Issues in Applying Probability Theory to AGI

Pei Wang

Temple University, Philadelphia PA 19122, USA
http://www.cis.temple.edu/∼pwang/

Abstract. This paper analyzes the three most common interpretations of probability (frequentist, subjective, and logical) in the context of AGI, and argues that probability theory cannot serve a central role in the theoretical foundation of AGI.

1 Probability and Its Interpretations

Probability theory is the most mature formal model for uncertainty representation and calculation, and has been successfully applied to solve many problems in various domains. In particular, it has played a central role in some representative models in AI [1, 2], Cognitive Science [3, 4], as well as AGI [5, 6]. Even so, there are some long-standing issues that have not obtained enough attention. Consequently, probability theory is often misused, and the research results are not properly justified. I have analyzed some related issues previously [7–10], and will approach the problem from a different angle in this paper, by focusing in the interpretation of "probability" with respect to AGI.

A probability function P is defined on a sample space Ω by assigning each event A (represented by a subset of Ω) a real number in $[0, 1]$. $P(A)$ is called "the probability that event A occurs", and the function is based on axioms $P(\Omega) = 1$ and $P(A \cup B) = P(A) + P(B)$ if A and B are disjoint [11, 12].

As a formal model, probability theory specifies the (valid) calculations within a probability function, without saying much about how to apply the theory to a practical situation. For this purpose, the notion of "probability" needs to be *interpreted*, by mapping the probability function into a concrete measurement in the domain. The best known interpretations of probability are [13, 14]:

Frequentist interpretation. $P(A)$ is the limit of the *occurrence frequency* of event A in a sequence of observations.

Subjective interpretation. $P(A)$ is the *degree of belief* of a cognitive system on belief A at a certain moment.

Logical interpretation. $P(A)$ is the *degree of support or confirmation* hypothesis A obtains from a body of evidence.

One common feature of all of the above interpretations is that $P(A)$ is not an intrinsic property of A, but is *relative*, with respect to a sequence of observations, a cognitive system at a certain time, or a body of evidence. Without such a "reference point", $P(A)$ cannot be meaningfully decided. One problem

in the discussions is that "probability" is used without a clear specification of the reference point, which often lead to implicitly changing of reference. Since different reference points often assigns different probability values to the same event, such changes are invalid in the discussions.

Even worse, in some discussions different interpretations are implicitly mixed with each other. Of course, different interpretations are related to each other, but it does not mean that they can be directly mixed together.

Even if an interpretation of probability is clearly specified and faithfully followed, there are still serious issues that prevent probability theory from playing a central role in AGI. In the following, I will analyze the issues encountered by each of the above three interpretations in AGI.

2 The Frequentist Interpretation

The frequentist interpretation of probability is the standard interpretation used in mathematical statistics. Now the problem is that whether an AGI system can be built using this idea. Since a major function of intelligence is to predict the future, some people may expect an AGI to find the probability of *any* event.

Now the task is to calculate $P_S(A)$, the limit of the occurrence frequency of an arbitrary event A in a sequence S. According to the usual expression, the sample space Ω is the set of all possible and distinct outcomes of some experiment, and each *event* is a subset of Ω. Though such a setting is natural for some simple and well-defined experiments (such as throwing dice or picking balls out of an urn), many experiments or observations do not have such well-defined results. Can we list all possible consequences of a government policy? What is the sample space of an AGI that has to deal with "unanticipated situations"?

One approach is not to categorize the phenomena in the outside world, but focus on what the system can perceive from the outside world. It is a common practice to represent the system's experience as a stream of *percepts* that are from a constant set Ω [2]. Considering the uncertainty in the environment, it seems natural to assign probability to the occurring of each perception. However, since the units of perception are not always the individual percepts, but often the *patterns* formed by the percepts, the probability values are no longer in a function defined on Ω but on Ω^n (if a pattern can be simply represented as a Cartesian product of n percepts). When the environment is described at multiple levels of granularity, the descriptions will correspond to n values that differ in magnitude. It will be hard to get the probability values at all these levels.

There is a more fundamental question: how can we know that an event under discussion *has* a probability? After all, not all sequences converge to a limit, and in principle it is impossible to decide whether a sequence has a limit only from finite observations. For instance, the price of a stock changes from time to time, but to take it as a "random number" that follows a (unknown) distribution is not a self-evident postulate. If in a world all the events are indeed governed by (known or unknown) probability distributions, that world must be very boring,

since its future is the same as its past, stochastically speaking. I do not see any valid reason to assume such a world for us, or for AGI.

Another issue of this interpretation is that the sample space and the events are not always clearly defined. Whether a television set is a piece of furniture does not have a clear "yes/no" answer, and the uncertainty in the statement cannot be naturally taken as the frequency for a television set to be furniture in a sequence of experiments. A related issue is the probability of a unique and unrepeatable accident, such as the "chance" for a specific patient to have a specific disease. The common practice is to take the patient as an instance of a reference class containing similar cases, but in that way a single case can often be taken as an instance of multiple reference classes, leading to inconsistent probability assignments [15, 16]. Similarly, it is unclear how to assign probability, under a frequentist interpretation, to abstract statements like "The world is a Turing machine", since it is unclear that what experiment should be taken, and how to interpret the outcomes.

In summary, the frequentist interpretation can only be used in limited problems, so cannot serve a central role in an AGI system.

3 The Subjective Interpretation

Partly as an attempt to go beyond the restriction of frequentist interpretation, the "subjective interpretation" (also called "subjectivistic interpretation") of probability is more "liberal" on how probability can be decided. In its original form, this interpretation takes the probability of a statement as the degree of belief of a system on the statement. It is "subjective" since different systems can legally assign different probability values to the same statement, and there is no objective standard to judge which one is the "true probability" of the statement. Under this interpretation, a system can assign a probability value to any statement.

Here "subjective" does not mean "arbitrary". This interpretation does not specify where these degrees of belief come from, but demand them to form a *consistent probability distribution*, as well as to be manipulated according to probability theory. In particular, when new information become available, the degrees of belief should be updated according to Bayes' Theorem, so this approach is often referred to as "Bayesianism". A well-known justification of this approach is provided by [17], where the probability axioms are derived from a set of intuitively reasonable assumptions, so it suggests that a rational agent should manage its beliefs according to probability theory.

Compared to the other interpretations, the subjective approach demands the least — it allows a system to believe in anything to any degree, as far as the beliefs are consistent (also called *coherent*), in the sense specified in probability theory. Remarkably, from the viewpoint of AGI, this position can be attacked from the opposite directions: the demand can be criticized as either "too weak" or "too strong", for different reasons.

Since an AGI system must be implemented in a computer, the design specification cannot say "the system can hold any beliefs as far as they are consistent", but has to provide a concrete way to assign probability to statements. Especially, the Bayesian updating process must start from a prior distribution, which the subjective interpretation does not provide. Several solutions have been proposed. One is to start with a "non-informative prior" as a place-holder, then let the system "learn" from incoming "evidence" by revising its beliefs accordingly. This approach will be analyzed in the following section, because it effectively treats probability as degree of evidential support.

Another approach that is especially influential in the AGI community is to use the "universal prior" proposed by Solomonoff [18] that assigns a prior probability to a statement according to its "algorithmic complexity". This idea is often justified as a form of "Occam's Razor" — though a Solomonoff prior does not "choose" a single hypothesis, it does give it the highest probability, among the hypotheses consistent with the observations. One problem of this approach is that intuitively "probability" is not a measurement of "preference" or "priority" in general, but "likelihood" or "plausibility" in particular. When choosing among competing hypotheses, we do prefer one that is both plausible and simple, but it does not mean that these two dimensions are correlated or exchangeable. Occam's Razor should be taken as "Simple hypotheses are preferred", but not "Simple hypotheses are preferred because they are more likely to be confirmed". It is fine for an AGI system to be based on the latter assumption, but it is not a self-evident postulate, but an *assumption* that may be rejected. An arguable analogy is that we prefer products that are both cheap and with high quality, but it does not mean we believe in general cheaper products will have higher quality — actually, if we have to use price to predict quality, many people may assume the opposite, that is, cheaper ones usually have lower quality.

Since the subjective interpretation only demands the consistency of the degrees of belief, to say it is "too strong" means to doubt the necessity of the consistency requirement among beliefs. Traditionally, this requirement is justified by the "Dutch Book" argument, which says that inconsistent beliefs lead to sure loss in certain betting situations [19]. However, this argument only suggests that consistency is *highly desired*, but not that it is *practically achievable*. It can be argued that for a system with finite information-processing capacity while still opening to novel situations and making real-time responses, it is impossible for the system to achieve and maintain the consistency of its beliefs as demanded by probability theory [9]. This issue will be further explored in the following.

4 The Logical Interpretation

The logical interpretation of probability is sometimes confused with the subjective interpretation, since according to both of them, different systems may legally have different opinions on the probability of the same statement. However, according to the subjective interpretation this difference does not need an explanation, while according to the logical interpretation this difference must be

caused by the systems' different evidence collections — if the systems all have the same evidence for a given statement, they should assign the same probability to the statement. According to the logical interpretation, the relation between belief and evidence is a logical relation that does not depend on any system's opinion, so there is no "subjectivity" beyond different evidential basis.

With this clarification, probability under logical interpretation seems to be exactly what AGI needs: it is a degree of belief indicating the extend of evidential support the statement has, or "Believing on the basis of the evidence" [20]. It has neither the restrictive nature of the frequentist interpretation nor the arbitrary nature of the subjective interpretation. Now the problem becomes: can such a probability be defined and maintained among the beliefs of an AGI system?

Carnap proposed a logical foundation for probability, in which probability value $P(H|E)$ indicates the extent of "$E \rightarrow H$" as the degree of "partial implication" [21]. Though mathematically elegant, this approach can only be applied to certain highly idealized domains, so cannot be depended on by an AGI.

A more common attitude can be found in the "Neo-Bayesianism", where probability is simply referred to as degree of belief that is based on available evidence and is revised by new evidence via Bayesian conditioning [22]. However, it does not explicitly specify how the prior probability raises from evidence. To use Solomonoff prior here is problemic, since, as argued previously, it is not based on a logical relation between evidence and hypothesis.

As I have argued in detail in [7, 8], a common conceptual confusion of this approach is between the background knowledge K on which a probability function $P_K(x)$ is established and the condition C of a conditional probability $P(x|C)$. The former is often written as $P(x|K)$, which suggests that the background knowledge can be treated as a condition, so can be learned through Bayesian conditioning. This treatment is not always valid, because background knowledge often cannot be explicitly represented as statements on which $P(x)$ is defined. In general, the background knowledge behind a probability distribution cannot be learned or revised using Bayesian conditioning. For this reason, the Bayesian approach cannot satisfy the learning requirements of AGI systems.

To fully discuss this issue is impossible within the length restriction of this paper. Here I can only briefly mention a solution proposed in my NARS system. In NARS, the truth-value of each belief is a pair of real numbers $\langle f, c\rangle \in [0, 1] \times (0, 1)$, and it indicates the evidential support of the statement [9, 10]. Based on a definition of the evidence of a statement, the first component of the truth-value, *frequency*, is defined as the proposition of positive evidence among all evidence, and the second component, *confidence*, is defined as the proportion of current evidence among all evidence at an *evidential horizon*. Based on this "experience-grounded" definition of truth-value, the inference rules are defined, with truth-value functions that decides the truth-value of a conclusion according to the evidence provided by the premises.

Defined in this way, the *frequency* factor in the truth-value is intuitively similar to probability under the frequentist interpretation, except in NARS the *frequency* is completely about the *past*, and no assumption is made about the

existence of a limit. Furthermore, in the system each belief has its own evidential basis, so the truth-values do not form a consistent probability distribution. Instead, the system is designed under the assumption that inconsistencies among beliefs are *inevitable*, though they are not desired. Whenever an inconsistency is recognized, the system always tries to resolve it by pooling the evidence in a summarizing conclusion, though it cannot guarantee the consistency of all the beliefs, nor that the system will never make mistake for this reason.

Consequently, even though a truth-value in NARS has certain similarities with a probability value, the truth-value functions are not derived from probability theory (though some of them agree with certain formulas in probability theory), but are designed as extended Boolean operators [9]. It is the case because probability theory does not tell us how to handle the situation where each belief has its own *distribution function* (not its own *value* within the same distribution function). In other words, the logical interpretation of probability assumes all the degree of beliefs are values of $P_K(x)$, while in NARS they are like values of $P_{K_i}(x)$, where each evidential basis K_i defines a separate probability distribution.

For this reason, the truth-value in NARS shares certain common properties with probability, though the system as a whole is not "based on probability theory". There are AGI systems that mix probabilistic calculations with "heuristics", and are proposed and justified as extensions or approximations of probability theory [6]. In general, I agree that an application of a formal model is usually more or less an approximation, but for the application to be valid, an *approximation ratio* must be provided. In the current case, I do not think a system can be considered as "based on probability theory" if even the axioms of the theory are not followed. Furthermore, if some axiom of a theory is not applicable to a situation, then the theory should not even be taken as an ideal model to be approximated.

5 Conclusions

After analyzing all the three common interpretations of probability, my conclusion is that an AGI system cannot be mainly based on probability theory.

This conclusion does not mean there is anything wrong in probability theory as a branch of mathematics, and nor does it exclude the possibility for probability theory to be used in an AGI system for some tasks here or there (e.g., NARS uses probability theory in resource allocation). However, it challenges the approaches that treat the world as consisting of events following a probability distribution, or the mind as consisting of beliefs following a probability distribution.

As I have argued in [9] and many other places, "intelligence" means *to adapt with insufficient knowledge and resources*. Probability theory does not fully assume this insufficiency. In probabilistic models, whether a future event will occur is uncertain, but its "occurring probability" is certain. Furthermore, all the probability values must be consistent, and the system can always afford the expense to carry out the calculation demanded by probability theory (though the resource

issue is not discussed in this paper). In an AGI system, these assumptions cannot be satisfied, since it usually has neither the knowledge nor the resources to maintain a consistent probability distribution as a summary of evidence.

References

1. Pearl, J.: Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann Publishers, San Mateo, California (1988)
2. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. 3rd edn. Prentice Hall, Upper Saddle River, New Jersey (2010)
3. Anderson, J.R.: The Adaptive Character of Thought. Lawrence Erlbaum Associates, Hillsdale, New Jersey (1990)
4. Tenenbaum, J.B., Kemp, C., Griffiths, T.L., Goodman, N.D.: How to grow a mind: Statistics, structure, and abstraction. Science **331**(6022) (March 2011) 1279–1285
5. Hutter, M.: Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability. Springer, Berlin (2005)
6. Goertzel, B., Iklé, M., Goertzel, I.F., Heljakka, A.: Probabilistic Logic Networks: A Comprehensive Framework for Uncertain Inference. Springer, New York (2008)
7. Wang, P.: Belief revision in probability theory. In: Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, San Mateo, California (1993) 519–526
8. Wang, P.: The limitation of Bayesianism. Artificial Intelligence **158**(1) (2004) 97–106
9. Wang, P.: Rigid Flexibility: The Logic of Intelligence. Springer, Dordrecht (2006)
10. Wang, P.: Formalization of evidence: A comparative study. Journal of Artificial General Intelligence **1** (2009) 25–53
11. Kolmogorov, A.N.: Foundations of the Theory of Probability. Chelsea Publishing Company, New York (1950)
12. Dekking, F.M., Kraaikamp, C., Lopuhaa, H.P., Meester, L.E.: A Modern Introduction to Probability and Statistics. Springer, London (2007)
13. Kyburg, H.E.: The Logical Foundations of Statistical Inference. D. Reidel Publishing Company, Boston (1974)
14. Hájek, A.: Interpretations of probability. In Zalta, E.N., ed.: The Stanford Encyclopedia of Philosophy. (2011)
15. Kyburg, H.E.: The reference class. Philosophy of Science **50** (1983) 374–397
16. Wang, P.: Reference classes and multiple inheritances. International Journal of Uncertainty, Fuzziness and and Knowledge-based Systems **3**(1) (1995) 79–91
17. Cox, R.T.: Probability, frequency, and reasonable expectation. American Journal of Physics **14** (1946) 1–13
18. Solomonoff, R.J.: A formal theory of inductive inference. Part I and II. Information and Control **7**(1-2) (1964) 1–22,224–254
19. Ramsey, F.P.: Truth and probability. In Braithwaite, R.B., ed.: The Foundations of Mathematics and other Logical Essays. Brace & Co. (1926) 156–198
20. Kyburg, H.E.: Believing on the basis of the evidence. Computational Intelligence **10** (1994) 3–20
21. Carnap, R.: Logical Foundations of Probability. The University of Chicago Press, Chicago (1950)
22. Pearl, J.: Jeffrey's rule, passage of experience, and Neo-Bayesianism. In Kyburg, H.E., Loui, R.P., N., C.G., eds.: Knowledge Representation and Defeasible Reasoning. Kluwer Academic Publishers, Amsterdam (1990) 245–265