Philosophers' Imprint

volume 20, no. 3
january, 2020

THE PRINCIPLE OF STABILITY

Samuel C. Fletcher

Department of Philosophy
University of Minnesota, Twin Cities

Abstract

How can inferences from models to the phenomena they represent be justified when those models represent only imperfectly? Pierre Duhem considered just this problem, arguing that inferences from mathematical models of phenomena to real physical applications must also be demonstrated to be approximately correct when the assumptions of the model are only approximately true. Despite being little discussed among philosophers, this challenge was taken up (if only sometimes implicitly) by mathematicians and physicists both contemporaneous with and subsequent to Duhem, yielding a novel and rich mathematical theory of stability with epistemological consequences.

1. Introduction

Much philosophical work on "first principles" in science has focused on what might be called *first-order* (first) principles, which provide the possibility for or constrain the content of the theory's subject matter. For instance, one of the first modern systematic monographs on the subject, Northrop's *Science and First Principles*, took up "dissecting the given scientific theories which our technical scientists have verified, to determine what concepts or principles are taken as primary or undefined" (1931, p. ix).[1] By contrast, I aim here to exposit a *second-order* (first) principle — the titular *principle of stability* (PS) — which delimits not what a theory can say about the phenomena it concerns, but what one, as a user of a theory, can justifiably infer about phenomena given a theory that models it.[2] Here is an informal rendering:

Principle of Stability (informal) An inference from the statement that a property of a model holds to the statement that the property of phenomena (or some possible world) it represents holds is

[1] For earlier and later expositions, see also Whewell (1847) and Dilworth (1994, 2007), and Ivanova and Farr (2015) for a collection of recent papers on conventional and constitutive principles in science, especially physics.

[2] Throughout, I understand 'inference' to denote a particular (generic) token of drawing a conclusion, rather than the type denoted by 'entailment relation' or 'inference rule/schema.'

> justified only if all sufficiently similar models also have that property.

Such a property is said to be *stable*, from which the PS derives its name. (Except in the inset statements of the PS, I will abbreviate 'an inference from the statement that p holds to the statement that q holds' to 'an inference from p to q' to eliminate the grammatically ungainly propositional bracket.)

What are the origins of the PS, and how does it compare to other similar principles? Can it be made more precise? What arguments support it, and what consequences follow from it? Answering these fourfold historical, comparative, formal, and epistemological questions occupy the sequel. In particular, I begin in section 2 showing that the PS can in fact be traced definitively at least as far back to Duhem (1954/1914). In this, he presages the reinvention of the principle under various guises and levels of generality by about a half-century. (Comparison with these versions, articulated by philosophers of science and epistemologists, occupies appendix A. The upshot is that there is a sense in which Duhem's PS, in modern language, is still distinct and powerful.)

The PS must not, however, be mistaken for a metaphysical principle, as some early workers in dynamical systems theory and recent commentators have. Showing this in section 3 provides a natural segue to a discussion of how the PS interacts with that theory. The far future states of dynamical systems, it is well known, can exhibit sensitive dependence on the details of their initial states. This presents what might seem to be a puzzle: Duhem presents an example of an unstable property of a system that is taken to be stable in another sense by his contemporary Hadamard. I argue that this is not in conflict with the PS, but to show this in more detail requires making the concepts of stability and similarity more precise. Following work by mathematical physicists, in section 4 I do so through topological structure on models. This structure is not itself a part of a theory, but contextually depends on the interface with experiment, independently specified empirical knowledge, and the aims of inquiry.

I explore in section 5 one argument for the PS, according to which the PS is justified in part as a "contingent a priori principle" — one to which one commits solely in virtue of engaging in most types of representational model-based reasoning. After addressing two objections to this justification, the PS appears to be of unusually wide scope, applying in principle to inferences from properties of models in any science when one can describe those models with sufficient precision. Whether this justification is tenable, acceptance of the PS has implications for other questions in epistemology and philosophy of science. I adumbrate these for future research and summarize the foregoing sections in the concluding section 6.

2. Pierre Duhem's Principle of Stability

In part II, chapter III ("Mathematical Deduction and Physical Theory") of *The Aim and Structure of Physical Theory*,[3] Pierre Duhem (1954/1914) argued that inferences from mathematical models of phenomena — what he called "mathematical deductions" — must satisfy an additional criterion if they are to be justified as predictions or descriptions. The basis of this criterion arises from the nature of the representational relationship between theory and phenomena. Duhem (1954/1914, p. 133) observed that

> at both its starting and terminal points, the mathematical development of a physical theory cannot be welded to observable facts except by a translation . . . which replaces the language of concrete observation by the language of numbers; . . . But translation is treacherous,

for the completely precise theoretical facts of mathematical models belie the vague and imprecise practical facts of actual measured phenom-

[3] Despite being included in Duhem's most influential contribution to philosophy of science, this chapter is little discussed among philosophers. An exception is Schmidt (2011, 2017), whose different interpretation of Duhem I critically discuss in section 3.

ena.

The metaphor of "translation" for *representation* thus emphasizes its ability to distort and the gap between the abstract, mathematical models of a theory and the concrete, physical phenomena of the world. Consequently, in general, "An infinity of different theoretical facts may be taken for the translation of the same practical fact" (Duhem, 1954/1914, p. 134). For instance, the theoretical facts of 10°C, 9.99°C, and 10.01°C in a mathematical model of temperature could equally well represent the same practical fact of 10°C as determined by a thermometer with an accuracy of only 0.2°C. So when one reasons with a theory, one must translate practical facts into the appropriate variety of theoretical facts as inputs, and in deducing from the theory consequences, one must then translate its theoretical consequences into practical facts. This final translation yields one or more practical facts, but only when it yields exactly one does the theory successfully justify inference: "A mathematical deduction ... may therefore be useful or otiose, according to whether or not it permits us to derive a *practically definite* prediction of the result of an experiment whose conditions are *practically given*" (Duhem, 1954/1914, p. 137). Moreover, the extent to which it is so given is not fixed a priori, but rather is relative to the sensitivity of the experimental apparatus, and more generally, just what properties can be measured.

These considerations reveal that only certain deductions from mathematical models to physical phenomena can be justified, as such deductions must be faithful to the translation between phenomenal practical facts and a "bundle" of theoretical facts within the model. As a tool for making justified inferences about phenomena, therefore,

> a mathematical deduction is of no use to the physicist so long as it is limited to asserting that a given *rigorously* true proposition has for its consequence the *rigorous* accuracy of some such other proposition. To be useful to the physicist, it must still be proved that the second proposition remains *approximately* exact when the first is only *approximately* true. (Duhem, 1954/1914, p. 143)

Duhem's invocation of "usefulness" refers to what he takes to be the goal of physical theory as an economical representation of experimental laws that describe and predict phenomena to facilitate precise surrogative reasoning about that phenomena. Ideally, the ranges of these approximations should be quantified, delimited, and controlled, even if only probabilistically,[4] but even when these ranges are unknown, one can still demand that the above criterion hold for *some* degree of approximation or other. Understanding approximation as a kind of similarity, one can then see this statement as a clear expression of a version of the PS.

It seems that Duhem was motivated to formulate this criterion by the work of Poincaré and Hadamard on celestial and rational mechanics, particularly the problems of the long-term trajectories of bodies and the "stability" of the solar system — not to be confused with the sense of stability found in the PS. Rather, the central question of the "stability" of the solar system is whether each planet will remain within a bounded area for the whole of its future evolution, or will instead escape to infinity; hence, it might be rechristened the problem of the *boundedness* of the solar system. Hadamard (1898) had demonstrated how to calculate the exact trajectories of a particle sliding frictionlessly on a surface of negative curvature, shaped not unlike that of a bull's head but with the ears and horns extended to infinity. He showed that arbitrarily small variations in the initial position and velocity of the particle could change whether it would remain in a bounded area for all time or eventually shoot off to infinity. Duhem (1954/1914, p. 141) explicitly cites the question of the boundedness of a particle's trajectory in this example as one whose mathematical answers could never be utilized in physics. For the same reasons, regarding the question of the boundedness of the trajectories of the planets of the solar

[4] The discussion of the interaction between the PS and probability theory — or measure-theoretic considerations more generally — is beyond the scope of this essay. However, see the suggestions for further research in the concluding section 6.

system,[5]

> If such a circumstance analogous to the one offered by Hadamard's problem should turn up here, any mathematical deduction relative to the stability [boundedness] of the solar system would be for the physicist a deduction that he could never use.
>
> One cannot go through the numerous and difficult deductions of celestial mechanics and mathematical physics without suspecting that many of these deductions are condemned to eternal sterility. (Duhem, 1954/1914, pp. 142–143)

The "eternal sterility" of a deduction is its failure to meet his criterion for any degree of approximation: Not only does the present degree of measurement accuracy fail to license the deduction, but any further improvement of it would as well.[6] For Duhem, then, the question of whether inferences from mathematical models satisfy his criterion is not confined to fanciful theories unintended to represent actual phenomena, but applies substantively to actual successful scientific theories of interest.

3. Stability, Instability, and Dogma

It is important to distinguish the PS, as one finds its expression in Duhem (1954/1914) in particular, from other, superficially similar principles regarding idealization and the so-called *stability dogma* in the theory of dynamical systems. As I discuss in section 3.1, while the PS has a similar form to various principles concerning the interpretation of idealized models in science, it is distinct from them in that it is an epistemological (i.e., inferential) principle rather than a metaphysical one, and it provides a necessary rather than a sufficient condition for inference to be justified. Explaining what the stability dogma is, and why Duhem's position and the PS are definitely distinct from it (in section 3.3), requires a short informal excursion into the motivations of that theory and some of its concepts (in section 3.2). This allows one to show that a particular type of stability discussed in that theory — structural stability — falls under the auspices of the PS, and that modern attitudes towards it among practitioners mirror the PS. These considerations from dynamical systems theory in turn raise a prima facie puzzle about the toy example by Hadamard that Duhem discussed, which I outline in section 3.4; the following section 4 goes on to develop formalism that helps to resolve the puzzle.

3.1 Principles Regarding Idealization

As I outlined in section 2, Duhem's motivation in "Mathematical Deduction and Physical Theory" to constrain which inferences from models were physically justified originated in the imprecision of empirical data that one provides to the models to make predictions. In more contemporary terminology, one might say that one *idealizes* the inputs into the model to make the model's use feasible. But the fact that it is the input data which is idealized, rather than (or even in addition to) some other aspect of the model, isn't so essential: The PS constrains inferences from *any* idealized models.

Philosophers of science have indeed long recognized that there must be some connection between idealization and justification of in-

[5] In fact, Hadamard (1901, p. 14) had already made this comparison — that is, between the question of the long-term boundedness ("stability") of the solar system and the trajectories of the particles in Hadamard (1898) — but only concludes that this could lead one to draw erroneous predictions. He stops short of drawing any general epistemological conclusions. (See also footnote 15.)

[6] This example also illustrates that Duhem took his concerns to apply equally to qualitative predictions — here, whether the solar system will remain bounded — as to quantitative predictions, e.g., the exact future location of the planets. This is in contrast to claims by Franceschelli (2014, p. 122; my translation) that for Duhem, "the guarantee of validity of a physical theory is its ability to make experimentally verifiable quantitative predictions." Qualitative predictions are equally important for Duhem, even if much of the mathematical apparatus now used to demonstrate their stability was developed afterwards — cf. sections 3.2 and 4.

ferences from models, but what that connection is supposed to be has varied. For example, van Fraassen (2008, pp. 52–53) calls the following formulation the Principle of Approximation, which he attributes to Reichenbach (1956, pp. 93–95): "if certain conditions follow from the ideal case, then approximately those conditions will follow from an approximation to the ideal case." He notes that it cannot apply to all conditions; the ones to which it will apply will be context-dependent. This principle is epistemological, like the PS, but provides a *sufficient* condition for inferences for other models rather than a *necessary* one for inferences from the (perhaps idealized) model under consideration.

However, one can find an expression of a similar principle much earlier, in James Clerk Maxwell's *Matter and Motion*:[7]

> "That like causes produce like effects." This is only true when small variations in the initial circumstances produce only small variations in the final state of the system. In a great many physical phenomena this condition is satisfied; but there are other cases in which a small initial variation may produce a very great change in the final state of the system. (Maxwell, 1925/1877, pp. 13–14)

Like van Fraassen, Maxwell stresses that the applicability of this principle cannot be completely general. Unlike van Fraassen (and Duhem), Maxwell took this to be a metaphysical principle rather than an epistemological one. Indeed, not only does he use causal language to formulate it, but he sees it as a constraint on the formulation of physical law and determinism:[8]

> [O]nly in so far as stability subsists that principles of natural law can be formulated: it thus perhaps puts a limitation on any postulate of universal physical determinacy such as Laplace was credited with. . . .
>
> We may perhaps say that the observable regularities of nature belong to statistical molecular phenomena which have settled down into permanent stable conditions. In so far as the weather may be due to an unlimited assemblage of local instabilities, it may not be amenable to a finite scheme of law at all. (Maxwell, 1925/1877, pp. 13–14)

Thus Maxwell views stability — here understood in the sense of "small variations" from the foregoing quotation — as a constraint on the formulation of physical law itself.[9]

More recent authors have also expressed similar principles constraining the interpretation (if not the construction) of physical theories using idealized models. For example, what Jones (2006) and Landsman (2013) call Earman's Principle states that "no effect [predicated by a model] can be counted as a genuine physical effect if it disappears when the idealizations [of the model] are removed" (Earman, 2004, p. 191). Similarly, there is what Landsman (2013) calls Butterfield's Principle, that for genuine emergent properties of idealized models, "there is a weaker, yet still vivid, novel and robust behaviour that occurs before we get to the [idealized] limit . . . And it is this weaker behaviour which is physically real" (Butterfield, 2011, p. 1065). Unlike Maxwell's (and van Fraassen's) stated principles (but like the PS), these are necessary conditions rather than sufficient ones. But like Maxwell's (and unlike the PS), they are metaphysical or interpretive principles rather than purely epistemological or inferential ones: Instead of merely constraining what one can infer about phenomena

[7] In fact, one might even go further back to Leibniz's Law of Continuity (d'Alembert, 2008/1754) and its precursors in Kepler and Cusanus, but one finds vague, striving, and unqualified expressions in these authors (Boyer, 1959/1949); to my knowledge, Maxwell appears first to give a clear and definite statement to the principle.

[8] Although the contemporary distinction between determinism and predictability was not available to Maxwell, charitably he would mean the latter in this case (Earman, 1986, 2007).

[9] The connection between determinism and laws of nature was perhaps more immediate to those in the nineteenth century, as determinism was widely (though not exclusively) taken to be a necessary feature of physical theory (van Strien, 2014). (Cf. footnote 8.)

from one's models, they delimit what physical possibilities those models can describe.[10]

3.2 *Dynamical Systems and the Stability Dogma*

The work of Hadamard (1898) that Duhem (1954/1914) discussed was a precursor to the modern theory of dynamical systems, which concerns the description of the changes of the state of a system over time. In abstract terms, a dynamical system consists of a state space representing the possible states in which the considered system could be, and a dynamical rule that prescribes how that state changes over time (Meiss, 2007). A curve in the state space determined by that dynamical rule then depicts the history of the dynamical system, its trajectory through the state space over time. For example, in simple mechanical systems, the state space consists of the position and velocity of the mechanical components, such as those of particles sliding on a surface. The equation of motion for those particles is given by Newton's second law, and the collection of all histories in state space — the so-called flow for or phase portrait of the system — depicts how the positions and velocities of any particle changes over time.

An important concept in dynamical systems theory is that of structural stability, first formally introduced by Andronov and Pontrjagin (1937) (but under the name "rough system"). Informally, a dynamical system is structurally stable just when all systems with sufficiently similar dynamical rules produce qualitatively similar phase portraits.[11] For example, if there are any fixed points in the phase portrait of a particular structurally stable dynamical system — states at which the system is unchanging under the dynamical rule — they and their qualitative features (such as how trajectories approach or recede from them) will be the same for all sufficiently similar dynamical systems, even though other particular trajectories in the phase portrait may be perturbed. In this sense, structural stability is a type of stability in the sense described by the PS, specifically regarding the qualitative features of a dynamical system's phase portrait.

By the lights of Andronov and Pontrjagin (1937) and other early dynamical systems theories, "structurally unstable systems were seen as somehow suspect. . . . [So] structural stability was *imposed* as an *a priori* restriction on 'good' models of physical phenomena" (Guckenheimer and Holmes, 1983, p. 259). This is the *stability dogma*. However, Guckenheimer and Holmes (1983, p. 259) argue that "The logic which supports the stability dogma is faulty," because it confuses the epistemological significance of stability for physical (or metaphysical) significance, that is, what we can infer for what we can describe as being possible. They suggest that a proper replacement for the stability dogma would

> state that the only properties of a dynamical system (or a family of dynamical systems) which are *physically relevant* are those which are preserved under perturbations of the system. . . . This is quite different from the original statement that the only good systems are ones with *all* of their qualitative properties preserved by perturbations. (Guckenheimer and Holmes, 1983, p. 259)

By "*physically relevant*" they mean "*verifiable physical properties*" (Guckenheimer and Holmes, 1983, p. 259) that are limited by the precision of measurement and specification of the system in question, and by "good systems" they simply mean those that genuinely represent physical phenomena. This replacement is merely a statement of the PS in the context of qualitative features of dynamical systems, and seems to

[10] A confusing feature of these principles is that they treat models as being either idealized or not; but surely idealization comes in degrees and is relative to the phenomena one wants to represent. In a slogan: no idealization without (mis-)representation. Further analysis of these problems for principles concerning idealization are beyond the scope of the present essay.

[11] More precisely, given the set of self-diffeomorphisms of a compact smooth manifold M, a particular such diffeomorphism $f : M \to M$ is structurally stable when there is a neighborhood of f in the C^1 (compact-open) topology, such that, for each diffeomorphism g in that neighborhood, there is a homeomorphism $h : M \to M$ such that $h \circ f = g \circ h$ (Pugh and Peixoto, 2008).

be the new consensus among dynamical systems theorists.[12]

3.3 *Duhem and the Stability Dogma*

Why can't Duhem be read as an early advocate for the stability dogma, rather than the PS? Indeed, Schmidt (2011, 2017) has interpreted Duhem as proposing to constrain which models can represent physical phenomena to those with stable properties, because without such a restriction, scientific methodology would be undermined:

> Duhem believed . . . that exact sciences are threatened by instabilities and, based on such considerations, he argued in favor of stability, formulated by a stability requirement, and pursued what later has become known as a *stability dogma*. (Schmidt, 2017, p. 432)

In particular, he takes Duhem to have held that

> Inasmuch as only stable theories (with non-diverging deductions) are assumed to represent a *physical* phenomenon, a phenomenon is seen as a *physical* phenomenon if, and only if, it is stable (Schmidt, 2017, p. 424).

A stable theory in this sense is one whose models have only stable (empirical) properties — "non-diverging deductions".[13] Thus, Schmidt

(2011, 2017) places Duhem within the metaphysical tradition of Andronov and Pontrjagin (1937) and Maxwell (1925/1877) before him (for which, see section 3.1), who take stability to be a constraint on what is possible to represent in a physical theory.

There are at least two reasons why this interpretation is difficult to maintain. In the first place, Duhem did not take physical theory to be in the business of providing constraints on what phenomena count as physical; the goal of physical theory is rather concision, completeness, and accuracy in representing measured experimental laws as they reflect the natural classification of the world. (See also footnote 13.)

The second reason is that this reading is hardly compatible with Duhem's discussion of mathematical astronomy. If Duhem really believed that just the theories whose models only have stable properties could represent physical phenomena, then his speculations about the boundedness ("stability") of the long-term evolution of the solar system within Newtonian mechanics would have led to a consequent skepticism towards the ability of that theory to represent phenomena at all, a skepticism wholly absent from his writings. Moreover, Duhem never suggests that the boundedness ("stability") of the solar system in the far future could be an unphysical phenomenon, only that astronomers may have no means to answer this question, which arises from

> the rigorous conditions that we are bound to impose on mathematical deduction if we wish this absolutely precise language to be able to translate without betraying the physicist's idiom, for the terms of this latter idiom are and always will be vague and inexact like the perceptions which they are to express. (Duhem, 1954/1914, p. 143)

In contrast, "The problem of the stability [boundedness] of the solar system certainly has a meaning for the mathematician" (Duhem, 1954/1914, p. 142) because the mathematician need only concern him-

[12] While this consensus is not complete — see, e.g., Gyenis (2013) for a recent novel defense of the stability dogma — it does seem to be the dominant position. See also Batterman (2002, ch. 4.4), especially pages 58–59.

[13] In this context, Schmidt (2017, p. 424) continues: "methodology constitutes and constructs reality" for Duhem because he is a scientific "conventionalist" for whom "the presupposed mathematical structure of theories always plays a central role in the constitution of physical phenomena." It isn't clear how to reconcile this position on Duhem's philosophy of science with statements such as: "A physical theory . . . is a system of mathematical propositions, deduced from a small number of principles, which aim to represent as simply, as completely, and as exactly as possible a set of experimental laws" (Duhem, 1954/1914, p. 19), or "We have proposed that the aim of physical theory is to become a natural classification, to establish among diverse experimental laws a logical coordination serving as a sort of image and reflection of the true order according to which the realities escaping us are organized" (Duhem, 1954/1914, p. 31). See also Ariew (2014, §2).

self with theoretical facts, not the imprecise practical facts to which the empirically minded astronomer is limited.

3.4 *Stability and Instability*

There is another connection between the PS and dynamical systems theory besides the former's superficial similarity with the stability dogma. Around the same time that Duhem was writing the manuscript for the first edition of *The Aim and Structure of Physical Theory*, Hadamard (1902, p. 49) introduced the idea of a *well-posed* ("bien posé") problem. Such a problem, which in the context of a dynamical system would be to find its trajectory (the solution) given its state at a time (the data), is well-posed just when:[14]

1. it has a solution;
2. that solution is unique; and
3. solutions depend continuously on the data.

The last condition is sometimes referred to as the *stability* condition for well-posedness, for it guarantees that sufficiently small changes in the data induce only small changes in the solution. Indeed, it is an example of stability in the sense of the PS: The (approximate) empirical properties of solutions to well-posed problems are stable, for sufficiently similar data yield sufficiently similar solutions. Hadamard (1923, p. 38), much like Maxwell (as described in section 3.1), took being well-posed to be a necessary condition for the mathematical problem to represent well a natural law; those that did not would be better described as stochastic.

Remarkably, Hadamard's problem from 1898, the one which Duhem considered, is well-posed![15] But this presents a prima facie puzzle: How can the trajectories of sliding particles be both stable in the sense of Hadamard, yet unstable in the sense relevant for Duhem's argument? The answer, of course, is that the two are applying *different* notions of approximation or similarity to the particle trajectories. I develop this resolution in the next section (4) using the mathematics of topology, showing how a topological structure on a set encodes in a way how its elements are similar to one another; different such structures correspond to different specifications of similarity, which in turn have different specifications of continuity and stability.

This mathematics was developed only after the primary work of Hadamard and Duhem under discussion was completed, and indeed, was in part inspired from it, for under its influence,

> analysts were then obliged to examine, as [Hadamard] says, the "different types of neighborhoods and continuity," which led unavoidably to functional spaces, general topology and functional analysis. (Mandelbrojt and Schwartz in Maz'ya and Shaposhnikova, 1999, p. 453)

Indeed, with uncanny prescience, Duhem recognized that demonstrating how inferences from models satisfy the PS, though epistemologically necessary, might require a level of mathematical sophistication "transcending the methods at the disposal of algebra today" (Duhem, 1954/1914, p. 143).

[14] Initially (Hadamard, 1902), the definition only included the first two conditions, as it was only later clear to mathematicians that they did not guarantee the third. To illustrate this, Hadamard (1923, pp. 33–34) uses the example of the Cauchy problem for the Laplace equation which he had first discussed in a lecture to the Swiss Mathematical Society in Zurich in 1917. (There, the term is rendered as "correctly set.") But even by then, the third condition is not included, and Hadamard would only do so himself late in life after others, such as Courant and Hilbert (1962/1937), had done so (Maz'ya and Shaposhnikova, 1999, p. 457).

[15] In fact, Hadamard (1901, p. 14) had already introduced the idea of an ill-posed ("mal posé") question to describe, potentially, the boundedness ("stability") of the solar system! However, he does not define this term explicitly but only in reference to the results of his 1898. (See also footnote 5.) At this point, the concepts were still vague, and indeed the definition he goes on to suggest in his 1902 would not make his 1898 ill-posed.

4. Stability, Topology, and Similarity

Since Duhem (1954/1914) wrote, physicists — especially those working in the general theory of relativity, the relativistic heir to Newtonian astronomy — have taken up a formalized version of the PS using topology, in particular topological structure on the models of the theory with which one is working:

> It is a general feature of the description of physical systems by mathematics that only conclusions which are stable, in an appropriate sense, are of physical interest. . . . To obtain a precise notion of stability in general relativity we must say what "sufficiently small perturbation" means, i.e., we must find a suitable topology on the space of solutions of Einstein's equations. (Geroch, 1971, p. 70)

These solutions are just the models of general relativity, the relativistic spacetimes. Similarly, Hawking (1971, p. 395) writes that "the only properties of space-time that are physically significant are those that are stable in some appropriate topology." Fletcher (2016, §2) has argued convincingly that when these authors ascribe "physical significance" to a property of a model, they mean that the property of phenomena or some possible world it represents can be inferred from the model. This is indeed the epistemological or inferential reading that one finds in the PS, which is to be distinguished from the metaphysical reading found in the stability dogma and better attributed to the variant phrase, "physical reasonableness," found in the literature on which models to exclude from a theory (Manchak, 2011).

I begin in section 4.1 with the general, abstract features of topologies relevant to the current discussion, leading to a formulation of the PS in these terms. This explication of the PS makes it more precise and applicable to the mathematical sciences. I also address what it means for a topology to be "appropriate" — cf. the foregoing quotations. This in turn reveals the means (in section 4.2) to resolve the puzzle between Hadamard's and Duhem's senses of stability adumbrated at the end of the previous section (3).

4.1 Topology as Similarity

Topology is often known informally as "rubber sheet geometry" because it concerns the properties of spaces that are preserved under continuous deformations — stretching and bending, but not tearing or puncturing. Apart from their geometrical applications, though, the notions of "closeness" among elements that a topological structure thereon describes serve well as abstract notions of similarity.

The standard definition of a topology is formulated in terms of an algebra of open sets, but for present purposes, it will be more illuminating to define them in terms of neighborhood systems (Willard, 2004/1970, §4). A neighborhood system for a set X is an assignment to each $x \in X$ of a nonempty set $N(x)$ of subsets of X satisfying the following conditions:

1. For every $U \in N(x)$, $x \in U$.
2. For every $U, V \in N(x)$, $U \cap V \in N(x)$.
3. For every $U \in N(x)$, if $V \supset U$ then $V \in N(x)$.
4. For every $U \in N(x)$, there is some $V \in N(x)$ such that for all $y \in V$, $U \in N(y)$.

A set equipped with such a neighborhood system can be called a topological space. Each $U \in N(x)$ is called a neighborhood of x, which makes sense in light of the first condition, which for our purposes is the most important, for it allows us to interpret a neighborhood of x as those elements that are qualitatively similar to x.[16] For example, the

[16] It is also all one needs in order to specify a neighborhood system. Given, for each $x \in X$ an assignment of a set of subsets satisfying only the first condition, the resulting *neighborhood system subbasis* generates a unique neighborhood system, namely the smallest neighborhood system whose $N(x)$ contain the corresponding elements of the neighborhood system subbasis. If one does not demand the fourth condition, one arrives at a neighborhood system for a *pretopological* space. Essentially all the conclusions drawn in the sequel about topological structure hold as well for pretopological structure, which also has some advantages for representing finite-precision similarity, but I elide the dif-

elements of $U \in N(x)$ might be said to be "U-similar" to x; those of $V \in N(x)$ "V-similar" to x; and those of $U \cap V$ "U-and-V-similar" to x. The second condition guarantees this lattermost conjuctive similarity set is in the neighborhood system. (The third and fourth, which are less important for present purposes — see footnote 16 — guarantee, respectively, that in a sense, one can always find sufficiently coarse and sufficiently fine similarity sets.) Hence, a topology on a set, understood as a neighborhood system, defines a class of qualitative notions of similarity amongst its elements.

Although this is a structurally weak formalization of similarity, it is strong enough to formalize concepts such as continuity and stability.

Continuity Formally, a function $f : X \to Y$ between two topological spaces is continuous at $x \in X$ when for all $V \in N(f(x))$, there is some $U \in N(x)$ such that $f[U] \subseteq V$; it is continuous (simpliciter) when it is so at all $x \in X$. This is just to say that for all points arbitrarily similar to some point in the range of the function, there is a set whose elements are similar to the corresponding point in the domain. In other words, similar points in the domain of the function map to similar points in the range.

Stability Formally, a property of an element $x \in X$ is stable just when that property holds of all elements in some neighborhood of x. This is just to say that the property holds of all elements sufficiently similar to x.

These formalizations also explain why mathematicians would call the third condition for being well-posed from section 3.4 — that solutions to a dynamical system depend continuously on the data — a stability condition. For, given topologies on each of the sets of solutions and data encoding how the members of each respective set are similar to one another, the map f from data to solutions is continuous at a particular point d just in case for all the stable properties P_s of $f(d)$, there is a stable property P_d of d such that f maps all data with P_d to solutions with P_s.

Thus, one can give a topological formalization of the PS as follows.

Principle of Stability (topological) An inference from the statement that a property of a model holds to the statement that the property of phenomena (or some possible world) it represents holds is justified only if that property is stable in an appropriate topology on an appropriate class of models.

The two questions that remain regarding the application of this version of PS are: What are appropriate classes of models and topologies on them?

The appropriate class of models should be all and only those of the theory one is applying, or more generally, just those models representing states of affairs that one deems possible in the context of application. This entails that the application of the PS is always relativized to such a class; judgments of the stability of a property of a model may therefore vary depending on of which class of models one is considering that model to be a member.[17] The topology appropriate to a context of application, meanwhile, should encode amongst those models all and only the properties that make a difference for that context. Often, these properties will consist in the relevant empirical descriptions or predictions that a model entails, although in principle, non-empirical properties could be included. Just as with the class of models, judgments of the stability of a property of a model depend on the way in which one considers models to be relevantly similar. Indeed, in the sequel (section 4.2), I argue that this relativity explains the apparent differences between Hadamard's and Duhem's conclusions regarding the boundedness ("stability") of dynamical systems like the solar system.

ference here for simplicity.

[17] This is in contrast with analogous uses of similarity using possible-worlds semantics, where the "worlds" are supposed to be all those metaphysically possible; here, one might say instead that the models represent all the scientific possibilities (or in the case of a physical theory, physical possibilities). (Cf. appendix A.2.)

4.2 *Two Topologies on Particle Trajectories*

Consider a dynamical system consisting of a particle smoothly sliding on a surface, much like the example of Hadamard (1898) that Duhem (1954/1914) discusses, and say that two possible positions or velocities on the surface are within ϵ of each other (or ϵ-close) just when the magnitude of their difference is less than ϵ. Then, consider the following two subbases for neighborhood systems[18] that determine different histories of the system to be similar:

1. Given a history x, let $N_{\epsilon,T}(x)$ be all the histories whose positions and velocities are each within ϵ of those of x for each time during the time interval $[0, T]$, where T is finite.
2. Given a history x, let $N_\epsilon(x)$ be all the histories whose positions and velocities are each within ϵ of those of x for each time during the time interval $[0, \infty)$.

According to the first topology, two histories are similar just when their positions and velocities are so for some finite interval of time starting with the initial condition. By contrast, according to the second topology, two histories are similar just when their positions and velocities are so for *all* time after the initial condition.

These two topologies render different verdicts regarding whether an initial-value problem for the dynamical system is well-posed, because they differ regarding whether the map from initial conditions to histories is continuous: The first topology only requires that histories be similar (in position and velocity) for (arbitrarily) finite times, while the second requires them to be so for all times. In other words, according to the first topology, the evolution of the dynamical system is continuous at some initial state just when, for any finite $T, \epsilon > 0$, all states sufficiently close to that initial state yield trajectories whose position and velocity are within ϵ of that of the initial state for non-negative times through T. According to the second topology, stability requires the same, but for *all* non-negative times.

To see how the two differ even in the simplest cases, consider a particle constrained to fall in one direction with constant acceleration a, so that its position is given by $x(t) = x_0 + v_0 t - at^2$, where x_0 and v_0 are its initial position and velocity, respectively. For any finite $T, \epsilon > 0$, one can find δ_x and δ_v sufficiently small in magnitude so that, letting $x'(t) = (x_0 + \delta_x) + (v_0 + \delta_v)t - at^2$ be the position for quite similar initial conditions, both $|x(t) - x'(t)|$ and $|v(t) - v'(t)|$ are less than ϵ for $t \in [0, T]$.[19] But one cannot find such δ_x and δ_v to satisfy these conditions for *all* times — eventually, particles with slightly different velocities will grow arbitrarily far apart from one another.

The answer to the puzzle of section 4, then, is that, implicitly, Hadamard used the first topology, while Duhem's question of the long-term behavior of the solar system — whether the planets will stay in bounded orbits — is a property to which only the second topology is sensitive, of the two.[20] In other words, the sense in which the evolution of the dynamical system under discussion is continuous is that sufficiently similar initial states of the system produce sufficiently similar trajectories for bounded time; the sense in which it is discontinuous is that this is not so for unbounded time. Thus, the bounded-time state properties of the system — i.e., those that are (continuous) functions of the bounded trajectories — are stable, and are therefore candidates for inference about the physical system being represented, while some unbounded-time state properties are not stable, hence are not candidates for such inference, according to the PS.

There is a sense, however, in which both Duhem and Hadamard would have agreed that the first topology is more appropriate in this

[18] The neighborhood system determined from a subbase is simply the smallest neighborhood system that includes the elements of the subbase.

[19] Here, $v = dx/dt$ and $v' = dx'/dt$.

[20] The topologies on the space of histories, or trajectories, that I have defined are called the C^1 compact-open and open (or Whitney) topologies, respectively. Depending on the initial value problem, one might use slightly different topologies that are more sensitive to "average" differences rather than maximum ones. For more technical details on the topologies used in initial value problems, see, e.g., Lavrent'ev et al. (2003).

context of investigation. Both of them would have judged two states of the solar system to be similar when the initial positions and velocities of the planets were similar and that our judgments of the similarity of future states would be based on the same, not on similarity for all time, to which we have no empirical access. So the first topology better captures the continuous connection between similarity of initial states and similarity of near-future states; if one accepts the PS, near-future states are thus candidates for justified inference from astronomical models, while this is not so according to the second topology. Therefore, this agreement does not conflict with Duhem's judgment that the perpetual boundedness of the planets' orbits could be an unstable property, for based on what we can justifiably infer from the present state of the solar system, the first topology better captures the ways in which the planets' trajectories are similar.

5. Epistemic Status of the Principle

5.1 Contingent Transcendental Arguments for Metaphysical Principles

Implicit evocations of the PS or the patterns of reasoning it supports seem to be common in the scientific and mathematical modeling literatures, but I have found few instances in which an author questions it, furnishes an argument for it, or provides a reference or citation either for it or for arguments for it. This provides some (weak) evidence that it could be a candidate conclusion for an *epistemological* version of what Chang (2008, p. 113) has called a contingent transcendental argument for a *metaphysical* principle,[21] one with the following form: "*If* we want to engage in a certain epistemic activity, then we *must* presume the truth of some particular metaphysical principles." The resulting principles are a priori not in the sense that they can be known independently of empirical input or that they prescribe unconditional categories of concepts for all thought, but rather that they are "necessary conditions for carrying out certain epistemic activities" (Chang, 2008, p. 121). He stresses that, in contrast with other accounts of the conditions of cognitive judgment, there is no assumption that anyone engage in any particular sort of activity at any particular time, so these principles are really contingent on one's commitment to doing so; once one does make such a commitment, it entails commitment to the principles, because the latter are necessary in order to render the former well-defined, intelligible, or possible.

Chang (2008, p. 122) illustrates this idea with the activity of counting: "If we want to engage in the activity of counting, then we have to presume that the things we are trying to count are discrete." The metaphysical principle, of course, is that the type which is the object of the count must be presumed to be discrete, i.e., consist in absolutely distinguishable units corresponding with some subset of the natural numbers. There is no presumption that one must engage in counting, but if one does, discreteness of the counted must be assumed, for such an assumption is (partly) constitutive of that activity; without it, counting is unintelligible and pragmatically impossible. Although Chang (2008, p. 127) suggests a list of other candidate pairs of simple epistemic activities and their (at least partially) constitutive metaphysical principles, he is ultimately interested in how complex activities comprise and combine, perhaps holistically, more basic ones.

I would like to suggest the PS as the principle partly concomitant with the complex activity of representational modeling — any modeling practice predicated on explaining, predicting, or controlling variable phenomena. This activity is compatible with empiricist approaches to science: To represent with models does not entail commitment to believing that the model represents non-observable aspects of reality.[22] In particular, I shall suggest that it follows from three other principles — **Justification**, **Similarity**, and **Imprecision**, each de-

[21] Here, Chang is influenced by the work of Friedman (2001) and C. I. Lewis (1956/1929) on the relativized and pragmatic a priori, respectively.

[22] Thus, I do not take representation to be solely for the goal of truth-apt explanation, insofar as it must appeal. Nevertheless, the listed activities do not necessarily include all modeling practices, such as how-possibly explanations or the heuristic generation of new hypotheses or research questions (Isaac, 2013).

scribed below — that are constitutive commitments of certain aspects of representational modeling.

In representational modeling, one uses the model as a kind of surrogate to reason about real world phenomena (Swoyer, 1991; Suárez, 2004; Contessa, 2007).[23] Insofar as one is committed to doing so, it seems that one is committed to the following principle:

Justification One has justification to infer of phenomena only those properties licensed by the model(s) adequately representing that phenomena.

Here, a representation is adequate when it satisfies the representational context's purposes or objectives, such as meeting a certain standard of accuracy (Suárez, 2004; van Fraassen, 2008). If several models represent the phenomena equally well, then only those properties which *all* those models have are licensed.[24] This principle indeed just states one's commitment to representational modeling, for if one had justification to infer some other properties not licensed by the model(s), one just wouldn't be engaging in that activity.

The kind of representational modeling under consideration concerns *variable* phenomena, phenomena whose proper description may differ according to circumstance. Accordingly, it typically is often used not just to reason about an individual token of phenomena separately from and without consideration of its relations to all others, but also about how similar tokens of phenomena relate to one another. For example, one would like to be able to describe the circumstances under which a prediction of a quantity from a model is verified when the measured value of the quantity is very close to but not exactly equal to the predicted value. So if one commits to reason about the similarity of phenomena using models as surrogates, then from **Justification** we have the following principle:

Similarity As representations of phenomena, models are relevantly similar according to the similarity of the phenomena they represent.

In other words, models are relevantly similar in virtue of their representational similarity, for it is this similarity which allows one to use justified similarity of models as a surrogate for justified similarity of phenomena in one's reasoning. This doesn't prevent one from judging models to be similar in other ways in other contexts; it only requires that, *as* representational models, models are to be considered similar in just the ways the phenomena they represent are. A measurement model for the phenomena, explicit or implicit, determines how phenomena are similar according to the modeler's investigative interests.

Another common aspect of representational modeling is that the process of modeling is imperfect: Variable phenomena may not be accurately represented in various ways, leading to the general defeasibility of model-based reasoning. The sources of this imperfection are diverse: it could arise from uncertainty regarding the variable phenomena itself, as Duhem stressed regarding measurement error in the astronomical sciences, or from idealizations that deliberately distort or abstract from the phenomena. But ultimately they arise because of the variability of the phenomena themselves and our imperfection in representing it in our models:

Imprecision If a model-based inference of a property of phenomena is justified, then that property obtains for sufficiently similar phenomena.

This principle enshrines a kind of inevitable degree of inaccuracy in justifiably determining, measuring, and representing phenomena through modeling. It bears on modeling directly because of the central goal of

[23] No assumption is needed that a model represents independently of the users of a model — this sort of representation is compatible with viewing models as epistemic tools in the sense of Knuuttila (2011).

[24] This aspect of **Justification** depends on what Chang (2008, p. 123) calls "the *principle of single value* . . . that a real physical property can have no more than one definite value in a given situation." In contrast with Chang (2008, pp. 124–125), I see this as a necessary commitment of ascribing the usual sorts of properties to phenomena rather than the practice of testing by overdetermination, but this difference is not essential for present purposes. (See also the argument below for the PS based on **Justification** and the following two principles, **Similarity** and **Imprecision**.)

modeling to represent phenomena for surrogative reasoning. The selection of a model to represent measurable phenomena must then be only up to some sufficiently similar class.

The PS follows from these three principles. For, suppose that some inference of a property from a model to that of phenomena is justified. By **Imprecision**, that property obtains also for a class of sufficiently similar possible phenomena. Those sufficiently similar possible phenomena are represented by sufficiently similar models, according to **Similarity**. Since all those models could have represented the phenomena equally well, they must all have the property originally inferred, for otherwise, **Justification** would have not licensed it; one cannot consistently infer incompatible values of a property.[25]

At least the first two principles are both sequentially dependent on one another and are each constitutive of a contingent activity. One needn't engage in surrogative reasoning about phenomena using representational models, but if one does, one commits to **Justification**. Even if one does, that reasoning needn't use similarity concepts, but if one does, one commits to **Similarity**. By contrast, **Imprecision** seems more directly a hypothesis explaining a widespread empirical generalization, namely that our ability to measure and represent phenomena comes always with some degree of error or imperfection. Perhaps it could be given a contingent transcendental justification via the inevitable imperfect representation of phenomena in a model — for perhaps, this is just a part of what representational modeling *is* — but I shall not pursue that further here. My goal in this section, after all, has been just to indicate one possible argument for the PS as being (partly) constitutive of the practice of representational modeling itself.

5.2 Two Objections

While **Justification** seems quite fundamental to representational modeling, one could consider apparent modeling contexts in which **Similarity** or **Imprecision** does not hold. In this subsection, I thus consider two objections, or seemingly problematic cases, for these latter two principles, respectively. If these cases were counterexamples to these respective principles, they would undermine the argument of the previous subsection in support of the PS. One could still maintain the PS in light of this, but not in virtue of that argument. Nevertheless, I will argue that these cases, properly understood, are not such counterexamples.

The first case I wish to consider involves a measurement without any apparent attendant uncertainty; the measurement is thus of a phenomenon that is completely distinguishable from other phenomena of the same type. It concerns, essentially, discrete-valued phenomena. For vividness, consider a pilot experiment regarding a treatment designed to increase the fertility of rabbits. The control group, consisting of those rabbits which did not receive the treatment, had an average litter size of 6 kits. Meanwhile, the treatment group, consisting of those rabbits which did receive the treatment, had an average litter size of 10 kits. There is no vagueness about the number of kits in each case; although a statistical model of the treatment effect size — the change in average litter size arising from the treatment — would allow for uncertainty thereof, there is no imprecision regarding the phenomena themselves. Are measurements of such discrete phenomena therefore counterexamples to **Imprecision**?

They are not if one recalls that the notion of similarity evoked in **Imprecision** must be a sort relevant to the context of investigation. The relevant way in which phenomena should be judged to be similar should depend on what we can actually measure and determine of those phenomena. In the case of real-valued measurements, such as those of position, velocity, or mass, the relevant notion of similarity can be formalized with the standard topology on the real line, that which takes as a neighborhood basis at a point the ϵ-balls, i.e., all the points within ϵ distance of that point. This is because arbitrarily close but non-identical values for these quantities cannot be distinguished. But in the case of discrete-valued measurements, often the *discrete* topology on the measurement values will be appropriate. This is the topology

[25] This uses the "principle of single value," for which see footnote 24.

whose neighborhood basis at a point can be specified by just the singleton set containing that point. According to the discrete topology, each point therefore has a neighborhood consisting only of itself; there is a smallest set of elements sufficiently similar to a point, which is just that containing only the point. This encodes the idea that discrete-valued measurements can be completely distinguished from one another in a way entirely compatible with **Imprecision**: A discrete value at the very least is indistinguishable only from itself.

The second case I wish to consider involves an apparent mismatch between model similarity and the similarity of the phenomena the models represent. I have in mind in particular the class of models involved in so-called *inverse problems* (Lavrent'ev et al., 2003). The name is in reference to standard initial value problems, previously discussed in section 3.2. In those problems, one begins with an initial state of a system and a fixed dynamical rule on the system's state space, whereby one calculates a future state for the system. In inverse problems, however, one begins with both an initial and final state of the system, whereby one attempts to calculate relevant features, such as parameter values, for the dynamical rule leading to that evolution. Inverse scattering problems, for instance, are a common inverse problem in which one attempts to infer the composition of an object according to how particles or waves travel through it. An example inverse scattering problem encountered by geophysicists is to determine the composition of the earth by measuring the amplitude and frequency of different types of seismic waves produced in an earthquake. They know that the dynamical law for the waves is some type of wave equation, but the way that equation describes the propagation of waves through a medium depends on the density of that medium. The inverse problem here is thus to infer a density function for the earth from seismic measurements.[26]

This inverse problem presents a difficulty typical of that class: It is an ill-posed problem in the sense discussed in section 3.4. In addition to failing the uniqueness criterion — i.e., having more than one density function as a solution — it fails the stability criterion: Arbitrarily small changes in the data can entail arbitrarily large changes in the inferred densities. In other words, similar phenomena do not seem to correspond to similar models. Is the existence of such ill-posed inverse problems a counterexample to **Similarity**?

It would be, only if scientists could be shown to engage in reasoning invoking claims about the similarity of phenomena using their solutions to the ill-posed inverse problems. But in fact that does not occur. Rather, scientists slightly *change* the problem using so-called *regularization* techniques that restore uniqueness and stability before engaging in similarity reasoning. How do they do this? Well, one essential reason for the ill-posed nature of an inverse problem is that there are many more degrees of freedom available in the target of inference — the density function for the earth in the case above — than in the initial and final data used to make that inference. This allows for density function solutions that, for example, vary wildly over very short distances. Such variations are not just physically implausible, but lead to high predictive error.[27] Regularization techniques restrict the class of solutions considered so that much more data would be needed to infer high frequency density variations. In other words, these techniques bias inference methods towards solutions that are more similar to expected phenomena (e.g., reasonable density variations) and away from solutions that are less similar (e.g., wild density variations). So, far from being counterexample to **Similarity**, inverse problems in fact provide paradigmatic examples of how the PS influences scientific methodology: In order to draw inferences about phenomena, scientists introduce further assumptions in order to make satisfaction of the consequent of the PS possible.

[26] I have simplified many aspects of this problem for ease of presentation; see, e.g., Aster et al. (2011) for more on inverse problems in the geosciences.

[27] This can be seen as related to or even an example of over-fitting in statistical inference.

6. Conclusions, Implications, and Future Work

In this paper, I have introduced the PS in its historical, comparative, formal, and epistemological dimensions. After tracing some of the history of the PS as it is found in the work of Duhem (1954/1914) in section 2, I compared, in section 3, my own formulation of it with related notions regarding idealization in science and stability in dynamical systems theory (as well as safety and robustness in appendix A) that have perhaps received more attention. The PS emerges as a distinct epistemological principle among these. Then, in section 4, I showed how the work of mathematicians and physicists both contemporaneous with and subsequent to Duhem took up his challenge (if only sometimes implicitly) to develop the "mathematics of approximation" with topology, yielding a novel and rich mathematical theory of stability. This allowed for a precise resolution of the seeming tension between different senses of stability invoked by Duhem and Hadamard regarding dynamical systems like the solar system. Finally, I suggested in section 5 that the PS might be justified as a principle partially constitutive of the activity of representational modeling itself.

There are implications of the PS that deserve further scrutiny. First, it shows that what we can justifiably infer in much of model-based science depends on what we think is possible, and how we think those possibilia are similar. Changing the class of models under consideration can change whether a property is stable, as the example in section 5.2 of regularization for inverse problems showed. Changing the topology on the models can also change whether a property is stable, as the example of the two topologies on particle trajectories from section 4.2 evinced. Thus, the PS provides a focal point for more detailed study of the role of both of these in scientific modal epistemology.

Second, accepting the PS also entails that there will in general be properties of a domain of phenomena described by a theory that one is never justified in inferring[28] — for Duhem, this could include the (un)boundedness ("(in)stability") of the solar system! This sort of unknowability deserves comparison with other sorts that philosophers have investigated. For example, many take the conclusion of the Church-Fitch knowability paradox, that there are some unknowable truths, to be counterintuitive (Brogaard and Salerno, 2013). But in model-based science, the PS seems to entail just this, and moreover, give a sort of explanation of why: Some truthful ascriptions of properties, as described by a model, may not be stable and thus not candidates for inference, hence knowledge, at least through that model.

To take another example, the PS could provide examples of the failure of transmission of justification (Moretti and Piazza, 2013). For one might be justified in believing that a particular model represents a phenomenon adequately, yet not be in a position to justifiably believe all the properties of the phenomena that the model describes — justification is not transmitted from model of phenomena to property of phenomena for unstable properties. The PS thus also provides the possibility of a bridge between traditional epistemology and the literature on modeling in the sciences.

This bridge could include further consideration of the possible connections between stability and probability. (See further footnote 33.) Because the stability of a property at a model requires that there be a neighborhood of that model *all* of whose members have that property, a natural probabilistic (or, more generally, measure-theoretic) liberalization would be to allow for exceptions of probability (resp. measure) zero. What about more probable or widespread exceptions? Perhaps this will depend on the appetite for risk that the model's user can stomach; in any case, a principled argument for or against such a modification would be interesting.[29]

[28] For example, fix any model M and consider the property that M is exactly the right model to represent the phenomena. Formalized on the space of models under consideration, this property can be represented as the characteristic function for M, χ_M, which takes on the value 1 at M and 0 otherwise. Unless M is an isolated point in the topology — i.e., it has a neighborhood consisting of just itself — this property will be unstable at M. Under these circumstances, the PS entails that it is never justified to infer that phenomena are represented by exactly the right model.

[29] Earlier versions of this manuscript were presented in Madrid, Pittsburgh,

Appendix A. Comparisons

Many find Duhem's reasoning, the concept of stability, and the PS quite natural and similar, even, to other patterns of reasoning, concepts, or principles found in the philosophy of science and epistemology literatures. However, this similarity can misleadingly suggest conflation. In section 3, I already compared the PS with principles regarding idealizations in philosophy of science and with the stability dogma. In this appendix, I compare stability with two further kinds of similar concepts: *robustness* as discussed in philosophy of science (section A.1) and *safety* as discussed in epistemology (section A.2). The upshots of these respective comparisons are as follows:

1. The concept of stability is related to some concepts of robustness, most closely to derivational robustness, but an analogous principle concerning the latter has not hitherto been formulated.
2. The concept of stability is also closely related to the concept of safety, and the PS to safety principles used in the analysis of knowledge, although the latter bears on justified belief rather than justified inference.

A.1 Robustness

The concepts of "stability" and "robustness" are not always distinguished in the literatures that use them, and even when they are, their relation is not always consistent. For example, the philosophers Hansson and Helgesson (2003, p. 221) describe stability as a label for a family of constancy, invariance, or persistence concepts, of which robustness is "the tendency of a system to remain unchanged, or nearly unchanged, when exposed to perturbations." By contrast, the complex systems theorist Jen (2003, p. 17) takes robustness to be the more general concept of the two, it being "an approach to feature persistence in systems for which we do not have the mathematical tools to use the approaches of stability theory," which concerns only "fluctuations in external inputs or internal system parameters" (Jen, 2003, p. 13). Additionally, Jen (2003, p. 13) emphasizes that questions of robustness direct researchers *away* from the specified and delimited class of models they had been considering. By contrast, the concept of stability used in the PS is always made relative to a fixed class of models. Thus, each of the more specific concepts just mentioned — robustness for Hansson and Helgesson (2003) and stability for Jen (2003) — is closely related to but less general than the concept of stability as used in the PS: The reference to tendencies or fluctuations imputes a causal or temporal dimension to the sort of invariance being described that stability — as I am using it — need not. Indeed, the models one is considering when applying the PS need not even invoke any temporal or causal concepts, so the notion of similarity the PS invokes is more abstract and general.

Woodward (2006) gives a classification of robustness concepts into four types: inferential, derivational, measurement, and causal. The more specific concepts discussed above are examples of causal robustness, while measurement robustness refers to the concurrence of different, usually quasi-independent procedures yielding a common (or at least similar) measured value of a property or quantity. Derivational robustness is the closest analog of the stability concept under discussion here (although I return to inferential robustness below):[30]

> Suppose that we have a model or theory that allows for the derivation of observed facts *P*. Suppose the model contains some

Doorn, Istanbul, San Sebastián, Munich, Cambridge, Bristol, and Seattle, whose audiences I would like to thank for their comments, especially Ben Feintzeig and Conor Mayo-Wilson. This work was supported in part by a NSF Doctoral Dissertation Improvement Grant and by a Marie Curie International Incoming Fellowship (PIIF-GA-2013-628533).

[30] Calcott (2011), building on work by Wimsatt (2007) since the 1980s, delineates three types of robustness: robust theorems, robust phenomena, and robust detection. Putting these into Woodward's typology, roughly: Robust phenomena are causally robust, robust detection is measurement robustness, and robust theorems are inferentially robust propositions; Woodward's robust derivations don't have a clear counterpart (the last paragraph of this subsection notwithstanding). For this reason, I focus on Woodward's terminology instead. (See also Soler et al. (2012) for a variety of other proposals regarding the classification and analysis of robustness concepts.)

> assumption A, which might concern, e.g., the value taken by a some parameter x, . . . Suppose A was replaced by a different assumption, e.g., a different (perhaps only slightly different) value for x. Would (a) it still be possible to derive the same result P (or a result very close to it) or (b) would the model instead predict a very different outcome P' or perhaps make no relevant prediction at all? To the extent that (a) is the case, the model might be thought of as providing a 'robust' derivation of P. (Woodward, 2006, pp. 231–232)[31]

In his discussion of the boundedness ("stability") of the solar system, Duhem asked analogs of these questions, where the model is a Newtonian model of the solar system, P is the eternal boundedness of the system, and the assumption A consists in the initial positions and velocities of the planets.[32]

Despite the close similarity of stability and derivational robustness, there are at least three notable differences between them and their relations to the models or theories to which they are applied. First, the stability of a property in a model requires that *all* sufficiently similar models display that property, while its derivational robustness is relative to a *particular* similar model. This bears on the second difference: Woodward suggests that derivational robustness comes in degrees to the extent that one can tally different assumptions (i.e., different similar models) yielding the same prediction, while stability, as I have described it so far, does not.[33] There is thus a sense in which the degree of derivational robustness could provide *evidence for* stability. But third, Woodward is primarily concerned with whether derivational robustness is a theoretical virtue or lack thereof a vice — the answer to which, he maintains, "depends in part on what else is true of the model and the evidence for it" (2006, p. 232) — while the PS concerns the justification of inferences (or "derivations") from models.

The fourth type of robustness Woodward considers, inferential robustness, concerns inferences from a *fixed* data set that are invariant under a variety of incompatible modeling assumptions, among which one has no preferential evidence. There is a sense in which this could be an example of stability, too, in which all the relevantly similar models agree on the data exactly, although the range of modeling assumptions considered is typically much broader than what Duhem has in mind. For, as Duhem's remarks discussed in section 2 made clear, it would be an unusual application that took the exact values of input data as inviolably certain; even for astronomical models, conclusions drawn therefrom cannot depend on the exact values of the input data (i.e., the initial conditions) to be justified. Thus, stability applies not just to contexts of theoretical modeling that have been appropriately detached from data — a restriction of derivational robustness suggested by Woodward's remarks — but also to models that do depend more directly on data, such as statistical models for parameter estimation. Like Woodward, Yu (2013) describes stability with respect to appropriately similar data as a non-essential *virtue* of an estimation procedure, but this is again distinct from the PS, which, in this case, would place a necessary *constraint* on which inferences from estimation procedures were justified.

A.2 The Safety Condition

Modal conditions figure in many contemporary accounts of justified belief, particularly in understanding or analyzing the concept of knowledge. The modal concept most similar to stability is known as *safety*, and the necessary condition for knowledge most similar to the PS is

[31] See also Schwartz (1992, p. 21): "The physicist rightly dreads precise argument, since an argument which is only convincing if precise loses all its force if the assumptions upon which it is based are slightly changed, while an argument which is convincing though imprecise may well be stable under small perturbations of its underlying axioms."

[32] An apparent but unsubstantial difference: Despite using the phrase "observed facts," Woodward's discussion (2006, pp. 232–233) makes it clear that P represents predictions or descriptions, as with Duhem's analogs.

[33] In his discussion of inferential robustness, Woodward considers the possibility of measuring the degree thereof with probabilities; a probabilistic version of the PS is worth further investigation but is beyond the scope of this discussion. See also the discussions in sections 6 and A.2 .

In def. of Safety		In def. of Stability
belief in P	$\Leftrightarrow$	property p
is true at world W	$\Leftrightarrow$	holds of an element x
close worlds to W	$\Leftrightarrow$	elements similar to x

Table 1: Substitutions that map between the definitions of safety and stability reveal their common structure.

accordingly the safety condition (Rabinowitz, 2017). Sosa (1999, p. 146) originally defined a belief in a proposition P of agent S to be safe just when "If S were to believe P, P would be true." One can then understand this counterfactual conditional in terms of possible worlds: A belief in P that is true at a world W is safe at W just when all (sufficiently) close worlds to W in which belief in P obtains is also true.[34] When formulated in this way, the analogy between safety and stability is obvious: One obtains the definition of the latter from that of the former through the substitutions listed in Table 1. By the same substitutions, one obtains a necessary condition for justified inference, the PS, from a necessary condition for justified belief (or knowledge), the safety condition. An important difference between inference and belief (or knowledge) here is that the sort of inference under consideration is always conditional on the model, so it will not entail belief if one does not also believe the model. The PS applies even to instrumental uses of model-based inference, and is not intended as an antidote for skepticism as the safety condition has sometimes been. Despite this difference, the close parallels between the PS and the safety condition make it striking that Duhem has not been recognized as formulating a version of this condition over 80 years prior to Sosa.[35]

Williamson (2000, p. 146) has formulated a similar version of the safety condition for knowledge as follows:[36] "If one knows, one could not easily have been wrong in a similar case." Again, if one substitutes (sufficiently) similar models for similar cases, making a justified inference from a model to phenomena for knowing, and making a different inference for being wrong, one obtains a version of the PS. One potential difference between Williamson's and Sosa's versions of safety is that Williamson allows for some chance of error — to emphasize, "one could not have been *easily* wrong." Although it likely goes beyond Williamson's intent, this does suggest a probabilistic qualification to the PS.[37] (See also footnote 33.) As interesting as it is, I shall not pursue that qualification here; I consider some directions in which it could be understood for future work in section 6. (See also footnote 4 for Duhem's invocation of probability in approximative reasoning.)

References

Andronov, A. and Pontrjagin, L. (1937), 'Systèmes grossiers', *Dokl. Akad. Nauk., SSSR* **14**, 247–251.

Ariew, R. (2014), Pierre Duhem, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Fall 2014 edn, Metaphysics Research Lab, Stanford University.
URL: *https://plato.stanford.edu/archives/fall2014/entries/duhem/*

Aster, R. C., Borchers, B. and Thurber, C. H. (2011), *Parameter Estimation and Inverse Problems*, 2nd edn, Academic Press, San Diego.

Batterman, R. W. (2002), *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*, Oxford University Press, Oxford.

[34] For simplicity of presentation, I have omitted the reference to a method M by which belief is acquired, which Sosa requires be held constant across the worlds in which the consequent of the conditional is evaluated.

[35] The first edition of *The Aim and Structure of Physical Theory* dates from 1906.

[36] Williamson rejects the *analysis* of knowledge as a misguided project, and so does not see the safety condition as playing a non-circular definitional role. But this and other variant attitudes one might have toward the role of the safety condition in epistemology do not matter so much for my present *comparative* goal.

[37] See also Pritchard (2007, 2008, 2009), who presents a version of the safety condition related to those of Williamson and Sosa in connection with his account of epistemic luck.

Boyer, C. B. (1959/1949), *The History of the Calculus and its Conceptual Development*, Dover, New York.

Brogaard, B. and Salerno, J. (2013), Fitch's paradox of knowability, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Winter 2013 edn, Metaphysics Research Lab, Stanford University.
URL: *https://plato.stanford.edu/archives/win2013/entries/fitch-paradox/*

Butterfield, J. (2011), 'Less is different: Emergence and reduction reconciled', *Foundations of Physics* **41**(6), 1065–1135.

Calcott, B. (2011), 'Wimsatt and the robustness family: Review of Wimsatt's Re-engineering Philosophy for Limited Beings', *Biology and Philosophy* **26**(2), 281–293.

Chang, H. (2008), 'Contingent transcendental arguments for metaphysical principles', *Royal Institute of Philosophy Supplement* **63**, 113–133.

Contessa, G. (2007), 'Scientific representation, interpretation and surrogative reasoning', *Philosophy of Science* **74**(1), 48–68.

Courant, R. and Hilbert, D. (1962/1937), *Methods of Mathematical Physics, Vol. II: Partial Differential Equations*, Wiley, Singapore.

d'Alembert, J.-B. l. R. (2008/1754), Continuity, law of, *in* 'The Encyclopedia of Diderot and d'Alembert Collaborative Translation Project', University of Michigan Press, Ann Arbor. Trans. John S. D. Glaus of "Continuité, loi de," Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, vol. 4. Paris, 1754, pp. 116–117.
URL: *http://hdl.handle.net/2027/spo.did2222.0000.868*

Dilworth, C. (1994), 'Principles, laws, theories and the metaphysics of science', *Synthese* **101**(2), 223–247.

Dilworth, C. (2007), *The Metaphysics of Science: An Account of Modern Science in Terms of Principles, Laws and Theories*, 2nd edn, Springer, Dordrecht.

Duhem, P. (1954/1914), *The Aim and Structure of Physical Theory*, 2nd edn, Princeton University Press, Princeton. Trans. Philip P. Wiener.

Earman, J. (1986), *A Primer on Determinism*, D. Reidel, Dordrecht.

Earman, J. (2004), 'Curie's principle and spontaneous symmetry breaking', *International Studies in the Philosophy of Science* **18**, 173–198.

Earman, J. (2007), Aspects of determinism in modern physics, *in* J. Butterfield and J. Earman, eds, 'Philosophy of Physics', Elsevier, Amsterdam, pp. 1369–1434.

Fletcher, S. C. (2016), 'Similarity, topology, and physical significance in relativity theory', *British Journal for the Philosophy of Science* **67**(2), 365–389.

Franceschelli, S. (2014), La déduction mathématique et la théorie physique. Exemple de solutions numériques physiquement utiles, *in* F. Varenne, M. Silberstein, S. Dutreuil and P. Huneman, eds, 'Modéliser & simuler – Tome 2', Editions Matériologiques, Paris, pp. 109–135.

Friedman, M. (2001), *Dynamics of Reason*, CSLI Publications, Stanford.

Geroch, R. (1971), 'General relativity in the large', *General Relativity and Gravitation* **2**(1), 61–74.

Guckenheimer, J. and Holmes, P. J. (1983), *Nonlinear Oscillations, Dynamical Systems, and Bifurcation of Vector Fields*, Springer, New York.

Gyenis, B. (2013), Well Posedness and Physical Possibility, PhD thesis, University of Pittsburgh, Pittsburgh.

Hadamard, J. (1898), 'Les surfaces à courbures oppossées et leurs lignes géodésiques', *Journal de Mathématiques pures et appliquées, 5^e série* **4**, 27–74.

Hadamard, J. (1901), *Notice sur les Travaux Scientifique de M. Jacques Hadamard*, Gauthier-Villars, Paris.

Hadamard, J. (1902), 'Sur les problèmes aux dérivées partielles et leur signification physique', *Princeton University Bulletin* **13**(4), 49–52.

Hadamard, J. (1923), *Lectures on Cauchy's problem in linear partial differential equations*, Yale University Press, New Haven.

Hansson, S. O. and Helgesson, G. (2003), 'What is stability?', *Synthese* **136**, 219–235.

Hawking, S. W. (1971), 'Stable and generic properties in general relativity', *General Relativity and Gravitation* **1**(4), 393–400.

Isaac, A. M. C. (2013), 'Modeling without representation', *Synthese* **190**(16), 3611–3623.

Ivanova, M. and Farr, M. (2015), 'Conventional principles in science: On the foundations and development of the relativized a priori', *Studies* *in History and Philosophy of Modern Physics* **52**, 111–113.

Jen, E. (2003), 'Stable or robust? What's the difference?', *Complexity* **8**(3), 12–18.

Jones, N. J. (2006), Ineliminable Idealizations, Phase Transitions, and Irreversibility, PhD thesis, Ohio State University, Columbus.

Knuuttila, T. (2011), 'Modelling and representing: An artefactual approach to model-based representation', *Studies in History and Philosophy of Science* **42**(2), 262–271.

Landsman, N. (2013), 'Spontaneous symmetry breaking in quantum systems: Emergence or reduction?', *Studies in History and Philosophy of Modern Physics* **44**(4), 379–394.

Lavrent'ev, M. M., Avdeev, A. A., Lavrent'ev, Jr., M. M. and Priimenko, V. I. (2003), *Inverse Problems of Mathematical Physics*, VSP, Zeist.

Lewis, C. I. (1956/1929), *Mind and the World Order: Outline of a Theory of Knowledge*, Dover, New York.

Manchak, J. B. (2011), 'What is a physically reasonable space-time?', *Philosophy of Science* **78**, 410–420.

Maxwell, J. C. (1925/1877), *Matter and Motion*, Sheldon Press, London.

Maz'ya, V. and Shaposhnikova, T. (1999), *Jacques Hadamard, A Universal Mathematician*, American Mathematical Society, Providence.

Meiss, J. (2007), 'Dynamical systems', *Scholarpedia* **2**(2), 1629. Revision #121407.

Moretti, L. and Piazza, T. (2013), Transmission of justification and warrant, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Winter 2013 edn, Metaphysics Research Lab, Stanford University.
URL: *https://plato.stanford.edu/archives/win2013/entries/transmission-justification-warrant/*

Northrop, F. S. C. (1931), *Science and First Principles*, Cambridge University Press, Cambridge.

Pritchard, D. (2007), 'Anti-luck epistemology', *Synthese* **158**(3), 277–298.

Pritchard, D. (2008), Knowledge, luck, and lotteries, *in* V. F. Hendricks and D. Pritchard, eds, 'New Waves in Epistemology', Palgrave Macmillan, London, pp. 28–51.

Pritchard, D. (2009), 'Safety-based epistemology: Whither now?', *Journal of Philosophical Research* **34**, 33–45.

Pugh, C. and Peixoto, M. M. (2008), 'Structural stability', *Scholarpedia* **3**(9), 4008. Revision #91834.

Rabinowitz, D. (2017), The safety condition for knowledge, *in* 'The Internet Encyclopedia of Philosophy'.
URL: *http://www.iep.utm.edu/safety-c/*

Reichenbach, H. (1956), *The Direction of Time*, University of California Press, Berkeley.

Schmidt, J. C. (2011), Challenged by instability and complexity: On the methodological discussion of mathematical models in nonlinear sciences and complexity theory, *in* C. Hooker, ed., 'Philosophy of Complex Systems', North Holland, Oxford, pp. 223–254.

Schmidt, J. C. (2017), Science in an unstable world: On Pierre Duhem's challenge to the methodology of exact sciences, *in* W. Pietsch, J. Wernecke and M. Ott, eds, 'Berechenbarkeit der Welt? Philosophie und Wissenschaft im Zeitalter von Big Data', Springer Fachmedien Wiesbaden, Wiesbaden, pp. 403–434.

Schwartz, J. T. (1992), The pernicious influence of mathematics on science, *in* M. Kac, G.-C. Rota and J. T. Schwartz, eds, 'Discrete Thoughts: Essays in Mathematics, Science, and Philosophy', 2nd edn, Birkhäuser, Boston, pp. 19–26.

Soler, L., Trizio, E., Nickles, T. and Wimsatt, W. C., eds (2012), *Characterizing the Robustness of Science: After the Practice Turn in Philosophy of Science*, Springer, Dordrecht.

Sosa, E. (1999), 'How to defeat opposition to Moore', *Philosophical Perspectives* **13**, 141–154.

Suárez, M. (2004), 'An inferential conception of scientific representation', *Philosophy of Science* **71**(5), 767–779.

Swoyer, C. (1991), 'Structural representation and surrogative reasoning', *Synthese* **87**(3), 449–508.

van Fraassen, B. C. (2008), *Scientific Representation: Paradoxes of Perspective*, Oxford University Press, Oxford.

van Strien, M. (2014), 'The Norton dome and the nineteenth century foundations of determinism', *Journal for General Philosophy of Science*

45(1), 167–185.

Whewell, W. (1847), *The Philosophy of the Inductive Sciences: Founded Upon Their History*, new edn, John W. Parker, London.

Willard, S. (2004/1970), *General Topology*, Dover, Mineola, NY.

Williamson, T. (2000), *Knowledge and its Limits*, Oxford University Press, Oxford.

Wimsatt, W. C. (2007), *Re-engineering philosophy for limited beings: Piecewise approximations to reality*, Harvard University Press, Cambridge, MA.

Woodward, J. (2006), 'Some varieties of robustness', *Journal of Economic Methodology* **13**(2), 219–240.

Yu, B. (2013), 'Stability', *Bernoulli* **19**(4), 1484–1500.