Certainty or Intelligence: Pick One!

Edward A. Lee
EECS
University of California at Berkeley
Berkeley, CA, USA
ORCID: 0000-0002-5663-0584

***Abstract*—Mathematical models can yield certainty, as can probabilistic models where the probabilities degenerate. The field of formal methods emphasizes developing such certainty about engineering designs. In safety critical systems, such certainty is highly valued and, in some cases, even required by regulatory bodies. But achieving reasonable performance for sufficiently complex environments appears to require the use of AI technologies, which resist such certainty. This extended abstract suggests that certainty and intelligence may be fundamentally incompatible.**

***Index Terms*—artificial intelligence, formal methods, Bayes rule**

Autonomous systems, by definition, act without immediate human control. When engineering such systems, we require assurance that bad behaviors cannot occur. To achieve such assurance, we combine careful engineering, rigorous testing, and, sometimes, formal methods. It has become clear, however, that many of the autonomous systems we are trying to deploy, such as self-driving cars, cannot work reliably enough without employing AI techniques such as deep neural networks. The engineering of such systems is, in the words of Janelle Shane, more like educating a child than like traditional software engineering [8]. Traditional methods of "careful engineering" prove inadequate. Moreover, rigorous testing is extremely challenging because the scenarios we most care about, which are typically anomalous ones, are difficult to anticipate and extremely difficult to test for. And while formal methods have shown some progress for small AI systems, their applicability to the kinds of systems being deployed is questionable. They certainly can play a role, for example by systematically synthesizing scenarios for training and testing [1]. But they show little promise of providing the sorts of "proofs of correctness" that formal methods strive for, except at rather small scales.

Consider the example where, in San Francisco in October, 2023, a pedestrian was hurdled in front of a self-driving taxi by another accident and ended up under the taxi. The taxi detected the pedestrian and braked hard, but then decided to move out of traffic, dragging the pedestrian pinned under the car some 20 feet and stopping with the rear wheel on top of the pedestrian. The California DMV suspended Cruise's driverless permits and Cruise recalled all its driverless vehicles for redesign.

Legg and Hutter say that "intelligence measures an agent's ability to achieve goals in a wide range of environments." [7]. City roads certainly qualify as a "wide range of environments." Legg and Hutter assume that each environment can be modeled as a computable function, they assign a probability proportional to the complexity involved in computing that function, and then define intelligence as the average degree to which goals are met in a given environment weighted by the probability of that environment occurring.

Here, Legg and Hutter are relying on a foundational assumption, pervasive today among scientists and engineers, that real world environments are, at a fundamental level, computable. I have shown that this hypothesis is untestable by experiment [5], and that it leads to models of the physical world that are awkward and inconvenient [6].

The intuition behind Legg and Hutter's model is nevertheless valid, that intelligence requires adaptability to a variety of environments. Adaptability is a form of learning, an ability to improve based on observations. Cruise's recall represents a form of learning, and here, the AIs have a distinct advantage over humans. The experience of single autonomous system can improve the performance of a whole fleet.

Formal verification promises certainty about a *model*, not about a physical system [4]. The usefulness of that certainty, of course, depends on the likelihood that the physical system will actually conform with the model. But more importantly, it requires us to have a model of the environments in which the system will operate.

Here, we have to distinguish between two uses of models [5]. For a "scientific model," we demand that the model conform with the physical system. For an "engineering model," we demand that the physical system conform with the model. A piece of software is usually an engineering model. Ultimately, it specifies how we would like electrons to flow, and the goal of the electronics industry is to deliver that flow with very high (but never perfect) reliability. Models of an uncertain environment, however, are necessarily scientific models. We cannot design those environments, at least not completely. Because such models will always be approximate and incomplete, it seems we have to incorporate learning into our systems. This means that the system will evolve, it will change in the field, further undermining our certainty that it will not do bad things.

The discipline of formal verification is all about providing assurance in a mathematically rigorous way. A model that is amenable to formal verification is a formal system with clearly stated axioms and assumptions, and assurance is a proof that a certain well-defined class of bad behaviors is inconsistent with the model and the assumptions. The proof is based on the axioms and the rules of some logic. Such a proof provides certainty, meaning that if the assumptions hold, and the logic and axioms are sound, then the model cannot exhibit a behavior in the defined class of bad behaviors.

979-8-3503-4859-0/DATE24/© 2024 EDAA

Certainty, as expressed here, has a plethora of caveats. What makes us believe in the model, the assumptions, the axioms, and the rules of logic? An axiom, by definition, is something we take to be self evidently true. And most of us trust at least some set of rules of logic. If we accept these, we are left with the assumptions and the model as sources of doubt.

Digital electronics has proven astonishingly reliable at carrying out logical operations. A modest microprocessor performs billions of operations per second and can work flawlessly for years. The software that we write for autonomous systems executes on such microprocessors, and itself can be abstracted and analyzed using rules of logic. The software can encode its own models [5], which therefore leaves only the assumptions outside the rigor of logic. In principle, we can have very high confidence in our software, as an engineering model. It will do what is specified with high probability. But this is not enough because of the incomplete modeling of the environment in which it operates.

Any remaining doubt may be captured with probabilities. Here, I take a distinctly Bayesian approach, where probability is a model of what we don't know, not a model of some intrinsic randomness in a system (see [5, Chapter 11]). Suppose that A and B are events. For example, suppose that A is the event that there is a pedestrian under the car, and B is the event that the car hits a pedestrian. Suppose our sensors have told us that the car has hit a pedestrian, i.e., that B has occurred. Then we are interested in $P(A|B)$, the probability that there is a pedestrian under the car given that we have hit a pedestrian. According to Bayes' rule,

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}.$$

$P(A)$ is called our "prior probability," and this formula gives us a way to update that prior to a "posterior probability," $P(A|B)$, the probability that there is a pedestrian under the car given that we have hit a pedestrian. In a Bayesian approach, probabilities are subjective. But notice a key feature of Bayes' rule: if we are *certain* that there is no pedestrian under the car, then our prior probability $P(A) = 0$ (perhaps we have formally verified this, proving for example that no pedestrian would fit under the car). Then the posterior probability $P(A|B)$ is also zero (unless $P(B) = 0$, i.e. we are certain that the car could never hit a pedestrian, an unreasonable choice). Certainty, in this case, makes it impossible to update our prior! The posterior is always equal to the prior when we are certain, regardless of observations. Certainty makes it impossible to learn, impossible to adapt, and hence impossible to be intelligent.

Through Bayesian learning, we can steadily improve our models of the environments in which our systems operate. Does this lead, in the limit, to a model in which we can reach a reasonable level of confidence? The optimism that it does is based on an assumption that many of us take as an axiom and rarely question. We assume that the physical world can be modeled to arbitrary precision.

There is a background paradigm here, one almost universally held, that every action has a cause. Modeling is all about associating causes with results. Fundamentally, this assumption reduces to a deterministic model of the real world [3]. We refuse to accept results without causes. Is this a faith or a truth?

I have previously shown that any set of deterministic models that is rich enough to include Newton's laws and also allows discrete behaviors is incomplete [2]. This is shown by constructing a sequence of deterministic Newtonian models of elastic collisions (hence the discrete behaviors), where each model in the sequence is the deterministic, the sequence is Cauchy in a metric space, and the limit of the sequence does not exist within the set of deterministic models. As a consequence, such a set of deterministic models has "holes" that exhibit nondeterministic behavior.

An immediate corollary is that the best models we have of the physical world have to admit the possibility of actions that have no cause. The "holes" can only be filled in with models where any single cause can lead to any of a multiplicity of behaviors. Although this result is only about models, not about reality, it suggests that our models of reality should, in fact, admit the possibility of uncaused action. Yet we resist this conclusion, even in the face of everyday empirical evidence. The reality is that we cannot identify causes for much of what happens, and yet we dogmatically assume that there is a cause.

The doubt about causation is confirmed by recent theories and experiments towards reconciling quantum mechanics and relativity [9]. In a system subject to both quantum mechanics and relativity, one can construct a scenario where an event A causes another event B, and, simultaneously, B causes A. Causality is not as simple as we thought. And yet, causality is required for certainty.

My (admittedly radical) stance here is that certainty is achievable only at the expense of intelligence. It requires us to reject learning, to reject adaptability, and to accept a model of the world that counters experience.

We demand certainty from engineered, safety-critical systems. We also demand intelligence, that they should behave reasonably even in unanticipated scenarios. These two requirements may be fundamentally incompatible.

References

[1] Fremont, D.J., Kim, E., Dreossi, T., Ghosh, S., Yue, X., Sangiovanni-Vincentelli, A.L., Seshia, S.A.: Scenic: a language for scenario specification and data generation. Machine Learning **112**, 3805–3849 (2022). https://doi.org/10.1007/s10994-021-06120-5

[2] Lee, E.A.: Fundamental limits of cyber-physical systems modeling. ACM Transactions on Cyber-Physical Systems **1**(1) (2016). https://doi.org/10.1145/2912149

[3] Lee, E.A.: Determinism. ACM Transactions on Embedded Computing Systems (TECS) **20**(5), 1–34 (2021). https://doi.org/10.1145/3453652

[4] Lee, E.A., Sirjani, M.: What good are models? In: Formal Aspects of Component Software (FACS). vol. LNCS 11222. Springer

[5] Lee, E.A.: Plato and the Nerd — The Creative Partnership of Humans and Technology. MIT Press (2017)

[6] Lee, E.A.: The Coevolution: The Entwined Futures of Humans and Machines. MIT Press, Cambridge, MA (2020)

[7] Legg, S., Hutter, M.: A universal measure of intelligence for artificial agents. In: International Joint Conference on Artificial Intelligence (IJCAI). pp. 1509–1510. Lawrence Erlbaum, http://www.ijcai.org/papers/post-0042.pdf

[8] Shane, J.: You look like a thing and I love you. Hachette, United Kingdom (2019)

[9] Wolchover, N.: Quantum mischief rewrites the laws of cause and effect. Quanta Magazine (2023), https://www.quantamagazine.org/quantum-mischief-rewrites-the-laws-of-cause-and-effect-20210311/