THE MICHELSON-MORLEY EXPERIMENT: A CASE STUDY IN VALIDATION

The Michelson-Morley experiment can be examined, using standard mathematical concepts of error analysis, as a case study in validation. A few philosophical principles focusing on scientific confirmation help elucidate the formal logical concepts that result in a foundation for scientific validation. The author proposes two research directions.

The loss of the Mars Polar Orbiter to a sophomore physics error regarding physical units reminds us that a major focus of computational science and engineering must be the development of correct models and simulations. Two terms are commonly used to describe correctness in simulations: *verification* and *validation*. The US Defense Modeling and Simulation Office offers a central information repository on the V&V of models and simulations.[1,2] According to the DMSO's Web site (https://www.dmso.mil/public/resources/glossary), verification is

> the process of determining that a model or simulation implementation accurately represents the developer's conceptual description and specification. Verification also evaluates the extent to which the model or simulation has been developed using sound and established software engineering techniques.

The site defines validation as

> the process of determining the degree to which a model or simulation is an accurate representation of the real-world from the perspective of the intended uses of the model or simulation.

The ideas of *conceptual description and specification* and *real world*, taken from these definitions, are not formally defined, yet most readers would have a clear idea of what the terms mean—to them. Since these terms are not formally defined, the DMSO's definitions of V&V are effectively meaningless.

To establish the meaning of V&V, I will produce a philosophical basis for the terms and then develop formal frameworks within this philosophical setting. Finally, I will provide two useful formal approaches for reasoning about V&V as one problem. One approach uses the concept of evidence based on mathematical evidence as formulated by Shafer.[3] The second model is based on concepts of information. I define both systems from the idea of a probability knowledge graph, which I develop from the verification (proof) of system properties.

Proof is not an absolute concept; it is defined within epistemology. In computational science

1521-9615/02/$17.00 © 2002 IEEE

D.E. STEVENSON
Clemson University

and engineering, there are three epistemologies: confirmation in science, proof in mathematics, and construction in computer science. To understand formal systems and their justification, we must start at the top: philosophy. The role of philosophy in science, mathematics, or other study is to criticize current practice. But what is current practice? To illustrate it, I apply current mathematical tools—conditioning, uncertainty, error, and sensitivity—to a case study, the famous Michelson-Morley experiment.

Principles to keep in mind

Understanding validation begins with understanding how people think about knowledge. The sum of all our beliefs is called a *worldview*. Groups as well as individuals have worldviews. We also hold certain ideas to exist (ontology) that we call *metaphysics* or *categories*. The concept of *knowledge* falls within the categories view. *Epistemology* is concerned with the nature, sources, and limits of knowledge. Validation is part knowledge and part epistemology (I will clarify this later). I provisionally define *validation* as justifying new knowledge based on a verified model and experiments in the real world. Later I will try to make *justifying* and *real world* meaningful to V&V. Just as I claim that *conceptual model and specification* and *real world* have no meaning, so too *justifying* has no meaning at this point. Thus far, I have defined neither *justification* nor *real world*.

I begin by informally introducing concepts to apply to the Michelson-Morley experiment.

Scientific theories consist of several "languages," two of which are a theoretical language (the usually mathematical language in which the theory is written) and an observational language (such as numbers and statistics). Theoretical language and observational language are connected through *reduction sentences* and testing procedures.[4] Reduction sentences assign meaning to theoretical symbols; testing procedures relate to such things as acceptable statistical procedures on data.

Computational science and engineering has three kinds of participants: application researchers (science and other fields), algorithm researchers, and computational and architecture researchers (software and hardware).[5] Each brings his or her own language, methods, and beliefs. Put into philosophical terms, individuals have their own worldview, categories, knowledge, and epistemology. To analyze the Michelson-Morley experiment, we need all three kinds of worldviews: application (physics) to elucidate meaning, algorithms (mathematics) to derive expressions, and the hardware and software to evaluate those algorithms.

Mathematics uses the term *error* to indicate the difference between true and calculated solutions.[5] We can use three principles relating to error to guide evaluation of models and simulations:

P1. *Physical exactness.* We strive to identify nonphysical (mathematically convenient) assumptions and eliminate them.
P2. *Computability.* We must identify noncomputable relationships. Most mathematical relationships in scientific computation turn out to be approximate.
P3. *Bounded errors.* No formulation is acceptable without a priori error estimates or a posteriori error results.

The errors generated by computation must be less than the uncertainty in the observed values (more on this later). This must be so, otherwise decisions would be made on computational and mathematical artifacts rather than on the model.

Conditioning and error analysis

Condition describes a measure of our ability to compute values for y given x in the equation $y = F(x)$, any or all of which might be multivariate. Wilkinson introduced *condition numbers*,[6] with the generalities relying on tensors. Briefly, let $x_1, x_2, \ldots, x_m$ be parameters to a problem and $y_1, y_2, \ldots, y_n$ be the solution. There are m equations in n unknowns. The mn terms

$$k_{rs} = \frac{\partial y_r}{\partial x_s} \quad (r = 1, \ldots, n, \ s = 1, \ldots, m)$$

constitute the complete information on the solution's sensitivity with respect to perturbations in the parameters. The conditioning of the problem $y = F(x)$ is a measure of the total sensitivities. Because most problems have a large mn and because there are different ways to consider the total sensitivities, various approximations to the problem's condition are used. The Michelson-Morley experiment has few variables, so conditioning analysis can be carried out in full.

Now let us look at how we might use the three principles listed earlier. We consider two types of errors: scientific errors, denoted δ_s, and numerical errors from approximations and finite arithmetic, denoted δ_a. From the mathematical perspective, δ_s is unknown and consequently ignored. Let F be the function received from the science, and suppose (as a gross simplification)that F is linear.

For our problem, one measure of merit for δ_a is *relative error*, which is

$$\frac{\left|y-\hat{y}\right|}{y}=\frac{dF/dx}{F(x)},$$

where y is the *true value* and $\hat{y}$ is the *computed value*.

Suppose F is a linear operator. For the *forward problem*, where F and x are known and we want $y = F(x)$, we can determine the relative error in y:

$$(F+\delta_a F)(x+\delta_a x)=y+\delta_a y$$

$$F(x)+F(\delta_a x)+(\delta_a F)(x)+(\delta_a F)(\delta_a x) = y+\delta_a y, \qquad (1)$$

since F is linear. We can organize Equation 1 as a series of fractions

$$\frac{F(\delta_a x)}{F(x)}+\frac{(\delta_a F)(x)}{F(x)}+\frac{(\delta_a F)(\delta_a x)}{F(x)}=\frac{\delta_a y}{y}. \qquad (2)$$

The first term in Equation 2 measures the conditioning, the second measures the method's *stability*, and the third term measures the *finite arithmetic errors*. I interpret Equation 2 this way:

1. If the condition number is large, the function is extremely difficult to compute. Therefore, I need a different formulation of the problem.
2. If the condition number is small, the function is easy to compute. Stability now comes into play. Stability is the measure of how approximation errors are propagated. If the stability number is large, the method probably will not be satisfactory; find another method.
3. If both conditioning and stability are good, then the errors due to finite arithmetic determine the computed value's accuracy.

Uncertainty and sensitivity

From a computational standpoint, we must determine the effect of using finite numerical representations instead of infinite ones. These measures help quantify the uncertainty caused by calculations. Clearly, relative error is one measure of uncertainty.

In any theory, we would expect to see mathematical relationships among the theory's objects. When we subject a theory to observation, we introduce observational errors. A natural question would be, "How large can these observational errors be and still not affect the decision?" In other words, how sensitive is our result to errors in the observations? We can formulate this by asking, how does change in a parameter's error change the function's value? Using our arbitrary $F(x)$ again, we want to study how $F(x + \Delta x)$ changes with respect to the individual occurrences of Δx_i.

$$\text{Sensitivity of } \Delta x_i = \left.\frac{\partial F}{\partial \Delta x_i}\right|_{\Delta x_j=0} \text{ where } j \neq i.$$

This particular definition describes the function F's sensitivity to changes in the perturbations Δx_i.

I have presented these measures in their simplest fashion—what you might call first-order analysis. Clearly, we can expand these ideas; for example, we can perform conditioning in higher-order derivatives. These concepts are known by different names in different disciplines; for instance, in economics, conditioning is known as *elasticity*.

The Michelson-Morley experiment

In 1886, Albert Michelson and Edward Morley published disconfirming results on the existence of ether.[7] After others criticized their experiment, they published a second article in November 1887 that again disconfirmed the ether theory.[8] Interestingly enough, Michelson and Morley never completed their experimental plan—they collected only two complete sets of observations of a year-long protocol. The ether idea lingered for almost 30 years; it did not die but simply disappeared from mainstream physics.

I cannot overemphasize this last point. The concept of ether is not "disproven" or "disconfirmed." Absolute assurance that something is valid or invalid is impossible. Therefore, one clear issue validation theory must address is determining how much evidence is needed for a particular model or simulation.

Historical background[9,10]

The framework was originally set by Aristotle in the 4th century BC when he introduced the concept of an invisible, frictionless fluid called "ether" that transmitted light. Over time, this term became a catchall phrase for "action at a distance." Christiaan Hugyens introduced a *wave theory* of light transmission in 1678. Descartes and Newton weighed in with a *corpuscular theory* that posited that the nature of light was explained by vibrations in the ether, which was made up of small discrete particles. Because of Newton's immense standing in the scientific community, the wave theory was largely disregarded.

In the period from 1727 to 1732, Samuel Molyneux and James Bradley performed several parallax measurements on the star γ-Draconis but were startled to find that in the three days between measurements, the star had shifted to the south by one arc second. This was in the wrong direction and by far too great a distance to be explained by the currently understood theories of light. Bradley, analyzing the situation, called the ratio of the speed of the Earth's orbit to the speed of light in a vacuum (v/v_e) the *aberration constant.* The last step of interest involves Thomas Young, François Arago, and Augustin-Jean Fresnel. As a result of their interactions, Fresnel published a paper in 1818 that set the expressed aberration constant in terms of the index of refraction in an equation known as the *partial dragging coefficient* (PDC).

Refraction is the term given to the "bending" of light at the interface between two materials with different optical characteristics. The index of refraction is related to the speed of light in the two materials. Let n be the index of refraction of light in ether with respect to light in a vacuum (no ether). The Fresnel PDC is the same as the aberration constant and is

$$\frac{v}{v_e} = 1 - \frac{1}{n^2} \ .$$

The concept for the Michelson-Morley experiment[7,8] is derived from Fresnel's derivation. If the ether has the properties ascribed to it, then an *interferometer* would be able to detect a difference in time to travel in the ether (see Figure 1). A monochromatic light source (sodium) is shone on a half-silvered mirror (1). The light is split into two beams (2 and 3). The fixed and movable mirrors reflect the light back to the half-silvered mirror and then to the detector. The movable mirror is used to calibrate the distances. Because the two beams travel the same distance but at different velocities, they should *interfere* with one another.[11] The interference phenomenon is often seen on lakes when two wakes come together and the amplitudes add together or cancel one another depending on frequency and phase. In light interferometer experiments, this phenomenon plays out as a set of dark and light rings. The rings' location can be measured accurately and can be used to determine the difference in time.

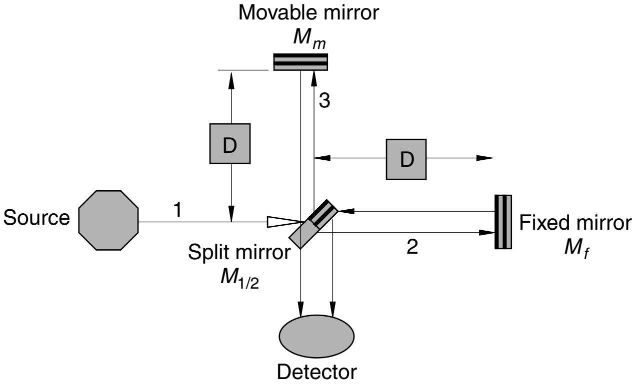

Figure 1. The Michelson-Morley interferometer.

Derivation

In terms of the observational–theoretical distinction, thus far we have considered the PDC in theoretical language. In fact, we run into trouble as soon as we try to use the observational–theoretical model. The PDC's value is not connected to any observable measurement: the observable v is eliminated. Let V be the velocity inside the ether. We cannot measure the density difference between air and ether. The ratio we can measure is given by v/V. This is the key to the refutation.

Let us now derive the equations to support the Michelson-Morley experiment. We must find the time difference to predict the fringe interference pattern difference. Figure 1 shows the experimental setup, and Figure 2 gives the variables' definitions, the assumptions, and values.

The times for each leg are $T = D/(V - v)$ and $T_1 = D/(V + v)$. The round-trip time is clearly

$$T + T_1 = 2D\frac{V}{V^2 - v^2} \cdot$$

(The 1887 paper illustrates the breaking of principle P1 listed earlier: "and the distance traveled in this time is

$$2DV^2/(V^2 - v^2) = 2D(1 + v^2/V^2), \qquad (3)$$

Variable	**Definition**
v_e	Velocity of ether outside the material
V	Velocity of light inside the material
v	Velocity of the Earth with respect to the ether—that is, in orbit
D	Distance between the two points of the apparatus *ab* or *ac*
T	Time for light to go in direction *a* to *c*
T_1	Time for light to go in direction *c* to *a*

Value correspondence rules

$\lambda = 5.9 \times 10^{-7}\ m$	National Inst. of Standards and Technology
$V = (299.8524 \pm 0.0790105478190518) \times 10^6$ m/s	NIST Michelson Project
$v = 2.979 \times 10$ k/s $= 2.979 \times 10^4$ m/s	NASA National Space Science Data Center
$v/V = 0.99 \times 10^{-4}$	
$(v/V)^2 = 0.99 \times 10^{-8}$	

Figure 2. Variable definitions, assumptions, and values.

neglecting terms of the fourth order" [italics mine]. Having now crossed the line, Michelson and Morley must now continually refer to both the exact and approximate derivations.)

The experiment calls for one arm of the interferometer to be in line with the travel in orbit and the other to be perpendicular to the orbit. The arm moving parallel to the path of the Earth is actually undergoing a Lorenz contraction. So, the light on the other path must travel $2D\sqrt{1+v^2/V^2}$. To get the approximation compatible with Equation 3, we expand this by the binomial expansion $(a + b)^n$, with $a = 1$ and $n = 1/2$. We obtain a first-order approximation for the length as $2D(1 + v^2/2V^2)$. Subtracting the two distances, we get Dv^2/V^2. This distance is the one that provides the interference pattern. The distance Dv^2/V^2 divided by the wavelength of the light equals an integral number of wavelengths; that is, the interference bands occur at those intervals. Michelson and Morley ran the experiment using a sodium lamp, emitting at 5,900Å or 5.9×10^{-7} meters.

The protocol for the two experiment instances that Michelson and Morley recorded was to set up the apparatus and take a reading at noon, then rotate the apparatus 90 degrees clockwise and take the second reading at night. So, the total difference is twice that given in the derivation, or (using the approximations) $2Dv^2/V^2$.

Experimental results

As reported in the 1887 paper, Michelson and Morley recorded their observations from 8 to 12 July 1887. Quoting from the report:[9]

> It seems fair to conclude from the figure that if there is any displacement due to the relative motion of the earth and luminiferous ether, this cannot be much greater than 0.01 of the distance between fringes.
>
> Consider the motion of the earth in its orbit only, this displacement should be $2Dv^2/V^2 = 2D \times 10^{-8}$. The distance D was about 11 meters, or 2×10^7 wavelengths of yellow [sodium] light; hence the displacement to be expected was 0.4 of a fringe. The actual displacement was certainly less than the twentieth part of this, and probably less than the fortieth part. But since the displacement is proportional to the square of the velocity, the relative velocity of the earth and the ether is probably less than one-sixth the earth's orbital velocity, and certainly less than one-fourth.
>
> It appears … if there be any relative motion between the earth and the luminiferous ether, it must be small; quite small enough entirely to refute Fresnel's explanation of aberration.

The distances were to be measured using a graduated set screw that was reported to be able to measure 0.02 wavelength.

Applying the principles

We now look at the Michelson-Morley results in terms of the measures proposed earlier.

Can we compute accurately?

As mentioned earlier, the derivation gratuitously uses unwarranted approximations: We can compute the time difference without intermediate approximations, as I did earlier using the Maple computer algebra system. Without the approximations, the difference between the time along the orbit and the time perpendicular is

$$\frac{2D(\rho V^2+\rho v^2+V)}{-V^2+v^2}\ , \tag{4}$$

where

$$\rho=\sqrt{1+(v/V)^2}\ .$$

We define the condition number for a function $f(x)$ as

$$C_f^x=\frac{x}{f(x)}\frac{df}{dx}.$$

Using Maple and the values in Figure 2, we find (after a lot of work) that the condition number is 3×10^{-9}. This means we can compute the time difference very accurately. Using the definition of relative error, and again using Maple, we get a relative error of 3×10^{-13}; that is, the computed difference should differ from the real one in the 13th decimal place.

This means we can compute the values accurately enough to guarantee that observational errors are the dominant errors in the results. This means experimental errors will dominate any experiment, so we cannot conclude the wrong thing due to computational error.

The logical case

For convenience, Equation 4 is repeated here but with the values labeled as observational:

$$T_i=\frac{2D_o(\rho V_o^2+\rho v_o^2+V_o)}{-V_o^2+v_o^2}, \tag{5}$$

where $\rho=\sqrt{1+(v_o/V_o)^2}$ and T_i is the time difference with the auxiliary definition for ρ. The equation

$$m_o\lambda_o=T_i \tag{6}$$

relates the spacing of the interference fringes to the time difference. m_o is actually what was calculated as 0.4 in the report on results. The Michelson-Morley report says that D is "approximately 11 meters" or 2×10^7 times the wavelength of yellow light. Using the values from Figure 2 and Equation 5, we find that D is 11.8 meters. Solving Equation 6, we find that m_t's value should be 0.3971. Taking "certainly less than a fortieth" as 1/40, $m_o = m_t/40 = 0.00993$.

We can also solve the *theoretical* T_i equation for v_t. If we use the observed values and the theoretical m value, we compute that

$v_t = 29{,}789.999\underline{99}$ with relative error 0.3×10^{-9}

(with the last two digits in error), but the observed speed in orbit would have to be

$v_o = 4{,}710$ with relative error 0.8×10^0.

The zero exponent in the relative error says that the real and computed values of v_o do not even agree in magnitude.

We still cannot connect the observational and the theoretical unless we accept a correspondence rule that allows $v_o = v_t$. How does this logically generate the contradiction to Fresnel's PDC? We have both v and V by measuring the speed of light in a vacuum. The value is given as $V = 2.997 \times 10^8$. If we allow such a rule, then we can say that $v_o \neq v_t$; so, we have a disconfirming experiment.

Sensitivity and uncertainty

Equation 4 has no arbitrary constants and just three variables: D, V, and v. These variables' sensitivity values are as follows:

D: 2.0000
V: 3.1413×10^{-15}
v: -2.9157×10^{-11}.

Thus, the results are very *insensitive* to the computation and dominated by observational errors, because the distance is the most sensitive but the easiest to measure: we should easily be able to measure D to 10^{-3}. The other values are many orders of magnitude smaller, especially at the reported errors. The relative error is 1.517×10^{-16}, as calculated in Maple with 20 digits of precision.

These numerical considerations indicate that uncertainty about computed values is low. The sensitivity coefficients indicate that the value is quite stable within the errors stated.

Generalization

If computational science and engineering were only about computation in physics experiments, we could draw specific rules about validation from this experiment. But computational science is much broader and as much about modeling as it is about computation. We must develop general guidelines that make modeling and validation part of the same process.

Unified conduct

Let us examine the process of the whole experiment, paying particular attention to the use of knowledge.

1. Using current knowledge, we derive a model.
2. Using the proposed model, we derive implications of that model that are consistent with current knowledge.
3. Using the implications, we perform experiments that result in evidence for and against the new model.
4. Comparing the theoretical implications and our observations, we arrive at a conclusion concerning the model's coherence with current understanding.
5. Taking the changes to current knowledge arrived at thus far, we attempt to integrate the new knowledge into the old.

Items 1 and 2 are formal, logical steps using deductive arguments based on the current state of knowledge. We assume that the knowledge base is consistent, at a minimum.

Item 5 is the crux of the validation process: we must fit the new knowledge into the old knowledge base. This process is often difficult scientifically and in human terms, because the community must accept it. The usual slogan about validation is wrong: validation is not "Did you build the right thing?" but rather "Can you fit the model or simulation results into accepted knowledge bases using the allowed epistemological processes?" Thus, validation is about knowledge and fitting one's experiences into current knowledge. (See an interesting take on this process elsewhere.[10])

Philosophy in a nutshell

Knowledge is defined as "justified, true belief" or "warranted belief." It is consensual and communal, and not absolute; the community sets the limits of uncertainty. We generally mean *propositional knowledge* or *explicit knowledge*.[12] *Warrant* is both a noun and a verb, according to the *Oxford English Dictionary*. The verb is a synonym for "certify," but its technical meaning is to add something to experience to form knowledge.

Epistemology comprises rules for turning experience into knowledge. Warranting is only part of epistemology. How might these philosophical concepts be implemented?

Knowledge is belief that a community holds to be true. There are three major issues:

1. What ought we believe (implicit knowledge)?
2. What types of reasoning should be allowed for knowledge to be warranted?
3. What conditions are necessary and sufficient for knowing which and why the reasons in Issue 2 hold?

Several overarching rules have been proposed; two standard ones are *correspondence* and *coherence*. (I propose several others later.)

Correspondence rules fit this schema: A sentence is *true* if and only if the sentence corresponds to a state of affairs and that state actually occurs. This leads to considerations of information and evidence. In the Michelson–Morley case, the experimental values evaluated in the theory do lead to correct results.

Coherence has degrees. Coherence relates to the degree of interconnectedness or unification of the knowledge. There are three standard concepts of coherence:

- Logical coherence involves consistency.
- Inferential coherence involves soundness and probabilistic measures.
- Explanation involves relevance.

We can use the three types of coherence to judge the value of a proposed confirmation's properties. The Michelson-Morley experiment certainly fails the last point, because it does not explain the process.

Without further discussion, we accept here the current confirmation theory approach to science.

Confirmation theory

The *confirmation problem*, as the validation problem is known in the philosophy of science literature, was the focus of the philosophy of science from the mid 1930s until the mid 1970s. Rudolf Carnap, together with Carl Hempel and others, developed a "logical framework" for confirmation. Carnap's 1936–1937 article entitled "Testability and Meaning"[13] was the first in a series of publications on this. Karl Popper, in *The Logic of Scientific Discovery*, said that scientific theories were set up to fail—the so-called *falsification* theory.[14] Thomas Kuhn's *The Structure of Scientific Revolutions* was among the more influential works on the conduct of science over confirmation.[15] Kuhn tried to show that theories evolve owing to confirmation or disconfirmation.

Confirmation's central focus is the belief that science is lawful. There are three basic categories of constructs to consider: objects and measurements, science-theoretic definitions and functions, and logical axioms and operators. For physics and engineering, we have four sorts of objects to consider: idealizations, inferred entities, unobservable objects, and theoretical postulates. These idealized objects are placed in space–time. Our operators are idealized operators and relations over objects and space–time points. The question is how to relate observed objects to theoretical objects and how to understand idealized relations.

Any statement made about a system could be true, false, or neutral. But this valuation could vary over time: for example, better instruments and better measurements might resurrect the idea of the ether. There are two problems with trying to judge the correctness of models over time: evaluating hypotheses over time in light of an increasing amount of evidence, and deciding if the current evidence suffices. In other words, the warranting process appears to be a type of decision problem.[16]

Reasoning and risks

Others have proposed more specific rules than correspondence and coherence. For example, the DMSO proposed that validation concern itself with credibility, capability, and relevance with a focus on fidelity (accuracy, precision, and sensitivity).[1,2] Typical scientific studies have inherited rules from statistical inference.

The DMSO credibility criterion needs further amplification. In particular, many verification and validation efforts do not pay enough attention to the role people play. People hold knowledge, people approve rules, and people can make mistakes. Recall two incidents: the *Challenger* disaster in 1986 and the Cerro Grande fire at Los Alamos, New Mexico, in 2000. In each case, people misread the evidence. So, any validation approach must face the essential uncertainty of the enterprise and people's frailty.

Also, measurements are clearly necessary but not sufficient. We must reason about a system on the basis of measurements, but we should not think that they are a cure-all.

Confirmation and explanation

A useful standard is that a validation should ensure that the model explains the phenomena it models. Explanation is a logical concept that has two idealistic components regarding prediction:

- If a law is approximately true, then we should expect its predictions to be approximately true.
- If a law is approximately not true, then we should expect its predictions to be approximately not true.

The catch is the word "approximately." We can solve this problem using *counterfactuals*[4] with the following schema.

C explains E if all these are true:

1. C and E are events.
2. C occurs before E.
3. If C had not occurred, E would not have occurred, all else being equal.
4. If C had occurred in similar experimental situations, E would have occurred.

(An aside: I use a minimum of logical notation. The main connective symbol is → for implication; ¬ is "not." Logical rules are displayed in modern style, with hypotheses above the line and conclusions below.) In symbolic language, we might say (L and P) or (not L and not P). This leads us to a mode of reasoning known as *denying the consequent*:

$$\frac{L \to P \quad \neg P}{\neg L},$$

which is read, "If L implies P and if not P, then conclude not L." This is opposite to the usual forward direction of *affirming the antecedent*:

$$\frac{L \to P \quad L}{P}.$$

(In the logic and mathematics literature, affirming the antecedent is known as *modus ponens* and denying the consequent is *modus tollens*.)

A framework for integrated verification and validation

Formal Methods Europe defines *formal methods* as "mathematical approaches to software and system development which support the rigorous specification, design, and verification of computer systems" (www.fmeurope.org). NASA, for example, uses formal methods to develop projects such as a collision avoidance system called Airborne Information for Lateral Spacing. In the pure computer science context, again quoting FM Europe's Web site,

> The use of notations and languages with a defined mathematical meaning enable specifications, that is statements of what the proposed system should do, to be expressed with precision and no ambiguity. The properties of the specifications can be deduced with greater confidence and replayed to the customers, often uncovering facets implicit in the stated requirements which they had not realized. In this way a more complete requirements validation can take place earlier in the life-cycle, with subsequent cost savings.

My goal here is to develop formal methods for scientific simulations.

The paradigm

Earlier, we considered validation as a decision process: whether or not the model fits into the application knowledge base as augmented by mathematical and computational considerations. Here is a possible formulation of V&V based on the philosophical ideas developed in the Generalization section.

The first issue is to decide what we will try to fit into the knowledge base. In light of our focus on knowledge, we begin by realizing that models are formulated to answer questions. They are then developed to study various facets of a question. In the Michelson-Morley experiment, the question was the Fresnel theory's ontology. Using all the information (knowledge) at hand, the scientists derived a model that enabled theoretical conclusions about observations. In our case, the measures of merit were the observables. We used the correspondence rules to settle issues about which experimental values are mapped to which theoretical variables and how the values are manipulated.

After an experiment is run, each replicate becomes a warrant. The warranting process is to determine how the values observed match the known (and accepted) values. In our case, the results were disconfirming rather than confirming.

Verified model formalization

Let $\mathbb{Q}$ be the logical statement of the questions to be answered. The current knowledge base $\mathbb{K}$ is a system of logical statements. The initial question is "Is the question consistent with the knowledge base?" In symbols, this is $\mathbb{K} \to \mathbb{Q}$.

Simulations as computations can be logically expressed in *Hoare logics*.[17] If the input specification of the simulation is $\mathbb{S}_{\text{in}}$, the processing is S, and the output specification is $\mathbb{S}_{\text{out}}$, then we denote the logical statement that "Given the input state specification $\mathbb{S}_{\text{in}}$, the computation $\mathbb{S}$ terminates in the state $\mathbb{S}_{\text{out}}$" as

$$\mathbb{S}_{\text{in}}\{S\}\mathbb{S}_{\text{out}}$$

Therefore,

$$\mathbb{K} \to \mathbb{Q} \to \mathbb{S}_{\text{in}}\{S\}\mathbb{S}_{\text{out}}\,.$$

What we normally call a model is the derivation of the application ($\mathbb{D}$) to the questions. Therefore, we denote the *unverified* model as

$$\mathbb{M} = (\mathbb{K} \to \mathbb{Q} \to \mathbb{S}_{\text{in}}\{S\}\mathbb{S}_{\text{out}}, \mathbb{D}_{\mathbb{M}}). \tag{7}$$

Let $v\mathbb{M}$ denote the verified model. Logically, this is the model, the derivation, and a proof that the derivation is correct—in other words, $v\mathbb{M} = (\mathbb{M}, \text{Proof}(\mathbb{D}))$.

The decision formalization

We can develop the questions that led to the verified model into a decision process using another derivation. We write the decision model similarly to Equation 7:

$$v\mathbb{DP} = (v\mathbb{M}, \mathbb{D}_{\mathbb{DP}}). \tag{8}$$

The measures of merit *mm* are defined by $v\mathbb{DP}$. This verified decision process is formed from the verified model's logical conclusions. The decision process is the warranting process for this particular verified model.

Given the verified model and the correspondence rules, we can define the observations and the theoretical versions of the measures of merit. Let *cd* be the values incorporated into the verified model by the reduction sentences and *md* be the values from the solved model. A warrant is one triple (*mm*, *cd*, *md*). Therefore, the decision regarding validation is made on the set of warrants available at the time the decision is made.

We assume that the explanations defining the warranting process are coherent, law-like, consistent, organized, credible, predictive, justified, reasoned, and relevant. That is, the warranting process must ensure that the new knowledge fits the old knowledge in at least those properties.

The probability knowledge graph

One can depict any knowledge base as a forest of trees, not necessarily rooted. A proof is a tree. Suppose that we have a proof tree (actually, any verification tree should work) and that we can assign probabilities to some nodes. Then we can use a linear program to assign a consistent set of probability intervals to the rest of the trees.[18] This processed tree is a *probabilistic knowledge graph*. An important property of an individual PKG is the circumstances under which it exists:

- If the graph cannot be built, the warrant is disconfirming.
- If the graph can be built but lowers critical measures, the warrant is ambiguous.
- If the graph can be built and raises critical measures, the warrant is confirming.

This process of assigning probabilities can repeat as long as resources are available.

In terms of the warranting properties,

- Coherence, consistency, justification, lawfulness, organization, and reasoned content are all inherent to the justification process.
- Prediction comes from the measures of merit.
- Relevance comes from the appearance of a particular item in the PKG.

Using the PKG

The PKG, in and of itself, does not answer validation requirements because it captures only one situation—the relationship between theoretical language and one set of observations. We expect many experiments to be necessary. How do we arrive at a decision given these multiple experiments?

Two approaches come to mind. The first is a formulation of the Dempster-Shafer theory based on uncertain dynamical systems.[3] The second approach follows George Polya and Edwin Jaynes' work formalizing evidence.[19,20] Both approaches are open to further research.

Shafer structures. The Dempster-Shafer formalism is widely used in artificial intelligence and statistics to deal with systems with uncertainty using axiomatic definitions of belief functions. Briefly, the Dempster-Shafer theory posits *basic probability assignments*, *belief* or *support* functions, and *plausibility* functions. The latter two are defined on the basis of the related basic probability assignments. Support functions measure a hypothesis's support for a conclusion, and plausibility functions measure suitability.

Jürg Kohlas and Paul-Andre Monney proposed a Dempster-Shafer system related to *uncertain dynamical systems*.[20] This formulation is attractive because many models are of dynamical systems. Here is just a flavor of their approach.

Let Θ be a set of possible values (the resolution set) for variable X. Suppose there is conflicting evidence as to which interpretation of X is correct. Let Ω be the set of possible interpretations. Let $\omega \in \Omega$ be an interpretation and $\Gamma(\omega) \subseteq \Theta$ be the set of values that X would take on if ω were the true interpretation. In a probability setting, this situation induces a probability space $(\Omega, \mathcal{A}, P)$ where the probability measure P describes the likelihood of the different interpretations. We define *hints* as $\mathcal{H} = (\Omega, \mathcal{A}, P, \Gamma, \Theta)$ and interpret them as uncertain restrictions on X: $X \subseteq \Gamma(\omega)$.

Let $H_X \subseteq \Theta$. The question is, "To what degree is H_X supported by $\mathcal{H}$?" At this juncture, Kohlas generates two concepts in Shafer: *support* and *plausibility*.

1. The *degree of support* $sp(H)$ is the probability that a particular interpretation ω is the correct one:

$$sp(H) = P(\{\omega \in \Omega : \Gamma(\omega) \subseteq H\}).$$

2. The *degree of plausibility* measures the probability that H is not excluded:

$$pl(H) = P(\{\omega \in \Omega : \Gamma(\omega) \cap H \neq \varnothing\}).$$

We can use this formulation to complete the development of Shafer-like systems, but it also has uses in dynamical systems. Let $\mathcal{X} = \{X_i\}$, a sequence of variables, and let $\Theta_i = \Theta_0$ for all i be the state space. For uncertain dynamical systems, we can describe the uncertain transitions as

$$\mathcal{T}_i = (\Omega_i, \mathcal{A}_i, P_i, \Theta_{i-1} \times \Theta_i).$$

Uncertain transitions represent uncertain inclusions (that is, what elements are in the next state). We can similarly define uncertain observation relations. With this, control-theoretic formulations of systems are possible.

Information approach. The second approach offers a model of evidence accumulation.[19,20,22–24] Information is a measure of our knowledge of the system state after an event relative to our overall knowledge of the state. Information can also be described as measuring relevance.[16]

In the following, let G be global assumptions, H be hypotheses, and C be conclusions. Then $P(C \mid H \& G)$ is the conditional probability that the conclusion C obtains given that both the hypotheses H and global conditions G obtain. We first consider the *likelihood* function:

$$L(H{:}C \mid G) = \frac{P(C \mid H \& G)}{P(C \mid G)} , \tag{9}$$

where $L(H : C \mid G)$ means "the likelihood of H in light of event C given global knowledge G." Normal use of the term *information* leads to the interpretation that if the likelihood is 1, we would say there is no information. We define information I by

$$I(H{:}C \mid G) = \log_b L(H{:}C \mid G) . \tag{10}$$

The base b of the logarithms is immaterial, letting us use "natural" units such as decibels and bits.

We want to measure the evidence available on the basis of the information at hand.[20,21] In our case, this is the difference in the information for a state minus the information available for not being in that particular state. In other words, the weight of evidence is

$$\begin{aligned} W(H{:}C \mid G) & \\ &= I(H{:}C \mid G) - I\left(\overline{H}{:}C \mid G\right) \\ &= \log_b \frac{P(C \mid H \& G)}{P(C \mid \overline{H} \& G)} \qquad (11) \\ &= \log_b F(H{:}C \mid G). \qquad (12) \end{aligned}$$

where $\bar{H}$ is the complementary state of H. We introduce the abbreviation

$$\log_b F(H{:}C \mid G) = \log_b \frac{P(C \mid H \& G)}{P(C \mid \overline{H} \& G)} .$$

W is naturally stated as odds. (Recall that *odds* are the ratio of the probability for an event's occurrence to the probability against such an occurrence, and that Bayes theorem is $P(A \mid B) = P(B \mid A)P(A)$.) Using Bayes theorem in its odds form, we get

$$O(H{:}C \mid G) = O(H \mid G)F(H{:}C \mid G)$$

and the log odds as

$$\begin{aligned} &\log_b O(H{:}C \mid G) \\ &\quad = \log_b O(H \mid G) + \log F(H{:}C \mid G). \end{aligned}$$

We take this to be our computational rule for evidence:

$$\begin{aligned} e(H{:}C \mid G) &= e(H \mid G) + \log F(H{:}C \mid G) \\ &= e(H \mid G) + W(H{:}C \mid G). \end{aligned}$$

To make the summation start at zero, we set the evidence at a study's beginning to zero by setting the initial odds at 1; that is, confirmation and disconfirmation are equally likely.

A clear problem with this formulation is that we do not have a single variable but a vector of probabilities. If the matrix used to generate the probabilities were square, we could solve the defining equation for the logarithm. Because matrices are unlikely to be square, we need a pseudologarithm.

Now we can answer the question posed in the introduction: What do the terms *conceptual description and specification* and *real world* mean? I propose that

- A *conceptual description and specification* is a verified model.
- *Real world* implies the existance of a probabilistic knowledge graph.

But saying this does not make it so. I have raised more questions than I have provided answers for, and I continue to explore these concepts. Ultimately, I expect to develop a computer algebra system similar to Maple to manipulate formal model definitions according to these concepts.

Acknowledgments

Support for this work came from the Shodor Education Foundation and the National Science Foundation grant DUE-0127488. I thank one referee who invested a great deal of time in commenting on every aspect of the original draft. This new version is much different in its orientation and message. That same referee provided a second set of detailed criticisms. I take responsibility for all errors. Thanks to Robert M. Panoff, Shodor Education Foundation, Holly Hirst of Appalachian State University, and Daniel D. Warner of Clemson who originally posed the question of what a case study in verification and validation should look like.

References

1. *VV&A Recommended Best Practices Guide*, Defense Modeling and Simulation Office, US Dept. of Defense, Washington, D.C., Nov. 1996.
2. *VV&A Recommended Best Practices Guide*, Defense Modeling and Simulation Office, US Dept. of Defense, Washington, D.C., May 2000.
3. G. Shafer, *A Mathematical Theory of Evidence*, Princeton Univ. Press, Princeton, N.J., 1976.
4. D.E. Stevenson, "Foundations of Validation: The Michelson-Morley Experiment," *Proc. European Simulation Multiconf.*, Soc. Computer Simulation, San Diego, Calif., 1999, pp. 269–275.
5. D.E. Stevenson, "Science, Computational Science, and Computer Science: At a Crossroads," *Comm. ACM*, vol. 37, no. 12, Dec. 1994, pp. 85–96.
6. J.H. Wilkinson, *Rounding Errors in Algebraic Processes*, John Wiley & Sons, New York, 1963.
7. A. Michelson and E. Morley, "Influence of Motion of the Medium on the Velocity of Light," *Am. J. Science*, vol. 31, May 1886, pp. 377–386.
8. A. Michelson and E. Morley, "On the Relative Motion of the Earth and the Luminiferous Ether," *Am. J. Science*, vol. 34, no. 203, Nov. 1887, pp. 333–336.
9. L.S. Swenson Jr., *The Ethereal Aether*, Univ. of Texas Press, Austin, Texas, 1972.
10. K.F. Schaffner, *Nineteenth-Century Aether Theories*, Pergamon Press, Amsterdam, 1972.
11. W.E. Gettys, F.J. Keller, and M.J. Skove, *Physics: Classical and Modern*, McGraw-Hill, New York, 1988.
12. G. Ryle, *The Concept of Mind*, Barnes and Noble, New York, 1949.
13. R. Carnap, "Testability and Meaning," *Philosophy of Science*, vols. 3 and 4, 1936 and 1937. Reprinted in H. Feigl and M. Brodbeck, eds., *Readings in the Philosophy of Science*, Appleton-Century-Crofts, New York, pp. 130–195.
14. K. Popper, *Logik der Forschung (The Logic of Scientific Discovery)*, Basic Books, New York, 1935 (reprinted 1959).
15. T.S. Kuhn, *The Structure of Scientific Revolutions*, Univ. of Chicago Press, Chicago, 1970.
16. D.E. Stevenson, "An Evidence-Based Approach to Fidelity," *Proc. 2001 Summer Computer Simulation Conf.* (SCSC 2001), Soc. for Computer Simulation Int'l, San Diego, Calif., 2002.
17. R.A. Krzysztof and E-R. Olderog, *Verification of Sequential and Concurrent Programs*, Springer-Verlag, New York, 1991.
18. T. Hailperin, "Probability Logic in the Twentieth Century," *History and Philosophy of Logic*, vol. 12, no. 1, Jan. 1991, pp. 71–110.
19. G. Polya, *Mathematics and Plausible Reasoning*, Princeton Univ. Press, Princeton, N.J., 1968.
20. E.T. Jaynes, *Maximum-Entropy and Bayesian Methods*, Kluwer Academic, Dordrecht, Netherlands, 1990.
21. J. Kohlas and P.A. Monney, *A Mathematical Theory of Hints: An Approach to the Dempster-Shafer Theory of Evidence*, Springer-Verlag, New York, 1995.
22. I.J. Good, *Probability and the Weighing of Evidence*, C. Griffin, London, 1950.
23. I.J. Good, *Good Thinking: The Foundations of Probability and Its Application*, Univ. of Minnesota Press, Minneapolis, Minn., 1983.
24. H. Jeffreys, *Theory of Probability*, Clarendon Press, Oxford, UK, 1939.

For more information on this or any other computing topic, please visit our Digital Library at http://computer.org/publications/dlib.

PURPOSE The IEEE Computer Society is the world's largest association of computing professionals, and is the leading provider of technical information in the field.

MEMBERSHIP Members receive the monthly magazine **COMPUTER**, discounts, and opportunities to serve (all activities are led by volunteer members). Membership is open to all IEEE members, affiliate society members, and others interested in the computer field.

BOARD OF GOVERNORS

Term Expiring 2002: Mark Grant, Gene F. Hoffnagle, Karl Reed, Kathleen M. Swigger, Ronald Waxman, Michael R. Williams, Akihiko Yamada

Term Expiring 2003: Fiorenza C. Albert-Howard, Manfred Broy, Alan Clements, Richard A. Kemmerer, Susan A. Mengel, James W. Moore, Christina M. Schober

Term Expiring 2004: Jean M. Bacon, Ricardo Baeza-Yates, Deborah M. Cooper, George V. Cybenko, Wolfgang K. Giloi, Haruhisha Ichikawa, Thomas W. Williams

Next Board Meeting: 8 Nov 02, Boston, MA.

IEEE OFFICERS

President: RAYMOND D. FINDLAY
President-Elect: MICHAEL S. ADLER
Past President: JOEL B. SYNDER
Executive Director: DANIEL J. SENESE
Secretary: HUGO M. FERNANDEZ VERSTAGEN
Treasurer: DALE C. CASTON
VP, Educational Activities: LYLE D. FEISEL
VP, Publications Activities: JAMES M. TIEN
VP, Regional Activities: W. CLEON ANDERSON
VP, Standards Association: BEN C. JOHNSON
VP, Technical Activities: MICHAEL R. LIGHTNER
President, IEEE-USA: LeEARL A. BRYANT

IEEE COMPUTER SOCIETY

EXECUTIVE COMMITTEE

President: WILLIS K. KING*
University of Houston
Dept. of Comp. Science
501 PGH
Houston, TX 77204-3010
Phone: +1 713 743 3349 Fax: +1 713 743 3335
w.king@computer.org

President-Elect: STEPHEN L. DIAMOND*
Past President: BENJAMIN W. WAH*
VP, Educational Activities: CARL K. CHANG *
VP, Conferences and Tutorials: GERALD L. ENGEL*
VP, Chapters Activities: JAMES H. CROSS†
VP, Publications: RANGACHAR KASTURI†
VP, Standards Activities: LOWELL G. JOHNSON (2ND VP)*
VP, Technical Activities: DEBORAH K. SCHERRER(1ST VP)*
Secretary: DEBORAH M. COOPER*
Treasurer: WOLFGANG K. GILOI*
2001–2002 IEEE Division VIII Director: THOMAS W. WILLIAMS
2002–2003 IEEE Division V Director: GUYLAINE M. POLLOCK†
Executive Director: DAVID W. HENNAGE†

*voting member of the Board of Governors

IEEE

COMPUTER SOCIETY WEB SITE

The IEEE Computer Society's Web site, at **http://computer.org**, offers information and samples from the society's publications and conferences, as well as a broad range of information about technical committees, standards, student activities, and more.

COMPUTER SOCIETY OFFICES

Headquarters Office
1730 Massachusetts Ave. NW
Washington, DC 20036-1992
Phone: +1 202 371 0101 • Fax: +1 202 728 9614
E-mail: hq.ofc@computer.org

Publications Office
10662 Los Vaqueros Cir., PO Box 3014
Los Alamitos, CA 90720-1314
Phone: +1 714 821 8380
E-mail: help@computer.org
Membership and Publication Orders:
Phone: +1 800 272 6657 Fax: +1 714 821 4641
E-mail: help@computer.org

European Office
13, Ave. de L'Aquilon
B-1200 Brussels, Belgium
Phone: +32 2 770 21 98 • Fax: +32 2 770 85 05
E-mail: euro.ofc@computer.org

Asia/Pacific Office
Watanabe Building
1-4-2 Minami-Aoyama, Minato-ku,
Tokyo 107-0062, Japan
Phone: +81 3 3408 3118 • Fax: +81 3 3408 3553
E-mail: tokyo.ofc@computer.org

EXECUTIVE STAFF

Executive Director: DAVID W. HENNAGE
Publisher: ANGELA BURGESS
Assistant Publisher: DICK PRICE
Associate Executive Director: ANNE MARIE KELLY
Chief Financial Officer: VIOLET S. DOAN
Director, Information Technology & Services: ROBERT CARE
Manager, Research & Planning: JOHN C. KEATON

23-AUG-2002