Noname manuscript No.
(will be inserted by the editor)

A not-too-simple solution to Goodman's new riddle of induction in the age of AI

Luigi Scorzato

Abstract I review the works of Gärdenfors (1990) and Scorzato (2013) and show that their combination provides an elegant solution of Goodman's new riddle of induction. The solution is based on two main ideas: (1) clarifying what is expected from a solution: understanding that philosophy of science is a science itself, with the same limitations and strengths as other scientific disciplines; (2) understanding that the concept of *complexity of a model's assumptions* and the concept of *direct measurements* must be characterized *together*. Although both *measurements* and *complexity* have been the subject of a vast literature, within the philosophy of science, essentially no other attempt has been made to combine them. The widespread expectation, among modern philosophers, that Goodman's new riddle cannot be solved is clearly not defensible without serious exploration of such a natural approach. A clarification of this riddle has always been very important, but it has become even more crucial in the age of AI.

1 The old new riddle

Goodman's new riddle of induction (Goodman, 1946, 1955, 1983) occupies a prominent role in the philosophy of science. In fact, it was recognized very early as a major obstacle to any attempt to specify clearly the goals of science. But, it was also considered a problem that philosophy should be capable of solving (see in particular Putnam's preface to the 4-th edition of Goodman (1983)). However, today, 70 years after the publication of the first edition of (Goodman, 1955), no solution is widely accepted and most philosophers have no confidence that a solution is possible at all.

In this article, I start by examining the problem of scientific model selection (Sec. 1.1). Then, in Sec. 1.2, I emphasize that the problem of model selection is a reformulation of Goodman's new riddle, and both problems are very much related to other classic problems in philosophy of science. In Sec. 1.3, I discuss what should count as a solution of of Goodman's riddle. In Sec. 2, I use the ideas of Gärdenfors (1990) and Scorzato (2013) to formulate a simple (but not too simple) solution. This entails a clarification of the role of direct measurements (Sec. 2.1) together with a well defined notion of complexity (Sec. 2.2). In Sec. 3, I argue that a justification of the solution should not fall back into a quest to solve the old riddle (Sec. 3.1), but it should rather show the descriptive power of the solution (Sec. 3.2). Finally, in Sec. 4, I compare the proposed solution to previous proposals (Sec. 4.1), focusing especially on the analysis of Gärdenfors and Stephens (2017) (in Sec. 4.2).

1.1 The problem of scientific model selection

How do scientists decide that some scientific models are viable options while others should be disregarded? **Empirical evidence** is not enough. No matter how much data we have collected, there are always infinite models that fit the data. This observation is known as *underdetermination* of the theory[1] by the data (Duhem, 1954; Quine, 1975; Stanford, 2021).

Accenture AG, Genève, Switzerland. E-mail: luigi.scorzato@accenture.com

[1] Note that 'theory' and 'model' are used as synonyms in this paper.

These infinite options are not just theoretical possibilities without practical relevance. Consider statements like: *"was my experimental device malfunctioning on day X?"*[2]. This is the very concrete way in which infinite options of ad-hoc assumptions[3] can be exploited to fit any data. These kind of questions emerge very often in real scientific practice, and they can be very challenging to resolve. However, they do not seem to pose an *insurmountable* obstacle to the scientific practice. Which other tools do the scientists use to discriminate among these equally accurate infinite options?

Predictions are also not enough. A successful experiment might increase the probability of a model, in some sense. Carnap (1950) devoted enormous efforts to this problem (see Crupi (2021) for a review and Leitgeb (2024) for a recent contribution on this topic). But can we translate these results into a rule for model selection, even a very crude and approximate one? Unfortunately, we cannot.

The reason is the same behind the impossibility of determining the "right" *p-value* (even approximatively) that can be used to confirm a discovery (Goodman, 2008). In fact, such hypothetical *p-value* should be extremely small (say, at least $p \ll 10^{-10}$) to prevent us from rejecting our best scientific theories in favor of a crazy alternative that happens to correctly guess some very unlikely events. On the other hand, the same *p-value* should be $p \gtrsim 1/3$ to justify a vast amount of valuable model selections that enjoy the unshakable support of the scientific community (especially in domains where impressive predictions are rare).

Probability alone clearly cannot determine model selection. Predictions are valuable only if they are based on *good* assumptions. We obviously need to assume also some **non-empirical** (i.e. non evidence based) **epistemic value** to justify the scientific model selections that are regularly adopted by the scientific community.

Indeed, many scientists and philosophers have recognized that some non-empirical epistemic values play a formal role in model selection. Einstein, for example, famously said (Barnett, 1950): "The grand aim of all science is to cover the greatest number of empirical facts by logical deduction from the smallest number of hypotheses". Many others made similar statements.

1.2 The new riddle of induction and its pervasiveness

Unfortunately, Einstein's characterization of the goal of science has a major flaw: how do we *count* the number of hypotheses? In fact, one can always introduce a new symbol Ξ to express all her hypotheses as $\Xi = 0$, and the number of hypotheses would be just one! This fact was noted by many (Feynman et al., 1963; Kelly, 2007; Votsis, 2016), and it is, essentially, a reformulation of Goodman's new riddle of induction: for every crazy theory that agrees with the data, there is always a grue language that makes it a simple and/or natural assumption. The notion of simplicity seems to be irremediably subjective, hence unsuitable to define the goals of science.

While Goodman (1983) focuses on generalizing single *sentences*, these are clearly meant as simple examples of *scientific models*. And determining which sentences are legitimately generalizable (projectible) is just a simple instance of the general problem of determining which scientific models can be legitimately considered (i.e. the state of the art) in a given scientific domain. In fact, the new riddle of induction and the problem of scientific model selection are essentially the same problem. But I believe that the reformulation in terms of Ξ is useful to emphasize how general the problem is.

Other famous philosophical conundrums are closely related to what I just discussed. Goodman himself discussed in detail how his riddle is essentially equivalent to the problem of *counterfactuals* (Starr, 2022) (deciding what would happen in an hypothetical, unrealized, scenario corresponds to selecting a class of models that can be used to draw any conclusion). A vast literature (Choi and Fara, 2021) also connects the problem of counterfactuals to the problem of *dispositions*.

The problem of model selection is also strongly entangled with the problem of *confirmation* (Crupi, 2021). In fact, both aim at identifying which models should be used and which shouldn't, although scholars working on confirmation have historically followed different approaches to try to solve the same problem.

Importantly, all these problems appear in two versions, depending on whether they aim at justifying one choice in terms of likelihood of future success (the old riddle), or they simply aim at describing the scientists' actual choices (the new riddle).

[2] We can test some of these statements, but only a tiny fraction of the possible.

[3] Often scientists argue that these are not scientific models, but a definition of *scientific model* that excludes these options without excluding many other legitimate options is not available.

Goodman's riddle occupies a central place in all textbooks of philosophy of science, because it was recognized very early as a major obstacle to define almost any key concept in this domain. In fact, if you can write all your laws as $\Xi = 0$ (and you certainly can), and you can measure Ξ directly (which is the weak point, as we will see, but difficult to contest, as there is certainly a one-to-one correspondence with what we do measure), and if the law $\Xi = 0$ is accurate (which is true by construction), how can this theory be less than optimal by any standard?

As a consequence, Goodman's argument can be used to show the emptiness of any precise definition of *theory selection, goals of science, induction, demarcation, confirmation, scientific progress, explanation or understanding*. Indeed, no definition of any of these concepts has gained the wide endorsement of the philosophical community and the only definitions that gained some popularity inevitably refer (more or less explicitly) to the irreplaceable judgment of the scientific community. This necessarily falls short of identifying the hidden assumptions behind the scientists' decisions and cannot be used to assess anything beyond what is already supported by overwhelming scientific consensus[4].

Until at least the 80's, prominent philosophers held the firm belief that philosophy should be able to offer a satisfactory solution to Goodman's new riddle (see e.g. the preface of Goodman (1983)). However, in the past few decades, the philosophical community has increasingly accepted the idea that also the new riddle, like the old one, might remain insurmountable. But Goodnam's intuition remains valid: as I have just reviewed, the scientists obviously use some unspecified assumptions in their decisions of model selection. Then, there are only two possibilities, either (a) we can identify such assumptions, which is the only way to, perhaps, judge whether they are acceptable; or (b) we cannot. But, in the latter case we are saying that any scientific conclusion rely, inexorably, on fundamentally mysterious assumptions. This is the worst that one can say to discredit any scientific enterprise. In fact, a major task in any scientific work consists in identifying and clarifying every hidden assumption in the arguments used by scientists. Any effort in this sense would be completely futile, in the scenario (b). It does not help to formulate scenario (b) with narratives that obfuscate how much it fundamentally destroys the credibility of science. It does, inevitably. Fortunately, scenario (b) is just not plausible. But then we must first understand what is wrong with the formulation $\Xi = 0$. This must be possible: *If there aren't objective standards, [we must] construct standards!* (Goodman, 1983).

1.3 What does count as a solution?

A major obstacle in solving Goodman's riddle lies already in the confusion of what should count as a solution (Scholz, 2024). In particular, requiring that a candidate solution is justified in terms of likeliness of future successes is not legitimate, because it represents a relapse in the quest for a solution to the *old* riddle, that we cannot expect to solve. On the other hand, a mere enumeration of the scientists' actual choices is not satisfactory either: it would be useless to interpret any new model selection.

An ideal solution should describe all cases of model selection supported by broad scientific consensus, but it should do it by providing a *general rule*[5] and not a mere list. What is a general rule? It is a concise rule that covers most cases with few or no exceptions. But, as we just saw in the case of scientific laws, it is always possible to forge a general rule from a mere list of examples by using a Ξ trick. This makes it clear that **deciding what counts as a solution to Goodman's riddle** is equivalent to **solving Goodman's riddle in the special case when the scientific model under scrutiny is a model for scientific model selection itself**. This is not very surprising, if we accept that philosophy of science is a science itself, it should demand of itself what it demands of other scientific disciplines. This provides further evidence that nothing makes sense in philosophy of science (and in science itself) unless we understand what is wrong, exactly, with the formulation $\Xi = 0$. But everything changes once we can clarify that conundrum.

2 A simple, but not too simple, solution of the riddle

To understand what is wrong with $\Xi = 0$, it is not enough to observe that Ξ is probably hard to measure, if at all. We need a clear criterion that excludes the formulations that should be excluded, but not more. The key observation is that, **although we can express any model as $\Xi = 0$, we cannot, at the same time, expect measurements in**

[4] Moreover, the difference between scientists and non-scientists is significantly more blurred in the age of AI.

[5] See in particular (Goodman, 1983): *"what we want, indeed, is an accurate and general way of saying which hypotheses are confirmed by (...) any given evidence"*.

the form of a central value and a connected error-bar as $\Xi = \Xi_0 \pm \Delta$. In fact, if we could, we would know from the formulation of the model itself what is measurable and what is not and the expected precision of any measurement. But we do not have this information for most real modern theories. So, this representation cannot be logically and empirically equivalent to any realistic modern theory. Chaotic billiards (Scorzato, 2013), deep neural networks (Scorzato, 2024) and the measurement of grue at t_0 (later in this section) provide examples where the limitation comes from fundamental physical reasons. The measurement of grue at time $t \gg t_0$ and the example in Sec. 3.2 show that this is often not possible for practical reasons.

Can we use this idea to define model selection? We can do it as follows. We are looking for a notion of complexity of the assumptions (epistemic complexity, Def. 3) that can be combined with accuracy to determine model selection (Def. 4). To be plausible, such notion should be invariant by reformulation of the model. Invariance can be easily achieved by taking the minimum over all possible equivalent reformulations. But we have seen that, if we consider all *logically equivalent* reformulations, the minimum is trivial. The key is to realize that logical equivalence is not sufficiently restrictive.

In fact, two formulations are truly equivalent only if they are also *empirically equivalent* (Def. 2), which doesn't follow automatically from being logically equivalent. Two formulations are empirically equivalent if their measurements are consistent with the same precision in both formulations. We do not need to list all the measurable quantities, when we formulate the assumptions of a model. But we must at least enumerate a *basis of directly measurable concepts* that is sufficient to define, operationally, all the other measurable ones (not the theoretical ones!). Such basis must be part of the model assumptions to claim unambiguous interpretation of the model's empirical content (Def. 1). There is some freedom in the choice of the basis, but it is also constrained by the need of (i) enabling an operative definition of any other measurements and (ii) being plausibly directly measurable, which entails the minimal requirement identified in Post. 1. This is where convexity plays a crucial role. These constraints are now sufficient to ensure that the shortest formulation among all logically *and empirically* equivalent ones, is, in general, not trivial anymore.

2.1 A postulate on direct measurements

A necessary requirement of any direct measurement can be identified in the following:

Postulate 1 *The result of a valid single direct measurement of (k-dimensional) property Q is always expressed as a (k-dimensional) central value Q_0 and a (k-dimensional) convex set (error-box) that contains Q_0.*

Post. 1 reflects the scientific practice of quoting error-boxes as a k-cell[6]. For example, if we measured the temperature of a room once, it makes no sense to say that the result was "either $20 \pm 1C^\circ$ or $30 \pm 1C^\circ$". This outcome might potentially result from multiple direct measurements (e.g. taken under two types of conditions) or from a derived measurement (e.g. obtained as solutions of constrints from other measurements), but not from a single direct one. Another example is shown in Fig. 1. The error-box on the left is a legitimate outcome a single direct measurement in two dimensions. On the contrary, the black region displayed on the right cannot represent a legitimate outcome of a single direct measurement.

Note that I have not defined *direct measurement* explicitely. Post. 1—together with Def. 1 below—offers an implicit (partial) definition of direct measurements. In fact, all we require from direct measurements is that any measurement must be constructible from a basis of direct ones, which must also fulfill Post. 1. Under these constraints, it is up to the model to decide which are the direct measurements and the derived ones (see following sections).

Although Post. 1 is very intuitive, we should remark that error-boxes are merely short-hand notations for features of the expected probability distribution of a measurement. It is therefore worth introducing an alternative formulation in term of probability distributions, as in the following:

Postulate 1' *The contour sets[7] of the expected probability distribution $P(Q)$ of a valid single direct measurement of a property Q are convex sets, for each level l.*

[6] Because these regions are small, the difference between a convex set and a k-cell is not significant. In fact, any k-dimensional box is convex and in any convex set we can inscribe an k-dimensional box (Behroozi, 2022). For simplicity, we use the name error-box in the general case.

[7] The contour set of P at level l is the set $\{Q : P(Q) \geq l\}$.

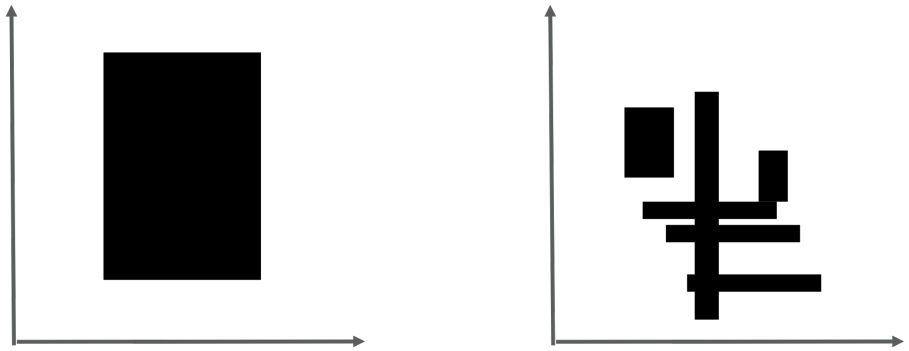

Fig. 1 Left: a legitimate error-box outcome of a single direct measurement in two dimensions. Right: not a legitimate one.

Fig. 2 Left: a legitimate probability distribution of a single direct measurement in one dimension. Right: not a legitimate one. Note that this is *not* the *empirical* distribution of *multiple* measurements, but the *expected* distribution of a *single* measurement.

In particular, the distribution on the left of Fig. 2 is a legitimate expected distribution of a single (one dimensional) measurement, while the distribution on the right of Fig. 2 is not[8]. It is important to emphasise that P is *not* the *empirical* distribution of *multiple* measurements, but the *theoretical expected* distribution of a *single* measurement. The former could very well be multimodal, but the latter cannot. In the following, I will refer to Post. 1, for simplicity. But one can easily (alghough tediously) verify that all important conclusions would be maintained under the more general Post. 1'.

Post. 1 represents a special case of the requirement proposed by Gärdenfors that natural properties form convex sets in conceptual spaces (Gärdenfors, 1990; Gärdenfors, 2000; Gärdenfors and Stephens, 2017). A deeper comparison with Gärdenfors' proposal is discussed in Sec. 4.2. Here, I only note that I do not introduce the concept of *natural properties*. Instead, I require Post. 1 only for directly measurable properties and only for the small regions that correspond to the uncertainty of a measurement.

What does the previous discussion imply for Goodman's grue?[9] If I measure the color of an emerald at the critical time t_0 (when the color of newly seen emeralds might be blue, according to Goodman's model), I have two uncertainties: one on the color wavelength and one on time (see left panel of Fig. 3). If I translate this observation from the blue/green representation into the grue/bleen one, I cannot be sure if the emerald is seen before or after t_0, so the error-box gets split, consistent with the fact that I measured directly color and time, not grue/bleen colors (see right panel of Fig. 3).

[8] It is not difficult to extend Post. 1' to measurable quantities defined on discrete sets. But it is quite instructive to review the options in detail. This is done in Appendix A.

[9] Note that in this paper, I adopt the definition of grue originally proposed in (Goodman, 1955): emeralds are grue iff they are green and they are *first seen* before t_0 or they are blue and they are *first seen* after t_0. So, individual stones do not change color. We could alternatively define grue emeralds as green before t_0 and blue emeralds after t_0 (Gärdenfors, 1990). In the latter model, individual stones could change color at t_0. As correctly noted by Gärdenfors (1990), the core idea of Goodman's new riddle and Gärdenfors' argument remain unchanged. In both cases, there is an uncertainty on the measurement of time, whether it is the time of first observation or the time of a subsequent observation, and an uncertainty on the measurement of wavelength. However, if we use Gärdenfors' model, we must distinguish two cases. If the stone does not change color at t_0, Fig. 3 remains unchanged, both right and left. If the stone is observed while it does change color (at around t_0), the horizontal errorbar necessarily increases (both right and left). In this case, the error-box doesn't completely split, but the measurement is less precise. This does not change the conclusion about non-convexity and lack of direct measurability of grue.
On the other hand, Goodman's original definition has an additional interesting twist for times $t \gg t_0$, because it is impossible to determine the grue/bleen property of an emerald at later time, unless we keep track of when the emerald was first observed. Gärdenfor's definition does not have the same problem for $t \gg t_0$.

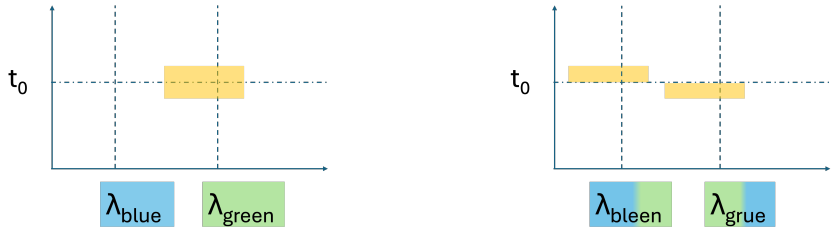

Fig. 3 Left: error-box (in yellow) associated to the measurement of the color of an emerald around the critical time t_0. Right: the same error-box appears split in the grue/bleen representation. Cfr Fig. 4 in (Gärdenfors, 1990).

Grue *is* measurable in the sense that I can estimate its value (I just need a colorimeter and a clock). A jump on its value is not a problem: many physical quantities display jumps. But when I measure directly a quantity with jumps, I see a *large* error-box, not a *split* error-box. Split error-boxes are incompatible with a direct measurement, although they are fully acceptable for indirect measurement. We can insist that we measure grue directly only at the cost of a decreased accuracy, which corresponds to replacing the split error-boxes in the right panel of Fig. 3 with at bigger box that contains both. In other words, the asymmetry between green and grue is due to the fact that, to measure grue, we also need a colorimeter (plus something else). The asymmetry can be characterized precisely only when considering the error-boxes.

Besides the violation of convexity around time t_0 discussed above, there are more problems even well after t_0. To appreciate it, consider a bunch of green emeralds at time $t \gg t_0$. Are they grue or bleen? They would be grue if they were first seen before t_0, and bleen if they were first seen after t_0. But, if the time of first detection *has not been recorded*, then we cannot tell: they are undetermined in the grue/bleen dimension. In other words, for the set of emeralds that we can practically collect, the grue property is not convex[10], even well after t_0. This example highlights a limitation of grue that is not traced back to fundamental physical laws—like the impossibility to measure time with arbitrary high precision—but it is nevertheless a practical limitation, because we do not record all possible information (even when we could), if we do not consider them useful. This scenario is very important in practice, as I will discuss in Sec. 3.2, and it is reminiscent of Goodman's own solution that relies on *entrenched* concepts. However, the logic is very different from the one proposed by Goodman, as I will clarify in Sec. 4.1.

Does the requirement of convexity solve Goodman's riddle? Not yet, although it is a key step in that direction. In fact, by itself, the requirement of convexity is both too weak and too strong to identify projectible laws. It is too strong because scientific theories rely fundamentally on concepts that are not even measurable (e.g. the quark wave function). For non-measurable concepts, the choice of the metric that determines convexity is too arbitrary to offer a useful selection. On the other hand, the requirement is also too weak because by using only natural and convex properties we can still build accurate but implausible models. For example, consider a model based on a simplistic general rule together with a long list of exceptions.

However, the idea of convexity does achieve an important result: it represents a well defined, general and detectable property that breaks the symmetry between green and grue. As a result, defining the grue model for emeralds is more *complex* than defining the standard green model. In fact, if we insist that any model must introduce, as part of its definition, all directly measurable properties that are necessary to derive its conclusions, then the concept of grue has an objective disadvantage with respect to the concept of green, because it needs an *additional* step to be defined. But, to make this statement precise, we now need to define a notion of complexity of (the assumptions of) a model. This is the goal of the next section.

2.2 Epistemic complexity and scientific model selection

The goal of this section is to define a philosophical model of scientific model selection that is based only on empirical accuracy and the complexity of the model's assumptions. To define the complexity of the assumptions of a scientific model I must first clarify what I mean by a scientific model. Requirements are kept to a minimum at this stage.

[10] Here I rely on Gärdenfors (1990) extension of the notion of convexity to discrete sets. See also the Appendix A.

Definition 1 *A* **model** *is a tuple* $\mathcal{M} = \{P, R, B\}$*, where:*

- P *is a set of assumptions*[11]*;*
- R *is a set of results, which are logically derived from* P*;*
- B *is a set of basic measurable quantities that enter the* P *and are assumed to be directly measurable with precision* $\Delta(b)$ *(*$\forall b \in B$*).*

$P(\mathcal{M})$ contains all the assumptions needed to deduce the results R to be compared with the experiments (including the rules of logic, all required mathematical assumptions, suitable model of the experimental devices, approximations, background science, initial conditions, tolerance Δ of all the quantities B that we assume to be directly measurable[12]). Any result in R that can't be derived from the assumptions must be part of the assumptions.

The above structure has some similarities to the much criticized received view (Feigl, 1970). Hence, it is important to stress the key differences. We can't assume observation sentences or even properties that are theory independent. The measurability of B, with a specific precision Δ, is part of the assumptions that can be tested only holistically (Quine, 1950, 1991). No general and neutral observation basis is assumed and none is needed. Still, theories can be tested against each other. This is possible as long as a directly measurable basis exists that *can be shared among those theories*. Although incommensurability (Oberheim and Hoyningen-Huene, 2025) remains a theoretical possibility, there is no evidence of two models dealing with the same topic where it is not possible to find a common sub-model that includes all necessary directly measurable quantities (see e.g. Fletcher (2024) for a recent study).

It is worth elaborating more on the differences between the present framework and the one of Carnap (1966). The idea of deriving basic observational properties from the observations of similarities between elementary perceptions has proved impossible both because n-ary (for any fixed n) similarity relations do not contain sufficient information (Leitgeb, 2007), and (even more fundamentally) because we cannot *ensure* that anyone will see similar objects in the same way. We can share prototypical examples[13], we can add narrative, but, no matter how much effort we put on clarifying a concept, we can always only *assume* that some similarities will be unambiguous. All that we can do is to test those assumptions (but only holistically, together with the other model assumptions) and analyse statistically any unexpected outcome.

Equialent formulations. As per Def. 1, two different formulations of the same model are seen as different models. It is therefore important to identify an equivalence relation between models.

Definition 2 $\mathcal{M}$ *and* $\mathcal{M}'$ *are* **equivalent formulations** *(*$\mathcal{M} \equiv \mathcal{M}'$*) iff there is a translation* J *between* $\mathcal{M}$ *and* $\mathcal{M}'$ *that:*

- *preserves logical structure and theorems (***logical equivalence**[14]*);*
- *for each measurable property*[15] p *of* $\mathcal{M}$*,* $J(p)$ *is also measurable for* $\mathcal{M}'$ *with the same precision and same outcome (via* J*). (***empirical equivalence***)*

It is well known that two models can be empirically equivalent while logically inequivalent (Mormann, 1995). For example, Einstein mechanics is empirically indistinguishable from classical Newton mechanics for phenomena whose velocities are much smaller than the speed of light. On the other hand, two models can be logically equivalent, but empirically inequivalent. This possibility is less discussed in the philosophical literature, but it is quite obvious to the scientific practitioner. For example, if I define the unit of length based on my foot, rather than the modern reference in (NIST, 2019), I obtain an alternative model that is logically equivalent to the original one (the assumptions of the model are exactly the same, except for what we chose to label as directly measurable), but significantly less accurate than (hence empirically inequivalent to) the original one. Note that the previous discussion lets us conclude, in particular, that a $\Xi = 0$ 'reformulation' of model $\mathcal{M}$ is not, in general, empirically equivalent to the original model $\mathcal{M}$ and it is therefore *not* just a reformulation.

[11] 'Assumptions', 'Hypotheses', 'Principles', 'Postulates' are used as synonyms in this paper.

[12] It might be convenient to distinguish core assumption, that we rarely change, from auxiliary assumptions that we often change (e.g. boundary conditions). Correspondingly, classic models like 'Newton Gravity' can be seen as a family of models as defined here.

[13] E.g. we can show many examples on how to use a yardstick under different circumstances.

[14] See the concept of bi-interpretability in (Visser, 1991, 2004).

[15] Measurable properties are the concepts of $\mathcal{M}$ that can be (operationally) defined from directly measurable properties.

Epistemic complexity. The equivalence class of models that results from Def. 2 is, finally, the object whose complexity we must define, to make precise Einstein's intuition of the *complexity of the assumptions*. Indeed I can now define the epistemic complexity of a model $\mathcal{M}$ as the minimum, over all equivalent formulations, of the length of its assumptions:

Definition 3 *The* **epistemic complexity** $\mathcal{C}$ *of a model* $\mathcal{M}$ *is the minimal length—across all possible equivalent formulations (in any language) of* $\mathcal{M}$*—of the assumptions* $P(\mathcal{M})$*. In other words:*

$$\mathcal{C}(\mathcal{M}) := \min_{\mathcal{M}' \equiv \mathcal{M}} \textit{length}\left[P(\mathcal{M}')\right].$$

The **epistemic simplicity or conciseness** *of a model* $\mathcal{M}$ *is the inverse of its complexity.*

This definition is inspired to Kolmogorov-Chaitin (KC) complexity (Kolmogorov, 1965; Chaitin, 1975; Zenil, 2020). However—and this is the key difference—KC complexity makes no reference to measurability. So, it must be defined for a fixed, externally given language. Otherwise, there is always a language (or Turing machine) in which KC becomes trivial (the $\Xi = 0$ formulation discussed before). The dependence on the language is fatal for epistemological applications of KC complexity, because different choices allow any conclusion.

But, if I restrict it to *logically and empirically equivalent formulations*, I ensure that $\Xi = 0$ is not anymore a legitimate version of my original model and I have a definition that is both **non-trivial** and **formulation independent** by construction (it only depends on the choice of what I can measure, which is given by the nature of the problem and not by an arbitrary choice). Moreover, Epistemic Complexity is **defined precisely** but **estimated approximatively** (as most quantities in science). As a result, it **justifies the use of ordinary scientific language to assess simplicity**. This is not a small feat: it is the the only example I know of a non-empirical epistemic value which is precisely defined, non-trivial and as much formulation-independent as one can possibly wish.

Model selection. Now that we have at least one well defined non-empirical epistemic value which is non-trivial and represent what we were looking for, can we build a model for model selection based on empirical accuracy and conciseness alone? This is detailed below.

How do scientists compare two scientific models, to decide whether any of them should be excluded? If they are not empirically equivalent, the scientists first identify the corrections (ad-hoc assumptions) that would be needed to make the least accurate model as accurate as the best one. If this requires too many or too complex corrections to the assumptions—beyond the estimated uncertainties in the assessment of the complexity—it makes sense to drop the most complex model, because it is just more complex for no empirical advantage. Note that there is no trade-off in this selection: it only eliminates models that are unambiguously worse than some other model with no advantage (the red area in Fig. 4). This selection is in fact uncontroversial among scientists. They do not even consider them as options (that's why they don't feel they are using epistemic complexity as a formal selection criterion at all). But these models are legitimate from a logical point of view, they are infinitely more than "good" models, and they are subtle to identify from the philosophical point of view. We can call them *ruled-out* models and defining them is the focus of this work.

On the other hand, the models that are not worse than any other model represent the *state-of-the-art* (SotA) models (the green surface in Fig. 4). Different SotA models represent different trade-offs between simplicity and accuracy in different applications. Choosing among these models is often the focus of the scientists, who may select a model based on the minimal precision required by a specific application or they may opt for a Bayesian model average. The selection between different SotA models is not the focus of the this work, nor it is the concern of Goodman's new riddle, because they are all legitimate (non-grue) extrapolations. In summary, model selection, state-of-the-art models and ruled-out models are defined as follow:

Definition 4 (Model selection) *Given a set of empirical questions (i.e. a topic* τ*), a model* A *is preferred to model* B *if* A *is neither more complex nor less empirically accurate than* B *on the topic* τ*, while being strictly better (beyond uncertainty) than* B *in at least one of these aspects. In this case, we say also that model* A *is* **better** *than model* B *and* B *is* **worse** *than* A*.*

Definition 5 *Given a topic* τ*, the* **State-of-the-art** *is the ensemble of models which are not worse than any other model for the topic* τ*. The models that are not state-of-the-art are* **ruled-out***.*

Note that model selection depends on what each model is able to describe accurately, but it does not depend on the specific basis chosen for the directly measurable concepts (within the same class of empirically equivalent models, which does not identify, in particular, models with different precision).

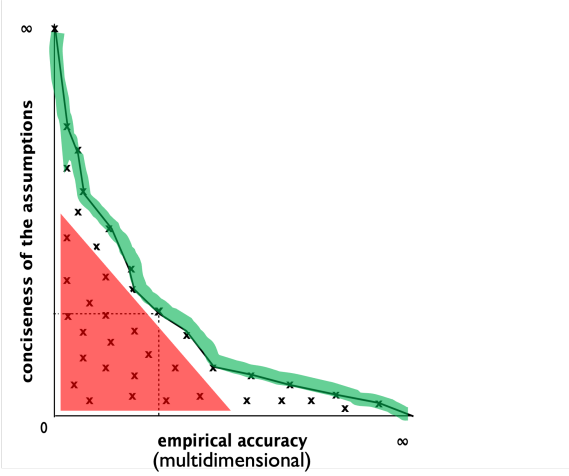

Fig. 4 Each 'x' represents a different scientific model, plotted by accuracy (which is multi-dimensional, but it is shown here as one-dimensional for simplicity) and conciseness. Ruled-out models appear in the red ares. State-of-the-art models are those in the green surface.

3 On the justification

A common misconception among modern philosophers of science is the idea that there are many possible definitions of simplicity and it is therefore arbitrary to pick one. If that were the case, we could select the definition(s) of simplicity that describe best the actual (consensual) choices of the scientists: those definition(s) would already provide a better solution to Goodman's riddle than anything else proposed. We do not have this option, because the real problem is not the *abundance* of definitions, but rather the *absence* of any definition that survives the Ξ trap. Any such definition is actually trivial under reformulation of the model and therefore totally non-descriptive[16].

After proposing a definition of simplicity which doesn't fall in the Ξ trap, the next step is not to try to justify this choice on the basis of some other (inevitably metaphysical) principle—which would be yet another attempt to solve the impossible *old* riddle of induction. The next step is to assess whether this definition of simplicity (and the associated model selection) provides a good description of those decisions that enjoy a broad support by the scientific community. Multiple examples supporting the accuracy of the present model have been published already, covering reductions and unifications (Scorzato, 2013), gravitational theories, quantum mechanics, theories of evolution (Scorzato, 2016), pseudo-science (Scorzato, 2015), and machine learning models (Scorzato, 2024). A new example is described in Sec. 3.2. To challenge the philosophical model examined in this paper one should find at least one counterexample, namely a model[17] that:

- either *should* be ruled-out and my philosophical model *doesn't*,
- or *shouldn't* be ruled-out and my philosophical model *does*.

Here, *should* and *shouldn't* refer to selections supported by broad scientific consensus. In other words, the proper test of this philosophical model must be performed with cases of scientific model selection that are undisputed. For those questions where there isn't a clear scientific consensus, the present model offers an original prediction.

Before analyzing a new example in Sec. 3.2, I elaborate more, in Sec. 3.1, on the rationale and the implications of the choice of epistemic complexity introduced in Def. 3.

3.1 On the choice of the notion of epistemic complexity

As already explained above, the choice of epistemic complexity in Def. 3 is motivated by the need to identify the hidden assumptions behind the actual scientific model selections supported by a broad scientific consensus. However, the question of why choosing Def. 3 is a very common one, and I elaborate on it further in this section.

[16] Unfortunately, this point is rarely discussed in the literature, where philosophers often criticize proposals as *merely descriptive*, while they are not. In fact, they are typically so vaguely defined that they might appear descriptive, but they are, strictly speaking, just undetermined.

[17] The model should cover a *relevant* domain. The present philosophical model does not try to determine which topics are relevant.

For example, one may wonder why I did not chose other classic measures of complexity (e.g. Akaike (1973); Schwarz (1978)). What distinguish these measures (on one side) from Kolmogorov-Chaitin and epistemic complexity (on the other side) is that the latter are very generally applicable: they can be used to compare models that make totally different assumptions (but describe the same phenomena). This is necessary to understand model selection across revolutionary times, or to address potential challenges from pseudo-science, or simply to compare models based on very different formulations. Measures like those of Akaike (1973) or Schwarz (1978) are only defined within a given parametrization, because only there it makes sense to compare the number of parameters.

Other notions of complexity do not capture the complexity of the assumptions and, therefore, they do not capture what matters for the scientists to select a model. For example, one could consider the complexity of *deriving results* from a model (such as 'proof complexity' (Krajıcek, 2004) or the computational complexity to derive a prediction from a Neural Network). This category definitely matters to assess the opacity of a model (Beisbart, 2021), but it does not introduce a new, independent, dimension valuable for scientific model selection. In fact, imagine that model $\mathcal{M}$ is as accurate as $\mathcal{M}'$, $\mathcal{M}$ has simpler assumptions, while $\mathcal{M}'$ enables simpler derivation of results. Would I ever select $\mathcal{M}'$ over $\mathcal{M}$? If the advantages of $\mathcal{M}'$ had enabled the derivation of more (accurate) results than $\mathcal{M}$, then $\mathcal{M}'$ would be more accurate than $\mathcal{M}$, but since this is not the case, by assumption, then the advantages of $\mathcal{M}'$ are only hypothetical and questionable. In other words, the simplicity of the derivations is already taken into account by the value of accuracy, for the extent that it is indeed a confirmed advantage. I don't know real cases that contradict this conclusion. Similar arguments can be made for other notions of complexity.

It is possible that other notions of complexity exist that are different from the epistemic complexity discussed here, but implies **the same conclusions** of model selection, within estimation errors. This would not challenge the present model: it would provide another perspective on the model discussed here, but it would be consistent with it. On the other hand, a different definition of complexity that leads to **different conclusions** for model selection is interesting only if one first identifies some cases where epistemic complexity leads to a model selection that differs from the scientific consensus. However, no such couter-example has been published, until now, to the model already published in (Scorzato, 2013). In fact, the claim presented earlier in this Sec. 3 remains unchallenged.

3.2 History as a scientific discipline

I claim that the characterization of model selection discussed in this paper is very general. A common objection is that epistemic complexity makes sense for highly formal domains, like physics, but less so in scientific domains where mathematics plays a less prominent role. To answer those criticisms, I consider, in this section, a field as removed as possible from highly mathematized ones: historical science. History is certainly a science and the task of the historian is to formulate conjectures whose likely effects agree with the available documents.

A major challenge for philosophy of science, when applied to history, is clarifying what's wrong with conspiracy theories. In fact, they are dismissed by the vast majority of historians, but their empirical accuracy is usually not the problem: they are often designed to agree with all the evidence and also to adapt quickly to any new evidence. Moreover, saying that conspiracy-theorists' assumptions are *unlikely* is also unsatisfactory. Indeed, they generally envisage circumstances that are very special by construction and, therefore, can't be declared unlikely, because there are no statistical data either in favor or against them.

Consider the example of the *Bielefeld conspiracy* (Wikipedia, 2025)[18]. It claims that Bielefeld does not exist. According to the theory, if you say that you have never been in Bielefeld, you confirm that it might not exist. If, instead, you say that you have been there, you must either be part of the conspiracy or have been deceived by it, which also confirms its widespread penetration. The theory always has an answer to any counter-evidence. The claim that it is *unlikely* that so many people lie or have been deceived about having been in Bielefeld is not justified. In fact, to explain its widespread diffusion, the conspiracy calls upon alien forces and extraordinary hidden organizations which are unique events by construction. No statistical evidence exists to either support or rule out the claim. Once again, accuracy and probability alone can't dismiss such conjectures. We can dismiss them only because they require *extra and complex assumptions*, which are not necessary to explain the evidence. Hence, we face again the problem of quantifying the amount of assumptions.

[18] A satirical theory plays here a useful role analogous to a *thought experiment* in physics: it allows discussing the essential features of a conspiracy theory without the interference of other complex factors that are inevitable in any seriously meant conspiracy theory.

Can I use the Ξ trick to make a conspiracy theory as concise as the standard one? It is very instructive to see what happens in this case. The proponents of the Bielefeld conspiracy implicitly make the following two assumptions (or similar ones):

- In general, people lie with a probability of $< 1\%$ [ordinary assumption].
- Except, people who allegedly lived in Bielefeld, who lie all the time [specific ad-hoc assumption necessary to justify the conspiracy claim].

We can hide the complexity of the second assumption if we say that Ξ-people stands for anyone, except those who allegedly lived in Bielefeld (who lie all the time). Then the assumption of the proponent of the conspiracy becomes as concise as the ordinary assumption:

- In general, Ξ-people lie with a probability of $< 1\%$.

But to gain any advantage from this reformulation, one should then use *Ξ-people* rather than *people* everywhere in history and the body of science. But to preserve empirical accuracy in this new formulation, one should ensure full convertibility between the concepts of *people* and *Ξ-people* in every measurement. In particular, any survey about any topic should also ask whether the respondent lived in Bielefeld! The same information should be verified about any person who is part of any studies or plays any role in any topic.

Such measurements are not prohibited by any fundamental natural law—as opposed to those involving chaotic systems discussed in Scorzato (2013)—but they are nevertheless not available. Proponents of the Bielefeld conspiracy cannot just claim that the notion of *Ξ-people* is, logically, as legitimate as the notion of *people*: they should provide evidence of measurements expressed in terms of *Ξ-people*. Just as a set of emeralds, whose first discovery was not recorded, is not convex under the grue property (see Sec. 2.1 and Footnote 10), a set of people whose stay in Bielefeld was not recorded is also not convex under the property of Ξ-people.

In conclusion, even for domains that rely on minimal mathematical background, gaining conciseness artificially is logically possible, but only by compromising accuracy, which means that the conciser model is not empirically equivalent to the original one.

4 Discussion

4.1 Relation to previously proposed solutions

Since Goodman posed his riddle 70 years ago, many solutions have been proposed. The goal of this section is not to review comprehensively the huge literature on this topic, but rather to compare the present model to those proposals that have interesting similarities and differences.

The idea that grue-like concepts are **not observable** has been put forward very early (Goodman, 1983), but it is not correct: to observe the grueness we simply need a detector of colors and a clock. Here, consistently with Gärdenfors (1990); Gärdenfors (2000), I have emphasized the importance of what distinguishes *direct* observations.

Fodor's idea (Piattelli-Palmarini, 1980) that some hypotheses are **innate** was already dismissed by Putnam as a potential solution (Goodman, 1983). In particular, scientists often introduce new hypotheses and concepts that cannot be considered innate, but they are nevertheless very successful (e.g. quarks). On the other hand, feasible and accurate *measurements* have the right degree of flexibility: they are neither fixed nor arbitrary. Scientists are able to design new experimental devices, but doing so is not as easy as introducing a new grue-like concept. The constraint of what is measurable is also a natural one for scientific models.

A vast literature (Bird and Tobin, 2008; Brzović, 2014) has tried to characterize the notion of **natural kinds** and the related notion of **similarity** (Quine, 1969). Identifying the right concept of similarity is fraught with issues (Fletcher, 2016). The original idea was that only natural kinds are used for projectible laws. The program attracted intense research over half a century, but it proved too ambitious. I will not review the multiple dead ends the program ran into, as it is done very well by Bird and Tobin (2008) and Brzović (2014). I will just explain verily briefly why I believe that the project itself is not a good idea. On one hand, it seems exceedingly difficult to be able to tell why the concept of *quark* (which lies at the heart of our best scientific theories) should be more natural than many grue-like concepts[19]. On the other hand, it is still possible to build accurate, but completely implausible

[19] Except if we *use* the fact that the concept of quark does appear in our best scientific theories. But then we can't use the concept of natural kind to tell what *should* be projectible, which misses the point of why natural kinds were introduced in the first place.

models based only on very natural kinds. This is easily accomplished by formulating very crude general laws, accompanied by a long list of exceptions. In this respect, Gärdenfors (1990) introduces a key refinement of the definition of natural properties, but he doesn't go beyond the idea that essentially identifies what is natural with what is projectible. This identification is problematic, because it leads to a criterion that is both too strong and too weak as explained above and in the end of Sec. 2.1.

My proposal can be seen as limiting the role of natural kinds to where it is strictly necessary: only for properties that we consider directly measurable. The criterion for admissibility is consistent with the one proposed by Gärdenfors (1990): natural properties are convex, at least locally[20], but natural (or directly measurable) properties are not enough to characterize what is projectible. The other essential component is conciseness. But directly measurable properties effectively introduce constraints on the language that enable a non-trivial definition of conciseness.

Goodman's own theory of **entrenchment** has been criticized by many authors (Stalker, 1994; Elgin, 1997; Cohnitz and Rossberg, 2024). Some of them (Cohnitz and Rossberg, 2024; Scholz, 2024) consider his solution merely descriptive. This is not true. If it were so, it would be exactly what Goodman was looking for. But it is not, as many others have pointed out (Teller, 1969). It is important to review why it is not.

First, it is very unclear how entrenchment is supposed to be assessed: (i) when does a past hypothesis count as the same hypothesis? The same sentence is not the same hypothesis, strictly speaking, when combined with different other assumptions, consistently with the idea of holism. If we adopt this strict view, we preclude any useful application of entrenchment. If we don't, we must define a similarity measure among different hypotheses, which was not addressed. Even if we succeed, (ii) how do we count how much an old hypothesis was used successfully? Do my home experiments (that I can repeat thousands of times per day) count as much as an experiment conducted at CERN after twenty years of preparation? This is a reformulation (not a solution) of the problem of confirmation (Crupi, 2021). Secondly, even if all these major uncertainties were settled, the model would still be wrong even in those cases where its interpretation is rather unambiguous. In fact, sticking with an old model and adding many ad-hoc corrections to it would be clearly preferable than introducing a completely new simple and accurate model.

There is, however, also some truth in Goodman's theory: past successful models do carry a legacy, but not in the sense that they should be preferred, ceteris paribus, to more recent ones. The legacy exists becuse past successful models determine which features we decide to record, and because old measurements remain a reference to asses any new model, as discussed in Sec. 2.1 and Sec. 3.2. Althought this might (or might not) be an advantage for the older model, it does not introduce arbitrariness into the comparison, because it is a fact that is hard to change.

Finally, one of the most popular modern approachs to confirmation theory (Crupi, 2021) is **Bayesianism** (Sprenger and Hartmann, 2019). The outcome of a Bayesian analysis depends on the choice of the prior probabilities (aka priors), which remain relevant for any realistic amount of data and any non-toy application (Scorzato, 2024). In turns, the priors can only be defined by relying on some non-empirical value, which is vulnerable to the Ξ trick, unless we adopt a non-trivial, reformulation independent measure of complexity, whose only published option is Def. 3. If we do that, it makes no sense to keep ruled-out models (according to Def. 4) in the Bayesian mix, because no evidence can ever prefer them over some state-of-the-art model, but they add unbounded perturbations to the Bayesian outcome. Hence, also Bayesianism can make sense only if its definition relies on epistemic complexity, which then makes it equivalent to the model defended here.

4.2 Conceptual spaces

I have already emphasized the deep relation between the present proposal and the one of (Gärdenfors, 1990; Gärdenfors, 2000). In this section, I elaborate more on this relation, focusing on the interesting analysis presented in Gärdenfors and Stephens (2017). The authors distinguish three types of knowledge: *knowledge-how*, *knowledge-that* and *knowledge-what*, that correspond, respectively, to three different types of memory: *procedural*, *semantic* and *episodic*.

I acknowledge that the distinction plays an important role in understanding human cognition. However, science differs significantly from natural human cognition: it may share the same basic functionalities, but it is far from instinctive, it involves additional deep conceptual elaboration and it is often unnatural for humans. Importantly, science strives to reduce all knowledge to knowledge-that, and it is mostly successful in this: the documentation

[20] Requiring convexity only in a small region around a measurement also answers the criticism that natural (or measurable) properties might not be globally convex (Hernández-Conde, 2017), although the specific example provided there is not correct (Gärdenfors, 2019). Note that the difference between a generic convex set and an error-box is inessential, for small regions.

of an experimental setup is an excellent example of removing any dependencies on any informal know-how and ensuring that the process is fully reproducible. Other authors (Williamson and Stanley, 2001) have argued that knowledge-how can be fully reduced to knowledge-that, and I agree with their conclusions.

Matters are different, however, when it comes to *knowledge-what*. While scientists still strives to reduce, as much as possible, any knowledge-what to knowledge-that, there are fundamental limitations that preclude a full reduction. For example, the correlation between the dimensions that characterize an apple can be expressed in terms of formal propositions that state statistical correlations. Moreover, whenever precision matters, the scientists will not rely on the intuitive notion of apple, but rather on genetic analysis. However, *some* basic measurements cannot be expressed in terms of propositions that can be verified as true or false (i.e. knowledge-that), except by introducing vicious circles. The determination of color can be reduced to a measurement of light wavelengths, but then we must assume a model for the spectrometer and what we measure directly is simply shifted from the relation eye-apple to the relation eye-spectrometer display (and many other direct measurements for set-up and calibration). In other words, there is an irreducible core of knowledge-what that is represented by the fundamental direct measurements that any model must assume somewhere. Science tries to reduce their scope to quantities that are as unambiguous as possible, but even if we could reduce every direct measurement to the reading of a digital display that prints either 0 or 1 on the screen, we should still assume that different people at different times will interpret those slightly different symbols in very different contexts in the same way. This is where the dream of a provably unambiguous observational language fails and must be replaced with assumptions about the appropriate knowledge-what and its learning mechanisms, which can be verified only holistically.

In other words, the present view is consistent with (Gärdenfors and Stephens, 2017), but with the caveat that the scope of knowledge-what, in science, is reduced to the essential. This reduces the need of convexity to small neighborhood around measurable points. This point of view also mitigates the implications of the criticism brought forward by Strössner (2022) to the applicability of conceptual spaces to natural multi-domain concepts.

5 Conclusions

The solution to Goodman's riddle of induction discussed in this paper is based on a combination of two main insights: (i) conceptual spaces (Gärdenfors, 1990)—applied only to directly measured concepts; (ii) the theory of complexity (Scorzato, 2013; Chaitin, 1975; Kolmogorov, 1965), which is used to characterize—in a formulation-independent way—the complexity of the assumptions of a model. In one sentence, **it is the constraint of convexity that enables a non-trivial notion of complexity**.

This provides a well defined model that makes it precise the informal idea that science always aims at *explaining more with less*. In the spirit of Goodman's *new* riddle, I do not try to explain *why science is successful*, I rather clarify *what we mean by science*, i.e., I identify the hidden assumptions behind the scientists' decisions about scientific model selection.

Most philosophers of science, today, believe that Goodman's new riddle cannot be solved. This is implausible for multiple reasons. First, it amounts to saying that the conclusions of science are based on fundamentally ineffable assumptions[21]. This claim implies the futility of trying to clarify the exact assumptions behind any scientific conclusion, which is one of the main focus of scientists of all discipleines. The only way to make sense of science is to admit that the scientists do use some hidden assumption, but they are identifiable.

Another reason why the skepticism is implausible is that some natural options had never been explored seriously before. In particular, although both *measurements* and *complexity* are classic topics in philosophy of science, I am not aware of any work (except for Scorzato (2013)) that tries to define them in combination. It is very natural to expect that complexity makes sense in science only after introducing a constraint of what is measurable with the acceptable precision. How can we claim that Goodman's riddle cannot be solved before exploring such a natural approach in full depth?

My philosophical model won't be the last word on the goals of science, but it is certainly the best description currently available: it achieves excellent accuracy (I am not aware of any counterexample) at the cost of a moderately higher level of sophistication than usual. To criticize it, or improve it, one should first identify at least one counterexample, namely a model that:

– either *should* be ruled-out (according to scientific consensus) and my philosophical model *doesn't*,

[21] This far worse than admitting that all conclusions of science are based on assumptions that we cannot conclusevly test, which is true and recognized by everyone, except for the philosophically most naive.

– or *shouldn't* be ruled-out (according to scientific consensus) and my philosophical model *does*.

In the age of AI, the role of philosophy of science is more crucial than ever to assess the reliability and to guide the evolution of models that represent a radical break with the tradition of scientific modeling (Scorzato, 2024). But this requires a philosophy of science that adopts for itself the same standards that it identifies for all scientific disciplines.

A Convexity and discrete measurements

It is not difficult to extend Postulates 1 and 1' to the case when the set of possible measurement outcomes Q is discrete. This is necessary, for example, to describe measurements whose possible outcomes are `true/false`, integer numbers or other finite set of categories.

A simple way to extend Post. 1 and 1' consists in chosing a metric $d(.,.)$ in the space of Q and define an error-box around a given value Q_0 as the set of all Q such that $d(Q_0, Q) < \Delta$. The level set $\{Q : P(Q) \geq l\}$ and the concept of convexity are still well defined on a metric space (Khamsi and Kirk, 2001) and Post. 1' is still meaningful and can be extended without change.

It is important to note that the choice of $d()$ is not arbitrary: $d(Q_1, Q_2)$ must represent how unlikely it is that a measurement device could read Q_1 while the target system is in the state Q_2. Both over-estimating and under-estimating $d()$ negatively impacts the accuracy of the model (either because the model claims lower precision than it actually has or because any natural fluctuation appears as significant model failure). The metric $d()$ forces us to embed the discrete measurement outcomes into a continuum space that represents more faithfully the underliying phenomena. For example, the integer digits on the display of the experimental device are typically discretizations of a continuum underlying process. In this case, the digit 7 must be assumed to be closest to digits 6 and 8. However, if the experimental set up includes a steps where the digits are handwritten, then we must also conside the possibility that digit 7 might be close to the digit 1. In this case, the relevant underlying continuum space is the one of all the possible handwritten digits.

In conclusion, discrete sets do not undermine the relevance of convexity in measurements, because even if the outcome is discrete, we must identify the continuous range of possibilities that might generate different outcomes in order to assign error-boxes to discrete measurements. This is how convex sets still play a fundamental role.

Disclosure and Disclaimer. The author works for a company that undertakes business in the deployment of AI systems as part of its commercial activities. The views expressed in this article are those of the author alone and do not necessarily represent the views of his employer.

References

Akaike, H. (1973). Information Theory as an Extension of the Maximum Likelihood Principle. In B. Petrov and F. Csaki (Eds.), *Second International Symposium on Information Theory*, pp. 267–281. Budapest: Akademiai Kiado.

Barnett, L. (1950, Jan). The Meaning of Einstein's New Theory – Interview of A. Einstein. *Life Magazine 28*, 22.

Behroozi, M. (2022). Largest inscribed rectangles in geometric convex sets.

Beisbart, C. (2021). Opacity thought through: on the intransparency of computer simulations. *Synthese 199*(3-4), 11643–11666.

Bird, A. and E. Tobin (2008). Natural Kinds. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Spring 2024 Edition ed.). Stanford University.

Brzović, Z. (2014). Natural Kinds. *The Internet Encyclopedia of Philosophy ISSN-2161-0002*, 1.

Carnap, R. (1950). *The Logical Foundations of Probability*. Chicago: University of Chicago Press.

Carnap, R. (1966). *Der Logische Aufbau der Welt* (3rd ed.). Hamburg, Germany: Felix Meiner.

Chaitin, G. J. (1975). Randomness and mathematical proof. *Scientific American 232*(5), 47–53.

Choi, S. and M. Fara (2021). Dispositions. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2021 ed.). Metaphysics Research Lab, Stanford University.

Cohnitz, D. and M. Rossberg (2024). Nelson Goodman. In E. N. Zalta and U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2024 ed.). Metaphysics Research Lab, Stanford University.

Crupi, V. (2021). Confirmation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2021 ed.). Metaphysics Research Lab, Stanford University.

Duhem, P. M. M. (1954). *The Aim and Structure of Physical Theory*. Princeton: Princeton University Press.

Elgin, C. Z. (Ed.) (1997). *The Philosophy of Nelson Goodman: Selected Essays*. New York: Garland.

Feigl, H. (1970). The "Orthodox" View of Theories: Remarks in Defense as well as Critique. In M. Radner and S. Winokur (Eds.), *Minnesota Studies in the Philosophy of Science*, Volume 4, pp. 3–16. University of Minnesota Press.

Feynman, R. P., R. B. Leighton, and M. L. Sands (1963). *The Feynman lectures on physics; New millennium ed.* New York, NY: Addison-Wesley Pub. Co.

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

Fletcher, S. C. (2024). On the Alleged Incommensurability of Newtonian and Relativistic Mass. *Erkenntnis*, 1–22.

Gärdenfors, P. (1990). Induction, Conceptual Spaces and AI. *Philosophy of Science 57*(1), 78–95.

Gärdenfors, P. (2000). *Conceptual spaces: the geometry of thought*. A Bradford book. MIT Press.

Gärdenfors, P. (2019). Convexity Is an Empirical Law in the Theory of Conceptual Spaces: Reply to Hernández-Conde. In M. Kaipainen, F. Zenker, A. Hautamäki, and P. Gärdenfors (Eds.), *Conceptual Spaces: Elaborations and Applications*, Volume 405, pp. 77. Springer.

Gärdenfors, P. and A. Stephens (2017). Induction and Knowledge-What. *European Journal for Philosophy of Science 8*(3), 1–21.

Goodman, N. (1946). A Query on Confirmation. *Journal of Philosophy 43*, 383–385.

Goodman, N. (1955). *Fact, Fiction, and Forecast* (2nd ed.). Cambridge, MA: Harvard University Press.
Goodman, N. (1983). *Fact, Fiction, and Forecast* (4th ed.). Cambridge, MA: Harvard University Press.
Goodman, S. (2008). A Dirty Dozen: Twelve P-Value Misconceptions. *Seminars in Hematology 45*(3), 135–140. Interpretation of Quantitative Research.
Hernández-Conde, J. V. (2017). A Case Against Convexity in Conceptual Spaces. *Synthese 194*(10), 4011–4037.
Kelly, K. T. (2007). Ockham's Razor, Empirical Complexity, and Truth-finding Efficiency. *Theoretical Computer Science 383*, 270–289.
Khamsi, M. and W. Kirk (2001). *An Introduction to Metric Spaces and Fixed Point Theory*. John Wiley & Sons, Ltd.
Kolmogorov, A. N. (1965). Three Approaches to the Quantitative Definition of Information. *Problems Inform. Transmission 1*, 1–7.
Krajıcek, J. (2004). Proof complexity. In *European congress of mathematics (ECM), Stockholm, Sweden*, pp. 221–231.
Leitgeb, H. (2007). A new analysis of quasianalysis. *Journal of Philosophical Logic 36*, 181–226.
Leitgeb, H. (2024). Vindicating the Verifiability Criterion. *Philosophical Studies 181*(1), 223–245.
Mormann, T. (1995). Incompatible empirically equivalent theories: A structural explication. *Synthese 103*, 203–249.
NIST (2019). SI definition of Meter. `https://www.nist.gov/si-redefinition/meter`.
Oberheim, E. and P. Hoyningen-Huene (2025). The Incommensurability of Scientific Theories. In E. N. Zalta and U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2025 ed.). Stanford University.
Piattelli-Palmarini, M. (1980). *Language and Learning: The Debate Between Jean Piaget and Noam Chomsky*. Harvard University Press.
Quine, W. v. O. (1950). Two Dogmas of Empiricism. *The Philosophical Review 60*, 20–43.
Quine, W. v. O. (1969). *Ontological Relativity and Other Essays*. New York: Columbia University Press.
Quine, W. v. O. (1975). On Empirically Equivalent Systems of the World. *Erkenntnis 9*, 313.
Quine, W. v. O. (1991). Two Dogmas in Retrospect. *Canadian Journal of Philosophy 21*(3), 265–274.
Scholz, S. (2024). Conceptual Spaces: A Solution to Goodman's New Riddle of Induction? *Philosophia 52*(4), 915–934.
Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics 4*, 461–464.
Scorzato, L. (2013). On the role of simplicity in science. *Synthese 190*, 2867–2895.
Scorzato, L. (2015). Science and Illusions. preprint: philsci-archive.pitt.edu/15570.
Scorzato, L. (2016). A simple model of scientific progress. In L. Felline, F. Paoli, and E. Rossanese (Eds.), *New Developments in Logic and Philosophy of Science*, Volume 3 of *SILFS*. College Publications.
Scorzato, L. (2024). Reliability and Interpretability in Science and Deep Learning. *Minds and Machines 34*(3), 27.
Sprenger, J. and S. Hartmann (2019). *Bayesian Philosophy of Science: Variations on a Theme by the Reverend Thomas Bayes*. Oxford and New York: Oxford University Press.
Stalker, D. F. (Ed.) (1994). *Grue!: The New Riddle of Induction*. Chicago and La Salle, IL: Open Court.
Stanford, K. (2021). Underdetermination of Scientific Theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 ed.). Metaphysics Research Lab, Stanford University.
Starr, W. (2022). Counterfactuals. In E. N. Zalta and U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2022 ed.). Metaphysics Research Lab, Stanford University.
Strössner, C. (2022). Criteria for Naturalness in Conceptual Spaces. *Synthese 200*(2), 1–36.
Teller, P. (1969). Goodman's Theory of Projection. *British Journal for the Philosophy of Science 20*(3), 219–238.
Visser, A. (1991). The Formalization of Interpretability. *Studia Logica 50*(1), 81–105.
Visser, A. (2004). Categories of theories and interpretations. *Logic Group Preprint Series 228*, 1–64.
Votsis, I. (2016). Philosophy of Science and Information. In L. Floridi (Ed.), *The Routledge Handbook of Philosophy of Information*. Routledge.
Wikipedia (2025). Bielefeld conspiracy — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/wiki/Bielefeld_conspiracy`. [Online; accessed 05-May-2025].
Williamson, T. and J. Stanley (2001). Knowing how. *Journal of Philosophy 98*(8), 411–444.
Zenil, H. (2020). A Review of Methods for Estimating Algorithmic Complexity: Options, Challenges, and New Directions. *Entropy 22*(6), 1–28.