Assessing the Foundations of Conscious Computing:
A Bayesian Exercise

Eric Horvitz

June 2001

Questions have long been posed with about whether we might one day be able to create computational systems that *experience* subjective states similar to those experienced by people. The challenge of understanding and explaining conscious experience extends well beyond discussions among philosophers. Many people, including AI scientists, pursuing principles of automated reasoning and decision making, and neurobiologists and psychologists, with interests in nervous systems and cognition, reflect about the nature of subjective states.

Given the absence of scientific methods for probing the principles and machinery of such conscious experience, beliefs about the nature and requirements of feeling and experiencing are diverse. When topics touching on *consciousness, awareness, subjectivity, feeling* and *qualia* arise in conversation, people are faced with a spectrum of beliefs about the foundations of subjective experience, and on related beliefs about the feasibility of one day developing computational algorithms and architectures that could achieve what we might call *conscious computing*. Such a diversity in beliefs appears extends to scientists with expertise in cognition, including decision scientists, cognitive psychology, neurobiology, and artificial intelligence.

To promote discussion, I've found it useful to assess the strength of belief that people have about different explanations for conscious experience. Rather than seek commitment to a single hypothesis, I find it valuable to take a Bayesian perspective on assessing beliefs under uncertainty. With such an approach, we lay out a set of distinct possibilities about the basis for conscious experience, engage people about clarifying and refining the hypotheses, and then assessing the likelihood of each hypothesis as being the true explanation. Such an assessment can be useful for revealing common patterns of belief and for focusing discussion.

The assessment of an explicit set of probabilities across a space of possibilities that provide conceptual coverage at a broad level of abstraction supports the admission of multiple possibilities with different degrees of belief, and helps to frame questions about the kinds of experiments, results, models, and observations that might be useful for updating the probabilities.

So, let's formulate a set of broad possibilities, attempting to define a candidate space of mutually exclusive and exhaustive set of possibilities. We'll work at a high-level of abstraction and talk about broad categories that could house more detailed hypotheses as further refinements. We'll force the set of possibilities to be exhaustive by considering an explicit *Other* category, containing all possible explanations that don't fit into one of the other broad hypotheses.

As an exercise, I'll ask readers to reflect about the probability that the correct explanation for conscious experience lays within each hypothesis. These assessed probabilities represent individuals' beliefs about the basis for consciousness, expressed as likelihoods of the ultimate truth of each of the categories. One way to view such probability assessments is to imagine that a

clairvoyant who knows the actual truth will soon reveal the nature of consciousness. The likelihoods represent the probabilities of each hypothesis being the one pointed out as being the right explanation by the clairvoyant.

Theological Foundations
Consciousness is founded on constructs drawn from theology. *E.g.,* Our awareness and sense of self extends from a notion of *soul* that might be said to reside outside of the scientific universe. Consciousness is linked to, extends from some grander religious goals, mission.

Other
Something else is going on, distinct from other explicit hypotheses, including various combinations, redefinitions, shifting of borders, interpretations of the other categories.

Information Processing
Our self awareness and subjective experience is generated by some aspect of information processing, potentially centered on real-time sensory integration and reasoning.

Mysterious Fabric
Consciousness arises via some fundamental properties of the universe, as yet poorly understood, that was leveraged through evolutionary processes. We may come to one day understand the properties via ongoing efforts in physics and neurobiology, etc. *E.g.,* It may be based in some fundamental aspect of quantum physics, or another deep aspect of energy and matter in space time.
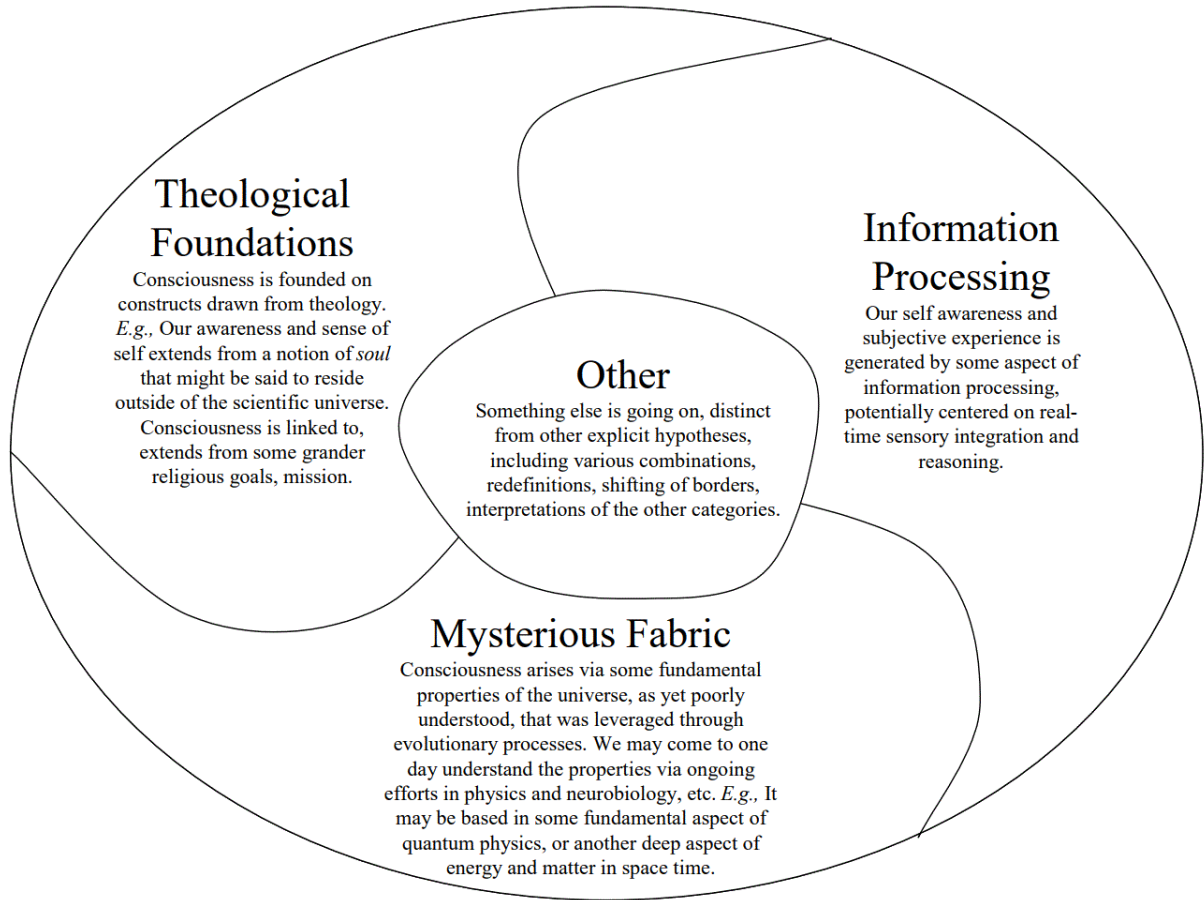
Figure 1. Formulation of a space of hypotheses about subjective experience. After seeking agreement on the essence of each hypothesis, we can assess beliefs, in terms of the differing likelihoods of a mutually exclusive and exhaustive set of hypotheses.

Let us now explore a set of candidate hypotheses. Figure 1 captures a potential space of possible explanations for conscious experience. These definitions are meant to conjure up a "conceptual centroid," capturing the essence of each explanation class. It's clear that there is some potential difference of interpretation of the hypotheses, and so, to be more precise, people should have the opportunity during communications about beliefs to refine and personalize the definitions of the hypotheses and to introduce new categories of explanation. Let's review each of a set of proposed hypotheses.

Theological Foundations includes the set of explanations that view conscious entities as having foundations in the spiritual world, outside of or in distinction to the mechanistic, scientific universe studied by physicists, chemists, and biologists. These explanations implicitly or explicitly pose conscious experience as arising as an essential property of souls or other experiencing entities founded on what one might refer to as classical theological or spiritual

constructs, as captured, for example, by the notions of individuals in Buddhist, Christian, Islamic, Hindu, Jewish, and other organized religions.

Information Processing covers the set of hypotheses that consider subjective experience as being generated by some aspect of information processing supported by a computational system. As an example, the hypothesis includes the assertion that subjective experience is created by, or associated with, patterns of coordinated information processing that support the real-time capabilities for sensory integration, reasoning, and action demonstrated by the human nervous system. People with strong beliefs about the *Information Processing* hypothesis tend to believe that conscious experience is based on some form of algorithmic activity, whether that form of algorithmic activity is performed by cell-based nervous activity, a set of silicon memory units and gates, a universal Turing machine, or a contraption built from Styrofoam balls, wood sticks, and rubber bands.

Mysterious Fabric spans the set of hypotheses that explain consciousness as hinging critically on some as yet poorly understand property or properties of the physical universe that was discovered and leveraged through evolutionary processes. The essential nature of the property and its role in subjective experience may one day be revealed via ongoing efforts in physics and neurobiology. Such a property may be metaphorically or in reality related to some fundamental aspect of quantum physics, or another property of energy, matter in space time.

Other refers to the set of all explanations that are not captured by the explicitly defined categories. That is, we ask people to assess the likelihood that "something else is going on," significantly distinct from the other, explicit categories. *Other* includes explanations that fall outside of the explicit hypotheses, include explanations that represent reformulations, extensions, or combinations of the definitions of the explicit categories—given personal interpretations of the definitions of the explicit hypotheses.

Notice that we can refine the hypotheses space, by breaking out as distinct hypotheses, more precise, finer-grained explanations. For example, as portrayed in Figure 2, we refine *Information Processing* into *IP-Competency* and *IP-Leap*.

IP-Competency is a refinement of *Information Processing* that explains conscious experience as being based in information processing, and asserts that such subjective states are an unavoidable aspect of increasing cognitive competency. That is, what we call *consciousness* is unavoidable as real-time sensory integration and information processing become increasingly competent. From the point of view of *IP-Competency*, searching for a special word or concept like *consciousness* is chasing an illusory "distinct" notion. Our self awareness and subjective experience is simply the way entities with our abilities necessarily must feel. Should an explanation in the *IP-Competency* category be true, there would be nothing explicit to discover about conscious experience with the increasing competency of reasoning systems.

IP Competency

Searching for a special word or concept like *consciousness* is chasing an illusory "distinct" notion. Our self awareness and subjective experience is simply the way creatures with our abilities necessarily have to feel. That is, what we call *consciousness* is unavoidable as real-time sensory integration and information processing become increasingly competent.

Theological Foundations

Consciousness is founded on constructs drawn from theology. *E.g.,* Our awareness and sense of self extends from a notion of *soul* that might be said to reside outside of the scientific universe. Consciousness is linked to, extends from some grander religious goals, mission.

Other

Something else is going on, distinct from other explicit hypotheses, including various combinations, redefinitions, shifting of borders, interpretations of the other categories.

IP Leap

Consciousness is generated by, or is aa emergent correlate of an interesting, specific innovation or set of innovations with information processing. When revealed, we would likely find the principles and machinery interesting and surprising as distinct, novel approach to information processing.

Mysterious Fabric

Consciousness arises via some fundamental properties of the universe, as yet poorly understood, that was leveraged through evolutionary processes. We may come to one day understand the properties via ongoing efforts in physics and neurobiology, etc. *E.g.,* It may be based in some fundamental aspect of quantum physics, or another deep aspect of energy and matter in space time.
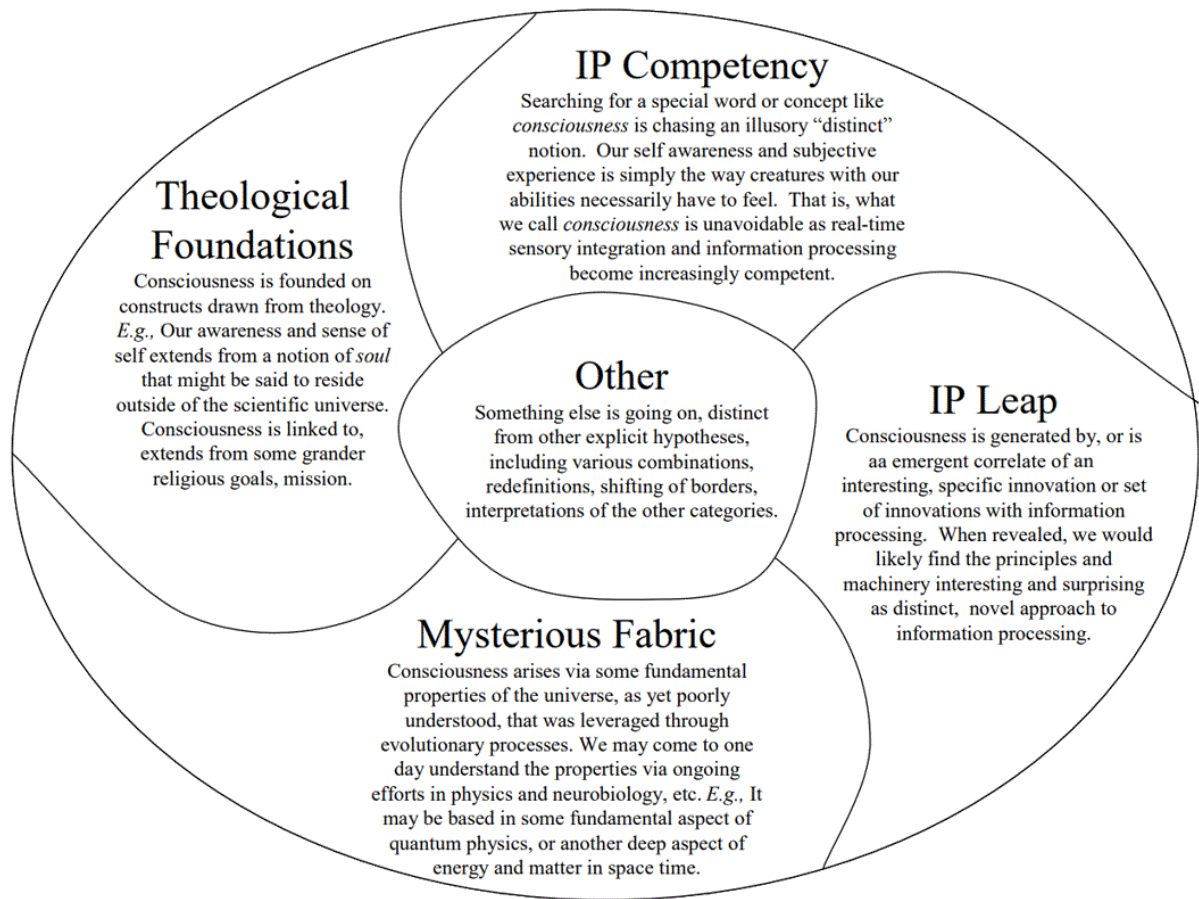
Figure 2. Refinement of the space of hypotheses. In this case, we refine *Information Processing* into the *IP Leap* and *IP Competency* classes

IP-Leap is a refinement of *Information Processing* that asserts that subjective experience is based on special, potentially interesting and surprising information processing. *IP-Leap* explanations assert that conscious experience is the result of some interesting novel information-processing analysis, coordination, integration, architecture, etc., that endows nervous systems with valuable capabilities that would not necessarily be obtainable or obtainable with the same overall architecture or resource limitations without such special processing or machinery. *IP-Leap* explanations suggest that there are some discoverable analytical and/or architectural aspects of cognitive systems that would enhance our insights about understanding and potentially synthesizing systems capable of supporting conscious experiences.

In performing assessment, we can identify different patterns of belief among the population. Figures 3 through 5 capture several classes of belief about the foundations for subjective experience.

Figure 3 captures the beliefs of an individual who finds consciousness mysterious, with an assessment that the explanation is likely founded on some as yet poorly understood property of the universe. I have found that many people with strong beliefs that consciousness likely is linked to some, as yet poorly understood, essential property of the physical universe, often cite an explanation that has "something to do with quantum physics." In this case, a person asserts that

the next most likely hypothesis is *Other*, or something outside of the explicit hypotheses, followed by *IP Leap* hypothesis, and then *IP Competency*, followed last by *Theology*.
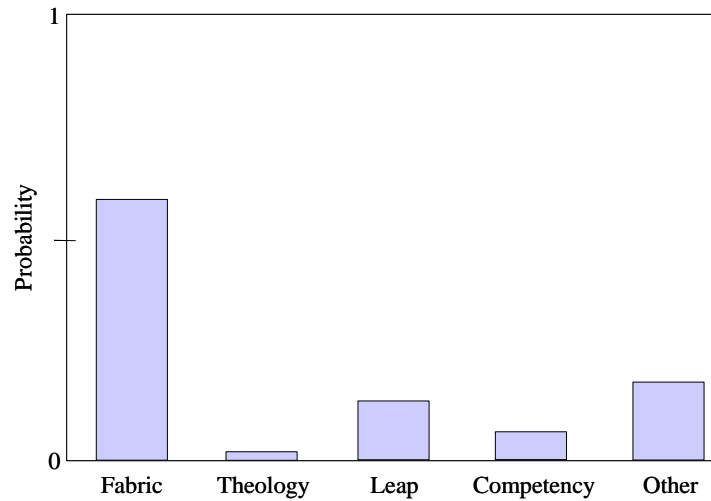
Figure 3. A pattern of belief with weight on *Mysterious Fabric* hypotheses.
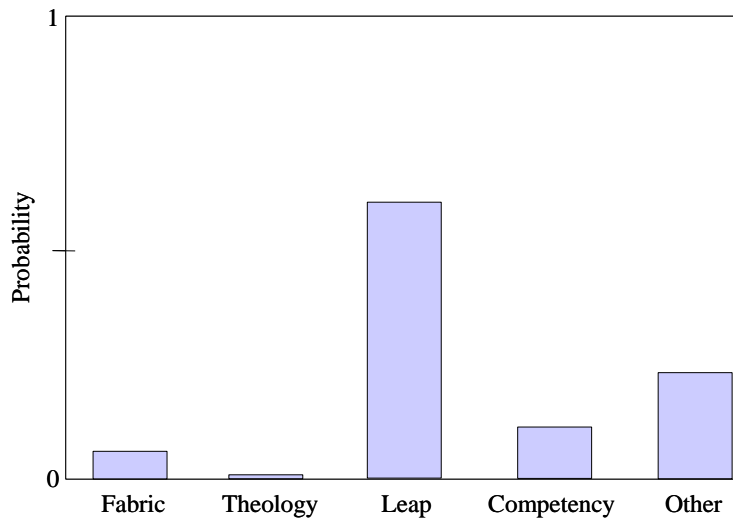
Figure 4. A pattern of belief with weight on the *Information-Processing* hypotheses, with a leaning toward *IP Leap*.

Figure 4 captures a pattern of beliefs leaning towards the conscious experience as a very special kind of information processing, followed by some probability that something else is going on, with an even less likelihood that the subjective world is an inescapable aspect of the intellectual competency associated with human (and potentially other) nervous systems. Although this person gives some probability to the *Mysterious Fabric* hypothesis, only a small probability is assigned to theological hypotheses.
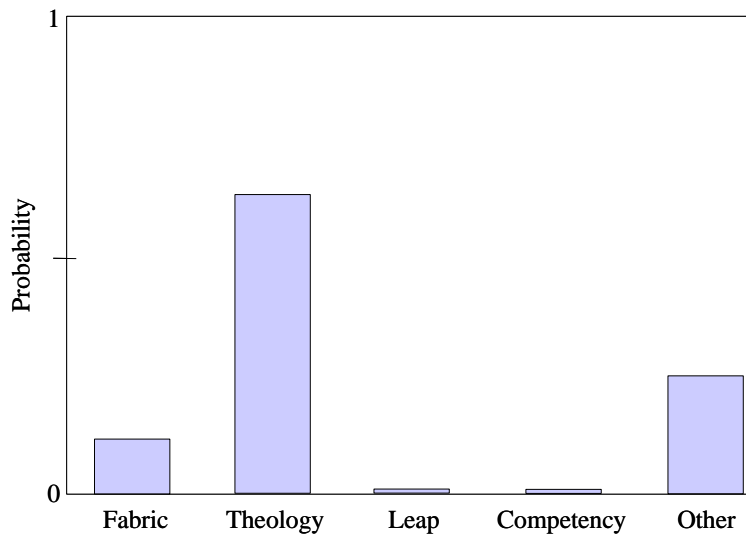
Figure 5. A pattern of belief with weight on hypotheses in the *Theological Foundations* category.

Figure 5 represents beliefs of a person with heaviest weight on the probability that subjective experience is founded in theology. Although, the person has assigned the highest probability to the *Theological Foundations* hypothesis, significant belief is also assigned to the *Mysterious Fabric* hypothesis, and to the *Other* category, expressing the probability that the explanation for conscious experience is rooted in a hypothesis outside of those expressed in Figure 2. This person largely rejects the *Information Processing* hypotheses.

Some readers may find that the categories of explanation do not cover the space of possibilities in a satisfying way. They may wish to modify the categories presented and recommend reformulating the categories or introducing new candidate explanations. Working to define a revised set of mutually exclusive and exhaustive categories can be a useful exercise.

What are your own beliefs about the foundations of our conscious experience? How would you apportion probabilities that each of the distinct explanations mentioned above (or a set of hypotheses of your own) above is the correct one? As we mentioned above, one approach to assessing the probabilities is to imagine that you encounter an omniscient clairvoyant who knows the truth about consciousness. For each category of explanation under consideration, assess the probability that the clairvoyant will soon tell you that the category is the correct explanation.

Appendix: Probability Assessment Tool

Here is a link to a tool created to enable readers to assess the probabilities of the hypotheses discussed. You have the option of sharing these assessments as part of a survey that aggregates the information into summaries of the probabilities that people assign to different explanations for conscious experience.