Understanding the foundations of measurement: Why a clock that ticks randomly is the best clock

Phil Maguire,[1a)] Philippe Moser,[1] and Rebecca Maguire[2]
[1]*Department of Computer Science, National University of Ireland, Maynooth, Co. Kildare, Ireland*
[2]*School of Business, National College of Ireland, IFSC, Dublin 1, Ireland*

(Received 1 April 2016; accepted 28 October 2016; published online 22 November 2016)

Abstract: In this article, we examine how a competition to find the world's most accurate clock might be run. How could the winning clock be identified if it outperforms every existing standard for timing? The intuitive view on time-keeping is that a good clock is one that keeps time consistently and hence agrees with other clocks. This view, we argue, is mistaken. Measurement is fundamentally about making high-quality predictions. Accordingly, the goal is not consistency, but independence between the clock and its environment. We propose that, counter-intuitively, the best clock is the one that ticks most unpredictably, making its predictions the most difficult to beat. The organizers of the clock competition should award the prize to the clock that ticks most randomly. © *2016 Physics Essays Publication*. [http://dx.doi.org/10.4006/0836-1398-29.4.574]

Résumé: Dans cet article, nous réfléchissons à la manière d'organiser un concours pour couronner l'horloge la plus précise du monde. Comment peut-on identifier l'horloge gagnante si elle surpasse tous les standards de mesure en vigueur? La définition habituelle est qu'une bonne horloge mesure le temps de manière cohérente, et par conséquent en accord avec d'autres horloges. Nous argumentons pourquoi nous pensons que ce point de vue est erroné. La mesure du temps consiste fondamentalement à faire des prévisions de haute qualité. Par conséquent, le but n'est pas la cohérence en soi, mais l'indépendance entre l'horloge et l'environnement dans lequel elle opère. Nous proposons que la meilleure horloge soit celle dont le tic-tac est le plus imprévisible, ce qui rend ses prédictions les plus difficiles à battre. Par conséquent, les organisateurs du concours de précision devraient attribuer leur prix à celui qui peut construire l'horloge qui fait tic-tac de la manière la plus aléatoire possible.

Key words: Metrology; Stability; Measurement Standards; Randomness; Prediction; Time.

I. INTRODUCTION

Why make measurements? Intuitively, people make measurements because measurements are useful. But what is it about measurements that make them useful? How could their value be justified to, say, an uncontacted tribe that has never thought of using measurements before?

When I ask a question such as "how long is this desk?" I can already see the desk in front of me, so its length is apparent. For example, I can stretch my arms from one end of the desk to the other and even *feel* how long it is. Yet this is not particularly useful. What I want to know in asking this question is how this length compares to the length of other objects: measurement is not about evaluating a single object in isolation, but about comparing things together. Wittgenstein[1] makes a pertinent observation, highlighting the meaninglessness of measurement without comparison:

"Imagine someone saying: 'But I know how tall I am!' and laying his hand on top of his head to prove it."

So, what is the value of relating things together? Tal[2] examines this question and suggests that at the heart of measurement-related activities lies the goal of prediction. When I measure the length of my desk I am effectively making a prediction about what will happen when it interacts with other measured objects (e.g., will my desk fit through that door?) Even if we imagine cases where measurement is carried out for its own sake, without any explicit expectations for prediction, the concept of reliable relationships still applies. For example, somebody who measures how fast they run around a race track expects those timings to enable comparisons involving other runners, suggesting who would win a hypothetical race between them. In sum, measurement readings can be regarded as predictions about the readings that would be elicited in different measurement contexts. Measuring my desk at 1.17 meters on Monday allows me to predict that, were I to measure my desk on Tuesday, Wednesday or Thursday I would obtain a reading[b)] of 1.17 meters (though, as we will see, this idealized assumption does not always hold).

[a)]pmaguire@cs.nuim.ie

[b)]When we refer to a measurement "reading" in this article we are referring to the final outcome of the process of measurement. For example, metrologists can often enhance the accuracy of indications taken from a particular measurement instrument through theoretical corrections, which "analyze away" some of the bias. We are referring to the end result of this multi-stage process, to the value which represents the best attempt at discerning an unbiased measurement.

In order for measurements to support accurate predictions, they must describe properties that are independent of other environmental variables. This independence allows a simplified abstracted model of the property to be developed, without needing to worry about the context in which the measurements were taken. This abstraction provides a very useful tool for negotiating the world. For example, I can measure my desk and use that measurement to predict whether the desk will fit through a given doorway, without needing to worry about what day of the week I am going to move it. If I measure my desk on Monday, and my office door on Tuesday, I expect to be able to relate the two measurements directly, despite the fact they were made on separate days: a reliable measurement of length should be independent of the day on which it was made.

If objects did not have certain properties which stood out as being independent of the environment, then measurement would not be possible. No generalization or abstract modeling would be possible, thus rendering prediction impossible: the world would appear as an inseparable chaos of complexity. Luckily for us, properties such as distance, time, and mass appear to stand out as being independent of the context in which they are measured, allowing predictions to be made about how objects will interact along these dimensions. The goal of measurement becomes that of identifying standards which achieve the greatest possible level of independence between measurement process and measurement context, a goal which occupies practitioners of the discipline known as "metrology."

II. METROLOGY

Metrology is the science of measurement and standardization, carried out by metrologists, who are experts in highly accurate and precise measurement. Professional metrologists are tasked with the job of maintaining, disseminating and refining high-quality standards. Under the guidance of the *Bureau International de Poids et Mesures* (BIPM), a worldwide network of metrological institutions is responsible for constantly comparing and adjusting standards to maximize a property known as "stability."[3]

Stability refers to the tendency of an apparatus to produce the "same" measurement outcome over repeated runs, as well as replicating the outcomes of similar instruments around the globe. Ideally, measurement readings should not be associated in any way with the location or moment in which they are taken. Standardization can be regarded as a process for ensuring independent agreement: despite being displaced in space and time, and having no causal interaction with each other, metrological laboratories can produce results which agree with each other. Under the guidance of the BIPM, a worldwide network of metrological institutions is responsible for comparing, adjusting, maintaining, disseminating, and refining these stable standards.[3]

One of the notable successes of these institutions is the standard measure of time used in almost every scientific context, known as Coordinated Universal Time (UTC).[3] UTC is regarded as overwhelmingly stable insofar as a variety of standardization labs around the world manage to closely reproduce it on an ongoing basis. In practice, what this means is that they are able to make highly accurate predictions about how independent dispersed clocks will behave in different circumstances across the globe each day.

A. Measurement error

The stability of UTC is a reflection of its very low predictive error, otherwise known as "measurement error." Metrologists identify two types of measurement error, namely, statistical error and systematic error.

Statistical error is the type of predictive uncertainty which exists between a single measurement and the average of a larger group of measurements. In other words, it is the type of uncertainty which can be reduced by taking many measurements and averaging them, rather than relying on a single one.

The foot, for instance, is an ancient unit of length based on the human body. As we know from experience, human feet vary in length: one person's foot only predicts the length of another person's foot to a limited degree of precision. In response, medieval surveyors came up with an ingenious idea. They would line up 16 randomly selected people, measure the combined length of their feet, and divide the total into 16 foot-long segments.[4] A 16-way average predicts another 16-way average much more precisely than a single foot predicts another single foot (the expected deviation is reduced by 75%). This is an early example of a powerful technique for reducing statistical error that statisticians refer to as "aggregation."

The other type of error, known as "systematic error," is the type of uncertainty which cannot be reduced by aggregation. Systematic errors cannot be detected through statistical analysis of repeated measurements, because they remain stable under repetitions. An example of a systematic error in the case of the medieval foot would be if the 16 people chosen to line up were not random samples of the population. For instance, just taking the 16 people in closest proximity might result in 16 adolescents or 16 women being used to estimate the foot length. Increasing the sample to 32 or 64 in this case would not improve the accuracy, because the error is due to a lack of diversity in the sample, not the size of the sample.

So how do metrologists estimate how much systematic error is present in a measurement? In this article, we advance a novel and counter-intuitive claim: We propose that the best indicator of low systematic uncertainty is high-quality statistical uncertainty.

B. Statistical uncertainty is desirable

Statistical uncertainty is a intrinsic feature of repeated sampling enshrined by the law of large numbers. According to this theorem, the average of a repeated set of measurements will always result in an improvement in precision, no matter how precise the underlying standard. For instance, in the medieval foot example, measuring 32 feet would provide a more precise measurement standard than using 16 feet, while 64 would be better again. Similarly, while super-precise UTC is currently based on approximately 400 atomic

clocks scattered around the world,[3] using 1000 clocks would be even better. The presence of statistical uncertainty should therefore not be viewed as a weakness: the issue of larger groups of measurements being superior to smaller groups of measurements is intrinsic to the concept of repeatable measurement. In contrast, a lack of statistical uncertainty implies a lack of informativeness such as, for instance, when a clock stops ticking and always shows the same time. In the absence of statistical uncertainty, no useful information can be derived.

Furthermore, the magnitude of statistical uncertainty is irrelevant to measurement stability as *any* required level of precision can be achieved by repeating the same measurement enough times and aggregating the results. According to the law of large numbers, the average of a large number of independent trials should tend toward an expected value, with the level of deviation falling with the square root of the number of trials. Thus, if a set of measurements reflects a level of statistical uncertainty that is unsatisfactory, those measurements can simply be repeated a thousand or a million times (whatever is needed) and the average taken. Assuming that the possibility of unrestricted repeated measurement is available, then the magnitude of statistical uncertainty for individual measurements is irrelevant.

In sum, it is inappropriate to regard statistical uncertainty as a source of inaccuracy: statistical uncertainty is desirable. Rather than minimizing statistical uncertainty, we propose that the goal of metrology is instead to minimize systematic uncertainty, leaving only statistical uncertainty behind. The ideal measurement standard is one whose quality can only be enhanced through further aggregation, and not by any other means. In other words, the ideal measurement standard is one whose uncertainty is purely statistical. When systematic uncertainty is decreased, the remaining uncertainty in a measurement standard becomes more and more statistical in nature, making the deviation between individual measurements appear more and more random.

Returning to the example of the medieval foot, we know intuitively that this measurement standard is weak and unreliable. However, the weakness of the standard does not stem from the fact that people's feet vary in length (statistical uncertainty). We have argued that unpredictable variations in individual measurements are actually a good thing. Instead, the weakness of the medieval foot standard stems from the fact that, first, it is awkward to assemble large numbers of people, and, more importantly, it is difficult to assemble random samples of people (systematic uncertainty). The problem with the system is not that human feet vary in length, but that the system is unscalable and prone to systematic bias.

C. The link between stability and randomness

Measurement standards with low systematic uncertainty (a.k.a. stability) demonstrate high statistical uncertainty (a.k.a. randomness). We can thus say that stability and randomness are effectively the same concept, separated only by repeated measurements.

This link is evident in the practice of metrology. For example, the BIPM currently defines the second as the duration of 9,192,361,770 cycles of radiation from the caesium-133 atom. Hyperfine atomic transitions are the most unpredictable event currently known to science, as enshrined by the highly successful theory of quantum mechanics. Each transition within the caesium-133 atom occurs at an entirely unpredictable (i.e., random) moment. The stability of atomic transition as a measurement standard is evidenced by its independence from all other earthly events. Given this link between stability and randomness, it is perhaps not a co-incidence that the same person, over the same three-month period in 1905, discovered that atomic photoemission provides both an immutable source of randomness in the form of individual photons (the quantum photoelectric effect) and, at the same time, immutable stability derived from a large aggregated set of emitted photons (the constant speed of light).

Any accurate measurement standard can be used to produce randomness (and vice versa). For example, computers generate high-quality randomness by measuring the drift between two internal clocks. The more stable the clocks, the higher the quality of the randomness in their drift. In the same way, randomness could be derived from the drift in relative weight of the copies of the International Prototype Kilogram (IPK), which are distributed around the world. The greater the stability of the copies, the more random (i.e., unpredictable) the drift between them will be (and vice versa).

The link between randomness and stability is also enshrined by Algorithmic Information Theory (AIT),[5] the discipline which unites theoretical computer science and information theory. The fundamental premise of AIT concerns the equivalence between likelihood and simplicity, or, in other words, between prediction and randomness. The idea of predicting sequences using compressed descriptions was first formulated by Solomonoff, who showed in 1964 that, for any predictable sequence of data, the optimal prediction of the next item converges quickly with the prediction made by the model which has the shortest description. Compressing a set of data is equivalent to removing patterns from it, and making the data more random.

Consider the following sequence: 4, 6, 8, 12, 14, 18, 20, 24… What number comes next? According to AIT, the best prediction is made by the hypothesis that is as concise as possible, in other words, as random as possible. The description "go up in an alternating pattern of +2, +4, skip the +4 if the number is a 6" is an unwieldy one, hence its prediction of 26 is low-quality. In contrast, the description "odds primes +1" is a well-compressed one, hence its prediction of 30 is high-quality. Vityányi and Li[6] show that data compression is almost always the best strategy, both in hypothesis identification and prediction. The more "random" the encoding of a hypothesis, the better the predictions it makes.

But why? In a nutshell, it has to do with independence. If we look at the unwieldy hypothesis, we can see it is "overfitted." There seems to be a relationship between the hypothesis and the sequence that it is intended to explain.

It looks as if somebody has crafted the hypothesis to fit the data that is visible. The two are not independent.

In contrast, the compressed hypothesis has had all the systematic patterns sucked out of it, leaving it statistically random. Because its encoding is close to random, there can be no connection remaining between the hypothesis and the sequence it is intended to describe. The model and data are independent, hence the prediction the model makes is more valid. Randomness acts like a guarantee of independence: phenomena which look random are likely to be independent, because they have no causal connections with the context in which they appear.

While working on nuclear weapons projects at the Los Alamos National Laboratory in 1946, Stanislaw Ulam and John von Neumann realized the importance of high-quality randomness for making predictions. Essentially, they wanted to "measure" an aspect of their nuclear model such as "will the neutrons breach the radiation shielding or not?" They initially found that they were unable to make predictions using conventional, deterministic mathematical methods, because the calculations were too complex. What they decided to do instead was to simulate random experiments, a novel technique they referred to as "the Monte Carlo method."

Although each result of a Monte Carlo simulation features statistical uncertainty, this uncertainty can be reduced by running more trials, so it is not a problem. Instead, the key to the success of the Monte Carlo method is ensuring that each simulation is independent of all the others, mitigating any systematic bias. Ulam and von Neumann wanted to run different, independent simulations, each making different mistakes, thus supporting a diversified overall perspective through aggregation. What they wanted to avoid was running different simulations, each making the same mistake (i.e., featuring systematic bias). The way to meet this criterion was to use high-quality randomness, thus ensuring the independence of successive trials, and yielding an accurate measurement.

Von Neumann struggled to find enough randomness to simulate the nuclear explosions, so he began using pseudorandom numbers, such as those produced using the middle-square method. The risk of using pseudorandomness rather than "true" randomness is that it raises the spectre of systematic bias: there might happen to be some relationship between the middle-square method and modeling nuclear explosions which would yield a type of uncertainty undiminished by repeated simulations. This would have led him to make the wrong predictions about the functioning of the hydrogen bomb.

In sum, Von Neumann needed a source of randomness to "illuminate" the properties of his bomb model in the same way that we need random photons to illuminate the length of a desk, random Caesium transitions to time the length of a second, or a random sample of voters to measure which politician will win the election. Randomness is the hallmark of independence between a property being measured and the context in which it appears: it is what allows us to reliably detach one property of an object from its other properties. Hence, we will argue, the most stable measurement standard is the one which behaves most randomly.

D. Measuring a desk

Let us explore this link between stability and randomness further, and show how it differs from common intuition in practice.

Intuitively, a measurement device is stable if it always produces the same reading when applied to the same object. For example, if I measure the width of my desk and obtain a reading of 1.17 m, then I expect to obtain the same reading if I re-measure it at any point in the future. This consistency, however, is not sufficient for stability. It may turn out that the measuring device is simply stuck on the reading of 1.17 m, and gives the same result for any measurement. In this case, the reading is simply uninformative and useless, lacking any statistical uncertainty.

Even when a measuring device gives different readings for different objects, and gives consistent readings for any given object, this is not sufficient for inferring stability. Imagine, for example, that my desk is located in a very hot room. Whenever I bring the measuring device into the room, the device expands significantly due to the heat, meaning that the reading of 1.12 m which it gives is too low. If I moved the desk to a cooler room and measured it again, I would obtain a reading of, say, 1.17 m. Thus, the measuring device is being affected by an external factor which is not an intrinsic property of the desk, namely, ambient temperature.

The reading is thus incomplete. It is not the desk that has a width of 1.12 m, but rather "the width of the desk when located in a room with temperature 34 °C". The fact that the measuring device fails to report the temperature in which readings have been taken means that a vital piece of information needed for accurate prediction is being left out. The omission of this contextual information is what we mean when we say the reading is unstable: the measurement readings are not fully independent of temperature. We have not succeeded in detaching the width of the desk from its background context.

In practice, an unstable measurement standard is one whose predictions a competitor can better. For example, if my desk is being moved through a building with different temperatures, and I have measured it at 1.12 m in width, then I expect it to fit through a doorway that is 1.15 m in width. However, if the doorway happens to be in a cooler part of the building, then my prediction will be incorrect. An opponent using a measurement standard that controls for temperature context can make better predictions, objectively demonstrating the superiority of his standard: whereas I make mistakes, he gets it right every time. His standard achieves greater independence between measurement reading and measurement context, leading to better predictions.

The problem with a measurement device whose readings vary with heat is that it introduces an association between measurement outcome and the process of measuring. When I measure my desk it is always located in the same room, with a temperature of 34 °C. I believe that I am taking separate independent measurements of the desk, thus the consistency of the readings of 1.17 m delude me into thinking that the device is stable. But consistency is not equivalent to stability: my predictions turn out to be wrong. The consistency I

observed was due to consistent contexts, not the stability of the measurement device across diverse contexts.

Rather than fixating on consistency, stability is better defined in terms of competitive success. What we want is for our measurements to support predictions that are hard to beat. The purer the statistical uncertainty evident in individual measurements (i.e., the more random they are), the harder it is for that measurement standard to be upstaged by competitors seeking to expose residual systematic error.

E. Overthrowing the krypton-86 standard

When a measurement standard is shown to be exploitable by competitors, it violates the assumption of independence between measurement readings and measurement context, and the standard is superseded. This "dethroning" is manifested as a failure of expected randomness.

The krypton standard for length, for example, was superseded by the current light-based definition in 1975, after its shortcomings were revealed by the enhanced stability of the wavelength of light emitted from a methane-stabilized helium-neon laser. Physicists using the laser were able to make predictions about the deviation of measurement readings carried out using krypton-86 lamps. The lamp-users believed the deviation of their readings to be random, that is, exhibiting pure statistical uncertainty. The laser-users, however, revealed a predictable bias in measurement reading depending on which point of the krypton-86 orange line was selected to standardize length. Because the 1960 krypton standard made no reference to choosing any particular point of the line, the existence of such systematic bias could not be expressed using the old definition.

The objective failure of independence between measurement reading and measurement context, as demonstrated through objectively superior predictions, is what led to the adoption in 1975 of the current light-based standard for length. The speed of light, derived from a large aggregated set of individually unpredictable photons, is grounded on a source of randomness so strong that it is embedded into the bedrock of physics. Anyone challenging the current BIPM standards for length or time will have to demonstrate that quantum events are not truly random. The ostensible difficulty of this task is what secures the stability of the current BIPM standards.

III. THE CLOCK COMPETITION

With these ideas in place, we can now turn our attention to how to identify the world's most accurate clock.

The Longitude Act was passed in July 1714, offering monetary rewards for anyone who could find a simple and practical method for the accurate determination of a ship's longitude. The bulk of the prize money was eventually claimed by John Harrison for his invention of the marine chronometer, revolutionizing long-distance sea travel. In this spirit, let us consider a similar competition, this time to develop a clock more accurate than any clock ever yet built. How would the winner of the competition be identified?

In the case of longitude, the proof required for winning the prize is immediately obvious: the system should enable ships to undertake long sea trips successfully. But in the case of the clock competition, how do we tell if a clock is ticking more accurately than any other clock on earth? What is it that timing is supposed to achieve? How do we recognize when the current standard for accuracy has been superseded?

In 1967, atomic time replaced ephemeris time as the standard for time. Up until then it had seemed as if the rotation of the celestial spheres was independent of any other events transpiring on earth. However, when Caesium atomic clocks became operational in 1955, it was quickly confirmed that, instead, the rotation of the earth fluctuates predictably relative to atomic transitions. For example, the particular day that we choose to measure a mean solar second can affect the reading we get, as predicted by the atomic second. Due to the presence of these predictable patterns, the mean solar second became unsuitable for carrying out the most accurate measurements. The goal of metrology is to deliver measurement readings that are as independent as possible from the context of measurement. Hence, the ideal source of timing is one whose fluctuations are unrelated to anything else on earth, or indeed, the universe.

As previously discussed, a common intuition regarding time is that clock accuracy can be defined purely in terms of readings consistently agreeing with each other. If my watch reads the same as your watch, and so does everybody else's watch, then this makes a pretty good case that we all have the right time.

This is acceptable if we are all synchronizing our watches to a more reliable source, but when it comes to setting the standard itself, the idea is flawed. Consider, for example, a set of supposedly accurate clocks that are placed into a vault. After 10 years this set of clocks is removed and found to still be ticking in perfect synchrony. This observation alone cannot be interpreted as evidence of stability. The clocks might all be making the *same mistake*, or as Wittgenstein puts it: "As if someone were to buy several copies of the morning paper to assure himself that what it said was true."[1]

For instance, the clocks might have a sensitive temperature detecting device that is capable of discerning night from day in the vault. At the coldest point of the night, they all reset to 00:00:00 midnight. Thus, what the clocks are agreeing on is the coldest moment of the previous night, not the passage of time over the last 10 years. What is needed to infer stability is evidence that the clocks are not simply making the same mistake (i.e., dependence on, or sensitivity to, some contextual influence which cannot be relied on to apply across all measurement contexts). Accordingly, the thing we need to focus on is not the extent to which clocks agree, but the manner in which they *differ*. What we want is for deviations in their behavior to be hard to anticipate. We want the clock drifts to demonstrate high-quality statistical uncertainty.

A. Running the competition

How can we identify a winner of the clock competition, without the verifying authority having to spend huge amounts of money on testing, and running the risk of having

competitors complain that the testing procedure was not fair or reliable? We propose that the most objective mechanism for identifying stability is to allow competitors to test each other's clocks and reach agreement among themselves.

We propose that the one thing that competitors will agree on is prediction. Let each competitor build a pair of clocks. Now, challenge them to predict the ticking of each other's clocks. Given a pair of clocks, which will tick first, clock A or clock B? (or, if that period is too brief to witness, which clock will be the first to tick, say, a billion times?) At chance, competitors have a 50% possibility of guessing right. However, if their understanding of time is superior then they can push these odds higher than 50%. Multiple repeats of the game quickly reveal who is making predictions above chance. The pair of clocks whose divergence is most difficult to predict (i.e., most random) is the winner; this is the clock with the purest form of statistical uncertainty and, hence, the lowest amount of systematic uncertainty. All competitors will agree on who the winner is, because the losers always get beaten in this game, despite their best efforts.

The clock competition thought experiment illustrates that measurement is an objective endeavour. For instance, the notion of timing stability is not something which is defined by any one institution: the BIPM do not have a set of arbitrary tests that a hyper-accurate clock should pass, on which other metrological institutions might potentially disagree. Instead, stability is grounded by an unrestricted competition to out-predict one's opponents by any means available. When a new, more accurate clock is built, it is easy to demonstrate this fact objectively by out-predicting everybody else. The most accurate clock is the one whose ticking is most random (i.e., more independent of measurement context than anyone else's clock). Measurement is not a theoretical issue that is subject to debate: it is a practical issue, with immediately overt consequences that everyone can appreciate.

If we are dissatisfied with the precision of the winning clock's ticking, we can reduce its statistical uncertainty to any level by cloning the clock. For example, we can build a million independent copies of the clock and define the second as the period of time it takes for 500,000 of those clocks to tick. Given a source of pure statistical uncertainty, there are no limits to the level of precision that can be achieved.

B. A common misunderstanding

The idea that randomness is the key to stability is very surprising. Intuitively, we think that a clock is accurate because it is in synchrony with "ideal" time, whatever that might be. We do not think that a clock is accurate because it wobbles randomly. Indeed, a clock that ticks randomly and unpredictably seems like the exact opposite of what is desirable in a clock.

When the media reports on new atomic clocks that improve on the accuracy of all existing clocks, it is common to see headlines of the type "NEW CLOCK ACCURATE TO WITHIN ONE SECOND IN A BILLION YEARS." What we do not see is headlines of the type "NEW CLOCK TICKS UNPREDICTABLY." And yet, the media focus on consistency as a standard for timing is a mistake. The everyday intuitive idea of judging the accuracy of our clocks relative to a more reliable standard only makes sense if there is a trusted source of authority. However, in the case of the world's most accurate clock, there *is* no more stable standard. Relative to what will the new clock lose only one second in a billion years? This seems to imply a comparison between the new clock and an ideal clock. But no ideal is available.

Assuming the drift represents statistical error, then the purported "one second" slippage over a billion years could be reduced to any arbitrary level (e.g., one millisecond; one microsecond) by simply building multiple copies of this new clock and taking the aggregated reading of the set. In this case the media headline makes no sense. On the other hand, if the drift represents systematic error, then it cannot be quantified at all, because we do not know what the error is.

If prompted—what does this media headline actually mean?—one might infer that, after a billion years, two of the new atomic clocks are expected to only have a discrepancy of one second between them. Yet, as previously discussed, consistency does not imply stability. For example, a group of farmers relying on a crude version of ephemeris time could also be expected to agree within one second in a billion years' time: they simply look up at the sky and define midday as the point when the sun is highest in the sky. In a billion years everybody will still agree on exactly what time midday is at, because they will all be looking at the same sun in the same sky, making the same mistake. Hence, claiming that a clock drifts one second in a billion years means absolutely nothing.

When a new hyper-accurate clock is introduced, the only claim it has to being more accurate than existing clocks is that it can out-predict them. Because the old clocks cannot predict the ticking of the new clock, the ticking of the new clock appears random relative to the old standard. Thus the headline "NEW CLOCK TICKS UNPREDICTABLY" is the appropriate one. Because there is no other reliable way to judge the clock competition, there is no other possible way of describing the winner of such a competition.

IV. CRITIQUE AND REBUTTAL

Given that the idea of the clock competition is so counter-intuitive, we will now address it again from an alternative perspective, namely, by responding to a series of arguments that a metrologist might raise.

Argument 1: *The relationship between measurement accuracy and predictability is the inverse of what is claimed in this paper. The behavior of measurement standards need not and must not be random. Indeed, the behavior of an accurate measurement standard should be the easiest to predict.*

Rebuttal: No, this is an important mistake in thinking about measurement. You cannot run a clock competition based on which clock is easiest to predict (i.e., a stopped watch), or whose behavior is most consistent. An atomic clock makes great predictions about the behavior of a

humble wrist-watch, but it does not work the other way around, and that's what makes the atomic clock superior.

For example, a wrist-watch and an atomic clock are going to slowly diverge in their timing. Wrist-watch adherents will be completely clueless as to which direction this divergence is going to move in. For them, it appears to trace out a random walk: if they gamble money on it, they are going to lose. Atomic clock adherents, on the other hand, have a much deeper understanding of time. They can model how the flaws in a quartz crystal vary with environmental context, and hence predict exactly how the wrist-watch is going to behave in its drift. Atomic clock behavior appears *random* relative to wrist-watch behavior, while wrist-watch behavior is *predictable* relative to atomic clock behavior. Because the atomic clock is less predictable, its users will win the clock competition every time. The best measurement standards are those whose behavior is hardest to predict, making them resistant to modeling.

Remember, randomness is relative. In a nutshell: if A outpredicts B, then A appears random (i.e., unpredictable) relative to B. If a given clock is capable of predicting worldly events, then the behavior of that clock will naturally appear random relative to those worldly events. Hence, the most accurate clock in the world is the one that behaves most randomly from our naive perspective.

Argument 2: *A crucial point of the existing literature on metrology is that the accuracy of a measurement standard is ultimately determined relative to a theoretical ideal rather than relative to other measurement standards.*

Rebuttal: No, this is another important mistake in thinking about measurement. The whole point of the clock competition is that there is no theoretical ideal available, so measurement cannot possibly work in this way. Humans have the intuition that measurement works like this, because we are used to delegating responsibility to trusted authorities. But when it comes to setting the standard, this intuition no longer works. The most accurate clock in the world attains its status not by matching an ideal, but by doing something that no other system can do. Its behavior cannot be anticipated or justified in any way by external observers: it overthrows every existing ideal. Instead, its superiority is manifested in practice, by defeating all competitors.

If measurement was based on a theoretical ideal, who would assume responsibility for setting the correct ideal? And what would give them the authority to do so? The problem here is one of justification. Successful measurement is not something that is decided by fiat. Instead, we need a practical means of demonstrating superiority, one that goes beyond theory, modeling and abstractions, one that everybody can participate in, one that everybody agrees on. Prediction is the objective process that meets that criterion. Whoever makes the most accurate predictions will win competitions in a manner that supports universal agreement. Forget about comparisons with nonexistent theoretical ideals, none of that matters. All that matters is winning.

Argument 3: *The source of accuracy of caesium clocks is not the randomness of the hyperfine atomic transition but the fact that all caesium-133 atoms have (under ideal conditions) the same frequency associated with that particular transition. Caesium fountain clocks should ideally "tick" as closely as possible to that frequency.*

Rebuttal: This statement is riddled with weasel words such as "same," "closely," and "frequency," which set up circularity.

For a start, what does it mean to claim that all caesium-133 atoms have the same property? How do we know that? What is the evidence? The only thing we can say is that it seems very hard to tell caesium-133 atoms apart based on their behavior. In other words, the behavior of caesium-133 atoms appears independent (i.e., unpredictable; random) relative to the environmental context in which they appear. Stating that caesium-133 atoms are good for timing because they are all the same is a circular argument, because it fails to define how the property of "sameness" is established. Instead, the genuine justification for using caesium-133 atoms is that, so far, their behavior has proved impossible to predict.

The use of the word "frequency" is another weasel word, because it assumes a pre-existing standard for time onto which events can be projected (frequency is defined as the rate per second of a vibration). Again, this harks back to the human intuition to defer to a trusted authority. When the standard for time itself is being set, the concept of frequency *does not yet exist* and cannot be used as justification for selecting a particular standard.

Finally, the assertion that caesium clocks should tick as closely as possible to an ideal perfect frequency has no practical implications. When we are setting the standard for measurement, the concept of "closely" cannot be relied on as a guide, since it is the very thing we are attempting to realize. In practice, clocks tick closely when it is difficult to discriminate between them based on their behavior, in other words, when they drift randomly from each other. The reason we use atomic clocks is not because caesium-133 has some apodictic God-given claim to stability, it is simply because atomic clocks are, to date, winning the clock competition.

In sum, we need to abandon the intuitive justification of unattainable ideal measurement standards just beyond the horizon, and embrace the fact that metrology is a discipline which delivers in practice. Measurement is ruthlessly objective, and this is what sets it apart from so many other kinds of human activity. For example, in subjective disciplines such as philosophy, a small elite group of practitioners decides what counts as good and bad practice; as a result much of the energy in the discipline is focused on behaving in certain ways which meets with the approval of the elite.

Metrology is completely the opposite; it is the ultimate objective discipline. When your measurements are inferior, you make inferior predictions and you start losing straight away, in a manner which is obvious to everyone. For this reason, metrology does not rely on an elite group of practitioners to dictate the kind of language that should be used, or to determine how metrologists should behave. All that matters is winning, by any means.

Better randomness always leads to better predictions.[6] Hence, if we accept that measurement is about prediction, then we also accept that a clock that ticks randomly is the best clock in the world. There is no need for debate: if

somebody found a way of building a clock that ticks more randomly than any other existing clock, then all metrologists would immediately abandon what they are doing and start building that clock. The goal of the clock competition thought experiment is to point out that measurement is fundamentally rooted in predictive success, and that predictive success depends on the realization of randomness, which is an infinitely difficult task.[5]

V. CONCLUSION

People have a mistaken intuition about measurement. In everyday life we are used to adjusting our clocks to that of a stronger authority on time, comparing the accuracy of our tools to those that are even more reliable. This attitude leads us to suppose that the route to stability is to identify immutable apodictic physical constants. This attitude toward measurement is a mistake. The key to enhancing measurement accuracy lies with understanding what it is that measurement is supposed to achieve, and the process by which superior measurements are recognized.

In this article, we have argued that, contrary to intuition, measurement is not about eradicating uncertainty. Instead, accurate measurement depends on having access to high-quality statistical uncertainty: in order to achieve independence between a measured property and its context, we must identify and leverage a source of even purer uncertainty than the one we are seeking to illuminate. The more random that source, the more stable a foundation it provides for supporting predictions. For example, Stanislaw Ulam and John von Neumann hunted out high-quality randomness and leveraged it to reduce their uncertainty about the functioning of the hydrogen bomb; they expressed one source of uncertainty in terms of a stronger source of uncertainty, rendering a failure of the hydrogen bomb equivalent to finding patterns in the middle-square method.

Why does this seem so counter-intuitive? We are accustomed to living in a world where most measurements are "sloppy." When I use a measuring tape, the things I measure with it are often just as good at maintaining length as the tape itself. For example, I could mark one meter on a stick, and then use the stick as a tool to measure the length of other objects. Here, the stick is just as effective as the original tape. Because our ordinary standards do not exceed the accuracy of intuitive manifestations of constancy in the surrounding environment, we are easily led to believe in the notion of an ideal "objective reality." We naively assume that the goal of measurement is to match the consistency of this objective reality.

For example, we do not have a pair of suns, only a single sun, so there is no obvious means of quantifying the randomness of its timing drift, as per the clock competition. Before the invention of atomic clocks, a time-keeping competition would have been determined based merely on how closely the competing clocks agreed with the sun. Because of this natural source of timing authority, the connections between stability, independence and randomness were, until recently, hidden. Such connections only became apparent once measurement capability transcended the efficacy of easily accessible environmental standards.

At the limits of accuracy, the goal switches from that of consistency (e.g., matching the accuracy of the sun), to achieving independence between measurement readings and measurement context. The more random (i.e., unpredictable) the drift in individual measurement readings, the greater the level of independence achieved, and the more successful the associated predictions. Understanding the role of randomness at the heart of measurement is the first step toward abandoning the flawed notion of an ideal objective reality.

[1]L. Wittgenstein, *Philosophical Investigations* (Blackwell, Oxford, UK, 1958).
[2]E. Tal, Philos. Sci. **78**, 1082 (2011).
[3]E. Tal, Br. J. Philos. Sci. **67**, 297 (2016).
[4]S. M. Stigler, *The Seven Pillars of Statistical Wisdom* (Harvard University Press, Cambridge, MA, 2016).
[5]M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications* (Springer, New York, 2008).
[6]P. Vitányi and M. Li, IEEE Trans. Inform. Theory **46**, 446 (2000).