of this flavor, which is obviously an unreliable argument. However, I do have the following concrete evidence. The International Congress of Logic, Methodology, and Philosophy of Science was in some respects the successor to the Symposium on the Axiomatic Method; it has met 15 times since its inception in 1960. Between 1960 and 1999 there was only one paper (Mehlberg, 1964) that presented an axiomatization of any physical theory (relativistic space-time), though there was a second paper (Ludwig, 1989) that argued in favor of axiomatizations.

Finally, PAVEL has the advantage of being AI, not philosophy. It therefore does not have to produce a theory that covers all cases, or to find its way down to the ultimate turtle or to characterize the whole chain of turtles; if it produces a useful partial answer, that is enough to justify the undertaking. We can set the starting point wherever we want, and get a theory that is more or less powerful and rich. For instance, rather than insisting on taking human perception as the grounding point and viewing the validity of experimental measurements as a hypothesis to be tested, we can take the experimental measurements as a given; that will give results that are in some respect more limited but could still be very enlightening. In my discussion of the BACON program, below, I am critical of BACON for using pre-digested data; but there is nothing wrong with taking that as a starting point, as long as one is aware of its limitations. Problem representations like the one in table 7 are also enormously pre-digested as compared to the actual sensor input, though much richer than the BACON input. The key point is to be aware of the many levels of abstraction that are ultimately involved, and to keep working toward realism.

5.4 Artificial Intelligence

Within AI, there is work of many kinds on physical reasoning (Davis, 2008a); there are AI programs that solve word problems e.g. (Gunning et al. 2010) (Khashabi, Khot, Sabharwal, and Roth, 2018) (Khot, Sabharwal, & Clark, 2018); that do qualitative reasoning (Bobrow, 1985); that design devices e.g. (Hornby, Globus, Linden, and Lohn, 2006), and that design experiments e.g. (Krenn et al. 2016), (Melnikov et al. 2017). Data mining and machine learning are now ubiquitous in scientific research. In this section I will limit the discussion to AI research on developing rich declarative theories of basic physics, and on inferring fundamental theories from data.

5.4.1 Knowledge-Based Physical Reasoning

The AI project closest to PAVEL was the GALILEO project (Lehmann, Chan, & Bundy, 2013). GALILEO used the Isabelle proof assistant to encode a number of models of physical theories and their experimental consequences, including: Joseph Black's theory of latent heat and heat capacity; the explanations of galactic orbital velocities by positing dark matter and by using Milgrom's proposed modification of Newtonian gravity; Roemer's (1676) measurement of the speed of light by delays and advances in the perceived eclipses of Io by Jupiter; the identification of the morning and evening star as the same planet, using observations and Kepler's theory (oddly unhistorical, since the identification was known to the Babylonians[13]); and Pythagoras' determination that the earth is spherical, based on its shadow on the moon during eclipses.

The primary objective of GALILEO was to characterize how ontologies and theories change as a result of disconfirming evidence, and the examples were used as illustrations of various techniques for changing theories. The details of the representation are therefore only developed necessary to illustrate these meta-level techniques. For instance, in the encoding of the speed-of-light example, the time delay on light coming from Jupiter is taken as a primitive measurement; there is no mention of Io or its revolutions.

A research programme, initiated by Patrick Hayes (1979, 1985) aims toward analyzing physical reasoning, particularly "naive" or "commonsense" physical reasoning, at the knowledge level (Newell, 1982) by formulating theories of physics in a logical form and demonstrating that simple inferences can be justified as inference within the logical theory. This is part of a more general project in AI of using logic to formalize the representation of commonsense knowledge and the process of commonsense reasoning (McCarthy, 1968) (van Harmelen, Lifschitz, and Porter, 2008), (Davis,

[13]In fact, despite its popularity as an philosophical example since Frege, there is little evidence that anyone who was aware of the existence of the planets has ever thought that Phosphorus and Hesperus were two different planets.

2017). I myself have continued this direction of research, axiomatizing elementary reasoning about cutting (Davis, 1993), carrying objects in boxes and containers (Davis, 2011), (Davis, Marcus, & Frazier-Logue, 2017), and pouring liquids (Davis, 2008b). The results are axiom sets and problem formulations similar in flavor to tables 1-7. The sample inferences in (Davis, Marcus, & Frazier-Logue, 2017) were automatically verified in SPASS (Weidenbach et al. 2009), a first-order theorem prover considerably less powerful and expressive than Coq or Isabelle, but easier to use.

Bundy et al. (1979) used logic in a quite different way for solving simple physics word problems. Using the "logic programming language" Prolog, they implemented a system that accepted a problem written in English; carried out a "semantic parse" to extract the content of the problem statement; used a rule-based system to find the appropriate equations; and then solved the equations. The program was supplied with schemas for translating categories of problems into equations, for example

```
schema(pullsys
    [Pull,Str,P1,P2], Time
    [ constacc(P1,Time),
      constacc(P2,Time),
      cue stringsys(Str,[Lpart,Rpart]),
      (tension(Lpart,T1,Time)
           <-- coeff(Pull,zero) &
               tension(Rpart,T,Time)  )
    ],
    [ coeff(Pull,zero),
      mass(Pull,zero,Time)
    ]
  )
```

The explanation is thus: "This schema asserts that in a standard pulley problem, the objects undergo constant acceleration, the tension in both parts of the string is equal if there is no friction, and that the friction and mass of the pulley default to zero if not otherwise specified." It is notable here that the general physical law about tension is placed subordinate to the class of pulley problems — that is, at least as done here, it would have to be restated separately in each class of problems where it is used; the general law is placed parallel to the defaults of zero mass and friction on pulleys, which are mostly just conventions about how exercises are written. In a more general knowledge base, it would be better to separate out these levels.

Friedman, Forbus, and Sherin (2017) develop a cognitive model of how student progress from incorrect to correct explanations of physical phenomena. The representation used in that model is a detailed knowledge-based (though not logic-based) structure (figure 3) that relates the observation that Chicago is warmer in summer than winter, both to the correct theory of the seasons (the earth's axis is tilted) and to a common misconception (the earth is closer to the sun in summer than in winter).

5.4.2 AI programs that induce scientific theories

AI programs that have induced broad or fundamental scientific theories from data are few. (There have of course been an enormous number of projects that have used data mining for scientific discovery for very specific projects.)

The largest project of this kind was the BACON project of Langley, Bradshaw, and Simon (1981, 1983) which modeled the induction of scientific laws from data. BACON, in its various incarnations, took as input data tables of results whose values are either numerical or uninterpreted

f_0 (isa earthPath EllipticalPath)

f_1 (spatiallyDisjoint earthPath TheSun)

f_2 (isa TheSun AstronomicalBody)

m_0 (isa ProximalPoint ModelFragment)

m_1 (isa DistalPoint ModelFragment)

m_2 (isa Approaching-Periodic ModelFragment)

m_3 (isa AstronomicalHeating ModelFragment)

m_4 (isa Retreating-Periodic ModelFragment)

f_3 (isa TheSun HeatSource)

f_4 (spatiallyDisjoint TheSun Planet Earth)

f_5 (isa APP-inst Approaching-PeriodicPath)

f_6 (isa AH-inst AstronomicalHeating)

f_7 (isa RPP-inst Retreating-PeriodicPath)

f_8 (i- (Dist TheSun PlanetEarth)
(Rate APP-inst))

f_9 (active AH-inst)

f_{10} (qprop- (Temp PlanetEarth)
(Dist TheSun PlanetEarth))

f_{11} (qprop (Temp PlanetEarth)
(Temp TheSun))

f_{12} (i+ (Dist TheSun PlanetEarth)
(Rate RPP-inst))

f_{13} (increasing (Temp PlanetEarth))

f_{14} (decreasing (Temp PlanetEarth))

f_{15} (qprop (Temp Australia) (Temp PlanetEarth))

f_{16} (qprop (Temp Chicago) (Temp PlanetEarth))

f_{17} (increasing (Temp Chicago))

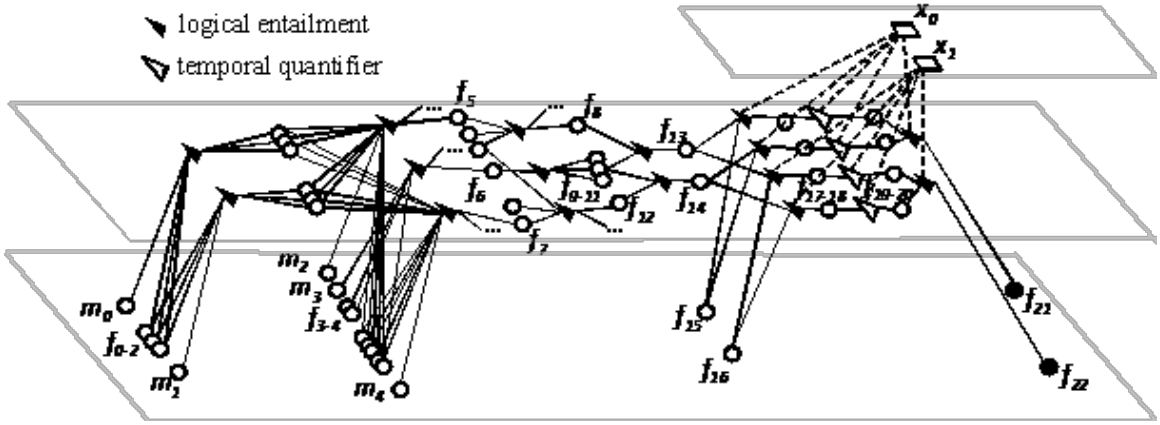
f_{18} (decreasing (Temp Chicago))

f_{19} (holdsIn (Interval ChiWinter ChiSummer)
(increasing (Temp Chicago)))

f_{20} (holdsIn (Interval ChiSummer ChiWinter)
(decreasing (Temp Chicago)))

f_{21} (greaterThan (M (Temp Australia) AusSummer)
(M (Temp Australia) AusWinter))

f_{22} (greaterThan (M (Temp Chicago) ChiSummer)
(M (Temp Chicago) ChiWinter))



This represents the structure of the explanation of the change of temperature over the seasons in terms of the false theory that the earth is closer to the sun in summer.

Figure 3: Network of explanations. From (Friedman, Forbus, & Sherin,2017)

ELEMENT	COMPOUND	w_E	w_C	v_E	v_C	w_E/w_C	w_E/v_E	w_E/v_C
hydrogen	water	10.0	90.0	112.08	112.08	0.1111	0.0892	0.0892
hydrogen	water	20.0	180.0	224.16	224.16	0.1111	0.0892	0.0892
hydrogen	water	30.0	270.0	336.25	336.25	0.1111	0.0892	0.0892
hydrogen	ammonia	10.0	56.79	112.08	74.72	0.1761	0.1338	0.1338
hydrogen	ammonia	20.0	113.58	224.16	149.44	0.1761	0.1338	0.1338
hydrogen	ammonia	30.0	170.37	336.25	224.16	0.1761	0.1338	0.1338
hydrogen	ethylene	10.0	140.10	112.08	112.08	0.0714	0.0892	0.0892
hydrogen	ethylene	20.0	280.21	224.16	224.16	0.0714	0.0892	0.0892
hydrogen	ethylene	30.0	420.31	336.25	336.25	0.0714	0.0892	0.0892

Table 8: Chemical data input to BACON (from Langley, Bradshaw, & Simon, 1983)

symbolic values. It had heuristics for formulating numerical laws which can depend on inferred intrinsic properties. For instance, if resistors A, B, and C each give rise to a linear relation between voltage and current, then BACON can formulate the rule $V = IR$, conjecturing that each of the resistors has a different value for R.

BACON's tabulated clean data is, of course, extremely remote from the realities of experiment interpretation that scientists had to deal with. For instance, table 8 shows the input from which BACON inferred Prout's law of definite proportion in chemical composition. This contrasts starkly with the actual situation of eighteenth and nineteenth century chemists (figure 4), who had to identify chemicals and elements and to distinguish them from mixtures using the techniques and methods available in the labs of the time. Langley, Bradshaw, and Simon do point out that Bacon had the advantage of using clean data, while the data available to the historical scientist used included both noise and significant errors; and that Bacon was presented with only the relevant variables, while a large part of the task facing the scientists was figuring out which variables were critical. But, despite a long historical discussion, they don't address the enormous epistemic gap between a table of numbers and a laboratory set up.

I argued above that in examining the relation of theory to data, it was reasonable to take the grounding data at any level of abstraction. So there is nothing inherently invalid with BACON having taken the data in table 8 as the starting point for theory construction; only, it is important to realize how much that leaves out, as a model of science.

More recently, Bridewell and Langley (2010) has been working on inducing process models characterized by differential equations from traces of parameters over time, across a wide range of domains, including aquatic eco-systems, biochemical kinetics, and molecular biology.

5.4.3 Bayesian inference of structure

Kemp and Tenenbaum (2009) implemented a program that quite directly follows the Bayesian program described in section 3.2 to infer theories from data. Their space of theories Φ is the space of graph structures. The prior $P(H)$ for $h \in \Phi$ is given by a generative process that generates graph structures with various kinds of regularities. The likelihood function $P(d|h)$ is a measure of how well the data fits the structure. The program uses heuristic search to approximately find the most probably structure given the data.

The program was applied to a variety of induction problems. As figure 5 illustrates, it inferred from a table of animal features that animal species conform to a tree structure; it inferred from

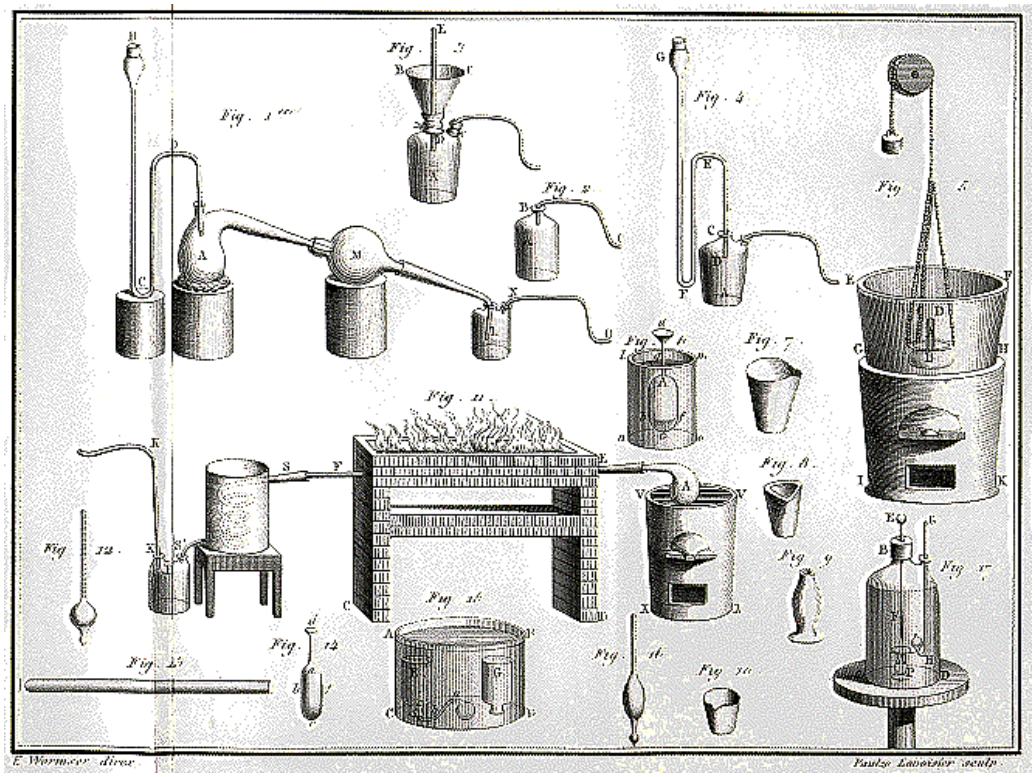
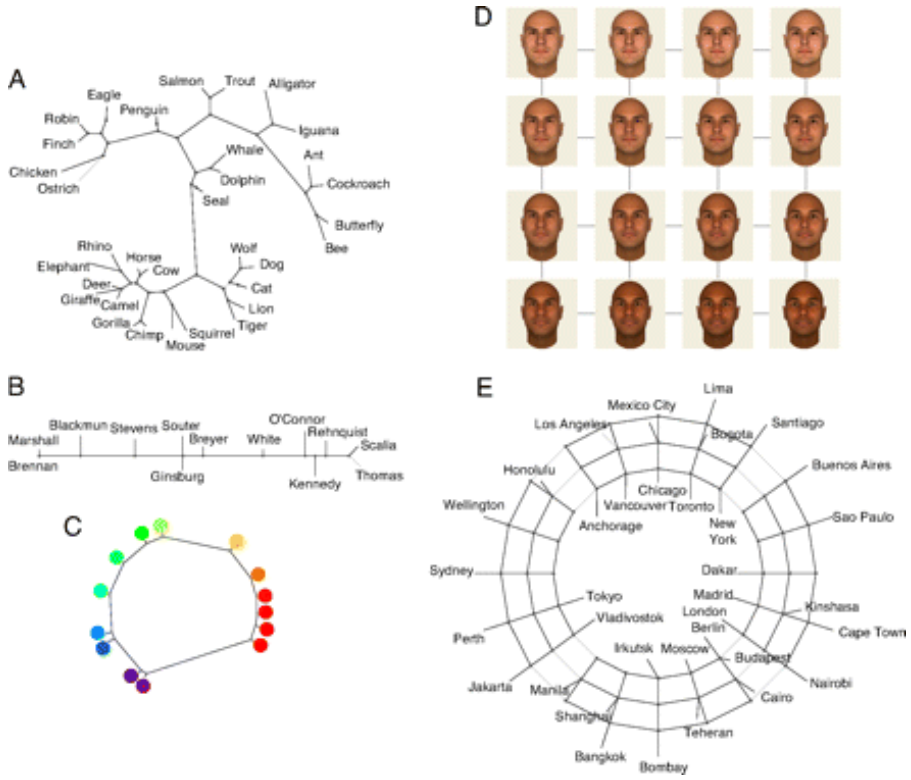


Figure 4: Lavoisier's equipment. From Lavoisier *Ouevres*, Paris, 1862.



From (Kemp and Tenenbaum, 2009)

Figure 5: Results of structure induction

a table of features of Supreme Court opinions that Supreme Court justices conform to a linear structure (conservative to liberal); it inferred from a table of similarity judgments over colors that that colors follow a ring structure; it inferred that a collection of images of faces varying along masculinity and race conforms to a two-dimensional grid; and it inferred from a table of distances between world cities that the position of cities corresponds to a graph that is the cross-product of a ring structure for latitude with a linear structure for longitude.

The last of these, however, inadvertently points out the dangers of using an inappropriate space of models in this kind of study. They write as follows:

> We applied the model to a dataset of distances between 35 world cities. Our model chooses a cylinder where the chain component corresponds approximately to latitude and the ring component corresponds approximately to longitude.

This outcome is so far from reality that one wonders why they would think it supports their theory. The correct model for the geodesic distances between cities on the globe, accurate to within the precision of measurement, is that they are points or small regions on the surface of a sphere; this model, however, is not even in the space of discrete models that they are searching over. Optimizing a model of the distance between cities is not, historically, how the shape of the earth was induced, or could have been induced. In general it is mathematically impossible to induce the concepts of

latitude and longitude from city distances, because the choice of the particular grid for latitude and longitude has essentially no connection to the position of cities, except insofar as there are no cities close to the poles, and that some major coastlines lie roughly north-south. There is no particular reason that a graph of cities should give one a cylindrical structure, rather than any other planar graph, since any planar graph can be embedded in the sphere. In fact, you will only get a cylindrical graph structure corresponding to latitude and longitude if you pick the cities rather carefully with that outcome in mind. If you actually look for structure in the distances between cities in the world, what will be most conspicuous is their tendency to cluster; cities are dense in some areas and very sparse in others — completely absent in the oceans that make up 7/10 of the earth's surface. The area in the South Pacific where there are no large cities is considerably larger than the areas around the North or South Pole.

In short, what Kemp and Tenenbaum did in this example is that they cherry-picked data to induce a structure that sound impressive but is actually meaningless in terms of the semantics of the data, using an inductive bias that bears no relation to the semantics of the data, searching through a space of models that does not contain models of the correct type.

5.4.4 Domingos and the Master Algorithm

The techniques of corpus-based machine learning that have recently been particularly successful, such as deep learning, are mostly highly specific in their focus and do not attempt to induce symbolic theories. Thus they are not directly relevant to Pavel. However, Domingos (2015), in his book *The Master Algorithm*, a survey of machine learning techniques, speculates as follows:

> The Master Algorithm is the germ of every theory: all we need to add to it to obtain theory x is the minimum amount of data required to induce it. (In the case of physics, that would be the results of perhaps a few hundred key experiments).

Domingos' "Master Algorithm" is a universal machine learning algorithm, which can optimally induce theories from data. He takes this as the Holy Grail of machine learning, and considers that it may well be found in the not very distant future. So his claim is that, in principle, one could choose a few hundred experiments that, given as input to the Master Algorithm, would enable the algorithm to induce all of physics.

I presume that Domingos is thinking here of something akin to the formulation in BACON; the input is a digested table of numbers, the target output is the foundational theories. Even so, "a few hundred" seems to me a huge underestimate. If the intended input is something close to a realistic description of the experiment, then the estimate of the number of experiments is surely off by at least a couple of orders of magnitude. (Not that it is always easy to individuate or count number of experiments; how are astronomical observations counted, for example?) Finally, cherry-picking only the evidence supporting the eventual theories is an unrealistic and ecologically invalid undertaking; a true logical reconstruction of science would have to take into account all the evidence that doesn't fit well, or is irrelevant. Still, in general what Domingos is suggesting here is somewhat comparable to Pavel.

6 Potential Philosophical Impact

It seems to me that implementing some part of Pavel might well yield insights that would be of interest to philosophers of science, on issues such as the nature of informal argumentation in physics, the sufficiency of physics as an explanation, the nature of the reduction of the other sciences to

physics, and the universalizing claims made by physics. Even if PAVEL takes an approach to issues that the philosophers found unacceptable from a philosophical standpoint, still the existence of one clear-cut approach to these issues would be valuable, if only as a point of comparison.

Another point where the construction of PAVEL might shed light is on the nature of the prior expectations. There seems to have been an enormous pressure over the centuries of the development of physics to find theories that are governed by a small, simple dynamic theory, at an almost arbitrary cost in the complexity of the boundary conditions; that are local in time and space; that are universal; that obey various kinds of symmetry; that conform to mathematically elegant equations;[14] and that are mechanistic. It would seem, moreover, that the preferences for these kinds of theories are stronger than be accounted for in terms of minimum description length or other such general principles. One evidence for this is that, historically, scientists were eager to claim the universality of physics long before, in retrospect, it would seem that the state of the data or theory came close to justifying it. For example, in 1814, Laplace contemplated

> An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies in the universe and those of the tiniest atom; for such an intellect, nothing would be uncertain and the future just like the past would be present before its eyes.

Laplace had every justification to say this of the solar system, having worked it out himself. But what evidence did he have that this applied to all the other motions in the universe, considering what a small fraction of motions the science of his time could actually explain or predict?

Assuming that this is right, are these preferences necessary, as prior preferences? Are they in any well-defined sense rational? Perhaps they are merely expressions of the existing power structure, in a Foucaultian sense.

At this point, I have to confess, I find myself seduced by the siren song of Bayesianism. It would be so wonderful to be able to assign a numerical confidence to the theory of gravity, or to Schrödinger's equation, or to the universalizing claims discussed in section 3.7! Or to determine to what extent any particular experimental finding should increase or decrease our confidence in any particular theory. It seems like it should be so close, comparatively speaking! The equation is sitting there, in section 3.2; all we have to do is to find well-founded values for the numbers.

7 Conclusions: Whither PAVEL?

There is no lack of things to do: there are easy things to do in the short term and harder things to do in the long term. The most important directions, it seems to me, would be:

- To increase the collection of physical theories we have in forms that can be used in a theorem prover.
- To develop techniques for choosing suitable idealizations, approximations, and abstractions for a given situation.
- To analyze the nature of the informal argumentations used in physics.

[14] Hossenfelder (2018) argues that the fetishizing of mathematical elegance is responsible for the stagnation of fundamental physics over the last few decades.

- To validate the approach by showing how word problems and experiments can be verified in these theories.
- To further validate the representation of word problems by developing natural language system that can translate verbal statements into formal representations.

Acknowledgments

Thanks for useful information and helpful feedback to Scott Aaronson, Alan Bundy, Ken Forbus, Tom LaGatta, Michael Strevens, David Tena Cucala, Peter Winkler, and the anonymous reviewer.

References

Appel, K. & Haken, W. (1977). The solution of the four-color-map problem. *Scientific American* 237(4):108-121.

Avigad, J., Donnelly, K., Gray, D., & Raff, P. (2007). A formally verified proof of the prime number theorem. *ACM Transactions on Computational Logic (TOCL), 9:*(1).

Bailey, D.H. & Borwein, J. (2015). Experimental computation as an ontological game changer: The impact of modern mathematical computation tools on the ontology of mathematics. In Davis, E. & Davis, P. (eds.) *Mathematics, Substance and Surmise: Views on the Meaning and Ontology of Mathematics,* Springer.

Beech, M. (2014). *The Pendulum Paradigm: Variations on a Theme and the Measure of Heaven and Earth.* Brown Walker Press.

Berger, M. (2010). *Geometry Revealed: A Jacob's Ladder to Modern Higher Geometry.* Springer.

Bhaskar, R. (1979). Realism in the natural sciences. In *Proceedings of the Sixth International Congress of Logic, Methodology and Philosophy of Science.*

Bobrow, D. (ed.) (1985) *Qualitative Reasoning about Physical Systems.* MIT Press.

Boender, J., Kammüller, F. & Nagarajan, R. (2015). Formalization of Quantum Protocols using Coq. In Huenen, C., Selinger, P. and Vicary, J. (eds.) *12th International Workshop on Quantum Physics and Logic,* pp. 71-83.

Bridewell, W. & Langley, P. (2010). Two Kinds of Knowledge in Scientific Discovery. *Topics in Cognitive Science,* **2:**36-52.

Bundy, A., Byrd, L., Luger, G., Mellish, C. & Palmer, M. (1979). Solving mechanics problems using meta-level inference. *IJCAI-79,* 1017-1027.

Cappellaro, P. (2012). *Quantum Theory of Radiation Interaction.* MIT Open Courseware. https://ocw.mit.edu/courses/nuclear-engineering/
22-51-quantum-theory-of-radiation-interactions-fall-2012/lecture-notes/

Corry, L. (2004). *David Hilbert and the Axiomatization of Physics (1898-1918): from* Grundlagen der Geometrie *to* Grundlagen der Physik. Kluwer.

Davis, E. (1988). A logical framework for commonsense predictions of solid object behavior. *AI in Engineering,* **3**(3):125-140

Davis, E. (1993). The kinematics of cutting solid objects. *Annals of Mathematics and Artificial Intelligence,* **9**(3,4):253-305.

Davis, E. (2008a). Physical reasoning. In van Harmelen, F., Lifschitz, V., & Porter, B. *The Handbook of Knowledge Engineering,* Elsevier. 597-620.

Davis, E. (2008b). Pouring liquids: A study in commonsense physical reasoning. *Artificial Intelligence, 172:* 1540-1578.

Davis, E. (2011). How does a box work? A study in the qualitative dynamics of solid objects. *Artificial Intelligence, 175*, 299-345.

Davis, E. (2017). Logical formalizations of commonsense reasoning: A survey. *JAIR* **59**:651-723.

Davis, E. (in prep.) *The Logic of Coal, Iron, Air, and Water: Representing Common Sense and Elementary Science.*

Davis, E. & Marcus, G. (2014). The scope and limits of simulation in cognition. https://arxiv.org/abs/1506.04956

Davis, E. & Marcus, G. (2016). The scope and limits of simulation in automated reasoning. *Artificial Intelligence,* **233:**60-72.

Davis, E., Marcus, G. & Frazier-Logue, N. (2017). Commonsense Reasoning about Containers using Radically Incomplete Information. *Artificial Intelligence,***248:**46-84.

Davis, P.J. (1993). Visual Theorems. *Educational Studies in Mathematics,* **24**(4): 333-344.

Dehaene, S. (1997). *The Number Sense: How the Mind Creates Mathematics.* Oxford University Press.

de Kleer, J. & Brown, J.S. (1985). A qualitative physics based on confluences. In Bobrow, D. (ed.) *Qualitative Reasoning about Physical Systems.* MIT Press.

Domingos, P. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World.* Basic Books.

Dwyer, Joseph & A. Uman, Martin. (2013). The physics of lightning. Physics Reports. 534. . 10.1016

Feynman, R., Leighton, R.B., & Sands, M. (1964). *The Feynman Lectures on Physics.* Addison-Wesley.

Friedman, S., Forbus, K. & Sherin, B. (2017). Representing, Running, and Revising Mental Models: A Computational Model. *Cognitive Science:* 1-36

Gonthier, G. et al. (2013). A Machine-Checked Proof of the Odd Order Theorem. *International Conference on Interactive Theorem Proving*, *Lecture Notes in Computer Science* vol. 7998, 163-179.

Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science,* **337**(6102), 1623-1627.

Gunning, D. et al. (2010). Project Halo Update — Progress Toward Digital Aristotle. *AI Magazine*, **31**(3), 33-58.

Gutierrez, T.D. (1999). Standard Model Lagrangian. http://nuclear.ucdavis.edu/ tgutierr/files/stmL1.html

Hales, T. et al. (2015). A formal proof of the Kepler conjecture. http://arxiv.org/abs/1501.02155

Hamel, G.K.W. (1912). *Elementare Mechanik.* Leizig and Berlin: Teubner.

Hamel, G.K.W. (1921). *Grundbegriffe der Mechanik.* Leipzig and Berlin: Teubner.

Harrison, J. (2006). Formal verification of floating point trigonometric functions. In *International*

Conference on Formal Methods in Computer-Aided Design 254-270.

Harrison, J. (2009). Formalizing an analytic proof of the prime number theorem. *Journal of Automated Reasoning, 43:*(3), 243-261.

Hayes, P. (1979). The naïve physics manifesto. In *Expert Systems in the Micro-electronic Age,* In Michie, D. (Ed.) Edinburgh University Press.

Hayes. P. (1985). Ontology for liquids. In Hobbs, J. & Moore, R. (Eds.) (1985). *Formal Theories of the Commonsense World.* ABLEX Publishing

Hendry, R.F. (1999). Chemistry and the completeness of physics. in Symons, J., van Dalen, D. & Davidson, D. (eds.) *Philosophical Dimensions of Logic and Science : Selected Contributed Papers from the 11th International Congress of Logic, Methodology, and Philosophy of Science, Kraków.* 165-178.

Henkin, L., Suppes, P. & Tarski, A. (1958). *The axiomatic method with special reference to geometry and physics.* North-Holland.

Hertz, H. (1894). *Die Prinzipien der Mechanik.* Leipzig: Johann Ambrosius Barth.

Hornby, G. S., Globus, A., Linden, D. S., & Lohn, J. D. (2006). Automated antenna design with evolutionary algorithms. *AIAA Space* 19-21.

Hossenfelder, S. (2018). *Lost in Math: How Beauty Leads Physics Astray.* Basic Books.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: the Bayesian approach.* Open Court Publishing.

Jaynes, E.T. (2003). *Probability Theory: The Logic of Science.* Cambridge University Press.

Jeannin, J. B., et al. (2017). A formally verified hybrid system for safe advisories in the next-generation airborne collision avoidance system. *International Journal on Software Tools for Technology Transfer,* **19**(6), 717-741.

Kemp, C. & Tenenbaum, J. (2009). The discovery of structural form. *Proceedings of the National Academy of Sciences,* **105**(31), 10687-10692.

Khashabi, D., Khot, T., Sabharwal, A. & Roth, D. (2018). Question answering as global reasoning over semantic abstractions. *AAAI-18.*

Khot, T., Sabharwal, A. & Clark, P. (2018). SciTail: A Textual Entailment Dataset from Science Question Answering *AAAI-18.*

Krenn et al. (2016). Automated search for new quantum experiments. *Physical Review Letters,* **116**(9), 090405.

Kushman, N., Artzi, Y., Zettlemoyer, L., & Barzilay, R. (2014). Learning to Automatically Solve Algebra Word Problems. *ACL-2014.*

Langley, P., Bradshaw, G. L., & Simon, H. A. (1981). BACON. 5: The discovery of conservation laws. *IJCAI-81,* 121-126.

Langley, P., Bradshaw, G. L. & Simon, H.A. (1983). Rediscovering chemistry with the BACON system. In Michalski, R.A. et al. *Machine Learning,* Springer-Verlag.

Laughlin, R. & Pines, D. (2000). The theory of everything. *Proceedings of the National Academy of Sciences,* **97**(1), 28-31.

Lehmann, J., Chan, M. & Bundy, A. (2013) A higher-order approach to ontology evolution in physics. *Journal of Data Semantics,* **2:**163-187.

Ludwig, G. (1985). *An Axiomatic Basis for Quantum Mechanics: Vol 1, Derivation of Hilbert Space Structure.* Springer-Verlag.

Ludwig, G. (1987). *An Axiomatic Basis for Quantum Mechanics: Vol 2, Quantum Mechanics and Macrosystems.* Springer-Verlag.

Ludwig, G. (1989). An axiomatic basis as a desired form of a physical theory. *Proceedings of the 1987 International Congress for Logic, Methodology, and the Philosophy of Science.* North Holland Publishing Co.

Martin, U. and Pease, A. (2015). Hardy, Littlewood, and polymath. In Davis, E. & Davis, P. (eds.) *Mathematics, Substance and Surmise: Views on the Meaning and Ontology of Mathematics,* Springer.

MathOverflow (2016). Is it possible to have a research career while checking the proof of every theorem you cite?
https://mathoverflow.net/questions/237987/is-it-possible-to-have-a-research-career-while-checking-the-proof-of-every-theor

McCarthy, J. (1968). Programs with Common Sense. In Minsky, M. (ed.) *Semantic Information Processing.* MIT Press.

McCune, W. (1997). Solution of the Robbins problem. *Journal of Automated Reasoning,* **19**3:264-276.

McKinsey, J.C.C., Sugar, A.C., & Suppes, P. (1953). Axiomatic foundations of classical particle mechanics. *J. Rational Mechanics and Analysis*, **2**, 253-272.

Mehlberg, H. (1965). Space, Time, Relativity. *Proceedings of the 1964 International Congress for Logic, Methodology, and the Philosophy of Science.* North Holland Publishing Co.

Melnikov, A. et al. (2017). Active learning machine learns to create new quantum experiments. *PNAS early edition.*

Montague, R. (1974). Deterministic theories. In Thomason, R. (ed.) *Formal Philosophy: Selected Papers of Richard Montague,* Yale University Press.

Newell, A. (1982). The knowledge level. *Artificial Intelligence,* **18**(1): 87-127.

Nipkow, T., Paulson, L. C., & Wenzel, M. (2002). *Isabelle/HOL: A Proof Assistant for Higher-Order Logic* Lecture Notes in Computer Science #2283, Springer.

Paleo, B. W. (2012). Physics and proof theory. *Applied Mathematics and Computation,* **219:** 45-53.

Rosenkrantz, R. (1977). *Inference, Method, and Decision: Towards a Bayesian Philosophy of Science.* Springer

Russell, B. (1903). *The Principles of Mathematics.* Cambridge University Press.

Sant'Anna, A. (1999). An axiomatic framework for classical particle mechanics without space-time. *Philosophia Naturalis,* **36**:307-319.

Schwartz, J. (1960). The pernicious influence of mathematics on science. *Proceedings of the 1960 International Congress for Logic, Methodology, and the Philosophy of Science.* North Holland Publishing Co.

Sigmund, K. (2017). *Exact Thinking in Demented Times: The Vienna Circle and the Epic Quest for the Foundations of Science.* Basic Books.

Slemrod, M. (2013), From Boltzmann to Euler: Hilbert's 6th problem revisited. *Computers & Mathematics with Applications,* **65**(10): 1497-1501.

Smith, K. A., Battaglia, P., & Vul, E. (2013). Consistent physics underlying ballistic motion prediction. *Cognitive Science*.

Sober, E. (2002). Bayesianism — its scope and limits. *Proceedings of the British Academy,* **113**:21-38.

Souyris, J., Wiels, V., Delmas, D., & Delseny, H. (2009) Formal verification of avionics software products. *International Symposium on Formal Methods,* 532-546.

Strevens, M. (2005). The Bayesian approach in the philosophy of science. In D. M. Borchet (Ed.), *Encyclopedia of philosophy (2nd ed.)* Detroit: Macmillan Reference.

Strevens, M. (2013). *Tychomancy: Inferring Probability from Causal Structure.* Harvard University Press.

Suppes, P., Luce, R.D., Krantz, D. & Tversky, A., (1974) *Foundations of Measurement*, Dover Pubs.

Tegmark, M. (2009). The multiverse hierarchy. arXiv preprint arXiv:0905.1283.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science,* **331**(6022), 1279-1285.

van Harmelen, F., Lifschitz, V. & Porter, B. (Eds.) (2008). *Handbook of Knowledge Representation.* Elsevier.

Vogt, A. (1997). Review of *An Axiomatic Basis for Quantum Mechanics: Vol 1, Derivation of Hilbert Space Structure* by Günther Ludwig, *SIAM Reviews,* **29**:3, 499-501.

Wallace, D. (to appear). The Logic of the Past Hypothesis. In Loewer, B., Weslake, B. & Winsberg, E. (eds.) *Time's Arrows and the Probability Structure of the World,* Harvard University Press.

Weidenbach, C., Dimova, D. Fietzke, A., Kumar, R. Suda, M. & Wischnewski, C. (2009). Spass Version 3.5. *Intl. Conf. on Automated Deduction (CADE), LNCS 5563,* 140-145.

Whitehead, A.N. & Russell, B. (1910). *Principia Mathematica.* Cambridge University Press.

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus.* Harcourt, Brace, and Company.

Yandell, B.H. (2002). *The Honors Class: Hilbert's Problems and Their Solvers.* A.K. Peters.