Definability of Truth in Probabilistic Logic
(Early draft)

Paul Christiano[*] Eliezer Yudkowsky[†] Marcello Herreshoff[‡]
Mihaly Barasz[§]

June 10, 2013

1 Introduction

A central notion in metamathematics is the *truth* of a sentence. To express this notion within a theory, we introduce a predicate True which acts on quoted sentences $\ulcorner\varphi\urcorner$ and returns their truth value $\text{True}\left(\ulcorner\varphi\urcorner\right)$ (where $\ulcorner\varphi\urcorner$ is a representation of φ within the theory, for example its Gödel number). We would like a truth predicate to satisfy a formal correctness property:

$$\forall\varphi : \text{True}\left(\ulcorner\varphi\urcorner\right) \iff \varphi. \tag{1}$$

Unfortunately, it is impossible for any expressive language to contain its own truth predicate True. For if it did, we could consider the liar's sentence G defined by the diagonalization

$$G \iff \text{True}\left(\ulcorner\neg G\urcorner\right).$$

Combining this with property 1, we obtain:

$$G \iff \text{True}\left(\ulcorner\neg G\urcorner\right) \iff \neg G,$$

a contradiction.

There are a few standard responses to this challenge.

The first and most popular is to work with meta-languages. For any language L we may form a new language by adding a predicate True, which acts only on sentences of L (i.e., sentences without the symbol True) and satisfies property 1. We can iterate this construction to obtain an infinite sequence of languages $L_1, L_2, \ldots$, each of which contains the preceding language's truth predicate.

[*] UC Berkeley. Email: paulfchristiano@eecs.berkeley.edu
[†] Machine Intelligence Research Institute
[‡] Google
[§] Google

A second approach is to accept that some sentences, such as the liar sentence G, are neither true nor false (or are simultaneously true and false, depending on which bullet we feel like biting). That is, we work with a many-valued logic such as Kleene's logic over $\{\text{True}, \text{False}, \perp\}$. Then we may continue to define True by the strong strong reflection property that the truth values of φ and True $(\ulcorner \varphi \urcorner)$ are the same for every φ. We don't run into trouble, because any would-be paradoxical sentence can simply be valued as $\perp$.

Although this construction successfully dodges the "undefinability of truth" it is unsatisfying in many respects. There is no way to test if a sentence φ is undefined (any attempt to do so will itself be undefined), and there is no bound on the number of sentences which might be undefined. In fact, if we are specifically concerned with self-reference, then many sentences of interest (and not just pathological counterexamples) may be undefined.

In this paper we show that it is possible to perform a similar construction over probabilistic logic. Though a language cannot contain its own truth predicate True, it can nevertheless contain its own "subjective probability" function $\mathbb{P}$. The assigned probabilities can be reflectively consistent in the sense of an appropriate analog of the reflection property 1. In practice, most meaningful assertions must already be treated probabilistically, and very little is lost by allowing some sentences to have probabilities intermediate between 0 and 1. (This is in contrast with Kleene's logic, in which the value $\perp$ provides little useful information where it appears.)

2 Preliminaries

2.1 Probabilistic Logic

Fix some language L, for example the language of first order set theory. Fix a theory T over L, for example ZFC. We are interested in doing "probabilistic" metamathematics within L, and so in addition to assuming that it is powerful enough to perform a Gödel numbering, we will assume that some terms in L correspond to the rational numbers and have their usual properties under T. (But T will still have models in which there are non-standard rational numbers, and this fact will be important later.)

We are interested in assignments of probabilities to the sentences of L. That is, we consider functions $\mathbb{P}$ which assign a real $\mathbb{P}(\varphi)$ to each sentence φ. In analogy with the logical consistency of an assignment of truth values to sentences, we are particularly interested in assignments $\mathbb{P}$ which are internally consistent in a certain sense.

We present two equivalent definitions of *coherence* for functions over a language L:

Definition 1 (Coherence). *We say that $\mathbb{P}$ is* coherent *if there is a probability measure μ over models of L such that $\mathbb{P}(\varphi) = \mu(\{\mathcal{M} : \mathcal{M} \models \varphi\})$.*

Theorem 1 (Equivalent definition of coherence). *$\mathbb{P}$ is coherent if and only if the following axioms hold:*

1. *For each φ and ψ, $\mathbb{P}(\varphi) = \mathbb{P}(\varphi \wedge \psi) + \mathbb{P}(\varphi \wedge \neg\psi)$.*
2. *For each tautology φ, $\mathbb{P}(\varphi) = 1$.*
3. *For each contradiction φ, $\mathbb{P}(\varphi) = 0$.*

Proof. It is clear that any coherent $\mathbb{P}$ satisfies these axioms.

To show that any $\mathbb{P}$ satisfying these axioms is coherent, we construct a distribution over complete consistent theories T which "reproduces" $\mathbb{P}$ and then appeal to the completeness theorem.

Fix some enumeration $\varphi_1, \varphi_2, \ldots$ of all of the sentences of L. Let $T_0 = \emptyset$ and iteratively define T_{i+1} in terms of T_i as follows. If T_i is complete, we set $T_{i+1} = T_i$. Otherwise, let φ_j be the first statement which is independent of T_i, and let $T_{i+1} = T_i \cup \varphi_j$ with probability $\mathbb{P}(\varphi_j \mid T_i)$ [1] and $T_{i+1} = T_i \cup \neg\varphi_j$ with probability $\mathbb{P}(\neg\varphi_j \mid T_i)$. Because φ_j was independent of T_i, the resulting system remains consistent. Define $T = \cup_i T_i$. Since each T_i is consistent, T is consistent by compactness. For each i, φ_i or $\neg\varphi_i$ will be included in some stage T_j (in fact at stage $j = i + 1$ at the latest), so T is complete.

Axiom 1 implies

$$\mathbb{P}(\varphi \mid T_i) = \mathbb{P}(\varphi \mid T_i \wedge \varphi_j)\, \mathbb{P}(\varphi_j \mid T_i) + \mathbb{P}(\varphi \mid T_i \wedge \neg\varphi_j)\, \mathbb{P}(\neg\varphi_j \mid T_i),$$

thus the sequence $\mathbb{P}(\varphi \mid T_i)$ is a martingale. Axiom 2 implies that if $T_i \vdash \varphi$, $\mathbb{P}(\varphi \mid T_i) = 1$ and if $T_i \vdash \neg\varphi$, $\mathbb{P}(\varphi \mid T_i) = 0$. Thus, this sequence eventually stabilizes at 0 or 1, and $T \vdash \varphi$ iff this sequence stabilizes at 1. The martingale property then implies that $T \vdash \varphi$ with probability $\mathbb{P}(\varphi \mid T_0)$. By axiom 3, $\mathbb{P}(T_0) = 1$, so $\mathbb{P}(\varphi \mid T_0) = \mathbb{P}(\varphi)$.

This process defines a measure μ over complete, consistent theories such that $\mathbb{P}(\varphi) = \mu(\{T : T \vdash \varphi\})$. For each complete, consistent theory T, the completeness theorem guarantees the existence of a model $\mathcal{M}$ such that $\mathcal{M} \models \varphi$ if and only if $T \vdash \varphi$. Thus we obtain a distribution μ over models such that $\mathbb{P}(\varphi) = \mu(\{\mathcal{M} : \mathcal{M} \models \varphi\})$, as desired. □

We are particularly interested in probabilities $\mathbb{P}$ that assign probability 1 to T. It is easy to check that such $\mathbb{P}$ correspond to distributions over models of T.

2.2 Going meta

So far we have talked about $\mathbb{P}$ as a function which assigns probabilities to sentences. If we would like L to serve as its own meta-language, we should augment it to include $\mathbb{P}$ as a real-valued function symbol. Call the agumented language L'.

We will use the symbol $\mathbb{P}$ in two different ways—both as our meta-level evaluation of truth in L', and as a quoted symbol within L'. However, whenever $\mathbb{P}$ appears as a symbol in L',

[1] $\mathbb{P}(\varphi_j \mid T_i) = \frac{\mathbb{P}(\varphi_j \wedge T_i)}{\mathbb{P}(T_i)}$. It is easy to verify by induction that $\mathbb{P}(T_i) > 0$.

its argument is always quoted. This ensures that references are unambiguous: "$\mathbb{P}(\varphi) = p$" is a statement at the meta level, i.e. "φ is true with probability p", while "$\mathbb{P}(\ulcorner\varphi\urcorner) = p$" is the same sentence expressed within L', which is in turn assigned a probability by $\mathbb{P}$.

3 Reflection

3.1 Reflection principle

We would now like to introduce some analog of the formal correctness property 1, which we will call a *reflection principle.*

This must be done with some care if we wish to avoid the problems with schema 1. For example, we could introduce the analogous reflection principle:

$$\forall\varphi \in L'\ \forall a, b \in \mathbb{Q} : (a < \mathbb{P}(\varphi) < b) \iff \mathbb{P}(a < \mathbb{P}(\ulcorner\varphi\urcorner) < b) = 1 \tag{2}$$

But this leads directly to a contradiction: if we define $G \iff \mathbb{P}(\ulcorner G\urcorner) < 1$, then

$$\mathbb{P}(G) < 1 \iff \mathbb{P}(\mathbb{P}(\ulcorner G\urcorner) < 1) = 1 \iff \mathbb{P}(G) = 1,$$

which contradicts $\mathbb{P}(G) \in [0, 1]$.

To overcome this challenge, we imagine $\mathbb{P}$ as having access to *arbitrarily precise* information about $\mathbb{P}$, without having access to the exact values of $\mathbb{P}$. Let (a, b) be an open interval containing $\mathbb{P}(\varphi)$. Then a sufficiently accurate approximation to $\mathbb{P}(\varphi)$ would let us conclude $\mathbb{P}(\varphi) \in (a, b)$, and so we have $\mathbb{P}(\varphi) \in (a, b) \implies \mathbb{P}(\mathbb{P}(\ulcorner\varphi\urcorner) \in (a, b)) = 1$. However, the converse is not necessarily true. If $\mathbb{P}(\varphi) = a$, then no possible approximation can distinguish between the cases $\mathbb{P}(\varphi) < a$, $\mathbb{P}(\varphi) = a$, and $\mathbb{P}(\varphi) > a$. So the reverse implication requires closed rather than open intervals: for any closed interval $[a, b]$ *not* containing $\mathbb{P}(\varphi)$, we could infer $\mathbb{P}(\varphi) \notin [a, b]$ from any sufficiently accurate approximation to $\mathbb{P}(\varphi)$, and thus we should have $\mathbb{P}(\varphi) \notin [a, b] \implies \mathbb{P}(\mathbb{P}(\ulcorner\varphi\urcorner) \in [a, b]) = 0$.

This suggests the following criteria:

$$\forall\varphi \in L'\ \forall a, b \in \mathbb{Q} : (a < \mathbb{P}(\varphi) < b) \Longrightarrow \mathbb{P}(a < \mathbb{P}(\ulcorner\varphi\urcorner) < b) = 1 \tag{3}$$

$$\forall\varphi \in L'\ \forall a, b \in \mathbb{Q} : (a \leq \mathbb{P}(\varphi) \leq b) \Longleftarrow \mathbb{P}(a \leq \mathbb{P}(\ulcorner\varphi\urcorner) \leq b) > 0 \tag{4}$$

We say that $\mathbb{P}$ is *reflectively consistent* if it satisfies these properties. Our goal is to show that starting from any consistent theory T we can obtain a reflectively consistent $\mathbb{P}$ which assigns probability 1 to each sentence of T.

The intuition behind the description "reflective consistency" is that such a $\mathbb{P}$ would not change upon further reflection. This is exactly analogous to the standard view of truth as a fixed point of a certain revision operation, and indeed we show that such $\mathbb{P}$ exist by constructing an appropriate revision operation and appealing to Kakutani's fixed point theorem.

In fact, property 3 and property 4 are equivalent. For suppose that property 3 holds, and that $\mathbb{P}(\varphi) < a$ or $\mathbb{P}(\varphi) > b$. Then either $\mathbb{P}(\mathbb{P}(\ulcorner\varphi\urcorner) < a) = 1$ or $\mathbb{P}(\mathbb{P}(\ulcorner\varphi\urcorner) > b) = 1$, and in either case $\mathbb{P}(a \leq \mathbb{P}(\ulcorner\varphi\urcorner) \leq b) = 0$. But this is precisely the contrapositive of propety 4. A similar argument can be applied to show that property 4 implies property 3. In light of this equivalence, we will focus only on property 3, and call this property the *reflection principle.*

3.2 Discussion of reflection principle

Given a statement such as $G \iff \mathbb{P}(\ulcorner G\urcorner) < p$, a reflectively consistent distribution $\mathbb{P}$ will assign $\mathbb{P}(G) = p$. If $\mathbb{P}(G) > p$, then $\mathbb{P}(\mathbb{P}(\ulcorner G\urcorner) > p) = 1$, so $\mathbb{P}(G) = 0$, and if $\mathbb{P}(G) < p$, then $\mathbb{P}(\mathbb{P}(\ulcorner G\urcorner) < p) = 1$, so $\mathbb{P}(G) = 1$. But if $\mathbb{P}(G) = p$, then $\mathbb{P}$ is "uncertain" about whether $\mathbb{P}(G)$ is slightly more or slightly less than p. No matter how precisely it estimates $\mathbb{P}(G)$, it remains uncertain.

If we look at the models of L' corresponding to reflectively consistent distributions $\mathbb{P}$, they have $\mathbb{P}(\ulcorner G\urcorner) \in (p - \epsilon, p + \epsilon)$, where ϵ is a non-standard infinitesimal which lies strictly between 0 and every standard rational number. That is, $\mathbb{P}$ is "mistaken" about itself, but its error is given by the infinitesimal ϵ. (These are the same infinitesimals which are used to formalize differentiation in nonstandard analysis.)

Although this paper focuses on Tarski's results on the undefinability of truth, similar paradoxes are at the core of Gödel's incompleteness theorem and the inconsistency of unrestricted comprehension in set theory. A similar approach to the one taken in this paper is adequate to resolve some of these challenges, and e.g. to construct a set theory which satisfies a probabilistic version of the unrestricted comprehension axiom.

We will prove that there are many reflectively consistent distributions, but our proof will be non-constructive. In fact any coherent $\mathbb{P}$ is necessarily uncomputable, whether or not it is reflectively consistent. But the mere existence of a reflectively consistent distribution implies that the property 3 will not lead to any contradictions if treated as a rule of inference which connects a reasoner's beliefs about $\mathbb{P}(\varphi)$ to its beliefs about φ itself. So even though we cannot construct any reflectively consistent $\mathbb{P}$, their existence may still have pragmatic implications for reasoning (in the same way that introducing the predicate True might usefully increase the expressive power of a language, even though the extension of that predicate is necessarily uncomputable).

3.3 Proof of consistency of reflection principle

The most important result of this paper is the following theorem:

Theorem 2 (Consistency of reflection principle)**.** *Let L be any language and T any consistent theory over L, and suppose that the theory of rational numbers with addition can be embedded in T in the natural sense. Let L' be the extension of L by the symbol $\mathbb{P}$ as above. Then there is a probabilistic valuation $\mathbb{P}$ over L' which is coherent, assigns probability 1 to T, and satisfies* *the reflection principle:*

$$\forall \varphi \in L' \; \forall a, b \in \mathbb{Q} : (a < \mathbb{P}(\varphi) < b) \implies \mathbb{P}(a < \mathbb{P}(\ulcorner \varphi \urcorner) < b) = 1.$$

Proof. Let $\mathcal{A}$ be the set of coherent probability distributions over L' which assign probability 1 to T. We consider $\mathcal{A}$ as a subset of $[0,1]^{L'}$ with the product topology. Clearly $\mathcal{A}$ is convex. By using the alternative characterization of coherence, we can see that $\mathcal{A}$ is a closed subset of $[0,1]^{L'}$. Thus by Tychonoff's theorem, we conclude that $\mathcal{A}$ is compact. Moreover, because T is consistent, $\mathcal{A}$ is non-empty—for example, let $\mathcal{M}$ be any model of T, in which each $\mathbb{P}(\ulcorner \varphi \urcorner)$ is assigned some arbitrary value, and let $\mathbb{P}(\varphi) = 1$ if $\mathcal{M} \models \varphi$ and 0 otherwise.

Given any $\mathbb{P} \in \mathcal{A}$, let $R_{\mathbb{P}}$ be the schema of axioms of the form $a < \mathbb{P}(\ulcorner \varphi \urcorner) < b$ where a, b are the endpoints of a rational interval containing $\mathbb{P}(\varphi)$. We will write $\mathbb{P}(R_{\mathbb{P}'}) = 1$ as shorthand for $\forall \psi \in R_{\mathbb{P}'} : \mathbb{P}(\psi) = 1$ (this notation is justified, since $R_{\mathbb{P}'}$ is countable). Define $f : \mathcal{A} \to \mathcal{P}(\mathcal{A})$ by $f(\mathbb{P}') = \{\mathbb{P} \in \mathcal{A} : \mathbb{P}(R_{\mathbb{P}'}) = 1\}$. If $\mathbb{P} \in f(\mathbb{P})$ then $\mathbb{P}$ is reflectively consistent and we are done.

Each $f(\mathbb{P})$ is convex and non-empty by the same argument that $\mathcal{A}$ itself is. We will show that f has a closed graph; we can then apply the Kakutani fixed point theorem to find some $\mathbb{P} \in f(\mathbb{P})$, as desired. (The Kakutani fixed-point theorem applies to $[0,1]^{L'}$ because it is a convex subset of a Hausdorff and locally convex topological vector space.)

Let $\{\mathbb{P}_i\}_i, \{\mathbb{P}'_i\}_i$ be sequences in $\mathcal{A}$ such that $\mathbb{P}_i \in f(\mathbb{P}'_i)$ for each i. We need to show that if $\mathbb{P}_i \to \mathbb{P}$ and $\mathbb{P}'_i \to \mathbb{P}'$ (in the product topology) then $\mathbb{P} \in f(\mathbb{P}')$. Since $\mathcal{A}$ is closed, we have $\mathbb{P} \in \mathcal{A}$. So it remains to show that $\mathbb{P}(R_{\mathbb{P}'}) = 1$. Pick any a, b, φ such that $a < \mathbb{P}'(\varphi) < b$. Since $\mathbb{P}'_i$ converges to $\mathbb{P}'$, for all sufficiently large i we have $a < \mathbb{P}'_i < b$, and thus $\mathbb{P}_i(a < \mathbb{P}(\ulcorner \varphi \urcorner) < b) = 1$. Since $\mathbb{P}_i$ converges to $\mathbb{P}$, we have $\mathbb{P}(a < \mathbb{P}(\ulcorner \varphi \urcorner) < b) = 1$. Thus $\mathbb{P}(R_{\mathbb{P}'}) = 1$ We conclude that f has a closed graph, so we can apply the Kakutani fixed point theorem to find $\mathbb{P} \in f(\mathbb{P})$, as desired.

□

3.4 Going meta

An important point is that the reflection principle is a *statement which is true about* $\mathbb{P}$, not an axiom to which $\mathbb{P}$ assigns probability 1. That is, each statement of the form $(a < \mathbb{P}(\varphi) < b) \implies \mathbb{P}(a < \mathbb{P}(\ulcorner \varphi \urcorner) < b) = 1$ is true. This is what we need in order for the reflection principle to meaningfully constrain the values of $\mathbb{P}$—we don't want $\mathbb{P}$ to assert that $\mathbb{P}$ is reflectively consistent, we want it to actually *be* reflectively consistent.

That said, we might also want $\mathbb{P}$ to assign probability 1 to its own reflective consistency. To this end, note that any reflectively consistent $\mathbb{P}$ also satisfies

$$\forall \epsilon > 0 \; \forall \varphi \in L' \; \forall a, b \in \mathbb{Q} \; : \; \mathbb{P}((a \le \mathbb{P}(\ulcorner \varphi \urcorner) \le b) \implies \mathbb{P}(\ulcorner a < \mathbb{P}(\ulcorner \varphi \urcorner) < b \urcorner) > 1 - \epsilon) = 1. \tag{5}$$

Indeed, for each a, b, ϵ, φ, if the antecedent of 5 is false we assign it probability 0, and if the consequent is true we assign it probability 1. Once again, we can't get *exactly* what we want, but we can get it up to some arbitrarily small error. The most unsatisfying thing about property 5 is not this infinitesimal error: it is that $\mathbb{P}$ appears *inside* the quantifiers.

We would really like $\mathbb{P}$ to assert its own reflective consistency *in general*, not just in each specific case. That is, we would like:

$$\mathbb{P}\left(\forall \varphi \in L' \; \forall a, b \in \mathbb{Q} : (a < \mathbb{P}\left(\ulcorner \varphi \urcorner\right) < b) \implies \mathbb{P}\left(\ulcorner a < \mathbb{P}\left(\ulcorner \varphi \urcorner\right) < b \urcorner\right) = 1\right) = 1,$$

but we do *not* prove that $\mathbb{P}$ has this property. Indeed, if $\mathbb{P}$ assigns probability 1 to property 3, then we have observed that $\mathbb{P}$ must also assign probability 1 to property 4. But then we have

$$\begin{aligned} \mathbb{P}\left(a \leq \mathbb{P}\left(\ulcorner \varphi \urcorner\right) \leq b\right) > 0 &\implies \mathbb{P}\left(\mathbb{P}\left(\ulcorner a \leq \mathbb{P}\left(\ulcorner \varphi \urcorner\right) \leq b \urcorner\right) > 0\right) = 1 \\ &\implies \mathbb{P}\left(a \leq \mathbb{P}\left(\ulcorner \varphi \urcorner\right) \leq b\right) = 1 \end{aligned}$$

which leads to a contradiction. In fact, a similar argument shows that $\mathbb{P}$ must assign *probability 0* both to property 3 and property 4.

In order to devise a reflection principle which is simultaneously *satisfied by* $\mathbb{P}$ and *assigned probability* 1 *by* $\mathbb{P}$, we need to weaken property 3. One alternative reflection principle is an approximate verison of the intuitively appealing identity $\mathbb{P}\left(\varphi \mid \mathbb{P}\left(\ulcorner \varphi \urcorner\right) = p\right) = p$, which formalizes the notion of "self-trust" rather than "self-knowledge." For example, we could consider the relaxation

$$\begin{aligned} \forall \varphi \in L' \; \forall a, b \in \mathbb{Q} \; &: \; \mathbb{P}\left(\varphi \wedge \left(a < \mathbb{P}\left(\ulcorner \varphi \urcorner\right) < b\right)\right) \leq b\mathbb{P}\left(a \leq \mathbb{P}\left(\ulcorner \varphi \urcorner\right) \leq b\right) \\ \forall \varphi \in L' \; \forall a, b \in \mathbb{Q} \; &: \; \mathbb{P}\left(\varphi \wedge \left(a \leq \mathbb{P}\left(\ulcorner \varphi \urcorner\right) \leq b\right)\right) \geq a\mathbb{P}\left(a < \mathbb{P}\left(\ulcorner \varphi \urcorner\right) < b\right) \end{aligned}$$

This is strictly weaker than our proposed reflection principle, so our result implies that there exist coherent $\mathbb{P}$ satisfying this principle. Moreover, no obvious analog of the Liar's paradox prevents this property from being both satisfied by $\mathbb{P}$ and assigned probability 1 by $\mathbb{P}$. (This property may be viewed as an approximate, asynchronous version of van Frassen's Reflection Principle. The exact version of van Frassen's principle is easily seen to be subject to the liar's paradox.)

It remains open whether it is possible to devise a principle which captures the important aspects of reflective consistency, and which can be simultaneously true of a distribution $\mathbb{P}$ and assigned probability 1 by $\mathbb{P}$. However, our work shows that the obstructions presented by the liar's paradox can be overcome by tolerating an infinitesimal error, and that Tarski's result on the undefinability of truth is in some sense an artifact of the infinite precision demanded by reasoning about complete certainty.