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

A not-too-simple solution to Goodman's New Riddle of induction

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.

Keywords theory change · information complexity · conceptual spaces · confirmation · induction · values in science · structure theory · pseudoscience · scientific progress

1 The old New Riddle

Goodman's New Riddle of induction (Goodman, 1946, 1955, 1983) occupies a prominent role in the philosophy of science (Henderson, 2024; Godfrey-Smith, 2021; Scholz, 2024). In fact, it was recognized very early as deeply entangled with many other major puzzles in philosophy of science[1], but also one 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 many philosophers have no confidence that a solution is possible at all.

In this article, after a brief review of Goodman's New Riddle (Sec. 1.1), I position it within the more general problem of scientific model selection (Sec. 1.2). Then, in Sec. 1.3, I emphasize that both problems are very much related to other classic problems in philosophy of science. In Sec. 1.4, I discuss what should count as a solution of Goodman's New 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) and the one of Douglas (2013) (in Sec. 4.3).

1.1 Review of Goodman's New Riddle

Goodman's New Riddle of induction (Goodman, 1955) builds on Hume's thesis (the Old Riddle, Hume (1739)) that past regularities do not, by themselves, justify predictions about the future. Goodman goes a step further because

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

[1] "*The urge to dispose of the problem as spurious or insoluble is understandable, of course, in view of the repeated failures to find a solution. The trouble is, though, that what confronts us is not a single isolated problem but a closely knit family of problems. If we set one of them aside, we usually encounter much the same difficulties when we try to deal with the others. And if we set aside all the problems of dispositions, possibility, scientific law, confirmation, and the like, we virtually abandon the philosophy of science.*" (Goodman, 1983)

he does not question the justification of why an inductive statement should (likely) be true about the future (he accepts that there isn't one), instead he questions the justification of why we regard some statements as legitimate conjectures (*projectible*), while others—equally supported by the data—are not considered such. Hence, the New Riddle is not a puzzle about our inability to form justified expectations about the future, which we might well have to accept. It is a puzzle about our inability to explain our own assumptions, which is much more disturbing.

To illustrate the riddle, Goodman uses the famous "grue" paradox. An object is grue if and only if "*it is observed before a specific future time t_0 and is green, or it is not observed before time t_0 and is blue*". Imagine that we have observed many emeralds and they are all green. As long as t_0 lies in the future, the two sentences (a) "All emeralds are green" and (b) "All emeralds are grue" are equally well supported by the data. So, the question is, why do we consider the statements (a) a legitimate extrapolation (aka *lawlike* or *projectible*), while (b) is not?

Before reviewing any attempt to solve the New Riddle, it is worth emphasising that the New Riddle is about philosophy of science. In Goodman's words: "*(...) although I talk of (...) the color of marbles, which are seldom discussed in books on chemistry or physics, what I am saying falls squarely within the philosophy of science.*" His interest is, of course, the implications for science. Because Goodman focuses on **identifying a problem**, he only needs to discuss one example that contains the minimal ingredients necessary to highlight the specific issue, without any distracting features. In this sense, the sentence "All emeralds are green" is a stylized example of scientific model (an isolated scientific law within mineralogy). But, if we want to **solve the problem**, we must, eventually, consider the general question of why some models are considered legitimate scientific options, while others—equally supported by the data—are not considered such by the scientific community. This general question is the well-known problem of scientific model selection, that I review in the next Sec. 1.2. Correspondingly, we can identify what Goodman calls *projectible* sentences with those scientific models that the scientists often call the *state of the art*. In both cases they represent those models (or sentences) that can be used for legitimate predictions[2].

Similarly, to truly solve Goodman's New Riddle, it is not sufficient to solve it in the specific case of the grue concept. It is necessary to show that the solution is robust against any trick that exploits similar ideas. To this end, I will introduce an extension of the idea of grue in Sec. 1.3.

1.2 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[3] by the data (Duhem, 1954; Quine, 1975; Stanford, 2021).

These infinite options are not just theoretical possibilities without practical relevance. Consider statements like: *"was my experimental device malfunctioning on day X?"*[4]. This is the very concrete way in which infinite options of ad-hoc assumptions[5] can be exploited to fit any data. This 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?

A good observation is that state-of-the-art scientific theories are typically **predictive**. However, 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

[2] Consistently with the *New* Riddle, the word *legitimate* here is granted by some assumptions, that we need to identify, not by any supposed likelihood of representing the truth.

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

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

[5] 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.

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 scientists and philosophers made similar statements.

1.3 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; Scorzato, 2013), and it is reviewed here in Appendix A.

As noted in Sec. 1.1, the problem of scientific model selection can be seen as a generalization of the New Riddle of induction. Along the same lines, the Ξ trick mentioned above can be seen as an extension of the idea behind Goodman's grue, and it emphasises the same lesson: for every crazy theory that agrees with the data, there is always an ad-hoc language that makes it a simple and/or natural assumption. The Ξ trick simply emphasises how general and deep the New Riddle is. The notion of simplicity seems to be irremediably subjective, hence unsuitable to define the goals of science.

If we cannot give a definition of syntactic simplicity that makes sense as an epistemic value, the status of other candidate non-empirical epistemic values is no better. For example, according to Baker (2022), a theory is more *ontological parsimonious* than another if it postulates less entity types. However, nothing prevents us to introduce a new entity type that stands for the union of all the types of the theory, and the new formulation has just one type. As another example, a long philosophical tradition, going back at least to Popper (1959) and recently reviewed by Schindler (2018), assigns special importance to whether a theory makes ad-hoc assumptions and how many. But if the only hypothesis of the theory is $\Xi = 0$, then the distinction between ad-hoc and non ad-hoc hypotheses cannot be defined. I review more candidate epistemic values in Sec. 4.3.

Goodman's New Riddle does not only hinder a characterization of induction and theory selection[6]. As emphasised by his quote reported in Footnote 1, the puzzle permeates much of philosophy of science. Goodman himself discussed in detail how the New Riddle is essentially equivalent to the problem of *counterfactuals* (Starr, 2022). In fact, deciding what would happen in a hypothetical, unrealized, scenario corresponds to selecting a class of models that are admissible to draw those conclusion. If any model that fits the empirical data is legitimate, almost any conclusion is possible, and counterfactual reasoning loses any meaning. A vast literature (Choi and Fara, 2021) also connects the problem of counterfactuals to the problem of *dispositions*.

The problem of induction (and, therefore, 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 for plausible predictions and which shouldn't. A solution to one problem would yield a solution to the other, but no solution enjoys widespread support. In particular, Bayesian confirmation theory (Howson and Urbach, 1993; Sprenger and Hartmann, 2019) had cherished the hope of establishing confirmation on empirical evidence and probability alone. However, also Bayesianism must rely on something more to be able to deliver any meaningful outcome (Boyd and Bogen, 2025; Norton, 2021). See also Sec. 4.1.

The classic problem of *demarcation* (Pigliucci, 2021) also depends on the problem of model selection. The former is often seen as the problem of excluding entire disciplines as non-scientific at all, rather than individual models. But, if we cannot rule out a model as implausible as "All emeralds are grue", it is obviously also hard to rule out, conclusively, any pseudo-scientific theory (and even less a discipline), provided that it has done its job of ensuring concordance with the empirical evidence, which is not too difficult, by introducing a sufficient amount of ad-hoc hypotheses.

[6] Importantly, many of the problems discussed in this section come in two versions, depending on whether they aim at justifying one choice in terms of likelihood of future success (e.g. the Old Riddle), or they simply aim at describing the scientists' actual choices (e.g. the New Riddle).

Furthermore, any idea of *scientific progress* that goes beyond a dull list of empirical results, must involve the idea that our set of state-of-the-art theories is somehow improving. It is difficult to imagine how this can be achieved without a definition of state-of-the-art or a definition of theory selection.

Finally, the analysis by Roche and Sober (2023) shows that no existing, purely probabilistic, measure of explanatory power is satisfactory. On the other hand, one of the most popular account of *explanation* (Woodward, 2005) relies crucially on counterfactual reasoning, which, as we have seen above, also relies on a solution of the New Riddle. Another, very influential account of explanation (Kitcher, 1989) relies on recurring patterns in scientific laws, an argument that loses any differentiating power, if any theory is expressed as $\Xi = 0$.

These deep interdependencies between some of the most fundamental philosophical issues are not surprising. The Ξ trick offers an interesting perspective into the core problem: 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? It is very hard to imagine any well-defined and meaningful comparison between different—equally accurate—models, if they can all be represented in this way. But such comparison is a precondition to address any of the philosophical problems mentioned above.

In conclusion, Goodman's New Riddle and its extensions seem to undermine any precise definition of *theory selection, goals of science, induction, demarcation, confirmation, scientific progress or explanation*. Indeed, no definition of any of these concepts has gained a consensual 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[7].

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 option (b) implies that any scientific conclusion rely, inexorably, on fundamentally mysterious assumptions. But an essential step in any scientific work consists in identifying every hidden assumption in the arguments used by scientists. Any effort in this sense becomes futile, in the scenario (b). It does not help to formulate scenario (b) with narratives that obfuscate how much it fundamentally undermines the value of science. 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).

1.4 What does count as a solution?

A major obstacle in solving Goodman's New 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*[8] 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 New Riddle** is equivalent to **solving Goodman's New Riddle in the special case when the scientific model under scrutiny is a model for scientific model selection itself**. This is not very

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

[8] 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"*.

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 little 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 New 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 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[9]. 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 the solutions of the constraints 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.

[9] A k-cell is a k-dimensional box. In the context of error-boxes, 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.

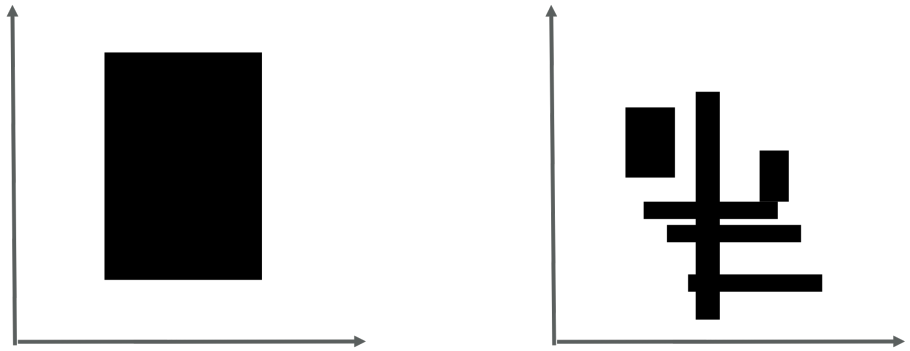

Fig. 1 Left: a legitimate error-box outcome of a single direct measurement in two dimensions. Right: not a legitimate one.

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.

Note that I have not defined *direct measurement* explicitly. Post. 1—together with Def. 1 below—offers an implicit (partial) definition of direct measurements[10]. On the other hand, any derived (i.e. non-direct) measurement must be constructible from a basis of direct ones. It is up to the model to decide which measurements are direct and which are derived, provided that Post. 1 and Def. 1 are fulfilled. The exact separation between direct and derived is largely conventional: in fact, I do not defend a unique objective cut between the empirical data and the theoretical concepts. But, it is important toensure that every measurements ultimately relies on something that fulfills Post. 1.

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

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[12]. 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[13]. In the following, I will refer to Post. 1, for simplicity. But one can easily (although 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

[10] One could define direct measurement explicitly as any measurement that fulfills Post. 1 and Def. 1. But this would designate as 'direct' also measures that, intuitively, are not quite so. I find it clearer to emphasise that it is the task of the theory to choose which measurements are directs, but—whatever they are—they must fulfill Post. 1 and Def. 1.

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

[12] It is not difficult to extend Post. 1' to measurable quantities defined on discrete sets. This is done in Appendix B.

[13] Note that even the double-slit experiment, that plays an key role in Quantum Mechanics (QM), does not contradict the claim. What is measured directly is the interaction of single electrons with the photographic plate. These individual dots have a convex form. Even if the QM explanation of the double-slit experiment implies that the wave function of a single electron traverses both slits at the same time, the wave function itself is not measured at all (not even indirectly), and the interference pattern is only visible after multiple hits.

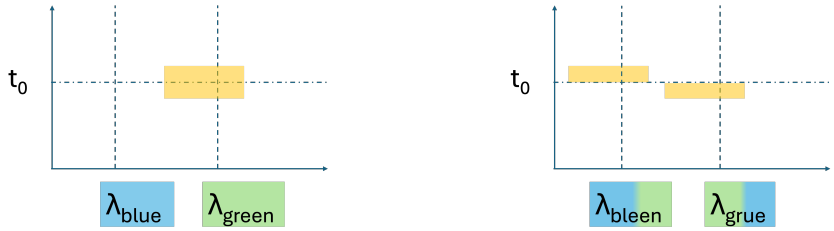

Fig. 3 Left: error-box (in yellow) associated to the measurement of the color of an emerald around the critical time t_0. Right: the same error-box appears split in the grue/bleen representation. Cfr Fig. 4 in (Gärdenfors, 1990).

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[14].

What does the previous discussion imply for Goodman's grue?[15] 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).

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 does not undermine measurability: 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 measurements. 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. But then, the grue/bleen model is not equivalent anymore to the green/blue one: it is less accurate. In synthesis, the convexity of error-boxes pinpoints a fundamental, well-defined, asymmetry between the green/blue and the grue/bleen models.

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[16], even well after t_0, as depicted in Fig. 4. 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.

[14] A related idea has been proposed also by Schurz (2015), because *ostensively learnable* concepts must be measurable and hence convex, according to Post. 1. However, Schurz (2015) characterizes ostensive learnability only empirically, which is not sufficient to identify the implicit assumptions in model selection.

[15] 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 error-bar 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 each emerald was first observed. Gärdenfor's definition does not have the same problem for $t \gg t_0$.

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

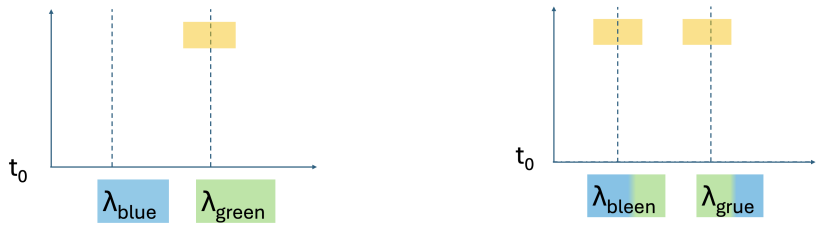

Fig. 4 Left: error-box (in yellow) associated to the measurement of the classical color of an emerald well after the critical time t_0. Right: the same error-box gets split in two disconnected boxes in the grue/bleen representation, if we cannot tell whether the emerald was first seen before or after t_0.

Does the requirement of convexity solve Goodman's New 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 (the introduction of the design criteria as in (Douven and Gärdenfors, 2020) does not change this conclusion).

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.

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

- P *is a set of assumptions*[17]*;*
- 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[18]). 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*

[17] 'Assumptions', 'Hypotheses', 'Principles', 'Postulates' are used as synonyms in this paper. Of course, there is no claim that they should be self-evident, consistently with the interpretation of, e.g., Jacobi, Mach, Riemann and many others (Pulte, 1998).

[18] 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.

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[19], 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 partially and holistically, together with the other model assumptions) and analyse statistically any unexpected outcome.

Equivalent 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**[20]*);*
- *for each measurable property*[21] *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, special relativistic 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.

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

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

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

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

and not by an arbitrary choice). This is not a small feat: it is **the the only proposal I know of a non-empirical candidate epistemic value which is precisely defined, non-trivial and as much formulation-independent as one can possibly wish**.

Note that epistemic complexity is **defined precisely** but **estimated approximatively**, as it is the case for any empirical quantity in science). As a result, it also explains scientific disagreement as different estimates of the epistemic complexity of different models (see also the discussion in the next paragraphs, as well as Sec. 4.3 and Scorzato (2026)). This also explains why disagreements eventually settle, which is a major puzzle for other accounts building on (Kuhn, 1977).

Moreover, real scientific models are often too complex (having dependences with multiple domains) to make a radical reformulation practical. But, for the same reason, it is also doubtful whether a radical reformulation will be able to achieve much greater simplicity. That explains why the scientists assess the epistemic complexity of a model by referring to the existing formulation in common scientific language. Hence, it **justifies the use of ordinary scientific language to assess simplicity**.

Model selection. Now that we have at least one well defined non-empirical epistemic value which is non-trivial and represents what we were looking for, can we build a philosophical model for scientific 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 equally empirically accurate (i.e. both models describe—or fail to do so—all *existing* empirical evidence, with the same precision[22]), then the scientists can directly compare the epistemic complexity of the two models and discard the more complex model, if the difference is larger than the estimated uncertainties in the assessment of the complexity (otherwise, both models are kept).

If the two models are not equally accurate, 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, it makes sense to drop the most complex model, because it is just more complex for no empirical advantage. For example, imagie that models M_1 and M_2 are equally empirically accurate for most observations, except that M_1 explains accurately experiment e_1 and M_2 doesn't, while M_2 explains accurately experiment e_2 and M_1 doesn't. The necessary ad-hoc assumptions in this case are A_1 (A_2): "the experiment e_1 (e_2) is not reliable". Hence, we can and must compare the simplicity of $M_1 + A_2$ vs $M_2 + A_1$. A single ad-hoc assumption is rarely decisive (unless it is the only difference between M_1 and M_2, which is also often the case), but the accumulation of evidence on one side, will make the cost of the multiple ad-hoc assumptions unambiguously higher. These difference will also motivate the scientists to devise new experiments which cannot be conclusive by themselves, strictly speaking, but they become conclusive if they succeed in forcing more and more complex ad-hoc assumptions predominantly on one side.

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 at all (the red area in Fig. 5). This selection is in fact uncontroversial among scientists. When too many ad-hoc assumptions are required, 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. 5). 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, and already well-defined criteria exist to chose among them: the scientists may select only some models based on the minimal precision required by a specific application and/or they may opt for a Bayesian model average. So, even if the State-of-the-art may be still very large, the selection between different SotA models is not conceptually problematic and it 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:

[22] Note that 'empirical equivalence' defined in Def. 2 refers to all conceivable *measurable* properties, while the expression 'equally accurate' refers to the existing, already *measured* quantities.

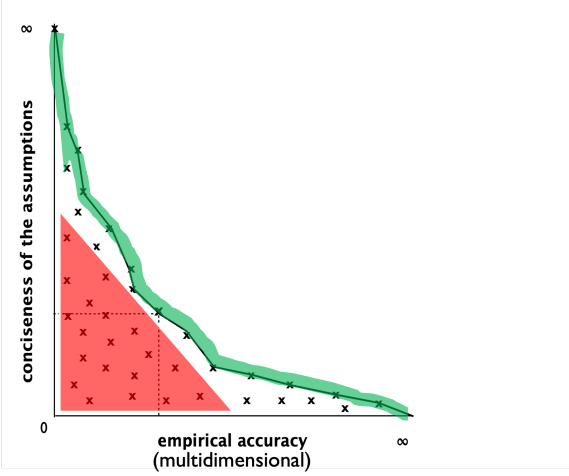

Fig. 5 Each 'x' represents a different scientific model, plotted by accuracy (which is multi-dimensional, but it is shown here as one-dimensional for simplicity) and conciseness. Ruled-out models appear in the red ares. State-of-the-art models are those in the green surface.

Definition 4 (Model selection) *A model A is preferred to model B if A is neither more complex nor less empirically accurate than B, 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 *The* **State-of-the-art** *is the ensemble of models which are not worse than any other model. The models that are not state-of-the-art are* **ruled-out**.

Note that the **scope** of a model is taken into account by its (multi-dimensional) accuracy in the following sense. If we assume that the accuracy of model M on topics outside its scope is minimal, then an approximate theory with a broader scope is more accurate, on some sub-domains, than a precise theory with narrow scope. In this way, we do not need to introduce scope as an independent epistemic value, because it is implicitly included in a multi-dimensional notion of accuracy.

Philosophy of science as a science itself. It is important to observe that the model presented here is automatically *self-consistent*. Namely, it is itself a simple and accurate description of the domain that it tries to describe: scientific model selection. This is very important to argue that philosophy of science is a science itself (science of science[23]). This property is also far from obvious. For example, saying that "Scientific model selection is decided by the scientists" is not self-consistent, because the rule quoted here is fixed, and it does not leave room to the philosophers to change it. Self-consistency is not sufficient, of course[24], but, if we accept that the Old Riddle cannot be solved, self-consistency, besides accuracy, is the best we can aim for to justify the choice of the non-empirical values. The empirical accuracy of the present model is discussed in the next Sec. 3.

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 New 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[25].

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 superior principle, which will never be superior for everyone and would be yet another

[23] In a sense that is complementary to (Fortunato et al., 2018).

[24] The rule "Good theories start with the word *Good*" is also self-consistent, but not accurate.

[25] Unfortunately, this point is rarely discussed in the literature, where philosophers often criticize some 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.

attempt to solve the *Old* Riddle of induction. The next step consists in assuming the epistemic value of conciseness (Def. 3) as a working hypothesis (just as the validity of every scientific theory is assumed as a working hypothesis) and assess whether the concept of model selection based on it provides a good description of those theory choices 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), machine learning models (Scorzato, 2024) and the oxidative phosophorylation controversy (Scorzato, 2026). 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[26] 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 analysing 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.

For example, one may wonder why I did not chose other classic measures of complexity (e.g. Akaike (1973); Schwarz (1978); Sober (2015)). 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 counter-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.

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

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 unwarranted 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)[27]. 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.

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 16), 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.

[27] 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.

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[28]. On the other hand, it is still possible to build accurate, but completely implausible 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[29], 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.

[28] 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.

[29] 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.

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 because 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. Although 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.

One of the most popular modern approaches to confirmation theory (Crupi, 2021) is **Bayesianism** (Howson and Urbach, 1993; Sprenger and Hartmann, 2019). However, 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 (Boyd and Bogen, 2025; Norton, 2021; Scorzato, 2024). In turns, the priors can only be defined by relying on some non-empirical criteria, 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.

The very influential **material theory of induction** (Norton, 2021) accepts that all inductions must rely on additional specific assumptions that depend on the domain under consideration (i.e. they are *local*). This aligns with the present emphasis that we are typically interested in extrapolating entire scientific models rather than individual sentences. But the analysis fails to identify those assumptions that the scientists do not make explicit, but are essential to select meaningful laws.

Because I have emphasised that the problem of **theory choice** is the proper generalization of the New Riddle of induction, any attempt to solve the former is also relevant to solve the latter. However, in recent decades, the philosophical literature on theory choice has largely swept the New Riddle under the rug. For example, a very influential tradition, going back at least to Kuhn (1977) (see (De Benedetto and Luchetti, 2024) and reference therein for recent developments), argues that theory choice is determined by a bunch of vaguely defined and evolving values. However, these values are never fully defined. In particular, non-purely empirical values like simplicity, parsimony, external coherence and explanatory power always appear in such lists, but in absence of a precise definition, it is not clear how they could possibly avoid the Ξ trap. Hence, all those accounts remain fundamentally undetermined and therefore non-descriptive. This topic is discussed further in Sec. 4.3.

A further instructive example, that lies outside the main stream, but still suffers from similar shortcomings, is the proposal by Dawid et al. (2015). In this case, a theory is selected if it has 'no alternatives'. However, this idea makes sense only if we assume some non-empirical constraints, otherwise we always have infinite alternatives (see Sec. 1.2). This fact is indeed recognized be the authors (Dawid et al., 2015). But, without a clear definition of the constraints, also this model remains vulnerable to the Ξ trap and undetermined.

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 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.

4.3 Do we need other independent epistemic values, besides accuracy and conciseness?

In this section I compare the present proposal to (Douglas, 2013), which represents a significant step towards rationalization, with respect to the vagueness of (Kuhn, 1977). I fully agree with (Douglas, 2013) that **internal coherence** and **empirical accuracy** (concerning the past) are not negotiable. Concerning internal coherence, the matter is not undisputed. So, it is worth reminding that *ex falso quodlibet*, which includes any statement in conflict with the evidence. Internal coherence must and can be recovered[30]. This is typically done by adopting more complex assumptions. For example, if X and Y are inconsistent, in general, we may say that in context A we assume X, in context B we assume Y. So, potential contradictions are eliminated at the cost of higher epistemic complexity.

While there is little doubt that these values are the top priority for science (let's call them the first class, which is a slight modification of Douglas' first group), they are far from sufficient to account for the scientists' consensus on theory choice, as discussed in Sec. 1.2. The second class of values recognized in (Douglas, 2013) includes: simplicity, scope, explanatory power, fruitfulness, unification, external coherence, novel predictions and precision. Note that the values of **scope** and **precision** are included in my (multi-dimensional) definition of empirical accuracy. So, I see them in the first class. In fact, I do not want to choose between a model that is more precise and a model with a broader scope: I keep both in the state-of-the-art.

Before analysing the remaining values, it is worth discussing how we expect them to be used. One possibility is to imagine that the scientists unconsciously keep weights in their heads, which they use to prefer one model over another and to quarrel with their colleagues. In this scenario, the discussions and further evidences lead to adjustments of the weights until they converge. This scenario does not explain why this social dynamics eventually converges and does not move much afterwards. Nor it explains why the scientists stubbornly perceive this evolution of their own personal weights as a driver of rather 'objective' conclusions. But, this seems the dominant view among modern philosophers of science.

Another possibility—defended here—is that, while there is certainly tension between the first and the second class of values, nevertheless (a) there isn't much tension among the values within the second class, when properly understood, and (b) once we recognize the few independent values that matter, theory selection becomes rather unambiguous. Both (a) and (b) are essentially defended also in (Douglas, 2013), but both claims remain necessarily vague there, because all the values in the second class (except novel predictions; more on this below) are vulnerable to the Ξ trap and it is impossible to recognize how they are all proxies of a well defined framework. But, if we accept the solution to the Ξ trap discussed in this paper, they can all be defined and, as suggested in (Douglas, 2013), they do not pull in many different directions.

Let us consider in more detail why the second class does not bring any value that is necessary and truly *independent* from accuracy and conciseness. First consider **fruitfulness**. If a model has *proved* to be more fruitful, it must have already achieved some other advantage, either in accuracy or something else, that must be specified

[30] In this paper, I mostly don't emphasis internal coherence, but it is assumed in Def. 1.

and would refer to some other value discussed here. Otherwise, its fruitfulness is only a conjecture that does not have to be included in the current assessment of the model.

Let us consider **external coherence**[31]. Logical coherence between *different* models in the state-of-the-art is not required and it is often violated: different models in the state-of-the-art are often competitors. We can define external coherence as follows. Two model are largely externally coherent, if then have a large common core of assumptions. Then, the model of scientific model selection presented here favors models that have larger external coherence with other state-of-the-art models, as it is illustrated by the following example. Imagine that we need a model that describes precisely the class of phenomena E_1 (e.g. rockets), but it should also be approximatively consistent with the broad class of phenomena E_0 (e.g. general phenomena in physics and chemistry). Imagine that model M_0 is generally used to describe the broad class E_0, while the models M_1 and N_1 are two competing alternatives to describe E_1. Imagine that M_1 is obtained from M_0 by adding just one simple assumption, while N_1 is very different from either M_0 or M_1. Although M_1 is more complex than N_1, because it includes M_0, N_1 alone is not sufficient to describe E_0, and we can use it for our purposes only if we also assume M_0 anyway, to describe E_0. So, M_1 is simpler than the combination of N_1 and M_0, because it is more externally coherent with M_0 and it is therefore preferred. Hence, external coherence does not add any valuable *independent* dimension that is not already included in the values of the first class plus conciseness.

Furthermore, **unifications** necessarily represent either an increase in the scope of a model or a reduction of its overall assumptions (or both). Finally, **explanatory power** would require a much longer discussion, given the variety of meanings that have been associated to it (Woodward and Ross, 2025), but at least one of its most influential interpretations (Kitcher, 1989) is fully aligned with the idea of unification and with the reduction to concise assumptions.

In conclusion, I do not see any evidence of the need to include other values in a model of theory choice that are not already sufficiently taken into account by the values of the first class plus conciseness[32]. Moreover, this model explains well why the scientific controversies mostly reach consensus: the debates are due to different, legitimate, estimates of the complexity of hypotheses. The complexity gap becomes increasingly clear under the growing constraints arising from the accumulation of empirical evidence. Complexity of the assumptions and empirical evidence are also the topics that the scientists actually discuss.

5 Conclusions

The solution to Goodman's New 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. I do not pretend to explain why those decisions must be right.

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[33]. This claim implies the futility of trying to clarify the exact assumptions behind any scientific

[31] The value of **fundamental justifications** is not discussed in (Douglas, 2013), because, I believe, it is a special case of external coherence.

[32] Pne more note about **novel predictions**. The model described in this paper does not attribute extra value to the fact that a theoretical prediction was formulated *before* the first measurement of the predicted phenomenon. Novel predictions have certainly played a role, historically, to accelerate the adoption of a new theory. But, are they necessary to define theory selection? Defining exactly *when* a phenomenon was first predicted and when the *same* phenomenon was first observed, is fraught with subtleties. The question is, do we need to enter these subtleties? To argue that we must, one should provide an example where accuracy and conciseness alone are not sufficient to justify a consensual case of theory selection, but the time order of the respective proposals and observation achieves that goal. I am not aware of any such case and I doubt there are, because such theory selection would be very fragile: the conclusion ought to change in case one would discover that some scientists was aware or was not aware of early experiments. Novel predictions are certainly impressive in the transition phase that leads to new theory selections, but when a new theory has firmly established its new role, it rarely depends on the time ordering of these events.

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

conclusion, which is one of the main focus of scientists of all disciplines. 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 some stated precision. How can we claim that Goodman's New 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*,
- 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 Legitimacy of the $\Xi = 0$ formulation

A few details are appropriate to explain why I claim that we can always reformulate the assumptions of any model as $\Xi = 0$, why this formulation is logically equivalent to the original formulation and why it is legitimate to talk about measurements of Ξ and their associated error-boxes. See also Scorzato (2013).

The assumptions of any model are, in general, statements (assumed to be true), equalities or inequalities. Any (in)equality can also be expressed as a statement whose value can be `true` or `false`. In fact, the equation $A = B$ (or $A < B$) can be replace by a sentence Σ that is true if and only if, indeed $A = B$ (or $A < B$). On the other hand, we can also expressed any sentence Σ' as an equality, simply by defining A, such that $A = 0$ if S' is true and $A \neq 0$ if Σ' is false. Moreover, any set of equations can be brought to the form $\Xi = 0$, by bringing any non-zero term to the left and assigning the symbol Ξ to represent everything on the left. It might be challenging to interpret Σ (or Ξ), but we have no obvious clear-cut arguments to refute a supporter of the model who claims to be able to interpret Σ (resp. Ξ) directly. Developing that argument is the topic of this paper.

The claim that the new formulation is logically equivalent to the original one is also easy to justify. Above, I have shown how to translate the original formulation into the new one. To go back to the original formulation it is sufficient to introduce, in the new formulation, as many symbols as in the original formulation, identify them with the corresponding symbols in the original formulation and assume the relation between Σ (resp. Ξ) and the newly introduced symbols to reproduce the relations that we defined in the original formulation. In this way, the two formulations can be simply identified one-to-one. Note that the introduction of all these symbols in the new formulation does not increase the complexity of the new formulation, because they are only needed to translate into the legacy formulation and they are not necessary to formulate all the assumptions of the model in its most concise form.

If we claim that "Σ is `true`" or "$\Xi = 0$" represent all the assumptions of the model, we must also explain how are the measurements documented in this formulation. This is where things become challenging for a defender of the new formulation of the model. But let's see how far we can go. Because the only symbol available is Σ (resp. Ξ), measurements must be expressed in terms of that symbol alone. Real measurements contain more information than just a `true` or `false` outcome. So, simply assigning a truth value to Σ is not sufficient: we should at least assign a probability to each possible outcome. This leads us to introduce at least another variable $P(\Sigma)$ that represents the probability of the outcome and takes continuous values. Because P is continuous, we can discuss convexity. Crucially, we cannot prove that any measurement in the original representation is mapped into a convex interval in the new representation P. However, this is not the notation chosen in this paper.

In the Ξ formulation we can assume that Ξ itself can take continuous values. If Ξ represents a sentence, we can, e.g., assume that $\Xi = 0$ stands for certainly `true`, $\Xi = 1$ stands for certainly `false` and the values in between stand for intermediate estimates of the corresponding probability. If Ξ represents a set of equations, the value of Ξ represents how precisely the measurement has verified the relation. Using standard scientific notation, we can write $\Xi = \Xi_0 \pm \delta$ to represent the statement that a measurement has found Ξ to be within an interval of size 2δ around Ξ_0 with a given probability. Again, we cannot prove that convex intervals in Ξ represent convex intervals in the original formulation. In the main text, I refer to the formulation $\Xi = 0$ for simplicity, because I find it more intuitive and it does not require the introduction of new symbols.

One might also ask whether $\Xi = 0$ represents a single equation or a set of equations. Both options are possible, but both options have difficulties in representing measurements faithfully and concisely. In fact, I could certainly write the assumptions as $\Xi_n = 0$, for $n = 1, ...N$, which is only slightly more complex than just $\Xi = 0$. But, if I want to express the expected error-boxes in a way that is faithful to the actual measurements in the original formulation, I will be forced to write $\Xi_n \in \cup_j \Delta_{j,n}(\Xi_1, ...\Xi_N)$ with very complex set-valued functions $\Delta_{j,n}$ that won't be simpler than the original formulation.

B 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 choosing 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 underlying 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 consider 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.

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.

Baker, A. (2022). Simplicity. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Summer 2022 Edition ed.). Stanford University.

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.

Boyd, N. M. and J. Bogen (2025). Theory and Observation in Science. In E. N. Zalta and U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2025 ed.). Metaphysics Research Lab, 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.

Dawid, R., S. Hartmann, and J. Sprenger (2015). The No Alternatives Argument. *British Journal for the Philosophy of Science 66*(1), 213–234.

De Benedetto, M. and M. Luchetti (2024). Theory choice as niche construction: the feedback loop between scientific theories and epistemic values. *Philosophy of Science 91*(3), 741–758.

Douglas, H. (2013). The value of cognitive values. *Philosophy of science 80*(5), 796–806.

Douven, I. and P. Gärdenfors (2020). What are natural concepts? A design perspective. *Mind & Language 35*(3), 313–334.

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 90*, 3567–3588.

Fortunato, S., C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, A. Vespignani, L. Waltman, D. Wang, and A.-L. Barabási (2018). Science of science. *Science 359*(6379), eaao0185.

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.

Godfrey-Smith, P. (2021). *Theory and Reality* (2nd ed.). Chicago: The University of Chicago Press.

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.

Henderson, L. (2024). The Problem of Induction. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2024 ed.). Metaphysics Research Lab, Stanford University.

Hernández-Conde, J. V. (2017). A Case Against Convexity in Conceptual Spaces. *Synthese 194*(10), 4011–4037.

Howson, C. and P. Urbach (1993). *Scientific Reasoning: The Bayesian Approach* (2nd ed.). La Salle IL: Open Court Publishing Company.

Hume, D. (1739). *A Treatise of Human Nature*. Oxford: Oxford University Press.

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.

Kitcher, P. (1989). Explanatary Unification and the Causal Structure of the World. In *Scientific Explanation*, Volume 8, pp. 410–505. University of Minnesota Press.

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.

Kuhn, T. S. (1977). Objectivity, Value, and Theory Choice. In *The Essential Tension*. Chicago: Chicago University Press.

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`.

Norton, J. D. (2021). *The material theory of induction*. University of Calgary Press.

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.

Pigliucci, M. (2021). Pseudoscience and the Demarcation Problem. *The Internet Encyclopedia of Philosophy ISSN-2161-0002*, 1.

Popper, K. (1959). *The Logic of Scientific Discovery*. New York: Basic Books.

Pulte, H. (1998). Jacobi's criticism of Lagrange: the changing role of mathematics in the foundations of classical mechanics. *Historia Mathematica 25*(2), 154–184.

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.

Roche, W. and E. Sober (2023). Purely probabilistic measures of explanatory power: A critique. *Philosophy of Science 90*(1), 129–149.

Schindler, S. (2018). *Theoretical virtues in science: Uncovering reality through theory*. Cambridge University Press.

Scholz, S. (2024). Conceptual Spaces: A Solution to Goodman's New Riddle of Induction? *Philosophia 52*(4), 915–934.

Schurz, G. (2015). Ostensive learnability as a test criterion for theory-neutral observation concepts. *Journal for General Philosophy of Science 46*(1), 139–153.

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.

Scorzato, L. (2026). The oxidative phosophorylation controversy in the light of epistemic complexity. preprint: philsci-archive.pitt.edu/28769.

Sober, E. (2015). *Ockham's razors: a user's manual*. Cambridge University Press.

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.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford university press.

Woodward, J. and L. Ross (2025). 20th Century Theories of Scientific Explanation. In E. N. Zalta and U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2025 ed.). Metaphysics Research Lab, Stanford University.

Zenil, H. (2020). A Review of Methods for Estimating Algorithmic Complexity: Options, Challenges, and New Directions. *Entropy 22*(6), 1–28.