The scientific demarcation problem: a formal and model-based approach to falsificationism

Jeremy Attard[1,2*]

[1]Philosophy and History of Science Department, University of Mons, Avenue Victor Maistriau, 15, Mons, 7000, Belgium.
[2]Department of Sciences, Philosophies and Societies, University of Namur, Rue de Bruxelles, 61, Namur, 5000, Belgium.

Corresponding author(s). E-mail(s): jeremy.attard@umons.ac.be;

Abstract

The problem of demarcating between what is scientific and what is pseudoscientific or merely unscientific – in other words, the problem of defining *scientificity* – remains open. The modern debate was firstly structured around Karl Popper's falsificationist epistemology from the 1930's, before diversifying a few decades later. His central idea is that what makes something scientific is not so much how adequate it is with data, but rather to what extent it might not have been so. Since the second half of the century, and in the wake of criticisms, such as the Duhem-Quine thesis, that were raised against falsificationism(s), most approaches to the problem of scientific demarcation are now multicriteria and holistic. However, the approach presented in this paper does not follow the same guideline. The present work can be seen as an attempt to adapt Popper's (sophisticated) falsificationism to a model-based view of scientific knowledge. Using formalization and focusing on a particular epistemic unit of analysis (namely, *empirical models*) allows to properly define the popperian corroboration degree and to view scientificity as the maximization of this degree of corroboration over all available models and data. We eventually recover, in a natural way, well-accepted scientificity criteria: empirical adequacy, Lakatos' progressive problemshifts, balance between strength and simplicity, parsimony, and coherence as special cases of this general scientificity principle. From this viewpoint, the language dependency of our empirical knowledge no longer appears as a limitation of falsificationism but as one more reason to take it as a good epistemological framework.

1 Introduction

1.1 A short history of the demarcation problem

The scientific demarcation problem (SDP) can be broadly presented as the problem of finding a clear demarcation between science and non-science or between science and pseudo-science – for instance, a set of criteria that would be jointly sufficient and individually necessary to define scientificity. Generally, scientists are quite able to distinguish between scientific and non-scientific statements, models, or theoretical constructs[1] without being able to clearly explain *why* this distinction holds. In other words, we seem to have a definition of scientificity in *extension* but lack such a definition in *intension*. From an epistemological viewpoint, the SDP is about finding such a definition.

The modern debate about the SDP has been firstly structured around Karl Popper's falsificationist epistemology since his 1934's *Logik der Forschung* (Popper, 1934, 1959). Popper's proposal has the advantage of resting on a simple criterion for scientificity: falsifiability.[2] According to this view, a system of statements can be qualified as scientific if it may potentially be falsified experimentally, that is, if it is possible, at least in principle, to imagine an experimental or observational outcome that could falsify it.

His first rationale supporting falsifiability is that it is claimed to solve the problem of induction, as posed e.g. by David Hume (Hume, 1748). Indeed, scientific reasoning seems to be ultimately based on inductive reasoning, because general empirical claims can never be fully empirically justified; science would then be founded on a not totally logical basis. Popper then exploits the logical asymmetry between the verification and the falsification of a given empirical claim. Indeed, such a claim cannot be fully empirically justified, but a single counterexample suffices to refute it. The confirmation procedure is an inductive process whereas the falsification procedure is a deductive process *via* the *modus tollens*. The guiding idea of Popper is then to make our knowledge rest on a logically robust basis again: scientific progress would finally be based on deductive reasoning, getting closer and closer to the truth by *deductively* eliminating theories and hypotheses.

[1] At least in their own field.

[2] Actually, Popper's proposal is not precisely monocriterial as it also includes avoiding *ad hoc* procedures and more generally "conventionalist stratagems" to save a theory facing a contradiction.

In contrast, empirical systems that are not falsifiable seem to have extraordinary explanatory power, but this power is illusionary: they are empirically right without having possibly been wrong anyway. The central question, regarding scientificity,[3] is then, from this viewpoint, not "is this theory adequate with data?" but "could this theory not to be adequate with data?" As long as the theory being tested is not refuted but could have been so, it is corroborated. The more audacious the claims the theory has made and the more severe the tests it has passed, the more audacious is the theory.

Popperian falsificationism has been criticized for decades, and Popper improved his view from these criticisms, especially after the release of his first book in English in 1959. One important criticism is that it seems impossible to definitively refute a given claim. Indeed, facing a contradiction with experience, it is always possible to adjust our logical system (e.g., adding an *ad hoc* hypothesis to our theory) to neutralize the discrepancy. As the Duhem-Quine thesis claims (Harding, 1975), a single empirical statement $\mathcal{S}$ never faces experience alone but is always embedded in an entire system of statements that are tested together with $\mathcal{S}$. This observation is related to the fact that observations are always theory-laden: there is no pure empirical fact; we always need some hypothesis in order to generate empirical data that will serve to test other hypotheses.

However, Popper is aware of this fact, which is a reason among others why in his further books (Popper, 1962, 1972), he developed a sophisticated version of falsificationism, as explained by Imre Lakatos in (Lakatos, 1978):

> [Popper] agrees that the problem is how to demarcate between scientific and pseudoscientific adjustments, between rational and irrational changes of theory. According to Popper, saving a theory with the help of auxiliary hypotheses which satisfy certain well-defined conditions represents scientific progress; but saving a theory with the help of auxiliary hypotheses which do not, represents degeneration.

Popperian falsificationism should not be caricatured as claiming that we must abandon a theory each time it is falsified: it is quite obvious that a theory is a structure with some complex logical interdependency, and that empirical falsification only indicates that there is something wrong in the whole structure. As Quine noticed, however, logic alone cannot tell us which part of our knowledge is to be revised when it faces a contradiction with experience (Quine, 1951). Theory is indeed always underdetermined by experience.

This observation, as well as the fact that the distinction between analytic and synthetic statements is actually dependent on the language used[4] deeply weakened the original ambitions of logical empiricism. The latter used to consider scientific theories as axiomatic systems, that is, non-interpreted syntaxic structures, and its aim was to base our knowledge on the purely logical relationships within this structure, the whole resting in the last instance on unquestionnable and purely empirical facts related to direct observations.

[3] Let us precise that in the whole paper we talk about *empirical* sciences and not formal ones.

[4] It is always possible, from a logical point of view, to change some basic definitions and turns any empirically falsified synthetic statement into a (trivially true) analytic statement.

Despite their fundamental differences, falsificationism might somehow be connected to this "received" or "orthodox" view of theories, for at least two reasons. First, it also rests on the idea of basing scientific knowledge (and science progress) on logical rules. Second, falsificationism is easier to define within a framework in which theories are viewed as axiomatic structures or logical networks of statements. The conclusion is that the usual criticisms of these two points usually also reach falsificationism.

Moreover, the history of science and, more precisely, Thomas Kuhn's work (Kuhn, 1962) shows that not only is it logically possible to save a theory facing a refutation, but that it often turns out to be very fruitful, sometimes even the most rational thing to do. Lakatos, from Popper's sophisticated falsificationism, synthetized both Popper's and Kuhn's views on scientific theory development (Lakatos, 1978). A great observation from Lakatos has been to notice that the relevant unit of analysis was not a single theory, but rather a series of theories $T_1, T_2, ...T_n$ which is such that for any i, T_{i+1} results from the modification of an auxiliary hypothesis to T_i in order to explain an anomaly that T_i faces. From this viewpoint, Lakatos then distinguished *progressive and degenerating problemshifts* (Lakatos, 1978, p. 33-34):

> (...) a series of theories is theoretically progressive (or 'constitutes a theoretically progressive problemshift ') if each new theory has some excess empirical content over its predecessor, that is, if it predicts some novel, hitherto unexpected fact. Let us say that a theoretically progressive series of theories is also empirically progressive (or 'constitutes an empirically progressive problemshift') if some of this excess empirical content is also corroborated, that is, if each new theory leads us to the actual discovery of some new fact. Finally, let us call a problemshift progressive if it is both theoretically and empirically progressive, and degenerating if it is not . We 'accept' problemshifts as 'scientific' only if they are at least theoretically progressive; if they are not, we 'reject' them as 'pseudoscientific'. Progress is measured by the degree to which a problemshift is progressive, by the degree to which the series of theories leads us to the discovery of novel facts. We regard a theory in the series 'falsified' when it is superseded by a theory with higher corroborated content.

Thus, the condition to accept an adjustment is that it not only explains the anomaly but also leads to the discovery of novel facts. In contrast, a problemshift that does not fit this condition is said to be degenerating.[5] From this viewpoint, "scientific" then would refer to progressive problemshifts, while "pseudo-scientific" would refer to degenerating problemshifts.

A well-known critical work on falsificationism in particular and the SDP in general is that of Larry Laudan. In his famous article "The demise of the demarcation problem" (Laudan, 1983), his point is as follows: we have not yet found any satisfying set of collectively sufficient and individually necessary

[5] From these reflections, Lakatos then coined the concept of research program. According to him, a research program always comprises two main features: an irrefutable *hard core* and a *protecting belt of auxiliary hypotheses*. The hard core contains fundamental principles, definitions, etc.. The auxiliary hypotheses protect the hard core from refutation because potential *modus tollens* are directed in their direction. The hard core is made irrefutable by methodological decision, defining the very theoretical frame within which explanations take place.

criteria that would demarcate between science and pseudo-science and allow us to reconstruct our primary intuition; this question is not interesting, because "science" is so heterogeneous that it is doom to fail. What is more interesting is knowing what makes a belief well-founded, reliable, or fruiful, and not what makes it scientific.

However, it seems that interest in this question has revived in the last twenty years. In the context of the 2005 trial against creationism in the US, philosophers of science were asked why creationism is not scientific whereas evolution theory is. The first part of the handbook on the philosophy of pseudoscience (Pigliucci & Boudry, 2013) mentions this episode and provides a set of good rationales to consider this problem as an important one.

By the way, (Fernandez-Beanato, 2020, 2022) provides a much more comprehensive outline of the different sets of criteria which has been proposed over the last century. Damian Fernandez-Beanato's work, a multi-criteria approach to the problem of scientific demarcation, is in line with Martin Mahner's proposal (Mahner, 2007) in the wake of Mario Bunge's approach to adress this problem (Bunge, 1983a, 1991). The idea is threefold: first, capturing the whole definition of "scientific" within a single criterion (like "being falsifiable") seems to be impossible, given the heterogeneity of science and scientific practices; second, scientificity is seen as a question of degree: there is no clear demarcation scientific/not scientific but rather a continuous line of scientificity; and third, science sometimes shares some features which what is used to be called "pseudoscience". Under a given list of criteria, a given epistemic unit will then be qualified as scientific if it satisfies a certain number of criteria, and not (or pseudo) scientific if it does not satisfy enough of them. For example, (Fernandez-Beanato, 2021) used a list of 36 criteria to assess the scientificity of Feng Shui.

However, the formal approach developed in this paper does not follow this guideline, as will be explained in the following section.

1.2 Our approach

From the previous considerations, it appears that several important observations must be taken into account when it comes to work on the notion of scientificity:

- Scientific knowledge has a complex and interdependent logical structure.
- The distinction between synthetic and analytic statements, and thus between falsifiable and not falsifiable statements, is dependent on the basic empirical language in which these statements are written.
- This very empirical language is also a part of what is tested.
- Logical rules are not enough to distinguish between different possible adjustments of theories facing an empirical contradiction.
- Scientificity of a given theory is dependent on historical and social contingencies.

Concerning the last point, there are several types of historical contingencies. For example, it may happen that with respect to the empirical data and the different hypotheses available at a given time, scientists at that time have very good reasons to consider a certain hypothesis or theory as the best one. In addition, it may happen that even if a new hypothesis is known and objectively the best one, it takes time for this new idea to impose itself within a scientific community. This may occur for many different reasons, including structural or inter-individual power relationships in the community, social stereotypes and norms, or even industrial conflicts of interest. However, we consider that the study of these important features is the work of history and sociology of science and that the role of epistemology (at least as we view it) is to define scientificity criteria (whether they apply to theoretical constructs like models or to social groups like epistemic communities) such that their validity does not depend on the historical or social context. We do not enter into details here, but as we shall explain later, our methodological postulates do not prevent us from considering that scientificity does depend on socio-historical contexts while epistemological criteria do not.

We now introduce the approach followed in this paper and how it incorporates the observations listed above. Our strategy is threefold: a reductionist wager, the use of formalization, and the statement of a scientificity *principle*.

In contrast to the holistic way of tackling the problem, i.e. starting from the largest possible scale of analysis (e.g. Bunge's *epistemic fields*), we start from a smaller one, namely that of *empirical models*. The term *model* is adopted for a large range of quite different objects, and we define this unit of analysis in detail in section 2, both concretely in 2.1 (with illustrative examples taken from actual science) and formally in 2.2 (in a language inspired from probability theory). This is a reductionist move because we voluntarily restrict our analysis to a strictly circumscribed conceptual object; thus, from the start, we discard some features as not essential regarding scientificity, and we may be wrong. Moreover, it is not clear whether the connection between our small-scale analysis and a larger-scale analysis (considering the complex structures of scientific theories in more detail) will be easy or even possible to establish. Despite these important observations, this reductionist wager and the use of formalization appear to be worth adopting because of the higher degree of clarity and (cognitive) manipulability it allows, doing exactly the same kind of job modelization itself does in scientific inquiry.

The second step is to formalize the popperian notion of degree of corroboration of models, thanks to the clear and precise framework presented in the previous section. From our viewpoint, the most important lesson of falsificationism is that the fundamental question, regarding scientificity, is not: *are our models adequate with data?* but rather: *how much information do we get knowing that our models are at this point adequate with data?* The corroboration degree is precisely defined as this amount of information, and section 3 is dedicated to its formalization. Therefore, our strategy does not consist of searching for a set of criteria defining scientificity, but rather in characterizing scientific

models as models *maximizing* a certain quantity, namely, the degree of corroboration. Indeed, given that our empirical knowledge cannot be absolutely justified from logical considerations only, we need to add "by hand" a fundamental principle to be satisfied in order to explain how it is possible to select between different rival models or theories. Moreover, language-dependency issues can also be tackled with this approach: it suffices to treat the empirical language as *a variable of the maximization problem*, that is, as something that can also vary and influence the degree of corroboration. Then, demanding a maximized degree of corroboration constrains the language used, avoiding naive relativism. It is also possible to think as if a clear distinction between analytical and synthetic, or falsifiable and unfalsifiable statements, existed, while taking into account the fact that this distinction depends on the language used.

Finally, this approach allows us not to get stuck with historicity issues in defining scientificity – this is where we come back to our previous discussion about the socio-historical context dependency in epistemology. Indeed, scientificity is seen here as the result of an algorithm that takes (empirical) models and a set of data as inputs and gives their respective corroboration degrees as outputs. The scientificity of a given model (the output of the algorithm) then depends, indeed, on historical and social contingencies, for instance how much data are available at a given time, or what are the other models in competition at that time (the inputs of the algorithm). However, the intrinsic rules of the algorithm remain, by definition, anhistorical. *It is precisely these rules that are under study in this paper*. In other words, our way of escaping from the historicity issue is to see scientificity not as an intrinsic quality of a given theoretical entity (such as a model, or theory) but more as a feature of the relationship between some data and models aiming to recover them.

The loss of logical justification following from adding "by hand" a scientificity principle is hoped to be compensated by recovering a certain number of well accepted scientificity criteria on a unified way. In section 4, we show that from this principle, it is possible to naturally derive some well-known criteria: empirical adequacy (4.1), clarity of basic variables (4.2), Lakatos' progressive problemshifts (4.3), Lewis' balance between strength and simplicity (4.4), and also (partly) parsimony and coherence (4.5). In 4.6, we also discuss the difference between Karl Popper's definition of the degree of falsifiability/corroboration and ours.

In section 5 we finally point to some limitations and perspectives of our approach, in connection with the already mentioned drawbacks related to the restriction to a given scale of analysis and the use of formalization.

Nota bene: the general purpose of this paper and most of its content can be understood even without any deep mathematical knowledge. However, sections 2.2 and 3.3 extensively use some mathematical formalism and implies a minimal mathematical knowledge. Nevertheless, an effort was made to explain everything qualitatively as well.

2 Empirical models as elementary epistemic units

A first important question to be adressed, working on the SDP, is that of the epistemic unit (the "object of demarcation" (Hansson, 2021)) which the analysis bears on. Indeed, discussions on scientific demarcation can go really confusing if we do not precise what is going to be qualified as scientific (or non-/pseudo-scientific). For instance, as Lakatos noticed, fundamental principles and some other statements of a theory are not aimed to be directly confronted with experience, but rather are thought as a framework in which experience can be rendered intelligible. Thus, their epistemological value is not given the way as, say, empirical predictions which have a more direct empirical meaning. Therefore, depending on what precisely we are talking about - empirical claims? fundamental principles? whole theories? epistemic communities? - scientificity cannot be defined the same way.

In a quite consensual quinean viewpoint, we picture our knowledge as a giant logical and interconnected web of statements which "impinges on experience only along the edges" (Quine, 1951). Its structure is way too complex to be analytically described and studied as a whole. Defining a unit of analysis, from this viewpoint, means choosing a scale at which we look at this logical web. The discussion shares some analogy with that in biology about the most relevant scale for analyzing life evolution: genes? individuals? species? The choice of a certain scale necessarily erases some details which could be interesting to study in spite of this. Yet, it allows, as models in science exactly do, a conceptual clarification and a clear basis for further discussions.

We choose to take *empirical models* as the epistemic units of our analysis. They constitute parts of the quinean web which are the closest of its edges, that is to say which are in the most direct connection with empirical data. In the first part 2.1 of this section, we define it more precisely on a qualitative way, based on concrete examples. Then, in 2.2, we make a proposal for a formal definition using a mathematical language inspired from probability theory. Finally, in 2.3, we stress the importance of the empirical language, which is essential to overcome the drawbacks of falsificationism mentioned in the introduction.

2.1 Qualitative definition

2.1.1 Examples

Empirical models under consideration in this paper are of two possible kinds: nomological or causal.

Nomological empirical models

In classical physics, an example of a nomological empirical model is the ideal gas law:

$$PV = nRT \tag{1}$$

where P is the pressure in Pa, V is the volume in m^3, n is the amount of substance in *mole*, T is the temperature in K and R is a constant equal to $8,314$ J mole^{-1} K^{-1}. The variables (P, V, n, T) can be measured on some concrete gases and an empirical model is a hypothetical relationship between these variables. Another possible empirical model for this set of variables is the van der Waals state equation:

$$(P + \frac{an^2}{V^2})(V - nb) = nRT \tag{2}$$

where P, V, T and n are as in (1) and a and b are some parameters. (1) is a special case of (2) for which $b = 0$ and $n/V \longrightarrow 0$.

Another example of a nomological empirical model is that of fall with friction. The typical experiment associated to it is the fall of a spherical marble into a given fluid. The basic variables that we can measure are, for example, the mass m of the marble, the speed v and the time t. A possible empirical model is the following hypothetical relationship between these variables:

$$|v(t)| = v_\infty(1 - e^{-t/\tau}) \tag{3}$$

where v_∞ and τ are *a priori* free parameters and m does not appear explicitely. Another possible empirical model is:

$$|v(t)| = v_\infty(1 - e^{-t/\tau}),\ v_\infty = g\tau,\ \tau = \frac{m}{\alpha},\ g = 10\ m.s^{-2}. \tag{4}$$

In (4), the mass appears explicitely and the remaining free parameter is α. (4) is a special case of (3) for which v_∞ and τ are assumed to be linearly related.

Such empirical models are thus, somehow, directly comparable to empirical data for they are relationships between variables which have a direct empirical meaning. A variable (like P, T, or v) is said to have a direct empirical meaning if it is given with a set of concrete operations which are sufficient to give it a value.

In these examples, empirical data look like a set of points taking value in the space defined by the basic variables. For example, typical empirical data D for a single experiment of fall with friction look like figure 1.

It is assumed that there is always a way of assessing the adequacy of a given empirical model with a set of data, whether the data are produced from a single experiment or from a set of different and independent experiments.

Causal empirical models

A causal empirical model is merely a hypothetical causal relationship between two variables measured on a given statistical population. In this case, the basic variables can be of four different types: quantitative and continuous (their value is a real number, like height or IQ), quantitative and discrete (their value

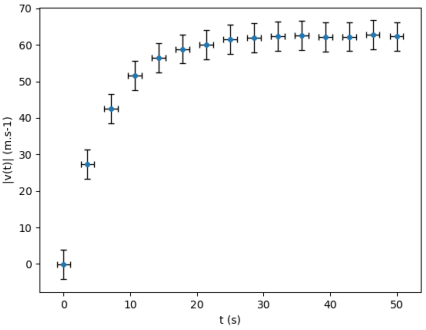

Fig. 1 Typical fall with friction data points.

is an integer, like the number of inhabitants in a city), categorical and ordinal (their value is an ordered category, like the level of pain in a visual analogue scale), or categorical and nominal (their value is an unordered category, like cities or countries).

Here we do not talk about the theoretical mechanisms that could explain the causal relationship: our analysis takes place at the statistical (i.e. empirical and not theoretical) level only. It corresponds to the macro-macro relationship in the Boudon-Coleman's diagram (Coleman, 1990) in sociology, or to observed statistical associations in terms of Bradford Hill's criteria (Bradford Hill, 1965) in epidemiology. More precisely, assuming a causal relationship between two variables A and B, denoted as $A \longrightarrow B$, means here that there is a statistical association between A and B and that this association remains *ceteris paribus*, i.e. everything else being equal.

For example, let A be the average cigarettes consumption per inhabitant and B the life expectancy at birth, both quantitative and continuous variables measured on the same set of countries. The Pearson correlation coefficient r or the determination coefficient r^2 are classical ways of assessing the strength of the association in this case. We would probably observe a *positive* correlation between A and B, that is to say: the more the cigarettes consumption in a country, the more the life expectancy at birth in this country. A possible empirical model is: "Smoking increases life expectancy" or:

$$A \longrightarrow B. \tag{5}$$

Another empirical model would include a third variable C as a possible confounder, C being the economical development of the country, measured e.g. as the GDP (but there are other possibilities). This alternative model is:

$$A \longleftarrow C \longrightarrow B. \tag{6}$$

That is to say, the association between A and B is predicted to disappear if we look at countries with similar GDPs. In this model, the economical development of the country is the cause of both the cigarettes consumption and that of the life expectancy.

Another example, classical in sociology (see e.g. (Boudon & Lipset, 1974)), is the relationship between the academic success of students (variable B) and the socioprofessional status of their parents (variable A). Both A and B are categorical and ordered variables, if we measure the academic success e.g. with the highest diploma obtained by the student (again, the operationalization of such a variable is not unique). The strength of the relationship between such variables can be assessed e.g. by the odd-ratio. A well known observed association is that the higher the socioprofessional status of their parents, the higher the academic success of their children.[6] A possible causal model is: $A \longrightarrow B$, i.e. the higher socioprofessional status causes somehow the academic success. This model implies that the association between A and B remains even if we control other potential confounders, like e.g. the school level as undergraduate, i.e. just before going to university (variable C). That is to say, for a group of students with the same school level in the last year before university, we would still observe that the higher the parents socioprofessional level, the higher the highest diploma obtained.

A causal empirical model is then a hypothetical causal relationship between two variables A and B, i.e. the prediction of a statistical association (assessed by a given statistical index like the Pearson coefficient or the odd-ratio)[7] between A and B which remains even if we fix the value of other possible variables. However, it is always possible to compute the Pearson coefficient, say, of two given quantitative and continuous variables. Assuming a positive correlation between such variables implies assuming a "high score" (i.e. close to 1) for their Pearson coefficient. Thus, it is a question of degree: every variables are trivially correlated (or more generally, associated) to each other, but assuming a causal relationship means predicting *a certain level* of correlation.

We assume, as in the case of nomological models, that it is always possible to associate to a causal empirical model a certain degree of adequacy with some empirical data D, taking into account the quite complicated structure of data and variables.

2.1.2 Components of an empirical model

From a more general viewpoint, an empirical model is the utterance of a relationship between a given finite set of variables measured on some empirical situations, such that it can be compared with empirical data.

An empirical model generally consists in several basic components:

[6] Let us stress again that this association is assumed to hold at the statistical (i.e. macrosocial) level and do not imply anything about what happens at the micro level, because this is not what we are interested in in this paper.

[7] A general overview of these different ways of assessing the prominence (usually called "size effect" in the literature) of statistical associations is given for instance in (Sullivan & Feinn, 2012).

- A finite set of variables with their operational definition. An operational definition of a variable is a set of concrete operations sufficient to give it a precise value. This is called *the empirical frame* and denoted as E.
- E can be seen as a space such that a piece of empirical data D can be represented as a set of elements of that space E.
- The proper empirical model is a relationship which is assumed to hold between these variables. It claims that given the empirical situations under consideration, the variables cannot take any possible value but are confined in some regions of E. Such empirical models can also be defined up to some parameters which are not variables in E and do not necessarily have a direct empirical meaning.
- It is assumed that there is always a way to assess the adequacy of the empirical model with the available data D, whether in the case of nomological models or in the case of causal models, although this way is usually not unique.

The set consisting in the four items above: an empirical frame, some data, an empirical model and a way of comparing the latter with data, constitutes the epistemic unit of our analysis.[8] This is also referred to as an *empirical model* by misuse of language.

2.2 Formal definition

2.2.1 Definition

We formally define an empirical model M as a set:

$$M = (E, \mathbb{P}, D, (m_\theta)_{\theta\in\Theta}, d), \tag{7}$$

where:

- E is a probability space, i.e. a set equipped with a $\sigma-$algebra $\mathcal{B}$[9] and a probability measure $\mathbb{P}$, which is a $\sigma-$additive map $\mathbb{P} : E \longrightarrow [0, 1]$ such that $\mathbb{P}(E) = 1$.[10]
- D is an element of $E \times E \times \ldots E = E^n$ with $n \in \mathbb{N}^*$. Since D can be seen as a finite set, we denote $n =\mid D \mid$ to remind that the integer n is associated to D.
- $(m_\theta)_{\theta\in\Theta}$ is a parametrized family of (measurable) subsets of E. Θ denotes a certain set on which the family of models is indexed.

[8] This is important to stress again that we are not focusing on models in general but on empirical models in particular. That is to say, we do not study in this paper the epistemological criteria bearing on what is a good theoretical explanation, but only on empirical relationships between measurable variables.

[9] As the $\sigma-$algebra does not intervenes explicitely in our work, we do not write it explicitely from now on. By misuse of language, $(E, \mathbb{P})$ refers to the probability space.

[10] We do not mean here the *base probability space* of probability theory usually denoted as $(\Omega, \mathcal{F}, \mathbb{P})$ the interpretation of which is not easy in general, and in our approach in particular. We may somehow consider the probability space $(E, \mathcal{B}, \mathbb{P})$ used in this work as the target space of a certain random variable $X : \Omega \longrightarrow E$ and the probability measure on E as that associated to X and implied from the probability measure on Ω. However, we do not explicitely use neither Ω nor X in the present work, and only talk about the space $(E, \mathbb{P})$.

- E is also equipped with a distance such that for any subset $B \subset E$, $n \in \mathbb{N}$ and $D \in E^n$, we can define a distance between B and D, denoted $d(B, D)$. The distance d can then be used to compare a model m_θ (with $\theta \in \Theta$) with data D as $d(m_\theta, D)$.

In what follows we explain in more details our choices of modelization.

2.2.2 Interpretation and justification of the terms

The empirical frame E

The space E and the associated probability measure $\mathbb{P}$ represent the empirical frame as we defined it above. The probability measure on E encodes somehow the background knowledge we have on the basic variables defining the empirical frame. For example, in the case of fall with friction, the empirical frame is defined from the basic variables (m, v, t). Thus, E could be represented, in this case, as $E = \mathbb{R}^3$. $\mathbb{P}$, defined on E, would represent what we know about m, v and t and their way of being measured, such that e.g. $v/c \ll 1$, c being the light celerity. In political science, poverty threshold in a given country is sometimes defined as a certain proportion of the median income. The poverty rate of any country, such defined, is then necessarily less than 0.5, by construction. This feature would be encoded by the corresponding probability measure on the space representing the empirical frame.

Models as subsets of E

This viewpoint obviously reminds the phase space approach of Frederick Suppe (Suppe, 1974) and Bas van Frassen (van Fraassen, 1980) in the context of semantic conception of theories. In this paper, E will sometimes be called "phase space"[11] for this reason. Notice that in our case, we define it for a particular epistemic unit and not for the whole theory, for we do not place our analysis at this scale. Let us now see some concrete examples.

The ideal gas law:

$$PV = nRT \tag{8}$$

is an example of a parametrized family of models. Here, the only free parameter is $\theta = R \in \Theta = \mathbb{R}$ (if we assume that we are in a case where its value is unknown). The case of fall with friction exhibits another example of a parametrized family of models:

$$|v(t)| = v_\infty(1 - e^{-t/\tau}), \tag{9}$$

with $\theta = (v_\infty, \tau) \in \Theta = \mathbb{R}^{+2}$.

For each $R \in \mathbb{R}$, equation (8) defines a 3-dimensional topological subspace of $E = \mathbb{R}^4$ (defined from variables (P, V, n, T)), while for each $(v_\infty, \tau) \in \mathbb{R}^{+2}$, equation (9) defines a curve in $E = \mathbb{R}^2$ (defined from variables (t, v)). This is why such models can always be seen as families of subsets of E.

[11] In physics, the phase space of a system usually consists in the whole set of states which the system can be in.

The same thing occurs in the case of causal models. A causal model $A \longrightarrow B$ between quantitative and continuous variables consists e.g. in the asumption of a set of linear correlations between A and B – the association of A and B being measured the other variables being fixed. For instance, let us consider three such variables A, B and C such that C can only take three values C_1, C_2 and C_3 (for instance, C is a categorical and nominal variable). In this case, $E = E_1 \times E_2 \times E_3$, where $E_i = \mathbb{R}^2$ for $i = 1, 2, 3$. Then, the correlation between A and B will be measured on three statistical subpopulations such that C is fixed on each of them and equal respectively to C_1, C_2 and C_3. For each subpopulation, a correlation is e.g. defined by two real parameters a and b of the affine relationship assumed to hold between A and B: $B = aA + b$. A causal relationship between A and B is then characterized by a set $\theta = \{(a_i, b_i)\}_{i=1,2,3}$. For such a θ, the causal model $A \longrightarrow B$ is represented by m_θ which can be seen as the union of a straightline defined by $(a_1, b_1) \in \mathbb{R}^2$ in E_1, a straightline defined by $(a_2, b_2) \in \mathbb{R}^2$ in E_2 and a straightline defined by $(a_3, b_3) \in \mathbb{R}^2$ in E_3. In short, such a causal model can also be seen as a family of subsets of E.

We thus make the assumption that any empirical model, however complicated it may be, can be seen as a parametrized family of subsets of E.

2.3 Importance of the empirical frame

A naive realism would make us think that the basic variables, for instance in physics, are nothing more than something which is directly given by Nature. However, a classical result in history of science (Hanson, 1958; Kuhn, 1962) is that the situation is not so linear. Consider temperature: in (Chang, 2004), we realize that the construction of a robust measure for temperature is something which took a certain time, and that nothing was directly given by Nature itself. At the beginning, its measure was mostly based on the observation of some empirical regularities like the relationship between the heating of a metal rod and its length variation. Hasok Chang shows that such a procedure can rapidly fall into circular reasonings, and that only some particular stable experimental conditions and a robust theoretical definition (that is to say, seing temperature as the mean kinetic energy of particles) allowed to go out of this circularity, ending up with a robust operational definition.

Some variables as temperature are now well-defined and do not cause any more problems, except when we try to define it in extreme conditions. However, this is not the case for all variables in physics. For example, in modern cosmology, measuring large scale distance is nothing but straightforward. Different methods are used depending on the distance scale which is probed: parallax measure (up to 10kpc),[12] Cepeid's measure of luminosity (up to 10Mpc) or

[12] A parsec (pc) is a usual astronomical unit which corresponds to the distance at which the distance Earth-Sun is seen under an angle of one arcsecond, and is approximatively equal to 3.26 light-years.

supernovae Ia for "distances" up to redshifts[13] around 2 (Czerny, Beaton, & Bejger, 2018). The use of these *standard candles* is resting on quite strong theoretical assumptions and the different methods have to be calibrated in order to give a coherent notion of distance along scales of several order of magnitudes. Therefore, given that the very basic variables in cosmology are thus theory-laden, the interpretations of the empirical data produced from these variables is far from obvious. An example is the acceleration of the expansion of the universe which comes from the observation of the relationship between redshift and distance of distant galaxies. However, as David Merritt noticed (Merritt, 2017), there exist alternative ways to explain this observation without appealing for an accelerating expansion, for instance "relinquish[ing] the assumption of homogeneity". Our purpose here is not to enter into this technical discussion, but only to stress that choosing to focus on the acceleration hypothesis is a choice, which may be defended with good arguments, but which is not an obvious conclusion that we could directly read in the empirical data.

In the social sciences and humanities, a fundamental question is that of defining the basic measured variables as intelligence, poverty, country's level of development, criminality, personnality's traits, etc. As in any other science, a change in the basic definitions of what is observed directly results in a different observation. For example, it has been recently shown (Stoet & Geary, 2018) that there were a *negative* correlation between the gender equality degree in a country (measured with the *Global Gender Gap Index* (GGGI) as defined in (WEF, 2022)) and its proportion of women graduates in science, technology, engineering and mathematics (STEM), while there is generally no difference in scientific skills between female and male students in school. In other words, the more equal women and men in a country, and thus the more the women are free to choose their academic career, the less they choose to graduate in STEM. This is called the "gender-equality paradox in STEM" for we would have *a priori* expected the contrary. Both this empirical result and its interpretation has been criticized and the controversy was in part focusing on the way of measuring the gender gap of a country. In particular, (Richardson et al., 2020) showed that if another index is used (namely, the Basic Index of Gender Inequality (BIGI)), the correlation just vanishes. This is, again, not our purpose here to enter into this controversy. The only aim of this interesting example is to illustrate the dependance of the empirical data, i.e. the way the world appears to us, on the basic variables used.

The empirical frame represents the basic language we need in order to say something about a particular part of the world. Its importance is capital: a change in the empirical frame results in a change in how we see the world, and thus conditionates the whole scientific inquiry. However, it takes time for a discipline to get robust and stabilized basic variables. It turns out that the detection of regularities in the empirical world *and* the construction of an empirical frame in which these regularities are detected are two historically

[13] The electromagnetic signals coming from distant galaxies appeared as redshifted, which gives information about their relative speed and motion.

concomitent and co-dependent processes. In our formalization, the empirical frame $(E, \mathbb{P})$ is made explicit to take into account the theoryladeness of observation and the fact that the very definition of basic variables and the way they are measured can vary through scientific inquiry – i.e. that these elements are also part of what is tested.

3 Maximized corroboration principle for empirical models

3.1 Outline of our approach

3.1.1 Karl Popper's legacy

According to Karl Popper, the adequacy of a theory with empirical data is not a sufficient condition for this theory to be scientific. Indeed, its scientificity is not only measured in the light of its empirical adequacy but also and above all in the light of *its degree of falsifiability*. As already mentioned in the introduction, the relevant question to be asked is thus not only "is this theory adequate with data?" but "would have been possible for this theory *not* to be adequate with data?" "Theories are not verifiable, but they can be 'corroborated'" (Popper, 1959, p. 248). According to Karl Popper and its followers, empirical data can never fully justify a theory, but it can corroborate it: the theory has to be right (with respect to empirical data) *whereas it could have been wrong*. In other words, a "prediction" which is true independently of the available data does not say anything about the relevance of the underlying theory. More precisely, any empirical language will exhibit some *endogeneous empirical regularities*, i.e. analytic *a priori* truths, or, in the language of Popper, unfalsifiable statements. This is not a bad thing *per se*, this is just a fact. The lesson from falsificationism is just that this kind of regularities is not of any strong epistemological support. As a scientific theory aims at saying something substantial about the world, the truth of its statements should, in the last instance, come from the world – that is, from the empirical data. Given a statement which is empirically adequate at a certain degree, the more the statement could have been refuted, the more information we get about the relevance of our theory in saying something about the world.

3.1.2 Corroboration degree as an amount of information

We start from the already mentioned idea that what really matters, regarding scientificity, is not how much our theories are empirically confirmed, but rather *how much information do we get when our theories are empirically confirmed.* This is precisely this amount of information that is called *degree of corroboration*. We formalize it in section 3.3, but let us give here a qualitative formalism-free description. Given our epistemic unit constiting in (see section 2.1.2): an empirical frame E, some empirical data D, a proper empirical model m and a way of comparing the data and the model, we define the following set:

the set of all data which are at least as much adequate with m as D is. This set, compared to the whole set of possible data that can be generated by E, measures at which point it is surprizing that m and D are at this point adequate. If this set identifies with the whole set of possible data, we are in the case where any possible data could have been at least as much adequate with the model as D. These are unfalsifiable models, true by construction and for which the gain of information is thus zero. On the contrary, the "smaller" this set, the more falsifiable and the more the information we get from empirical adequacy. Thus, we need a well defined way of measuring "how big" is this set compared to the whole set of possible data, which is rendered possible with the probability measure $\mathbb{P}$ defined on E, as we shall see below. We will eventually compare this approach with Popper's formal definition of corroboration degree in section 4.6.

3.1.3 Maximized corroboration principle

Once corroboration degree is defined on the epistemic units under study, it allows to compare them to each other on clear epistemological basis. We then state a general scientificity principle in section 3.4: given a set of comparable empirical models $\mathcal{M} = \{M_1, M_2, ..., M_N\}$, the most scientific is merely $M \in \mathcal{M}$ with the maximum corroboration degree. This simple principle allows to recover well-known epistemological criteria, as we shall see in section 4.

3.1.4 Theoryladeness and language dependency

Again, the fact that a given statement be falsifiable or unfalsifiable, and more generally the corroboration degree of a given empirical model M, does depend also on the language used.[14] Yet, this is not a dramatic limitation of our approach, on the contrary: the language being itself a variable of our problem, the maximized corroboration principle (see below) takes, by construction, this feature into account. The fact that genuine regularities *and* the language in which they are expressed are co-selected in the procedure of corroboration maximization makes it possible to define empirical languages which are, *ceteris paribus*, better than others. According to us, this is why language dependency is not, *per se*, a dramatic issue for a falsifiability-based definition of scientificity. Actually, from our viewpoint, this is even the other way around: theoryladeness of empirical data is actually *one more reason* to consider falsifiability as a relevant criterion for scientificity, for it allows to distinguish genuine regularities from regularities which are endogeneous to the empirical language.

3.2 Detecting regularities: endogeneous and genuine regularities

Detecting regularities in "Nature" is quite consensually viewed as being a basic aim of scientific inquiry. These regularities take different forms depending on

[14] Exactly like the distinction between analytic and synthetic statements.

the situation under consideration, yet the fundamental intuition - that we want to formalize - is the same: a regularity is something intringuing, interesting, which calls for an explanation.

As already mentioned, we are forced to express these regularities within a language represented by the empirical frame E. It turns out that this can generate spurious regularities, i.e. regularities which are mere artifacts of this language. These endogeneous regularities are of two possible forms: either they are just true by virtue of the definitions composing E, that is to say they are E–analytic,[15] or they are artifacts of our modelization hypotheses. We now take some concrete examples.

3.2.1 $E-$analytic statements.

Let us see two examples of empirical observed relations that are true directly because of the operational definitions of the terms they relate.

Basic analytical statements.

First, a quite artificial example to fix the idea: "All swans are birds" is an empirically meaningful statement which is true by the virtue of the operational definitions of this terms: the fact to be a bird is a necessary condition (by definition) in order to be a swan. Another such example is the one already mentioned in section 2.2.2: "poverty rate of a country cannot be greater than 0.5".[16] This does not tell us something deep about economical or social reality but is a mere artifact of the definitions.

Cobb-Douglas production function.

A more subtle example is the following. In neoclassical economic theories, an important concept is that of the aggregated production function, which relates the quantity of economic output (like the global production of a country) with a certain set of economic inputs (like labor, capital, ...) In particular, Charles Cobb and Paul Douglas proposed in 1928 such a function (Cobb & Douglas, 1928), relating the total production Q in a year with total amount of labor L and capital K measured in homogeneous units:

$$Q = AL^{\alpha}K^{\beta}, \tag{10}$$

where A is a parameter. They tested this function econometrically and found a very strong empirical adequacy with data. In particular, they empirically found that the parameters of this model are related by $\alpha + \beta = 1$, which constitutes an argument supporting some basic hypotheses of neoclassical economics. However, some criticism araised over time, and it has been recently claimed (Felipe & McCombie, 2013) that actually, this relation reduces to an *accounting identity*. In other words, the relation (10) together with $\alpha + \beta = 1$

[15] This notation is used to remind that the distinction between analytic and synthetic statements is dependent on the language used.

[16] Poverty threshold being defined as a proportion of the median income in a country.

can actually be mathematically deduced from an identity which is true by definition. This is a quite clear example of a regularity which is endogeneous to the language used - and still not trivial to detect.

3.2.2 Modelization artifacts.

Let us now see two examples where the tautology does not come directly from the definition of the basic variables but from the framework used to modelize the data.

Lagrange method of fitting.

A simple example of this is the following result: for any n data points $\{(x_i, y_i)\}_{i=1..n}$, there always exists a polynomial function P_n of degree at most $n-1$ (thus, with n parameters) which perfectly fits the data, i.e. such that:

$$\text{for any } i \in [1, n],\ y_i = P_n(x_i). \tag{11}$$

We could hardly imagine a better fit. However, for a given n-points piece of data D, the statement "there exists n parameters $\{a_i\}_{i=1..n}$ such that the polynomial fonction:

$$P_n(x) = a_0 + a_1 x + a_2 x^2 + ... + a_{n-1} x^{n-1} \tag{12}$$

perfectly fits D" is not a genuine empirical prediction for it is always possible to find such parameters, independently of any particular data. It seems to have great empirical success, but actually it is nothing but a mere reformulation of data D. The procedure to find the good polynomial from a given set of data is called the Lagrange method and an example of this is represented figure 2.[17]

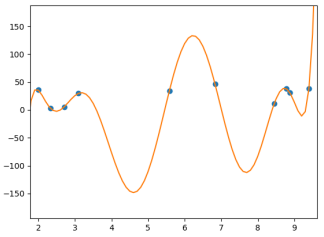

Fig. 2 Lagrange method applied to a set of n=10 randomly generated points.

More precisely, in the language of our formalization, we have a family of models $(m_\theta)_{\theta \in \Theta}$ where $\Theta = \mathbb{R}^n$, i.e. $\theta \in \Theta$ is under the form: $\theta = (a_0, a_1, ..., a_{n-1})$, and thus for $\theta \in \Theta$, $m_\theta = P_n$ as defined in (12). For any

[17] In this example we generated randomly ten points in a two-dimensional graph and then compute the unique 10-parameters polynome which perfectly fits them.

$D = \{(x_i, y_i)\}_{i=1..n}$, the statement "$\exists \theta \in \Theta$ such that m_θ and D are perfectly adequate" is thus true for all possible data $D \in \mathbb{R}^{2n}$. We gain no information from such an empirical adequacy, no matter it be a perfect adequacy.

Ptolemy's epicycloids.

Another more subtle example is that of Ptolemy's geocentric planetary model, which was predominant for centuries until the late XVI's. In this model, the Earth is at the center of the universe and the other planets together with the sun are rotating around it. Some of these planets, however, are not following strict circles around the Earth, but more complicated trajectories called epicycloids. An epicycloid is generated by the rotation of a point along a circle the center of which is also rotating along another circle, and so on. It turns out that this model had very great empirical adequacy with observational data available at this time. Moreover, each time a discrepancy appeared, it was fixed by adding another epicycle, i.e. a new circle the center of which is rotating along the previous one. This model seems to be quite robust, since it can accomodate any anomaly without modifying its basic hypotheses.

However, the empirical adequacy of this model is actually a mathematical artifact. Any closed trajectory in the Earth's frame can actually be described by an epicycloid with a finite number of circles, given that the observational precision is itself finite - which is the case. This is because any such closed trajectory in two dimensions can be decomposed as the addition of two periodic functions, and that any periodic function can be approximated by its Fourier's decomposition,[18] i.e. a finite sum of cosinus and sinus functions. In two dimensions, this decomposition (which only rests on the mathematical modelization of trajectories as closed, continuous and derivable curves) exactly gives finite epicycloids.

It does not mean that this model cannot be used to make accurate predictions, but just that the epicycloidal form is not something deep to be explained: it reduces to a mathematical artifact due to the way these trajectories are modelized. On the contrary, Kepler's elliptical trajectories in the Sun's frame does say something deeper which appeals for a genuine explanation, for not any set of data can be described by such a model.

3.3 Adequacy and corroboration degree

In this section we define the adequacy and the corroboration degrees formally within the framework presented in section 2.2. Let $M = (E, \mathbb{P}, D, (m_\theta)_{\theta \in \Theta}, d)$ be an epistemic unit under study. Remind that d allows to compare a given model m_θ, for $\theta \in \Theta$, with data D as $d(m_\theta, D) \geq 0$. The adequacy degree of M can then be defined as:

$$\alpha(M) = \inf_{\theta \in \Theta} d(m_\theta, D). \tag{13}$$

[18] More precisely, for any closed, continuous and derivable function f, its Fourier series uniformally converges to f. That means that if a non null precision interval ϵ is given, the Fourier series of f will get at a distance less than ϵ from f with a finite number of terms.

The adequacy degree is a positive real number, and the more $\alpha(M)$, the less adequate M. On the contrary, $\alpha(M) = 0$ means that the model is perfectly adequate, such as in the examples in section 3.2.2.

Once we have an empirical model M with a certain empirical adequacy degree $\alpha(M)$, we can define the set of all empirical data which are at least as much adequate with the empirical model $(m_\theta)_{\theta\in\Theta}$ as D is. That is to say,

$$\mathcal{D}[M] = \{D' \in E^{|D|} \mid \exists\theta \in \Theta, d(m_\theta, D') \leq \alpha(M)\} \tag{14}$$

$E^{|D|} = E^n$ is equipped with the product measure derived from $\mathbb{P}$ on E, that we also denote as $\mathbb{P}$ by misuse of language. $\mathcal{D}[M]$ is then a measurable subset of $E^{|D|}$ and its "volume" $\mathbb{P}(\mathcal{D}[M]) \in [0, 1]$ represents *how M might not have been as adequate with its data D as it is*, which is exactly what the notion of falsifiability aims at covering. More precisely, the more $\mathbb{P}(\mathcal{D}[M])$, the less falsifiable M. The degree of corroboration of an empirical model M, i.e. the amount of information that we get knowing that M is at this point empirically adequate, is then defined as:

$$\mathcal{C}[M] = -\log(\mathbb{P}(\mathcal{D}[M])). \tag{15}$$

By construction, $\mathcal{C}[M] \geq 0$ for all M. Moreover, $\mathcal{C}[M] = 0$ for M such that their empirical models $(m_\theta)_{\theta\in\Theta}$ correspond to empirical regularities which are mere artifacts of their empirical frame E, namely *endogeneous regularities* (see section 3.2). Indeed, a claim as "poverty rate of a country cannot be greater than 0.5" is true by virtue of the choosen definition of the poverty threshold. Therefore, any data produced thanks to this definition will empirically confirm that claim. Thus, for any such D, $\mathcal{D}[M] = E^{|D|}$ and thus $\mathbb{P}(\mathcal{D}[M]) = 1$, i.e. $\mathcal{C}[M] = 0$. In the case of the Lagrange polynomial regression method, the result is the same. Let $(m_\theta)_{\theta\in\Theta}$ be the polynomial defined from Lagrange method from data $D = \{(x_i, y_i)\}_{i=1..n}$, $n \in \mathbb{N}^*$. By construction, for any $D' \in E^n$, there exists $\theta = (a_0, ..., a_n) \in \mathbb{R}^n$ such that $d(m_\theta, D') = 0$. Therefore, in this case we also get $\mathcal{D}[M] = E^{|D|}$ from definition (14) and thus $\mathcal{C}[M] = 0$.

From this definition of corroboration degree we can now see scientificity as its maximization over available empirical models, as presented in the next section.

3.4 General scientificity principle for empirical models

In order to be comparable, two epistemic units like empirical models have to share at least some minimal features:

- They have to aim at covering the same set of phenomena – or at least, their respective set of phenomena must have a non null intersection.
- They have to be written in the same empirical language – that is to say, they have to share some common basic variables, even if their operational definitions are not exactly the same.
- Their respective data D must have a non null intersection.

Then we can state the following scientificity principle for empirical models:

Maximized corroboration principle: *Given a finite set of comparable empirical models* $\mathcal{M} = \{M_1, M_2, ..., M_N\}$*, the most scientific is the one which maximizes the corroboration degree* $\mathcal{C}[M]$ *over* $M \in \mathcal{M}$*.*

In the next section, we explore some direct entailments of this general principle and discuss how some well-known epistemological criteria turn out to be particular instanciations of it.

4 Epistemological criteria derived from the general principle

4.1 Empirical adequacy

Even if empirical adequacy is not a sufficient condition for scientificity, as falsificationism teaches us, it is still a necessary condition. It may seem strange at first glance that our general scientificity principle bears only on the corroboration degree $\mathcal{C}$: the latter does depend on the adequacy degree but does not reduce to it. Actually, the corroboration degree $\mathcal{C}$ as defined in section 3.3 turns out to be such that the general scientificity principle does encompass empirical adequacy as a necessary criterion.

The reasoning is as follows. Let M and M' be two comparable models such that M' is less adequate than M, i.e. $\alpha(M) < \alpha(M')$. What about $\mathcal{C}[M]$ and $\mathcal{C}[M']$? A greater α means, *ceteris paribus*, that a larger region of data is automatically adequate with the corresponding model. Thus, if α increases between M and M', it follows that $\mathcal{D}[M] \subset \mathcal{D}[M']$, and then $\mathcal{C}[M] > \mathcal{C}[M']$.

An increasing α (i.e. a less good adequacy) thus entails a decreasing corroboration degree. Therefore, an increasing corroboration degree necessarily entails an decreasing α, i.e. a better adequacy. A good empirical adequacy is then already encoded in the demand for a maximized corroboration degree.

4.2 Clear basic variables

Another criterion often mentioned as a good measure of scientificity is the clarity of the definitions of the basic variables used. This is directly related to the maximization of corroboration. Indeed, claims made out of fuzzily defined basic variables will do have an empiral meaning, but their adequacy conditions will be so loose that they will hardly be false. From this kind of basic variables it then seems to be hard to make any substantial empirical claim. In the field of psychology, a recent work (Scheel, 2022) even suggests that its current replication crisis is mostly due to the fact that most psychological claims are "not even wrong", being made out of fuzzy and not well defined basic variables.

A maximized corroboration principle then automatically demands for a sharply defined empirical frame E.

4.3 Lakatos' progressive and degenerating problemshifts

Hungarian philosopher and Popper's disciple Imre Lakatos developed his epistemology from a sophisticated version of falsificationism. According to Popper and Lakatos, it is fine to add a hypothesis to save a theory from refutation if and only if it increases the falsifiability of the given theory and eventually leads to new discoveries (Lakatos, 1978).

From our viewpoint restricted to empirical models, an empirical anomaly (or a "refutation") happens to an empirical model M when its adequacy degree $\alpha(M)$ is above a given threshold α_{crit}.[19] Then, an ajustment is a modification of M into a comparable empirical model M' aiming at resolving the anomaly, i.e. such that: $\alpha(M') < \alpha_{crit} < \alpha(M)$.

However, following sophisticated falsificationism and the general principle stated in section 3.4, a good adjustment is actually such that:

$$\mathcal{C}(M') > \mathcal{C}(M). \tag{16}$$

Indeed, in the Lakatos terms, the problemshift $M \longrightarrow M'$ such that:

$$\alpha(M') < \alpha(M) \text{ and } \mathcal{C}(M') \leq \mathcal{C}(M) \tag{17}$$

is a degenerating problemshift. The adequacy is higher, but somehow trivially. For instance, it is the same kind of adjustment consisting in changing the basic definitions to turn a false synthetic statement into a trivially true analytic one.[20] This way to go from M to M' is usually called an *ad hoc* procedure.

On the contrary, the problemshift $M \longrightarrow M'$ such that:

$$\alpha(M') < \alpha(M) \text{ and } \mathcal{C}(M') > \mathcal{C}(M) \tag{18}$$

is a progressive one, for solving the anomaly increases its corroboration degree.

Our formal constructions allow to make a graphic representations in a $(\alpha, \mathcal{C})-$diagram. A diagramatic representation of a typical progressive and degenerating problemshifts is given on the left hand side of figure 3.

This is obviously way too simple, for a problemshift is actually not only made out of two models, but a whole series of models. That is to say, a model, facing a discrepancy, can be modified by a hypothesis which could at first be seen as *ad hoc*, but which turns out to be eventually fruitful. For instance, imagine that M' in the degenerating case makes a prediction which is then empirically confirmed by new data such that the corroboration degree increases

[19] Remind from section 3.3 that $\alpha(M) = 0$ means that M is perfectly adequate with its data, and that the more $\alpha(M)$, the less empirically adequate M.

[20] As mentioned before, the empirical language used is itself a variable of our problem, so the modification of an auxiliary hypothesis can bear on the most basic observational definitions appearing in E. The only constraint is that the global degree of corroboration increases.

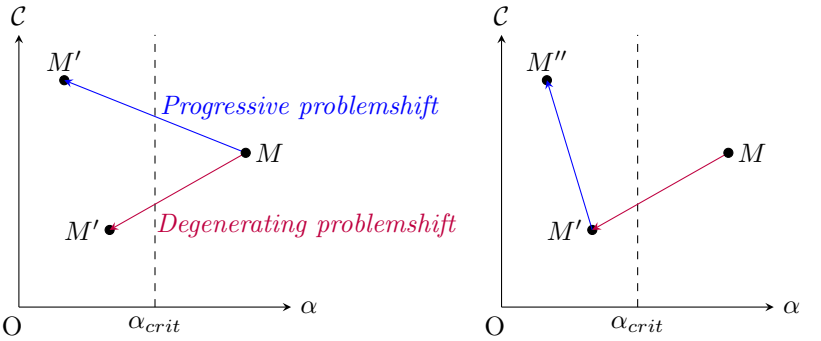

Fig. 3 A diagramatic representation of Lakatos' progressive and degenerating problemshifts. The origin O of the diagram represents endogeneous regularities: $\alpha = 0$ (they are perfectly empirically adequate) but independently of any data, i.e. their empirical adequacy carries no information ($\mathcal{C} = 0$). Left: difference between both. Right: a problemshift which seems to be degenerating at first but which is finally progressive.

even more. Then, we would have something like:

$$M \overset{\alpha\searrow \mathcal{C}\searrow}{\longrightarrow} M' \overset{\mathcal{C}\nearrow}{\longrightarrow} M'' \tag{19}$$

with $M = (m, D)$, $M' = (m', D)$ and $M'' = (m', D')$.[21] This adjustment is diagramatically represented in the right hand side of figure 3. This representation is still quite simplistic. However, it allows to see progressive problemshifts (i.e. with increasing scientificity) as particular trajectories in such a diagram: those for which the total corroboration degree globally increases. This is also why it can be difficult, at a given historical time, to easily distinguish between scientific and less scientific epistemic units: their trajectories in the diagram can be quite bumpy and locally (that is, temporarily) hard to distinguish.

We presented here the case of a change in the available set of data and the case of a modification of the model m. However, a problemshift may also be due to a change in a basic empirical definition (that is, a change of $(E, \mathbb{P})$) or in the distance d. The latter may correspond, for instance, in a new statistical method for assessing the quality of a fit or any statistical association. Again, no element of M is definitely given: they all are *a priori* variables of the problem. The demand for a maximized corroboration degree then allows a co-selection bearing on the *whole set* $(E, \mathbb{P}, (m_\theta)_{\theta\in\Theta}, D, d)$.

To summarize and conclude this section, we can schematically represents three important cases. Let $M \longrightarrow M'$ be a problemshift between two comparable models. From the same reasoning as in section **??**, we have:[22]

$$\Delta\alpha > 0 \text{ entails } \Delta\mathcal{C} \leq 0. \tag{20}$$

[21] Here we denote $M = (m, D)$ as a practical shortcut for: $M = (E, \mathbb{P}, (m_\theta)_{\theta\in\Theta}, D, d)$, where m stands for $(m_\theta)_{\theta\in\Theta}$.

[22] Denoting $\Delta\alpha = \alpha(M') - \alpha(M)$ and $\Delta\mathcal{C} = \mathcal{C}[M'] - \mathcal{C}[M]$.

This is due to the fact already mentioned that corroboration and adequacy are not entirely independent. More precisely, if α increases then, *ceteris paribus*, $\mathcal{C}$ necessarily decreases. Thus, it logically follows that:

$$\Delta\mathcal{C} > 0 \text{ entails } \Delta\alpha \leq 0. \tag{21}$$

That is to say, and still *ceteris paribus*, an increasing corroboration necessarily entails a better adequacy, as mentioned in section **??**. However, we could perfectly get:

$$\Delta\alpha \leq 0 \text{ and } \Delta\mathcal{C} \leq 0. \tag{22}$$

This case corresponds to a degenerating problemshift situation. The three corresponding trajectories in the $(\alpha, \mathcal{C})$–diagram: refutation (20), progression (21) and degeneration (22) are represented in figure 4. Notice that they are more to be taken as general tendencies than actual trajectories.

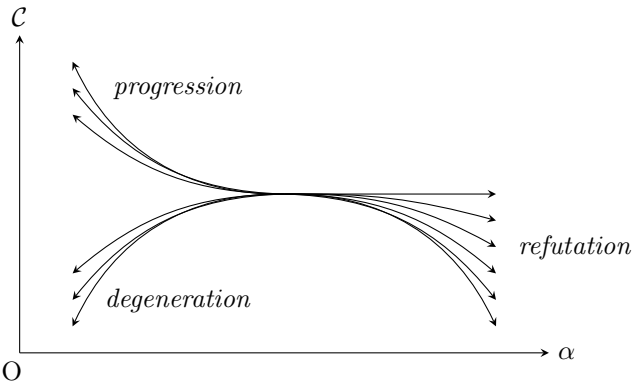

Fig. 4 Sketching three families of possible trajectories for empirical models.

4.4 Lewis' balance between strength and simplicity

In the continuity of our observation above, it seems to be relevant to make a connection with David Lewis' *best system account* (BSA) of laws (Lewis, 1983, 1994) even if Lewis' aim is not to solve the SDP, but rather to demarcate between genuine laws of nature and contingent truths. His idea is to consider theories as axiomatic systems with two main features: strength and simplicity. The strength of a theory is its empirical success, and this feature is in tension with simplicity. Indeed, a theory can be very simple, for instance if its rests on a single axiom, but for sure its strength will be quite low in this case. On the contrary, adding more and more axioms, the theory will explain a lot of different empirical cases, but at the price to become less and less simple.

Lewis then claims that there is a balance to be found between simplicity and strength, and that genuine natural laws are those axioms which allow to reach the best balance. The greatest difficulty in this approach, extensively

discussed since, is to precisely define the strength, the simplicity and thus the balance between them.

It is quite inspiring to transpose this viewpoint into the SDP context, and to see the notion of scientificity as a (best) balance between simplicity and strength of the corresponding epistemic units. In our approach, however, we see that the balance is directly given by the maximization of the corroboration degree. On the one hand, a "wrong" model (low strength) is a model for which the adequacy (and thus the corroboration degree $\mathcal{C}$, see above) is low. On the other hand, a "too complex" model is a model which is trivially true. For instance, it contains so many free parameters that any data could actually be described by it.[23] This case also corresponds, by construction, to a low corroboration degree $\mathcal{C}$. Thus, asking for a maximization of the corroboration degree seems to encompass both extreme cases and recover the intuition behind David Lewis' BSA.

4.5 Parsimony and coherence

The previous discussion about simplicity of theories regarding Lewis' BSA obviously remind the well-known and more general parsimony principle, or Ockham's razor. As noticed in (Sober, 2015), it is quite consensual in philosophy of science that parsimony is an epistemological virtue, but it is still quite hard to justify this principle *a priori* and to give parsimony a precise definition (as in the case of BSA's notion of simplicity).

We agree with Popper that parsimony is actually a particular case of falsifiability.[24] In other words, the most parsimonious theories are better because they are the most falsifiable too. In our words, the most parsimonious models are those which put more constraints on the phase space E, i.e. which are, given a compatibility degree $\alpha > 0$, compatible with less data. Parsimony can be directly connected with the number of free parameters, but not necessarily. What matters above all is to maximize the degree of corroboration $\mathcal{C}$.

Coherence is sometimes invoked to justify parsimony: a parsimonious theory can be seen as a theory which rests on as least unexplained features as possible, thus which is the most coherent with our current knowledge. Yet, coherence itself is a particular case of the general maximized corroboration principle. Indeed, asking for coherence put just more constraints on the phase space of data E (as parsimony does) and thus, *ceteris paribus*, such a model has a greater degree of corroboration that a model which allows, e.g., some parameters to have a value not coherent with what is already established, or considered so. Asking for a maximized global degree of corroboration as defined in our work will then probably entails a certain degree of coherence to be maximum too.[25]

[23] This has also an obvious connection with what is called "overfitting" in data science.

[24] (Falk & Muthukrishna, 2021) also makes a connection between parsimony and degree of falsifiability in the context of fit propensity in structural equation modelling.

[25] This is a conjecture that we should underpin in a future work. The main difficulty being to get a precise definition of coherence.

To illustrate our point, let us examinate the reflections of Henri Poincaré in the chapter IX of "Science and hypothesis" (Poincaré, 1902, 2017, p. 146) about the criterion to choose between different possible generalizations:

> The choice can only be guided by considerations of simplicity. Let us take the most ordinary case, that of interpolation. We draw a continuous line as regularly as possible between the points given by observation. Why do we avoid angular points and inflexions that are too sharp? Why do we not make our curve describe the most capricious zigzags? It is because we know beforehand, or think we know, that the law we have to express cannot be so complicated as all that. The mass of Jupiter may be deduced either from the movements of his satellites, of from the perturbations of the major planets, or from those of the minor planets. If we take the mean of the determinations obtained by these three methods, we find three numbers very close together, but not quite identical. This result might be interpreted by supposing that the gravitation constant is not the same in the three cases; the observations would be certainly much better represented. Why do we reject this interpretation? Not because it is absurd, but because it is uselessly complicated. (...) To sum up, in most cases every law is held simple until the contrary is proved.

Here, simplicity is seen as a guiding convention of scientific inquiry, without any other justification. Among different functions to fit (interpolate) some given data, we choose the "simplest", e.g. a linear one rather that one making complicated zigzags. In the case of the mass of Jupiter, we interpret the fact that the three results are not the same as being due to an observational imprecision and not to the fact that the gravitational constant G is not the same in the three situations. As Poincaré suggests, nothing really prevent us to do that, from a logical point of view. If we make this choice, it is in virtue of an independent principle, that Poincaré calls "simplicity", avoiding "useless complexity".

From our viewpoint, this choice is justified by the same principle, that of maximizing corroboration and not only adequacy: a model with the same gravitational constant for different experimental situations has a greater degree of corroboration (i.e. is *a priori* more falsifiable) than a model which allows the gravitational constant to vary – even if the latter is more adequate with data. Besides, Poincaré implicitely admits that it is not a sufficient epistemological virtue. It is not that it is fundamentally impossible for G to vary, but only that, everything else being equal, a model with G being constant put more constraints on the corresponding phase space than a model which allows G to vary. The same reasoning applies to the interpolation example: a function with complicated zigzags is likely to be based on more free parameters that a "simpler one", and thus, *ceteris paribus*, much more data are *a priori* compatible with it. We then gain less information knowing that it fits the data.

According to us, these considerations about the virtue of parsimonny have also something important to do with explanatory parts of theories, but as already mentioned this goes beyond the scope of this paper.

4.6 Karl Popper's degree of falsifiability

The present work can be seen as an attempt to adapt Karl Popper's sophisticated falsificationism to a model-based vision of scientific knowledge. Popper sees theories as set of statements, and defines their falsifiability from "the classes of their potential fasifiers". More precisely (Popper, 1959, p. 95-96):

> [If] we represent the class of all possible basic statements by a circular area, and the possible events by the radii of the circle, then we can say: At least one radius—or perhaps better, one narrow sector whose width may represent the fact that the event is to be 'observable'—must be incompatible with the theory and ruled out by it. One might then represent the potential falsifiers of various theories by sectors of various widths. And according to the greater and lesser width of the sectors ruled out by them, theories might then be said to have more, or fewer, potential falsifiers. (The question whether this 'more' or 'fewer' could be made at all precise will be left open for the moment.)

In the section 32 "How are classes of potential falsifiers to be compared", Popper then comes back to the last remark of the above citation: how to define properly that some theories are more falsifiable than others? That is to say, from his perspective: how to express the fact that some theories have more falsifiers than others, or that their classes of falsifiers are greater? The main issue is that, obviously, these classes are often infinite sets. His choice is eventually made for a "subclass relation" between classes of falsifiers. Two theories T and T' can then be compared with the help of their corresponding classes of falsifiers α and β, for a subclass relation can be defined on them. T is then more falsifiable than T' if and only if $\alpha \subset \beta$. As Popper notices it, this relation allows to formalize "the intuitive 'more' and 'fewer', but it suffers from the disadvantage that this relation can only be used to compare the two classes if one includes the other." (Popper, 1959, p. 98)

In this paper, we followed exactly this very idea, but from another perspective, another model of theories. From our viewpoint, the conditions for two epistemic units M and M' to be comparable are looser (see section 3.4). Indeed, the equivalent, in our formalism, of a popperian class of falsifiers is the set $\mathcal{D}[M]$ – or more precisely, $E^{|D|} \setminus \mathcal{D}[M]$: the set of all possible empirical data that are not as adequate with the model as D is. In order to be comparable, these respective sets just have to have a non null intersection – they do not need to be subsets of the other, as in Popper's way of defining degree of falsifiability. Thus, our approach allows to benefit from the central popperian idea of falsifiability without falling in the same kind of issues.

5 Limits and perspectives

This paper is a first outline of a more general formal approach to the SDP. A certain number of points remain unclear and have to be developed furthermore. As a conclusion, we analyse two possible criticisms which seem to be important: 1/ the superfluous nature of formalization and 2/ the restriction to empirical models. These limits give us the occasion to outline some further perspectives.

5.1 Formalization

The first criticism which could be raised about our approach is the level of formalization considered as superfluous. Are we just reformulating some well-known epistemological facts in an unecessarily complicated language? Does this formalization really bring something important and new?

Our main aim at formalizing is to clarify the underlying problem: on which epistemic unit our analysis applies and how to define it properly? Adressing these questions leads to the formulation of a clear scientificity principle, and allows to explore some possible implications of it. Formalizing in a mathematical language allows us to define clearly the entities under study. As in other fields where mathematics is used, its relative autonomy from our starting constructions allows us to be eventually guided by it. This process offers a cognitive support which seems to be salutary w.r.t. the intrinsic difficulty of the problem under study.

Moreover, this paper has two main and (to a certain extent) independent purposes. The first one is to outline a global strategy to adress the SDP in a certain way, which rests on the idea that scientificity can be seen as the maximisation of "something" over all possible models. The second one is a precise proposal for this "something" to be maximized. Thus, criticisms bearing on the too heavy formalization for the latter would not, *a priori*, reach the former.

5.2 Relevant scales of analysis

Another critical remark is that of the scale of analysis. In this paper, we restrict our analysis to a precise epistemic unit, namely empirical models. However, the epistemological justification of these models does not only rely on the corroboration degree (i.e. the quality of its connection with data) but also on theoretical supports. More precisely, the scientificity is not only something which can be assessed "locally" but also (and to some extent above all) at a larger scale, namely that of fundamental explanatory principles, i.e. theoretical and not only empirical models.

An important remaining question is thus how to connect the formal constructions made at the level of empirical models to the level of fundamental principles lying at the core of theories. The purpose is the same: clarifying the epistemological discussion about scientific explanations while aiming at recovering well-known criteria[26] from a more general principle.

Nevertheless, our approach explicitely relies on a certain methodological reductionism (studying the whole from the parts), and this is an open question to know to which extent this strategy is fruitful, and the task for future work.

Declarations

The authors have no relevant financial or non-financial interests to disclose.

[26] For instance, a part of the set of criteria used in current multi-criteria approaches.

References

Boudon, R., & Lipset, S. (1974). *Education, opportunity, and social inequality: Changing prospects in western society.* Wiley.

Bradford Hill, A. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, *58*, 295-300.

Bunge, M. (1983a). *Treatise on basic philosophy, vol. 5: Epistemology and methodology i: exploring the world.* (D.R. Dordrecht, Ed.).

Bunge, M. (1991). What is science? does it matter to distinguish it from pseudoscience? a reply to my commentators. *New Ideas in Psychology*, *9*(2), 245-283.

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress.* Oup Usa.

Cobb, C.W., & Douglas, P.H. (1928). A theory of production. *The American Economic Review*, *18*(1), 139–165.

Coleman, J.S. (1990). *Foundations of social theory.* Harvard University Press.

Czerny, B., Beaton, R., Bejger, M. (2018). Astronomical distance determination in the space age. *Space Sci Rev*, *214*.

Falk, C.F., & Muthukrishna, M. (2021). Parsimony in model selection: Tools for assessing fit propensity. *Psychological Methods*.

Felipe, J., & McCombie, J. (2013). *The aggregate production function and the measurement of technical change: 'not even wrong'.* Edward Elgar.

Fernandez-Beanato, D. (2020). The multicriterial approach to the problem of demarcation. *J Gen Philos Sci*, 375-390.

Fernandez-Beanato, D. (2021). Feng shui and the demarcation project. *Sci & Educ*, 1333-1351.

Fernandez-Beanato, D. (2022). *A working demarcation problem* (Unpublished doctoral dissertation). University of Bristol, Bristol, U.K.

Hanson, N. (1958). *Patterns of discovery.* Cambridge University Press.

Hansson, S.O. (2021). Science and Pseudo-Science. E.N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2021 ed.). Metaphysics Research Lab, Stanford University.

Harding, S. (1975). *Can theories be refuted?: Essays on the duhem-quine thesis.* Reidel.

Hume, D. (1748). *Enquiry concerning human understanding.* London: A. Millar.

Kuhn, T. (1962). *The structure of scientific revolutions.* University of Chicago Press.

Lakatos, I. (1978). *The methodology of scientific research programmes: Philosophical papers* (Vol. 1; J. Worrall & G. Currie, Eds.). Cambridge University Press.

Laudan, L. (1983). The demise of the demarcation problem. In R.S. Cohen & L. Laudan (Eds.), *Physics, philosophy and psychoanalysis: Essays in honour of adolf grünbaum* (pp. 111–127). Dordrecht: Springer Netherlands.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, *61*(4), 343-377.

Lewis, D. (1994). Humean supervenience debugged. *Mind*, *103*(412), 473–490.

Mahner, M. (2007, 12). Demarcating science from non-science. *General Philosophy of Science: Focal Issues*, 515-575.

Merritt, D. (2017). Cosmology and convention. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *57*, 41 - 52.

Pigliucci, M., & Boudry, M. (2013). *Philosophy of pseudoscience: Reconsidering the demarcation problem.* University of Chicago Press.

Poincaré, H. (1902). *La science et l'hypothèse.* Flammarion.

Poincaré, H. (2017). *Science and hypothesis.* Bloomsbury Publishing.

Popper, K. (1934). *Logik der forschung.* Julius Springer.

Popper, K. (1959). *The logic of scientific discovery.* Hutchinson and Co.

Popper, K. (1962). *Conjectures and refutations: The growth of scientific knowledge.* London, England: Routledge.

Popper, K. (1972). *Objective knowledge.* Clarendon Press - Oxford.

Quine, W.V.O. (1951). Two dogmas of empiricism. *Philosophical Review*, *60*, 20–43.

Richardson, S.S., Reiches, M.W., Bruch, J., Boulicault, M., Noll, N.E., Shattuck-Heidorn, H. (2020). Is there a gender-equality paradox in science, technology, engineering, and math (stem)? commentary on the study by stoet and geary (2018). *Psychological Science*, *31*(3), 338-341.

Scheel, A. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development*.

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

Stoet, G., & Geary, D.C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science*, *29*(4), 581-593.

Sullivan, G., & Feinn, R. (2012). Using effect size-or why the p value is not enough. *J Grad Med Educ.*, *4*(3), 279-82.

Suppe, F. (1974). *The structure of scientific theories.* Urbana, University of Illinois Press.

van Fraassen, B. (1980). *The scientific image.* Clarendon Press.

WEF (2022). *Global gender gap report 2022.* World Economic Forum.