The evaluation of measurement uncertainties and its epistemological ramifications

Nadine de Courtenay, Fabien Grégis

The Evaluation of Measurement Uncertainties and its Epistemological Ramifications

Nadine de Courtenay and Fabien Grégis

Université Paris Diderot, Sorbonne Paris Cité, laboratoire SPHERE UMR 7219, bâtiment Condorcet, case 7093, 5 rue Thomas Mann, 75205 Paris Cedex 13, France.

nadine.decourtenay@univ-paris-diderot.fr
fabien.gregis@etu.univ-paris-diderot.fr

———

Abstract
The way metrologists conceive of measurement has undergone a major shift in the last two decades. This shift can in great part be traced to a change in the statistical methods used to deal with the expression of measurement results, and, more particularly, with the calculation of measurement uncertainties. Indeed, as we show, the incapacity of the frequentist approach to the calculus of uncertainty to deal with systematic errors has prompted the replacement of the customary frequentist methods by fully Bayesian procedures. The epistemological ramifications of the Bayesian approach merge with a deep empiricist mood tantamount to an "epistemic turn": measurement results are analysed in terms of degrees of belief, and central concepts such as error and accuracy are called into question. We challenge the perspective entailed by this epistemic turn: we insist on the centrality of the concepts of error and accuracy by underlining the intentional character of measurement that is intimately linked to the process of correction of experimental data. We further circumvent the difficulties posed by the classical analysis of measurement by stressing the social rather than the epistemic dimension of measurement activities.

Keywords
Measurement, uncertainty, error, Bayesian statistics, frequentist statistics, social epistemology

Highlights
• Describes an evolution from frequentist to Bayesian statistics in measurement science
• Shows how measurement uncertainty is then understood as a statement of knowledge
• Argues that objective evaluation through error and accuracy remain central
• Insists on the intentionality of measurement related to a process of correction
• Stresses the social rather than the epistemic dimension of measurement activities

1. Introduction

Measurement science has been in a state of ferment in the past two decades. Catalysed by the scientific and technical advances of the last century, and by the requirements of economic globalization, it has experienced a period of clarification and reform. Two important guides have been published in order to harmonize the vocabulary and concepts of metrology (the science of measurement), as well as the measurement practices at the international level – the *International Vocabulary of Metrology* (*VIM*) and the *Guide to the Expression of Uncertainty in Measurement* (*GUM*) –, and a deep revision of the international system of units, the SI, is underway.

Part of the reason for this recent activity is the revamping of the statistical methods used to deal with experimental data, and, more particularly, with the calculation of measurement uncertainties. How to calculate uncertainty has been a major subject of discussions in metrology at least since the middle of the twentieth century. The *GUM*, published in 1993, aspired to resolve these discussions but in the end, failed to provide convincing probabilistic bases to the calculus. It did, however, generate a number of lively debates, which prompted today's profound transformation of the analysis of measurement data by replacing the classical frequentist methods with Bayesian approaches. The epistemological ramifications of the Bayesian approach merge with a deep empiricist mood pervading the metrological community to instigate a far-reaching revision of the way metrologists conceive of measurement tantamount to an "epistemic turn": measurement results, the traditional touchstones of scientific objectivity, are analysed in terms of degrees of belief, and central normative concepts such as error and accuracy are called into question.

After taking stock of the way measurement error and measurement uncertainty are introduced in the analysis of experimental data, we will explain how the transformation of the metrological conception of measurement originates in the attempt to provide a probabilistic treatment of systematic errors which are of paramount concern in measurement issues. Indeed, the epistemic interpretation of uncertainty, and measurement as a whole, is designed to avoid the difficulties encountered within the classical account of measurement when one contemplates assessing the correctness of a result by reference to an unknown and unknowable true value of the quantity one intends to measure. The determination to elude entities that cannot be given empirically, such as the true value of a quantity, results in dismissing the notion of error and replacing the requirement for accuracy with that of a rational expression of our knowledge. We will challenge the downgrading of error and accuracy and propose an analysis that stresses the pragmatic and social, rather than the epistemic dimension of measurement. Our approach will suggest that the difficulties attached to the objective evaluation of the quality of a measurement result, and therefore to the concepts of error and accuracy, can be circumvented when one thoroughly takes into account the intentional character of measurement and acknowledges that the expression of a measurement result involves the posit of a true value as a regulative idea guiding an activity of correction involving the interactive criticism of a community of agents that has a common target. It becomes then possible to conceive of accuracy in a new way; not as the impossible static appreciation of the closeness of the result to a true value, but as a feature related to the reliability of a process of correction anchored in the objectives, values and norms embedded in the social framework underlying measurement activities.

2. Analysis of the variability of measurement indications: measurement errors and measurement uncertainty

2.1 The singular nature of measurement data

A measurement datum is a singular entity. It is the result of a concrete interaction between a physical system bearing the quantity one wants to measure (the length of a particular end gauge, the velocity of light in the vacuum) and a particular experimental setup, in a particular environment, at a particular time, according to a particular procedure. The information derived from such an interaction on the quantity of interest is inevitably entwined with information pertaining to the setup, the environment, and the procedure followed. The question immediately arises of how this datum can be used to give adequate information on the quantity of interest when the quantity is set in a different experimental environment, in different circumstances. How can one obtain from such a measurement datum, information that is valid outside of the particular context in which the datum was produced? Here one is

confronted at the most basic level with the question of how to transform indications of a singular and local nature into measurement results that convey a general, public knowledge of quantities that can be meaningfully and reliably shared.

The first condition the information should meet is that of *communicability*. The chief agent of communicability is the unit of measurement. As Giordani and Mari (2011) have pointed out, measurement is an experimental process by which a concrete, empirically given quantity *Q*, known by acquaintance, which cannot always be shared, gets expressed by a quantity value *{Q}[Q]*, where *{Q}* is a number and *[Q]* a measurement unit, and thus turns out to be known also by description. Knowledge about the quantity can thus be communicated to distant operators. This description is accomplished by assigning the quantity to a class, identified by *{Q}*, within a classification determined by the publicly defined unit *[Q]*. The assignment is achieved by experimentally comparing the concrete quantity with a standard materializing the unit[1].

Our main concern, in this paper, will be with the other condition that the information must fulfil in order to be valid beyond the context of its production: a quantity value obtained in a given set of circumstances should be *comparable* with a quantity value of the same concrete quantity obtained in different circumstances; in other words, it should be projectable outside of the experimental context in which it was produced in order to be able to be compared with other evaluations of the same quantity obtained in different circumstances, with theoretical predictions or with technical specifications.

As already mentioned, rough indications obtained in a particular experiment do not satisfy this condition. Their singular nature, the fact that they are tied to a particular context, shows up in their variability: provided one operates with instruments having sufficient resolution, a measurement process will yield different indications when it is repeated. This variability is a straightforward obstacle to comparability; it can be analysed and rectified, but never entirely: it is not possible to completely do away with the context of production. We will see that the agent that makes it possible to deal with the remaining variability, and allows the handling of comparisons by giving the means to make judgements of sameness and difference, is the "uncertainty" associated with the measurement result and derived from the analysis of variability. In order to perform its function, the uncertainty must be quantified. As a consequence, public, usable measurement results should always be stated with their associated uncertainty.

2.2 From measurement errors to measurement uncertainty

The variability of measurement indications manifests itself in two very different ways. It appears, firstly, when one realizes a series of repeated measurements of the same physical system in identical conditions (one says, in "conditions of repeatability"): if the resolution of the instruments is good enough, the measurement indications obtained in these successive experiments will not be the same: they will show a dispersion. Another kind of variability appears when one undertakes to measure the same quantity in distinct experiments, differing either in the measurement principle applied, or in the instruments involved, in the environment or other circumstances (one talks then of "conditions of reproducibility"). Contrary to what happens in the first case, this kind of variability is not observed within the context of a single experiment; it only shows up when one confronts the indications gathered from a variety of different experiments.

The classical way to handle the problem posed by the variability of measurement data is to postulate the uniqueness of measurement results and, by so doing, to introduce the notion of

[1] For more on these issues, see Giordani & Mari (2011).

error of measurement.[2] The rationale for such a postulate lies in "our vague and general […] theory of physical objects"[3] which leads us to think that physical properties, and therefore the quantities measured, do not change in conditions of repeatability. We then *explain* the difference between the indications y_i (gathered in the case of a direct measurement) and the unique true target value *TV* of the quantity intended to be measured, called the "measurand"[4], by resorting to the concept of measurement error. The error e_i bearing on the indication y_i is then given by (1) $y_i = TV + e_i$. The true value of the measurand and the measurement errors introduced are thus theoretical concepts suggested by our expectations and theories, which play a normative role in orienting our analysis of measurement experiments. The theoretical status of errors of measurement appears even more salient when one considers the incompatibility between the indications obtained in conditions of reproducibility; one then draws on the hypothesis that quantities are conserved or can be reproduced in different places at different times, and can be determined by resorting to different laws and measurement principles.

Two kinds of measurement errors correspond to the two kinds of variability distinguished above. The dispersion of measurement indications is interpreted as the manifestation of deviations from the true value due to a multiplicity of variable, unknown causes that affect in an unpredictable and uncontrollable way the functioning of the instruments, the environment or the operator. These deviations are described as "random errors" because, as we will see below, the treatment that is applied to reduce them is a statistical treatment based on the assumption that the processes producing these errors are random. Thus, when one measures, for example, the length *L* of an end gauge by comparing it by means of a comparator to a standard gauge of the known length L_S, repeated direct measurements of the difference *D* between the two lengths will yield different indications d_1, d_2, …d_n. The length *L* of the end gauge is then given by applying the measurement principle: (2) $L = L_S + D$, where the value of *D* is obtained through a statistical analysis of the data $d_1, d_2, \ldots, d_n$, based on the hypothesis of random errors. In that case, equation (1) concerns the "input quantity" *D* and not directly the measurand *L*.

The other kind of variability exhibited by the indications is attributed to disturbing causes that are attached to each particular experiment; these causes remain constant in conditions of repeatability and always affect the indications delivered by the experimental setup in the same way. The resulting deviation from the true value is constant and therefore cannot be observed in the context of a given experiment; it can only be discerned when one confronts the indications obtained in one setup with those obtained in a different one. This deviation, which introduces a constant discrepancy between the numerical indications delivered by the experiment and the target value, is called a "systematic error." Contrary to random errors, systematic errors can be grasped and corrected only by calling on information that is not contained in the experimental observations themselves. For this reason they are quite difficult to identify. They can be due to physical influence factors that have an effect on the quantity of interest. In our example, the length of the end gauge depends on temperature; so, if the temperature of the room is not the one at which one wants to measure the length of the end gauge, one will have to take into account an error introduced by the actual temperature of the end gauge which modifies the length of the gauge with respect to the length one intends to measure. Other constant causes responsible for systematic errors are (i) the defects of the instruments used in the experiments – here the comparator and the standard gauge; the corrections then rely on the calibration certificates provided by the constructors; and (ii) the

---

[2] For a comprehensive account of the notion of measurement error, see Boumans & Hon (2014).

[3] Kyburg (1992), p. 77.

[4] Joint Committee for Guides in Metrology (2012), p. 17.

use of experimental results obtained in other experiments or of physical constants given in tables – like the coefficient of thermal expansion involved in our example. In order to take these factors into account, the measurement principle $L = L_S + D$ has to be extended. In the most general case, the key relationship is a mathematical expression *f*, called the "measurement function": (3) $Y = f(X_1, \ldots X_n, C_1, \ldots C_m)$, giving the measurand *Y* in terms of (i) the input quantities (X_p) that must be measured, and of (ii) other parameters involved in the correction of systematic effects (C_k).

It is only by submitting the measurement data to a thorough analysis, and correcting random and systematic errors, that one can arrive at a properly so called measurement result which can stand on its own and be used by other agents in different locations. As we will see below, the two types of error must receive a different treatment. However, in both cases, the corrections are limited, imperfect, so that the measurement result only supplies an *estimate* of the measurand. There always remains an element of doubt, an *uncertainty* surrounding the result as a consequence of the impossibility to perform complete corrections; as stated in the *GUM*:

> (W)hen all of the known or suspected components of error have been evaluated and the appropriate corrections have been applied, there still remains an uncertainty about the correctness of the stated result, that is, a doubt about how well the result of the measurement represents the value of the quantity being measured.[5]

It is therefore only the result, that is, the estimate together with its associated uncertainty that can have a claim to objective value.

As we just saw, there are two sources, two components of uncertainty – one coming from random, the other from systematic errors. The *GUM* set out to provide quantitative measures of these two components of uncertainty that could be combined and ensure comparability.

3. Quantifying uncertainty: the *GUM*'s proposal and its shortcomings

The *GUM*, published in 1993, is one of the two guides commissioned by the International Bureau of Weights and Measures (BIPM) to settle the host of different methods and practices that burdened measurement activities. The main purpose of the *GUM* was to survey the different ways of handling measurement uncertainties found in the scientific community and industry. The agenda of the *GUM* was twofold. It sought to express all the components of uncertainty in one unique term so as to make it easier for users, including scientists, industry and decision makers, to deal with measurement results. This pragmatic objective came with a theoretical one that was to put the calculation of uncertainty on sound probabilistic bases. As we will see, the *GUM*'s attempt to coordinate these two aims, far from clarifying the situation, turned out to exacerbate the debates about uncertainty issues that had been ongoing since the 1970s.

3.1 The frequentist account of measurement uncertainties: the stumbling block of systematic errors

The probabilistic treatment of the first source of uncertainty arising from the correction of random errors is straightforward; it stems from a tradition going back to the works of Gauss and Laplace. Considering that the parameters that influence the outcome of each measurement datum in conditions of repeatability are too complex to be analysed, the process by which each single datum is generated is viewed as a black box. The variability of the experimental data is modelled as if it were the result of a random process where every individual measurement indication is the product of the random "drawing" of a value from a statistical infinite "parent" population made-up of all the potential results that could possibly occur if

[5] Joint Committee for Guides in Metrology (2008), p. viii.

measurements were indefinitely repeated in conditions considered as the same. In our example, the measurement of the difference of length D between the end gauge and the standard gauge is represented mathematically by a random variable $\hat{d}$ describing the potential outcomes of the measurement. The distribution of probability of the random variable is the limit of the relative frequencies of occurrence of each possible value d_i in an infinite number of trials; this treatment of the data is known as the "frequentist approach".

Within such a probabilistic model of the measurement process, the dispersion of the data is ascribed to deviations introduced by a multiplicity of unidentifiable causes around the fixed value of the measurand, which, as we saw, is deemed to have no reason to change in repeatability conditions. In this model, associated with certain hypotheses concerning the deviations interpreted as random errors, it becomes possible to infer, through a collective analysis of the data, from the finite sample of the data actually collected back to the two parameters characterizing the entire population: its central value (expectation) μ and its standard deviation (dispersion) σ. This statistical inference provides empirical estimates of these parameters that can be calculated from the experimental data. According to the model, the estimate of the central value refers to the stable cause of the data, and thus supplies an estimate of the fixed value of the measurand. The arithmetical mean of the sample of data proves to be a good estimation of this value which tends towards μ when the size of the sample becomes infinite. The estimate of the standard deviation characterizes the amplitude of the tendency of the measurement process to produce variable results and reflects the fluctuation of the tiny, uncontrolled influence factors in the environment and experimental setup that are responsible for the scattering of the data around the value of the measurand.

One can show that the expected error made on the arithmetical mean (different means are obtained for different samples of data) is smaller than that of the raw data – the variability of the mean is smaller than the variability of the data. The random errors are therefore statistically reduced, but not corrected; there remains an unknown discrepancy between the mean, taken as the measurement result, and the true value of the measurand. As a consequence, there is an unavoidable *uncertainty* surrounding our knowledge of the latter. The standard deviation of the mean, denoted by *u(d)* in our example, is regarded as the measure of this uncertainty; it is a dispersion parameter that expresses the tendency of the mean to differ from μ for a sample of a given size n.

If one introduces another hypothesis, bearing on the probability distribution of the means (calculated for different samples), a probabilistic account of the uncertainty can be provided by constructing a "confidence interval" I_p containing μ with a probability, or confidence level, of p: $I_p = \left[\bar{d} - k_p u(d), \bar{d} + k_p u(d)\right]$, where k_p depends on the form of the probability distribution of the means and the confidence level chosen p. However, contrary to what this formulation might suggest, a confidence interval is not a statement of probability *about* μ (such as: μ belongs to I_p with a probability p). The level of confidence p does not stand for the probability that I_p contains μ. In tune with the frequency approach of probability, the above formulation expresses the *rate of success* with which the procedure of calculating I_p produces correct intervals, that is intervals containing μ: if one repeats the procedure of collecting different samples of size n, and constructs for each of them the mean and the corresponding interval I_p, the limit of the frequency with which the intervals I_p will contain μ is p. Confidence therefore hangs on an objective feature of the procedure. In practice, only one interval I_p is calculated, and $k_p u(d)$ is interpreted as a numerical evaluation of the confidence that one can have in the statistical inference performed and, therefore, in the measurement result to which it is associated.

A similar probabilistic treatment cannot be applied to the systematic component of error. Indeed, systematic errors remain fixed through measurement repetitions and cannot be

connected to a population; a statistical analysis is thus impracticable. Systematic errors are treated individually, on the basis of information that comes from outside the experiment, and yields corrections of the estimate obtained through the analysis of random errors. The partial character of these corrections results in another source of uncertainty concerning the value of the measurand. The relevant information can take the form of a physical model that allow for the correction of influencing factors, — like, for instance, the law of expansion of the end gauge under the effect of temperature. The information can also come from the workshops that have constructed the instruments involved in the experiment or that have carried out the calibration of the standard end gauge compared with the end gauge to be measured. In the latter case, in particular, the value of the length of the standard l_S is typically given by the constructor in the form of an interval such as: $l_0 - h < l_S < l_0 + h$, which is obviously not a probabilistic interval built on a statistical variance; it merely gives an estimate of the "credible limits of the error"[6]. In consequence, there is no way to combine mathematically such an interval with the uncertainty arising from the treatment of random errors, since the two evaluations are of quite a different nature. As a consequence, the measurement result is not associated with one, but two components of uncertainty: the probabilistic uncertainty component related to random errors, and the statement of accuracy pertaining to the systematic component.

3.2 From the *GUM*'s probabilistic account of systematic errors to a fully Bayesian account of measurement uncertainties

The *GUM* was specifically designed to overcome the inability of the frequentist approach to present the different components of uncertainty as one unique term. To this effect, it endeavoured to find a way to handle systematic errors probabilistically so as to be able to "(account) for both systematic and random errors on comparable footing."[7] The idea was to frame the uncertainty resulting from the correction of these errors in terms of a statistical variance that could be readily combined with the variances produced by the frequentist analysis of random errors. This aim was achieved by resorting to epistemic probabilities that afford a probabilistic treatment of a constant unknown parameter: this probabilistic account of the systematic components of uncertainty relies on the introduction of a random variable that describes the experimenter's incomplete knowledge of the systematic errors involved. The probabilities involved are no longer related to the factual variability of the measurement outcomes; instead they designate the state of knowledge of the experimenter by expressing the degree of belief with which he ascribes different possible values to the parameters in hand on the basis of the information that is available to him. Here, the quantification of uncertainty is achieved through the formulation of what amounts implicitly to an *a priori* distribution, very much in the Bayesian spirit. Thus, if he has no other information than that the value of the length of the standard end gauge l_S lies between $l_0 - h$ and $l_0 + h$, the experimenter will have no reason to believe that one value lying in the interval is more probable than another one, and he will take the distribution of probability of l_S, or rather, of his knowledge of l_S, to be a rectangular distribution centred on l_0 and of width $2h$. The standard uncertainty $u(l_S)$ of l_S will then be the mean square of the variance of this distribution.

Now, on the grounds that the frequentist and the epistemic probabilities both adhere to the same mathematical axioms of probability, the *GUM* proceeds to combine the uncertainties obtained by the two methods in one unique term. The total uncertainty associated with the measurement result – that is, with the estimate resulting from the reduction of the random errors, and corrected, although not completely, from the systematic effects – is calculated

---

6 Eisenhart (1963), p. 181.

7 Joint Committee for Guides in Metrology (2009), p. 3.

according to the mathematical rules of probability calculus, which are independent of the particular interpretation of probability adopted, by a formula called "propagation of uncertainties". In the case of our example, the combination is particularly simple; provided there is no correlation between the input quantities, it is the square root of the quadratic sum of the two standard uncertainties: $u(l) = \sqrt{u(d)^2 + u(l_S)^2}$. The resort to epistemic probabilities seems then to get round the shortcomings of the frequentist approach and afford a unification of the two components of uncertainty that the frequentist approach had to keep separate.

But, the solution advocated by the *GUM* hangs on purely technical considerations that have to do with the mathematical handling of the components involved. Various specialists of the field disagreed from the start with the general direction adopted:

> Several authors as well as the committee ISO/TC69/SC6/WG3, who also deal with the measurement uncertainties, are not satisfied with the BIPM recommendation, mainly because the BIPM uncertainty measure is not supported by conventional statistics.[8]

Many metrologists regarded the mixture of the two components of uncertainty, involving two quite different, indeed traditionally viewed as completely opposed interpretations of probability, as very problematic. The recommendations of the GUM

> […] (have) been of concern to many statisticians because it appears to combine frequentist performance measures and indices of subjective distributions in a way that neither frequentists nor Bayesians can fully endorse.[9]

For one thing, it was difficult to give a clear meaning to the final compound uncertainty. Moreover, there was the concern that possible contradictions might lurk in the results of such a combination of different kinds of probabilities. The solution recommended by the GUM was thus viewed with suspicion and, at any rate, deemed to be confused. It advanced an unsatisfactory compromise that failed to demonstrate how probabilities of different natures could be used simultaneously.

A way to respond to this criticism has gained considerable momentum among metrologists, to the point of being explicitly developed in the supplement 1 of the *GUM*, published in 2008. It has also influenced the revision of the *GUM* that is currently underway. The response to the criticism involves fleshing out the epistemic interpretation of probability used to handle systematic errors so that it becomes a genuine Bayesian approach applied to the treatment of all the components of uncertainty.

> The [GUM] recommends classical (frequentist) statistics for evaluating the [statistical] components of uncertainty; but it interprets the combined uncertainty from a Bayesian viewpoint. This is inconsistent. In order to overcome this inconsistency, we suggest that all […] uncertainties should be evaluated through a Bayesian approach.[10]

This means that the outcomes d_i of the repeated measurement of *D* are not described anymore as the realizations of a random variable generated by a measurement procedure. The experimenter starts by constructing an *a priori* distribution of probability δ (called prior distribution) of the input parameters, the difference of lengths *D* in our example, on the basis of all the information at his disposal. He then uses the experimental data gathered through repeated measurements (here the measurement outcomes d_i) as information to *revise* or update his knowledge of the probability distribution δ by calculating an *a posteriori* distribution thanks to Bayes' theorem. A standard uncertainty can be defined (here for *D*) as

---

8 Weise and Wöger (1992), p. 1. The BIPM recommendation is a short notice published in 1980 that establishes the groundwork for the *GUM*. In their article, the authors also refer to a draft of the *GUM*.

9 Gleser (1998), p. 277.

10 Kacker and Jones (2003), p. 235. The same kind of argument can be found in Bich (2012), D'Agostini (1996), and Lira and Wöger (2001).

the standard deviation of the posterior distribution. In this approach, the probabilities involved have become epistemic. They are not frequency limits of occurrence attached to the measurement outcomes d_i; they are the degrees of belief the experimenter holds concerning the different possible values he can ascribe to the parameter considered (here, the input parameter *d*). They are epistemic probabilities that bear directly on the unknown, fixed value one is trying to evaluate.

In this scheme, one and the same Bayesian method is applied to all the parameters involved in the measurement. Given the measurement function (3) $Y = f(X_1, \ldots X_n, C_1, \ldots C_m)$, the determination of the uncertainty associated to the value of *Y* proceeds in two stages. The experimenter first devises prior probability distributions for all the variables appearing on the right hand side of the relationship by making use of all available information, and possibly drawing on a principle of maximum entropy; he then updates the distributions corresponding to the variables for which he disposes of measurement data to obtain *a posteriori* distributions of probability. Combining these different distributions through a "propagation of distributions", often quite tedious to perform, the measurand *Y* is supplied in the form of a distribution of probability from which one can, eventually, extract a standard deviation.

4. The philosophical ramifications of the Bayesian approach and the "epistemic turn" of metrology

A natural question to ask is how the quantitative results from the frequentist and the Bayesian calculations compare. A comparison is, however, difficult to make since the two approaches do not express their results in the same form. If one limits the comparison to the random component, quantitative differences have indeed been pointed out.[11] Disagreements have also been brought to the surface regarding the outcome of the measurement function;[12] but for this paper we see greater benefit in a focus on the philosophical side of the matter. The Bayesian approach involves a major change of perspective that buttresses in many ways the point of view arising from the debates surrounding the ongoing revision of the *VIM* and the *GUM*. It is therefore worthwhile to gather a more comprehensive view of the epistemological gulf that separates the Bayesian from the traditional treatment of measurement.

One can start by emphasizing the contrast between the domains on which probabilities are defined in the two cases. Within the frequentist treatment, this domain is constituted by the infinite set of virtual outcomes of a data generating process, whereas, within the Bayesian approach, probabilities are defined on a domain constituted by *propositions* stating the possible values the experimenter thinks he can attribute to the systematic error or to the measurand. In the former case, probabilities describe the working of an *objective, physical operation displaying variability*; they are features of the external world. In the latter, they describe a *subjective attitude towards the possible value* of a certain parameter of interest[13]; they are not directly related to a physical trait (variability) but rather to the knowledge the experimenter believes he has about the quantity he wishes to evaluate – they are epistemic probabilities expressing degrees of belief as to the different values that, in the light of his incomplete knowledge, the agent can ascribe to the parameter. In the Bayesian approach, the focus of enquiry has therefore undergone a major displacement: it does not reside in what is

11 Kacker and Jones (2003).

12 See, for instance, Willink (2010a).

13 This does not mean that the agent is entirely driven by psychological factors. As both of our anonymous reviewers have underlined, the agent is controlled by rationality constraints. Indeed, Bayesianism stipulates that, according to Dutch book arguments, an agent's degrees of belief have to comply with the axioms of probability.

out there – the quantity carried by a concrete thing –, but rather in the state of knowledge one entertains about what is out there. It resides within the sphere of representations.[14]
This radical change of perspective from the objective to the subjective realm goes along with a marked modification in the way in which the result gets expressed. Indeed, the aim is no longer to *determine* the value of the measurand but to consider the range of possible values that can be *attributed* to the measurand on the basis of available information.[15] The measurement result is then a probability distribution displaying the degrees of belief associated with the possible values that the agent can assign to the measurand, not a point estimate associated with an uncertainty like in the frequentist approach for random errors. Out of this probability distribution, a "credibility interval" can very straightforwardly be extracted in order to state the result in probabilistic terms. In contrast to the frequentist confidence interval, the credibility interval is now a statement of probability about the true value of the measurand. This statement does not say anything regarding the localization of the true value; it is the expression of a degree of belief involving a probability which is of an epistemic character.

The Bayesian answer to the *GUM* is not the only approach to support an epistemic interpretation of uncertainty. The discontent raised by the *GUM* has nurtured a maze of other, often less structured currents that endorse similar epistemic standpoints. This means that the overall situation can be broadly characterized as an "epistemic turn". In this context, a number of metrologists have initiated what the VIM describes as a switch from an "error" to an "uncertainty approach."[16] According to this trend, measurement error is an antiquated notion that should be superseded by the concept of quantifiable uncertainty. The objective conception of a deviation between a measurement datum, or a measurement result, and an attribute of the external world ($y = TV + e$) gives way to epistemic concerns pertaining to the state of knowledge associated with the assignment of a range of values to the measurand. This move culminates in confronting the classical way to conceive of the goal of measurement:

> The change in the treatment of measurement uncertainty from an Error Approach (sometimes called Traditional Approach or True Value Approach) to an Uncertainty Approach necessitated reconsideration of some of the related concepts appearing in the second edition of the *VIM*. The objective of measurement in the Error Approach is to determine an estimate of the true value that is as close as possible to that single true value. […] The objective of measurement in the Uncertainty Approach *is not to determine a true value as closely as possible*. Rather, it is assumed that the information from measurement *only permits assignment of an interval of reasonable values to the measurand*, based on the assumption that no mistakes have been made in performing the measurement.[17]

Measurement is, in that case, less regarded as a means to grasp features of an independent reality than as a mode of expressing the state of our knowledge concerning situations delimitated according to certain objectives. This resonates with Mari's characterization of the transition that the *GUM* set in motion: "*ontology and the criterion of truth have been replaced by information and a criterion of adequacy*"; and measurement becomes "an evaluation

[14] As one of the reviewers of this paper has observed, although Bayesian probabilities do not model directly "in the world" entities and phenomena, but propositions that are *about* these "in the world" entities and phenomena, the Bayesian can still argue that his object of enquiry is what the propositions are about, not the propositions themselves.

[15] On the contrast between the determination and the attribution of values, see Mari (1997).

[16] Joint Committee for Guides in Metrology (2012), p. viii.

[17] Joint Committee for Guides in Metrology (2012), pp. viii-ix. We underline.

whose results are subjectively adequate to given goals."[18] Measurement is no longer driven by the ideal of representing an independent reality but rather by an instrumental conception of scientific activity as oriented towards applications and concrete goals – a conception that has gained considerable momentum with the tremendous amplification of technological requirements in science and industry.

It should be noted that the epistemic turn tries to address a genuine concern: the true value is completely out of reach. Indeed, the classical account of measurement stumbles upon a major problem as soon as one contemplates evaluating the closeness of the estimate to the true value: to compare the estimate and the true value in order to assess the size of the deviation one should already know the true value that one is trying to determine! The task of assessing the quality of the result seems therefore to involve an unworkable circularity. To avoid such difficulties, the shibboleth of the metrologists involved in the uncertainty approach has been to keep within the bounds of empirically accessible entities. The uncertainty approach turns away from the estimation of an unknown, and forever unknowable, deviation from the true value, the approximation of which seems to be impossible to assess, and instead sets out to determine a known, perfectly accessible state of knowledge concerning the measurand.[19] But this amounts to disconnect uncertainty from the assessment of accuracy, defined, in the *VIM*, as the "closeness of agreement between a measured quantity value and a true quantity value of a measurand".[20] The demand that our results be correct gives way to the requirement that our results be the rational expression of our knowledge. And yet, there seems to be a serious incentive to hold on to the notion of measurement accuracy[21]. In line with this demand, we will suggest in the following that a closer look at the classical account of measurement suggests quite a different analysis than the ones introduced above. Paying attention to issues of correction and confidence, this analysis reinstates the notions of error and accuracy, and at the same time opposes the way of thinking adopted in the traditional understanding of the classical account as well as in its Bayesian criticism.

5. Taking the intentional character of measurement seriously: the process of correction and the pertinence of the concept of error

Let us begin by considering, for the sake of simplicity, the case of a direct measurement. Both the standard understanding and the criticism of the traditional conception of measurement assume that, in the expression (1) $y = TV + e$, linking an actual measurement datum y with the true value TV and error e, TV and e designate values that already exist and that can enter with y into an actual, straightforward, direct relation. But on closer inspection, such an assumption appears quite misleading – y, on the one hand, and TV and e, on the other, are different kinds of entities. If one refrains from hastily interpreting the expression as the representation of a state of affairs to pay regard to how it is actually *used*, the expression comes across as being of a *prescriptive* rather than of a descriptive kind. The distinction between the directive and the depictive view tends all too often to be blurred since "(e)very instruction can be construed as a description, every description as an instruction"[22]; the difference can only be made on pragmatic grounds by considering the use of the sentence. The descriptive view is here inappropriate in so far as one is involved in the task of *acquiring* knowledge about a quantity

[18] Mari (2003), p. 25. Underlined by the author.

[19] Bich (2012).

[20] Joint Committee for Guides in Metrology (2012), p. 21. Accuracy can have many different meanings; Tal (2011) identifies five of them. The definition of the VIM corresponds to the one he labels "metaphysical accuracy".

[21] As Quinn insists on many occasions. See, for instance, Quinn (2002), p. 13.

[22] Wittgenstein (1998), §14, p. 10.

value, not of expressing a prior knowledge. Indeed, the expression outlines an *instruction* to embark on a certain kind of *activity*. The assignment encompasses the application of a complex *process of correction*[23] to an actual measurement outcome *y*, resulting from a concrete measurement operation, in order to obtain the value of the measurand, which is, as already noted, the quantity *intended* to be measured. This process of correction is the hallmark of the profoundly *intentional* character of measurement which gets overlooked in the descriptive interpretation. It manifests that measurement does not amount to the report of a fact; measurement proceeds assuredly from the world of experience, but its results are *inferred* from an analysis that draws on an abstract construction specifying the aim of the measurement. Indeed, expression (1) cannot be a description; it can only make sense, and be used, on the background of the *posit* of the measurand: the relation between the entities featured in the expression relies on the introduction of a conceptual framework, a model. Let us examine these issues more closely.

Whereas the measurement operations are carried out on a real, concrete quantity realized in the laboratory, the measurand is a conceptual, ideal entity given in a definition. As already pointed out by Duhem, the experimenter handles physically concrete entities and instruments set in a particular environment, but reasons about his task by substituting for his real setup an abstract, ideal construction devised on the basis of a theoretical symbolism which lends itself to conceptual and mathematical manipulations[24]. The measurand, the aim of the measurement, is delineated in the abstract realm of this conceptual framework, or model. In the case of a direct measurement of length made with a ruler, the conceptual framework comprises representations and concepts pertaining to geometry (instrumental in picking out the length of the measured object; in our example, it includes in addition, thermometry and, eventually, mechanics.

The model establishes a link between the realized quantity subjected to measurement and the measurand. It indicates how one goes from the realized quantity at hand to the quantity one wants to measure, and, in particular, the corrections that must be introduced. In the general case, where the measurement is indirect, the link involves the measurement principle as well as the corrections that must be performed. The corrections show how to generate an estimate of the value of the measurand out of the "material" of the data by entering specifications (concerning temperature, position in the gravitational field etc.) that accomplish the *identification* of the quantity as the one that the agent wants to measure.

The correctness of the result is difficult to assess for several reasons. One should mention, to begin with, an intrinsic source of difficulty that is due to the very nature of the measurand. Indeed, because of its conceptual nature, and of the necessarily limited amount of detail provided by its definition, the measurand is essentially a *general*, incomplete entity. This is to be contrasted with the individual, singular character of the real, complete, perfectly determinate entities that can realize the description of the measurand. There ensues an inherent indeterminacy attached to the measurand as a target since there are, in principle, not one but many quantities that can stand as realizations of the measurand, depending on the detail given in its definition. If, in our example, the measurand is defined as the length of the end gauge at temperature T_0, then different positions of the end gauge in the gravitational field correspond to different possible realizations of the measurand with different length values (due to the compression of the end gauge under the effect of its own weight) that will be compatible with the definition. This non-uniqueness of the possible realizations of the

[23] One shouldn't forget that the data are also treated statistically in order to reduce the random errors. This constitutes a first use of equation (1). In the following, we focus on the second use of equation (1), namely the correction of systematic errors.

[24] Duhem (1981), pp. 217-48. For more up to date and exhaustive analyses of the model-based character of measurement, see Giordani & Mari (2012), and Tal (2014).

measurand, and therefore also of the quantity values corresponding to these realizations, is a perfectly straightforward counterpart of the conceptual nature of any entity that is merely given by description. The *VIM* and the *GUM* have captured this feature under the notion of "definitional uncertainty" defined, in the *VIM*, as the "component of measurement uncertainty resulting from the finite amount of detail in the definition of a measurand"[25]. This results in the non-uniqueness of the true value itself, which is indeed defined as a "quantity value consistent with the definition of the measurand"[26]. Only fundamental physical constants are deemed to have a single true value[27].

The ideal character of the measurand is another fundamental reason why it is difficult to assess the correctness of the result. The true values compatible with the definition of the measurand cannot be gained from experience; as already mentioned, they are "in principle and in practice unknowable"[28]. This does not mean that *TV*s could be dispensed with, as some metrologists who profess to consider only empirical entities capable of being objects of knowledge seem to believe. Indeed, one can think of *TV*s as "properly […] heuristic, and not […] ostensive, conception(s)" which stand as *regulative ideas* in the Kantian sense. Although not relating directly to an object, they are, as a matter of fact, "useful only for the purpose of representing other objects to the mind, in a mediate and indirect manner, by means of their [these objects'] relation to the idea in the intellect" [29] – like in $y = TV + e$, where *TV* serves to get hold of an estimate of the measurand through the process of correction of *y*. They function not as objects of knowledge but as methods of scientific investigation that are instrumental in attaining certain ends in the theoretical as well as the practical domain: if they do "not give us any information respecting the constitution of an object, [they indicate] how, under the guidance of the idea, we ought to investigate the constitution and the relations of objects in the world of experience." [30] They are postulates that are indispensable to orient the activity of the agent since, in line with the normative character of measurement, they provide the process of correction with a horizon that guides the operations of the agent towards a goal. The corrections then identify the data as referring to a specific target. Under those circumstances, it can make sense to compare the resulting estimate to other, similarly corrected results, ensuing from experiments that purport to measure the same measurand in different conditions, since the corrections attribute the same meaning to all results.

Finally, as already mentioned in section 2, the target is bound to remain imperfectly known because of the partial nature of the corrections that take into account, beyond the model of the measurand, the model of the whole measurement process. These further corrections disentangle the data from the concrete, singular circumstances in which the measurement was performed so as to make the corrected result transportable, and usable elsewhere, in other contexts, by different agents. Such corrections handle the biases due to the instruments, the measurement procedure, the environment and the remaining influence factors, as well as, of course, the random errors. Of particular importance are the corrections pertaining to the measurement standards and, more generally, to the calibrated devices involved in obtaining the values of physical quantities. Indeed, these corrections relate the quantity values obtained

[25] Joint Committee for Guides in Metrology (2012), p. 25. Similar remarks are made by Teller (2013), although they are not couched in terms of definitional uncertainty. These remarks lead him to give up what he calls "measurement accuracy realism".

[26] *Ibid.*, p. 20.

[27] *Ibid.*

[28] *Ibid.*

[29] Kant (2015), Of the Ultimate End of the Natural Dialectic of Human Reason.

[30] *Ibid.*

to primary references and, ultimately, to the SI, through a network of institutions, thus making it possible to meaningfully compare the estimates with other results.
In the end, a measurement result appears to be the outcome of a complex set of inferences[31] epitomizing the intentional and normative character of measurement. The process of inference involves an ideal going beyond the empirical realm. For sure, neither the ideal nor error can be seized directly. But this doesn't mean that these concepts are inappropriate. The remaining error affecting the measurement result, the distance between an estimate and the ideal, is grasped *indirectly* through the inter-comparison of the *relative distances between different estimates* of the same measurand obtained in different experiments and conditions. And the inter-comparison itself is made possible, meaningful, by the relation each of the estimates has to the same ideal – one can indeed evaluate the distance between the various estimates since they all aim at the same *TV* and are all traceable to common references (the SI). This inter-comparison initiates a new phase of the process of correction that improves, through a non-vicious circle, the knowledge of the *TV*, the mere posit of which activated the first stage of the process of correction. This new phase of the process of correction is decidedly collective. It is carried out by agents belonging to different laboratories who examine the discrepancies between their respective estimates obtained while attempting to gain the same measurand under different descriptions (when different measuring principles are implemented). Through such a comparison, other, yet unknown systematic errors are tracked in order to make further, new and unforeseen corrections that go beyond the corrections already performed on the basis of known laws and models so as to acquire new knowledge. In this way, the *TV*, which had the status of a presupposition setting the framework of the inter-comparisons, becomes the horizon of an ongoing activity of interactive criticism and mutual corrections. The outlook has thus not only ceased to be descriptive but also individualistic. One should note, in addition, that the comparison of the different estimates is made possible by the quantitative account of the uncertainty that remains attached to the results because of the imperfection of the above mentioned corrections. Indeed, the uncertainty provides each estimate with a quantified margin of doubt that makes it possible to go beyond the simple record of the gulf that actually exists between estimates and make guarded judgements of sameness or discrepancy according to the way margins overlap or not (these judgements involve a "test of hypothesis"). The discrepancies between estimates aiming at the same target instigate the search for further systematic errors.
However, it still can happen that all estimates of a measurand agree within the margins of uncertainty, in other words, that the results appear reproducible, and that they are all affected by a common bias. The problem of accuracy remains pending and raises the issue of the confidence one can have in one's measurement results. We will now turn to this matter by coming back to the statistical methods of uncertainty evaluation.

6. Moving away from an individualistic account of measurement: the issues of confidence and accuracy

Within the epistemic approach, the measurement result takes the form of a probability distribution displaying the possible values of the measurand along with their degree of belief. As already mentioned, one can straightforwardly obtain out of this probability distribution a credibility interval attached to a probability (say, 0.95) that states to what extent the agent believes that the true value is contained in the interval. The construction of such a credibility interval seeks to make the best use of the information available to the agent in the present, and

[31] That measurement results are the product of an inference has already been pointed out by Bogen and Woodward (1988). However they mainly insist on the correction of random errors and the concrete corrections applied up-stream directly on the experimental set-up. They considerably downplay the corrections relying on models that are emphasized here as well as in Mayo (1996), pp. 128-73, and Tal (2014).

yields the agent's state of belief concerning a state of affairs in the world; but it doesn't involve any relation to other agents interested in the result. Now, confidence implies, over and beyond credibility (belief concerning a state of the world), a relation to other agents. And confidence is crucial in the domain of measurement where epistemic dependence is most pervasive: no agent can perform all the experiments that his own projects demand; agents must rely on the knowledge, the models, the expertise and the measurement results of other agents in order to go about their own measurements, take decisions, produce goods or extend their knowledge. As a matter of fact, a Bayesian credibility interval is not accountable to others since it states the belief of an individual agent[32]; it is not testable and is neither true nor false. Others can accept it only in the form of a testimony.

By contrast, a result given in terms of a confidence interval, in line with the frequentist approach, is accountable to other users because it is subjected to a criterion of performance[33]. Let us leave aside, for the moment, the crucial fact that the frequentist approach cannot deal properly with systematic effects. For sure, a given calculated confidence interval contains the true value of the measurand or not, although there is no way of knowing which is the case. However, that the confidence level of the interval is (say) 95% – or saying that there is a probability of 0.95 that the interval encloses the value of the measurand – means that when one calculates many intervals each corresponding to different samples of the data, the limit of the frequency with which the intervals thus generated will contain the true value is 0.95: on average 95 out of a 100 intervals calculated will contain the true value of the measurand[34]. The confidence level of 0.95 associated with the interval does not correspond to the probability that the interval contains the true value; it is attached to the rate of success of a technique. In contrast to the epistemic account, correctness is not stated in terms of the belief of an agent regarding a given interval, but rather to the "procedure generating it with regard to the stated probability"[35]. In this respect, the quality of a result given in the form of a confidence interval resides in the long-run performance of a process of correction (limited, here, to random effects). This provides users with *grounds* for accepting such a result and assuming responsibility for their decision: their confidence is justified on the basis of objective probabilities.

An important consequence is that, in this limited case, the assessment of the quality of a measurement result does not have to do with the impossible appreciation of the closeness of the estimate to the true value but, rather, with the reliability of the process of correction and of the building of a confidence interval[36]: one should think about accuracy in dynamical, not descriptive terms. However, only one interval, containing or not the true value, is calculated and handed out by an agent or a laboratory to other users. Reliability is therefore assessed not on the basis of long-run performances *actually* carried out (in the past, present or future) but of experiments that *could be* carried out. In this respect, the frequentist outlook goes beyond the exclusive concern with the actual epitomized by the epistemic approach. This implies embracing the point of view of a community of enquirers that could implement these would-be experiments rather than the point of view of individuals. The confidence interval of the

---

32 It can also be the belief of a group. In that case, the epistemic approach calculates an aggregate probability distribution.

33 We follow here Willink who strongly insists on this issue. See for instance Willink (2006).

34 This points to a tricky issue that we cannot discuss here. Strictly speaking, it is not possible to calculate a rate of success since, again, one cannot know if the target is contained in the intervals or not. However, according to Willink (2010b), p. 82, one can control the reliability of the procedure by testing it on known reference targets and calculating a theoretical rate of success.

35 Willink & Lira (2005), pp. 64-65.

36 Willink, R. (2010b), p. 82.

frequentist approach displays therefore features that are in tune with a dynamical as well as a social perspective.
It remains that, as we saw, the frequentist approach cannot provide a probabilistic account of the systematic component of error. The calculation of a frequentist confidence interval cannot accommodate a full analysis of measurement results. But, as we will see, the classical frequentist approach can be extended so as to enable the calculation of a "practical confidence interval" that fixes this limitation.

7. An extension of the classical frequentist approach: from an epistemic to a social turn?

An extension of the classical frequentist approach has recently been put forward by Willink to provide a probabilistic treatment of systematic errors[37]. Although this extension has, as yet, no practical ramifications, we will present it here for its critical and philosophical weight. The probabilistic treatment advanced runs along quite different lines than the Bayesian proposal since it leads to a randomization of systematic errors, in tune with the frequentist point of view. As we will try to show, this scheme puts the emphasis on the social rather than the epistemic dimension of measurement activities.
The extension rests on the primary idea that the process of measurement should be conceived in a more comprehensive way. It should include, beside the measurement experiment itself, all the background experiments previously carried out in the laboratories and industrial settings that provided the results required to realize the measurement and, in particular, those that generated systematic components of error. In our example, the process would then include the calibration of the standard end gauge and the experiment furnishing the coefficient of thermal expansion. An additional extension consists in considering that, *from the point of view of a user* performing a measurement experiment, who can choose to pick the results involving the systematic errors he needs in his experiment from a wide range of laboratories using different measurement methods and instruments, these diverse laboratories realize a kind of (unorthodox) repetition of the background experiment yielding a given type of systematic error. In view of all the laboratories from which the user can choose, systematic errors can be treated probabilistically: a given systematic error appears like a realization drawn from a population of similar systematic errors of the same kind produced by all the laboratories that were accessible to the user and that were handling the same type of problem[38]. Willink contends that random and systematic components of uncertainty can then be combined, and he shows how one can calculate a practical confidence interval that is relevant for the user[39].
However, this line of action is possible only if certain conditions are met: the user must be able to choose among a sufficiently large variety of laboratories dealing with the same kind of problem but working with different procedures and sets of apparatus. Hence, a unified probabilistic treatment of all components of error on an extended frequentist basis does not only draw on theoretical and experimental circumstances; we contend that it also involves the social and institutional environment that makes this variety of laboratories available. For one thing, the extension of the measurement process to include background experiments amounts to the explicit recognition that every measurement is bound to involve measurements made by others and thus openly manifests the collective nature of measurement activities. Issues of

[37] Willink (2013), chapters 4 & 5.

[38] As one of our reviewers points out, this supposes that it is possible to formulate a sound parent distribution. Willink notes indeed that "(t)he choice of distribution is not arbitrary. The distribution chosen should either reflect the parent population of systematic errors adequately or lead to a conservative procedure through overestimation of the variance." (Willink, 2013, p.70) However, Willink bypasses the problem by stating that in his book he will "work on the basis that an adequate distribution can in fact be proposed" (*ibid*, p.70).

[39] Willink (2013), chapters 4 & 5.

division of labour are also concerned since the distinction between the point of view (or universe of reference) of the user and that of the producer of a result (the laboratories supplying the result) is a major element in the analysis. The extension proposed can indeed be seen as a broadening of the frequentist outlook: in the classical frequentist approach, probability relates to potential outcomes in repeated experiments and hangs on the characteristics and the physical performances of instruments; in the extended view, they also depend on the material, institutional and social infrastructures that underpin these physical processes and procedures.

In this respect, it seems natural that the quality of a measurement result should not only reflect the long-term behaviour of the physical processes directly involved, but also the wider social organization that produces and maintains them and is responsible for their development in a given scientific and technical society. As a matter of fact, within this extension, the calculation of a confidence interval depends on the existence of certain social conditions ensuring that the users have access to a panel of adequately diversified laboratories applying different methods and devices. The agents engaged in measurement activities should therefore attend to these wider circumstances in a similar manner to the physical conditions pertaining to the conduct of measurement performances. The extended frequentist approach incorporates thus, at the very level of scientific conceptualization, issues explored by the different trends of social epistemology; in particular, the question of how scientific institutions should be designed in order to promote the development of knowledge and of scientific and industrial practices.

Explicitly acknowledging the social underpinnings of measurement activities thus supplies an alternative way to obtain a probabilistic treatment of systematic errors and to achieve a unified account leading up to the calculation of a confidence interval. The social and institutional conditions that must be met to calculate a confidence interval also prove to be instrumental in uncovering unknown biases both *physically* and *epistemologically*. Physically, because discrepancies between estimates obtained by different laboratories, and therefore possible systematic errors, become apparent primarily when the measurement setups involved are different from one another and explore a variety of measurement principles – the discovery of the Josephson effect and its use as a new measurement principle thus revealed the existence of significant biases in the previous evaluation of certain fundamental constants (such as e/h and α[40]). Epistemologically, because, as already mentioned, the search for systematic errors involves a community of enquirers organized in such a way that it can methodically compare its measurement results and subject them to an interactive kind of criticism.

If the different estimates obtained by different enquirers are indeed attained by distinct means, and are in agreement, they aim at the true value of the measurand under different descriptions and one could hope that the true value might thus be found at the intersection of these different perspectives. However, such a convergence doesn't prove anything since all the estimates and intervals obtained by the different enquirers (at a certain time) could be plagued by a common unknown bias. Such a bias wouldn't in the least prevent agreement – the agents would consider that their results are reproducible. The judgment that there is a bias is a possibility that will only become apparent in *future* evaluations. It can be made only if the overall organization of measurement activities in society is propitious to the realization of new physical situations, new setups liable to hit on unforeseen discrepancies.

This invites a rethink of what constitutes accuracy, not anchored in the actual – as when one professes to present to one's mind the deviation of an estimate from a true value – but rather as attached to potential features pertaining to society at large, features that encompass

[40] See Kose & Wöger (1986).

economic, political, epistemological, and moral virtues concurring in the capacity to embark on new, risky, demanding, and costly investigations prone to give access to original modes of evaluation. The assessment of the quality of a result does not rely on the evaluation of a deviation but on the long-run behaviour of a process of correction including, now, both random and systematic components. This requires that the technical as well as the social and institutional framework of the community of enquirers should be so devised as to promote, in due course, the removal of biases. The confidence bestowed on a given interval now explicitly takes into consideration, in an enlightened and responsible manner, the society in which measurements are performed.

Such a dynamical and social understanding avoids the difficulties of a static and individualistic account of accuracy and is thus able to withstand the dismissal of the concept of accuracy in the epistemic standpoint depicted in the so-called "uncertainty approach". Moreover, this dynamical understanding stands in marked contrast with the Bayesian approach in which beliefs are grounded in present, available information and in which no clue is given as to why revisions should be made. In particular, the Bayesian approach gives no indication about the conditions that must be met in the environment in order to guarantee the rationality of the process of revision – one should really examine the circumstances that can induce the experts to look for new information to update their results. This shows that the Bayesian approach still builds on an unquestioned trust in the providential character of the "natural" background in which its decisions are rooted. It does not adopt an attitude of "reflexive scientification"[41] going beyond the classical view of science (centred on the investigation of phenomena) to assume responsibility for the impact of scientific activity on the workings of society at large. The user who accepts information associated with a probability interpreted as a degree of belief accepts a testimony and endorses an unexamined epistemic dependence that makes him share the possible bias of the result as a passive member of the community which has abdicated his power of inspection.

8. A case study: the measurement of the "proton radius"

How do these remarks fare with the actual practice of researchers[42]? The discussions raised by recent results bearing on the proton distribution of charge radius, the "proton radius" for short, provide an instructive case study[43]. Until recently, the proton radius was thought to be fairly well determined by two different kinds of experiments involving hydrogen – precision spectroscopy experiments investigating the energy levels of hydrogen and experiments of electron scattering through a hydrogen gas. The uncertainties associated with the results of these two types of experiments show that these results are indeed compatible with each other. However, experiments realized in 2010 and in 2013 that explored the shift of energy levels (Lamb shift) in muonic hydrogen[44], which depends on the size of the proton, produced results that are not in agreement with the previous ones: the proton radius extracted from muonic hydrogen experiments differ by five standard deviations from the value of the radius calculated on the basis of the measurements involving hydrogen. No explanation for this difference in terms of systematic errors has been found as yet; the idea that the inconsistency could point to a new physics has been advanced and the problem is now addressed as the "proton radius puzzle".

---

41 Beck (1986).

42 We will leave aside here the practice of engineers and commercial manufacturers to which a special study should be devoted.

43 See, for instance, Bernauer & Pohl (2014).

44 In muonic hydrogen atoms, the electron of the hydrogen atom is replaced by a muon.

The problem posed by the value of the proton radius is of great concern to the Task Group on Fundamental Constants (TGFC) commissioned by the Committee on Data for Science and Technology (CODATA), established respectively in 1969 and 1966. Every four years, the latter international committee publishes the recommended values of the fundamental physical constants to be used by scientific and industrial laboratories all over the world. The international task group meets twice a year in order to analyse and discuss the experimental values of diverse quantities submitted by different top metrological laboratories around the world. These experimental values (hydrogen transition frequencies, various relative atomic masses, the electron anomalous magnetic moment, etc.) are related to the fundamental constants through theoretical expressions, and the values of the fundamental constants are obtained by a long-winded process of least-squares adjustments based on these expressions.[45] The value of the proton radius is of particular importance because it impinges on important fundamental constants such as the Rydberg constant R_∞ and the fine structure constant α.

A whole day was devoted to the discussion of the value of the proton radius at one of the last meetings of the task group in November 2014[46]. The task group was facing two options: include the muonic hydrogen results in the adjustment and, consequently, expand the uncertainty of the final result, or exclude them and limit the adjustment to the experiments involving hydrogen. Certain members put forward that since the muonic hydrogen results were taken to be valid – nothing could be found wrong with them and they had moreover a much lower uncertainty than the hydrogen results –, there was no reason to dismiss them. Someone argued bluntly that one did not seek to formulate the true values, one sought to express the status of our knowledge, and that therefore the muonic result should be included. But these Bayesian flavoured views were not endorsed by the majority of members who sided with the opinion that it made no sense to average the results since one of them was surely "plain wrong." It was contended that the discrepancy was "fruitful and had to be exhibited" instead of being covered up by averaging: one had to "wait for new results […] to resolve the discrepancy." In their final decision, the physicists excluded the muonic hydrogen results from the pool in order to "keep the puzzle as a puzzle" to be re-visited in the future[47]. By so doing they pressed for the search of corrections, either in the form of systematic errors, or in the form of a new theory (and therefore of a new way to model the entities involved).

The scientific community does not seem especially decided to embrace an epistemic stand. In confronting the problem of the proton radius, their solution is not to express their state of knowledge by combining the discrepant results. The choice of the TGFC is rather to declare and confront a state of inconsistency. In so doing, it demonstrates the relevance of the concept of error and stimulates an activity of correction and research, either of systematic errors, or of new theories. Moreover, the choice implies the posit of an unknown target: the results with their associated uncertainties are not interpreted as possible values with certain degrees of belief but are referred to a true value which gives a direction to evaluate the distance between the different results, and with respect to which discriminations can be made; one can think in terms of right and wrong. The judgment is of course fallible; it can be proved incorrect. And, indeed, the CODATA has revised, quite substantially, the values of certain fundamental constants on several occasions. This continued task of correction bears witness to the long-term character of the endeavour of the CODATA and the task group whose criticism of data and experiments is a social, interactive and institutionalized process.

[45] The latest available list of recommended values is published in Mohr, Taylor, & Newell (2012). The results of the 2014 adjustment should appear in the course of next year.

[46] We thank D. B. Newell and F. Nez for inviting us to attend the meeting of the task group.

[47] It was also suggested that such a decision had a practical import for the community of spectroscopists since it retains continuity of the value of the Rydberg constant.

9. Conclusion

As the discussion of the extended frequentist account has already made clear, one should acknowledge the existence of different groups of users with different objectives. The goals of the CODATA cannot be the same as those of average scientific or industrial users who need to have access to results *now* in order to go about their current affairs. There is no universal method: the *GUM* cannot pretend to provide recommendations secured on universal foundations for all groups of users. The Bayesian approach has undoubtedly its rationale, but it is a pragmatic one. It proposes a unified probabilistic interpretation of all components of uncertainty and fills the practical request to combine them in one term. It certainly gives expedient directives to deal with uncertainty if one considers a restricted group of agents – in the industry for instance – who should perhaps defer to experts and endorse their epistemic dependence. But the question of what justifies their confidence in these experts remains intact.

We have seen that, although technically appealing, the epistemic account of measurement uncertainties is not as philosophically compelling as many of its supporters contend. The epistemic criticism of the traditional account of measurement remains confined to a static and individualistic outlook[48] whereas our discussion suggests that the analysis should take another direction, acknowledging that measurement is an *activity* involving a *community* of agents and users. One should not mistake options ratified and put to use because they are technically tractable for outcomes of conceptual and epistemological reflection – although usage can indeed entail the entrenchment of conceptual and epistemological views.[49]

In contrast, the extended frequentist approach has scant practical import; but we hope to have shown that its philosophical and critical implications are noteworthy. It reveals that the quality of measurement results should incorporate, beyond the long-term behaviour of physical processes and instruments involved in measurement operations, the social and institutional background which produces and maintains them, and ensures their development. It shows the pertinence of an extended concept of confidence that encompasses the epistemic and organizational properties of the social infrastructure of research and measurement activities. Attending to institutions and background social issues is an intrinsic aspect of scientific rationality: it governs the ability to bring to light hidden biases.

Moreover, we have seen that the accuracy of measurement results is not a feature that can be assessed in actuality. One cannot know if a result is at present free from bias. Accuracy has to do with extensive, virtual comparisons. An accurate result is a result that will turn out to be stable, "constant in time", if the conditions of its production are varied in all possible ways. It is a *capacity* of the result to remain the same, within the limits of uncertainty, in the face of the changing conditions of its production. However, such a capacity can only be effective if certain external conditions prevail that depend on the objectives, values, norms, motivations embedded in the institutional and social framework in which the measurement activity takes place. Accuracy hangs on the power to correct results; it is tied to the correction process understood as an interactive criticism of data and experiments involving the activity of a community. Such a conception of accuracy is in deep consonance with the understanding of reality and truth advanced by C. S. Peirce:

---

[48] For sure, Bayesians introduce combinations of probability distributions in order to represent the degree of belief of groups. But the individual outlook remains primary since the probability distributions combined are those of individuals.

[49] See Humphreys (2004), p. 55 on the importance of tractable methods on the development of science.

> (I)f truth consists in satisfaction, it cannot be any actual satisfaction but must be the satisfaction that would ultimately be found if the enquiry were pushed to its ultimate and indefeasible issue.[50]

> (W)hat do we mean by real ? It is a conception which we must first have had when we discovered that there was a unreal, an illusion; that is, when we first corrected oneself. […] (T)he very origin of the conception of reality shows that this conception essentially involves the notion of community, without definite limits, and capable of an indefinite increase in knowledge.[51]

Acknowledgements
We wish to thank the participants of the conference "The Making of Measurement", held in Cambridge, on the 23-24 July 2015, where part of this paper was first presented, for helpful discussions, and especially Hasok Chang, Luca Mari, Daniel J. Mitchell, Terry Quinn, and Eran Tal. We are grateful to the two anonymous reviewers whose insightful comments have helped us to clarify many issues. We also thank our language editor Claire Neesham.

References

Beck, U. (1986). *Risk Society. Towards a New Modernity* (M. Ritter, Trans.). Los Angeles: Sage Publications Ltd. (First published 1992).

Bernauer, J. C. & Pohl, R. (2014). The Proton Radius Puzzle, *Scientific American*, 310(2), 32-39.

Bich, W. (2012). From Errors to Probability Density Functions. Evolution of the Concept of Measurement Uncertainty, *IEEE Transactions on Instruments and Measurement*, 61(8), 2153-59.

Bogen, J. & Woodward, J. (1988). Saving the Phenomena, *The Philosophical Review*, 97(3), 303-52.

Boumans, M. & Hon, G. (2014). Introduction. In M. Boumans, G. Hon, & A. C. Petersen (Eds.), *Error and Uncertainty in Scientific Practice* (pp. 1-12). London: Pickering & Chatto Publishers.

D'Agostini, G. (1996). A Theory of Measurement Uncertainty Based on Conditional Probability, *ArXiv:physics/9611016*, physics e-print. (Available at: arxiv.org/abs/physics/9611016)

Duhem, P. (1981). *La Théorie physique. Son objet – Sa structure* (2nd ed.). Paris: Vrin. (First published 1906).

[50] Peirce (1966b), p. 378.
[51] Peirce (1966a), p. 69.

Eisenhart, C. (1963). Realistic Evaluation of the Precision and Accuracy of Instruments Calibration Systems, *Journal of Research of the National Bureau of Standards – C. Engineering and Instrumentation*, 67C, 161-187.

Giordani, A. & L. Mari (2011). Quantity and Quantity Value, *Proc. TC1+TC7+TC13 14th IMEKO Joint International Symposium*. August 31st-September 2nd, Jena, Germany. Accessible at: http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-24414/ilm1-2011imeko-025.pdf

Giordani, A. & Mari, L. (2012). Measurement, Models, and Uncertainty, *IEEE Transactions on Instrumentation and Measurement*, 61(8), 2144-2152.

Gleser, L. J. (1998). Assessing Uncertainty in Measurement, *Statistical Science*, 13(3), 277-90.

Humphreys, P. (2004). *Extending Ourselves. Computational Science, Empiricism, and Scientific Method*. Oxford: Oxford University Press.

Joint Committee for Guides in Metrology (2008). *Evaluation for measurement data – Guide to the expression of uncertainty in measurement*. (Available at http://www.bipm.org/en/publications/guides/)

Joint Committee for Guides in Metrology (2009). *Evaluation of Measurement Data. An Introduction to the "Guide to the Expression of Uncertainty in Measurement" and Related Documents*. (Available at http://www.bipm.org/en/publications/guides/)

Joint Committee for Guides in Metrology (2012). *International Vocabulary of metrology – Basic and general concepts and associated terms (VIM)*, 3rd edition. (Available at http://www.bipm.org/en/publications/guides/)

Kacker, R. and Jones, A. (2003). On use of Bayesian statisitcs to make the *Guide to the Expression of Uncertainty in Measurement* consistent, *Metrologia*, 40(3), 235-48.

Kant, I. (2015). *The Critique of Pure Reason* (J. M. D. Meiklejohn Trans.) (2nd ed). (First published 1781/1787). eBooks@Adelaide (available at: https://ebooks.adelaide.edu.au/k/kant/immanuel/k16p/index.html). (Accessed 21 december 2015)

Kose, V. and W. Wöger, W. (1986). Fundamental constants and the Units of Physics, *Metrologia*, 22, 177-185.

Kyburg, H. (1992). Measuring Errors of Measurement. In C. W. Savage, & Ph. Ehrlich (Eds.), *Philosophical and Foundational Issues in Measurement Theory* (75-91). Hillsdale: Lawrence Erlbaum Associates Publishers.

Lira, I. and Wöger, W. (2001). Bayesian evaluation of the standard uncertainty and converage probability in a simple measurement model, *Measurement Science and Technology*, 12(8), 1172-79.

Mari, L. (1997). The role of determination and assignment in measurement, *Measurement*, 21(3), 79-90.

Mari, L. (2003). Epistemology of measurement, *Measurement*, 34, 17-30.

Mayo, D. (1996). *Error and the Growth of Knowledge*. Chicago: The University of Chicago Press.

Mohr, P. J., Taylor, B. N., & Newell, D. B. (2012). CODATA recommended values of the fundamental physical constants 2010, *Review of Modern Physics*, 84, 1527-1605.

Peirce, C. S. (1966a). Some Consequences of Four Incapacities. In P. P. Philip (Ed.) *C. S. Peirce: Selected Writings. (Values in a Universe of Chance)* (pp. 39-72) New York: Dover Publications. (First published 1868).

Peirce, C. S. (1966b). A Neglected Argument for the Reality of God. In P. P. Philip (Ed.) *C. S. Peirce: Selected Writings. (Values in a Universe of Chance)* (pp. 358-79) New York: Dover Publications. (First published 1908).

Quinn, T. (2002). Metrology, its Role in Today's World. In I. Lira, *Evaluating the Measurement Uncertainty. Fundamental and Practical Guidance*. Bristol: Institute of Physics Publishing.

Tal, E. (2011). How Accurate is the Standard Second? *Philosophy of Science*, 78(5), 1082-96.

Tal, E. (2014). Making Time: A Study in the Epistemology of Measurement, *The British Journal for the Philosophy of Science*. (Available at doi:10.1093/bjps/axu037)

Teller, P. (2013). Measurement Accuracy Realism, paper presented at Foundations of Physics 2013: the seventeenth UK and European Meeting on the Foundation of Physics. Available online: http://philsci-archive.pitt.edu/9740.

Weise, K. and Wöger, W. (1993). A Bayesian theory of measurement uncertainty, *Measurement Science and Technology*, 4(1), 1-11.

Willink, R. (2006). Principles of Probability and Statistics for Metrology, *Metrologia*, 43, 211-19.

Willink, R. (2010a). Difficulties Arising from the Representation of the Measurand by a Probability Distribution, *Measurement Science and Technology*, 21(1), 1-11.

Willink, R. (2010b). On the Validity of Methods of Uncertainty Evaluation, *Metrologia*, 47(1), 80-89.

Willink, R. (2013). *Measurement Uncertainty and Probability*, Cambridge, Cambridge University Press.

Willink, R. & Lira, I. (2005). A United Interpretation of Different Uncertainty Intervals, *Measurement*, 38, 61-66.

Wittgenstein, L. (1998). *Philosophical Remarks* (R. Hargreaves & R. White, Trans.). Oxford: Basic Blackwell. (First published 1964).