Questions of Reasoning Under Logical Uncertainty

Nate Soares and **Benja Fallenstein**
Machine Intelligence Research Institute
{nate,benja}@intelligence.org

Abstract

A logically uncertain reasoner would be able to reason as if they know both a programming language and a program, without knowing what the program outputs. Most practical reasoning involves some logical uncertainty, but no satisfactory theory of reasoning under logical uncertainty yet exists. A better theory of reasoning under logical uncertainty is needed in order to develop the tools necessary to construct highly reliable artificial reasoners. This paper introduces the topic, discusses a number of historical results, and describes a number of open problems.

1 Introduction

Consider a black box with one input chute and two output chutes. The box is known to take a ball in the input chute and then (via some complex Rube Goldberg machine) deposit the ball in one of the output chutes.

An *environmentally uncertain* reasoner does not know which Rube Goldberg machine the black box implements. A *logically uncertain* reasoner may know which machine the box implements, and may understand how the machine works, but does not (for lack of computational resources) know how the machine behaves.

Standard probability theory is a powerful tool for reasoning under environmental uncertainty, but it assumes logical omniscience: once a probabilistic reasoner has determined precisely which Rube Goldberg machine is in the black box, they are assumed to know which output chute will take the ball. By contrast, realistic reasoners must operate under logical uncertainty: we often know how a machine works, but not precisely what it will do.

General intelligence, at the human level, mostly consists of reasoning that involves logical uncertainty. Reasoning about the output of a computer program, the behavior of other actors in the environment, or the implications of a surprising observation are all done under logical (in addition to environmental) uncertainty. This would also be true of smarter-than-human systems: constructing a completely coherent Bayesian probability distribution in a complex world is intractable. Any artificially intelligent system writing software or evaluating complex plans must necessarily perform some reasoning under logical uncertainty.

When constructing smarter-than-human systems, the stakes are incredibly high: superintelligent machines could have an extraordinary impact upon humanity (Bostrom 2014), and if that impact is not beneficial, the results could be catastrophic (Yudkowsky 2008). If that system is to attain superintelligence by way of self-modification, logically uncertain reasoning will be critical to its reliability. The initial system's ability must reason about the unknown behavior of a known program (the contemplated self-modification) in order to understand the result of modifying itself.

In order to pose the question of whether a practical system reasons well under logical uncertainty, it is first necessary to gain a theoretical understanding of logically uncertain reasoning. Yet, despite significant research going back to Łoś (1955), Gaifman (1964) and before, continued by Halpern (2003), Hutter et al. (2013), Demski (2012), Christiano (2014) and many, many others,[1] this theoretical understanding does not yet exist.

It is natural to consider extending standard probability theory to include the consideration of worlds which are "logically impossible" (such as where a deterministic Rube Goldberg machine behaves in a way that it doesn't). This gives rise to two questions: What, precisely, are logically impossible possibilities? And, given some means of reasoning about impossible possibilities, what is a reasonable prior probability distribution over them?

This paper discusses the field of reasoning under logical uncertainty. At present, study into logically uncertain reasoning is largely concerned with the problem of reasoning probabilistically about sentences of logic.

Research supported by the Machine Intelligence Research Institute (intelligence.org). Published as Technical report 2015–1.

1. See, e.g., the work referenced by Hutter et al. (2013).

Sections 2 and 3 discuss the two problems posed above in that context. Ultimately, our understanding of logical uncertainty will need to move beyond the domain of logical sentences; this point is further explored in Section 4. Section 5 concludes by relating these problems back to the design of smarter-than-human systems which are reliably aligned with human interests.

2 Impossible Possibilities

Consider again the black box, with the Rube Goldberg machine inside. An agent reasoning using standard probability theory is environmentally uncertain; they do not know which Rube Goldberg machine is in the box. Probability distributions assign probabilities to some set of "possibilities," and when reasoning probabilistically under environmental uncertainty, the set of possibilities is all Rube Goldberg machines consistent with observation that could fit in the box.

What is the set of possibilities considered by a logically uncertain reasoner? They may know which Rube Goldberg machine is in the box, without knowing how that Rube Goldberg machine behaves (for lack of deductive capabilities). The machine, following the laws of logic and physics, deposits the ball in only one of the two chutes, but a logically uncertain reasoner must consider both output chutes as a "possibility." Logically uncertain reasoning, then, requires the consideration of logically impossible possibilities.

What sort of objects are logically impossible possibilities? What is the set of all impossible possibilities, to which probabilities are assigned? In full generality, this question is vague and intractable. But there is one setting in which it is natural to consider logical impossibilities, and that is the domain of formal logic itself: consider agents that are uncertain about the truth values of *sentences of logic*. Indeed, the study of logical uncertainty in the literature centers on reasoning according to assignments of probabilities to sentences of first-order logic (Gaifman 2004).

How does reasoning about logical sentences correspond to reasoning under logical uncertainty in the real world? Sentences of first order logic are extremely expressive: given a description of the Rube Goldberg machine in the black box, it is possible to construct a logical sentence which is true if and only if the machine deposits the ball in the top chute. A reasoner uncertain about whether that sentences is true is also uncertain about the behavior of that Rube Goldberg machine. Logical sentences can also encode statements such as "this Turing machine will halt," or "this function sorts its input and has time complexity $\mathcal{O}(\log n)$." Thus, while it is ultimately necessary to understand logically uncertain reasoning as it pertains to observation and interaction in the real world, it is reasonable to begin studying reasoning under logical uncertainty by studying probability assignments to logical sentences.

Picking any probability distribution over logical sentences does not automatically constitute "reasoning under logical uncertainty." Intuitively, logically uncertain reasoning must preserve some of the structure between sentences: if a reasoner assigns probability 1 to ϕ and deduces $\phi \to \psi$ (via some complex implication), then the reasoner must assign probability 1 to ψ thereafter. But clearly, not all of the structure between sentences can be preserved, for that would require logical omniscience. Which structure is preserved under logical uncertainty, and how?

It is illuminating to first consider the probabilities that a reasoner would assign to sentences of logic if they *could* preserve all the logical structure. How could a deductively omniscient reasoner assign probabilities consistently to logical sentences? It is not so simple as claiming that omniscient reasoners assign probability 1 to true sentences and 0 to false ones, because logical sentences are not simply "true" or "false" in a vacuum. It depends entirely on which logical theory obtains: if the domain is numbers and "$\times$" is multiplication, then the sentence $\forall a, b\colon a \times b = b \times a$ is true; but if the domain is vectors and "$\times$" is the vector cross product, then the same sentence is false. There are, in fact, two types of logical uncertainty: uncertainty about the logical theory, and uncertainty stemming from limited deductive capabilities.

The first type of logical uncertainty has been studied for many decades, with early work done by Gaifman (1964), Hacking (1967), and others. It is not merely a problem of defining symbols: in logic, there are many theories which are "incomplete," meaning that there exist sentences which are not necessarily either true or false according to that logical theory. Consider Peano Arithmetic ($\mathcal{PA}$), which formalizes the natural numbers. $\mathcal{PA}$ nails down the definitions of "$+$" and "$\times$," but there are still sentences which are true of the numbers but are not implied by the Peano axioms. The classic example is Gödel's sentence, which roughly claims "$\mathcal{PA}$ cannot prove this sentence" (Gödel, Kleene, and Rosser 1934). This statement is true, but it does not follow from the Peano axioms.

A consistent assignment of truth values to every sentence requires a *complete* logical theory. A complete theory of logic is a consistent set T of sentences such that for every sentence ϕ, either $\phi \in T$ or $(\neg\phi) \in T$. Incomplete theories can be "completed" by starting with the set of all consequences of the incomplete theory and then choosing arbitrary (consistent) assignments of truth for each independent sentence. For example, there are many different ways to complete $\mathcal{PA}$, only one of which is "true arithmetic." Identifying true arithmetic is uncomputable, as statements of true arithmetic include statements about which Turing machines halt.[2] Even a deductively unlimited reasoner—that always believes ψ whenever it believes both ϕ and $\phi \to \psi$, no matter how complex and obfuscated the implication—

2. And which oracle machines halt, and which meta-oracle machines halt, and so on.

may have uncertainty about which sentences are true or false, via uncertainty about which complete theory is the "real" one.

A logically uncertain but deductively unlimited reasoner—which knows all consequences of everything it knows, but does not know the "true" complete theory of logic—only entertains *consistent* logically impossible possibilities. It may not know which logical theory corresponds to how sentences act in "the real world," but each "possible world" is self-consistent. This provides a partial answer to the question of the nature of logically impossible possibilities: if a reasoner is deductively unlimited, an "impossible possibility" is any complete theory of logic. A "consistent" assignment of probabilities to sentences, then, corresponds to a probability distribution over complete theories of logic, where the probability of a sentence is equal to the measure of theories in which that sentence is true.[3] This is the standard model of logically uncertain reasoning throughout the literature, used e.g. by Gaifman (1964), Christiano et al. (2013), and many others.

This is a fine result for deductively unlimited reasoners, but the goal is to understand reasoning under deductive limitations. Deductively unlimited reasoners reason according to *consistent* impossible worlds, but detecting inconsistencies can be a computationally expensive task. Deductively limited reasoners must entertain *inconsistent* impossible possibilities. Recent study of reasoning under logical uncertainty has been pushing in this direction (Gaifman 2004).

Intuitively, deductively limited reasoners must reason according to "theories" (assignments of truth values to sentences) that seem consistent *so far*, discarding hypotheses as soon as a contradiction is deduced. An "impossible possibility," then, could be any assignment of truth values to logical sentences which does not allow a *short* proof of inconsistency. This intuition might be formalized as follows:

> Fix enumerations of sentences and proofs in first order logic. Consider all bit strings of some huge length, say Ackerman(10^{100}), and interpret these as assignments of truth to each sentence (a 1 in the n^{th} position claims that the n^{th} sentence is true, and a 0 claims it is false). Now search all strings with length of up to some far larger number, say Ackerman^2(10^{100}), for proofs in $\mathcal{PA}$ that one of these bit strings makes inconsistent claims, and reject any bit string which is found to be inconsistent by this procedure. Assign probabilities to each remaining bit string (according to some prior); this may be used to generate an assignment of probabilities to sentences of length less than Ackerman(10^{100}) according to the measure of bit strings which claim that the sentence is true.

This procedure is wildly impractical, but it does seem to allow for satisfactory logically uncertain reasoning using finite (if not limited) deduction. Clearly, the larger the number of sentences (and the larger the number of proofs searched), the more the remaining bit strings will resemble a collection of complete theories which are actually consistent. This process leads to intuitively "reasonable" logically uncertain beliefs: no "theories" that admit a short proof of inconsistency are considered, but "theories" with inconsistencies that are very difficult to deduce may remain. This process corresponds to a finite version of considering all complete theories: it considers bit strings assigning truth values to *many* sentences, for which there are no *reasonably sized* proofs of contradiction; deductively unlimited reasoning considers bit strings assigning truth values to *all* sentences, for which there is no proof of contradiction *at all*. However, for all that this technique seems intuitively nice, no precise statements about its performance have yet been proven.

These techniques shed light on the nature of impossible possibilities in the context of deductively limited reasoning: an impossible possibility is any assignment of truth values to logical sentences which has not yet been proven inconsistent, in some fashion. That is, practical agents may entertain contradictory possibilities, so long as the possibility is discarded once the contradiction is deduced; an "impossible possibility" is any assignment of truth to logical sentences which hasn't been found to be consistent so far. It is an open problem to develop more practical techniques than the one above which allow agents to reason as if according to truth-assignments which have "not yet been found to be inconsistent." For further discussion, see Christiano (2014).

While this answer is somewhat satisfactory, it only answers the question of impossible possibilities as it relates to uncertainty over *logical sentences*. Realistic reasoning under logical uncertainty will require more than just an ability to assign probabilities to logical sentences; discussion of this point is relegated to Section 4.

3 Logical Priors

In the context of uncertainty about logical sentences, deductively limited reasoners must approximate reasoning according to some probability distribution over complete theories. This gives rise to a second question: which probability distribution over complete theories should they approximate?

Of course, the answer is in part up to us: if we design a system which reasons about the probabilities of logical sentences, which takes questions in the form of sentences and outputs predictions in the form of probabilities, then the question of what logical theory it should

3. There are uncountably many complete theories of first order logic, but a probability distribution over them can be defined using the machinery of measure theory.

use depends entirely upon how we want the questions to be interpreted. For example, if a deductively limited system is built to help its designers reason about some Boolean algebra, then the "distribution over complete theories" might be some sort of simplicity prior which assigns probability 1 to the complete theory of Boolean algebras.

But what if the system is intended to reason according to some extremely powerful theory of logic (e.g. $\mathcal{PA}$) which is capable of expressing many questions about the real world (e.g. whether the Rube Goldberg machine deposits the ball into the top chute), but for which *we* do not know the preferred single complete theory (e.g. because the complete theory answers all halting problems)? Then what distribution over complete theories should be used? If the system is supposed to reason according to true arithmetic, then what initial state of knowledge captures *our* beliefs about that uncomputable theory?

This is the problem of *logical priors*. Intuitively, the problem may seem easy: just choose a maximum entropy (or otherwise weak) prior. Unfortunately, it is not obvious how to construct a weak prior over complete theories. Starting with a maximum entropy prior on logical sentences and refining towards consistency will not suffice: a prior which assigns 50% probability to every sentence places zero probability mass on the set of all complete theories, because there are infinitely many contradictory sentences, and so any infinite sequence of sentences generated by this prior is guaranteed to select at least one contradiction eventually.

Hutter et al. (2013) make an early attempt to answer the first question, by defining a logical prior in terms of a probability distribution over sentences which assigns positive probability to all consistent sentences and zero probability to contradictions. A probability distribution of this form allows for the definition of a satisfactory logical prior:

> **The Hutter prior:** For each sentence, select a model in which that sentence is true, and in which certain desirable properties hold (the "Gaifman condition" and the "Cournot condition" (Hutter et al. 2013)). Add the complete theory of that model to the distribution with measure in proportion to the probability of the sentence.

This prior has many desirable properties, but it cannot be computably approximated: the conditions that Hutter demands of each model (which yield the prior's nice properties) rely on the high-powered machinery of set theory, and it is not possible to computably approximate this prior. That is, there does not exist a computable process refining assignments of probabilities to sentences which converges on the assignments of Hutter's prior in the limit (Sawin and Demski 2013).[4]

The Hutter prior yields insight into what constitutes a desirable prior, but a study of logical uncertainty in deductively limited systems requires that the prior be approximable. Just as deductively limited reasoners must approximate reasoning about consistent theories (by entertaining inconsistent "theories" until a contradiction is deduced), so must deductively limited reasoners start with a prior that does not quite match the (inevitably intractable) intended prior, and then refine those probabilities as they reason.

But this process of starting with an incoherent prior (which places probability mass on inconsistencies) and refining it towards some desirable prior (eliminating inconsistencies as contradictions are deduced, and shifting probability mass to better correspond with the "true" prior) is precisely the problem of reasoning under logical uncertainty, entire!

That is, the approximation of a satisfactory logical prior exhibits, in miniature, all the problems of reasoning according to a probability distribution over sentences. Thus, the definition of a satisfactory approximable logical prior, and the study of its approximations, may yield solutions to the problem of reasoning under logical uncertainty more generally.

Unfortunately, it is not entirely clear what it would mean for an approximable logical prior to be "satisfactory," and naïve attempts at constructing computably approximable logical priors have all had undesirable properties.

Demski (2012) proposes a computably approximable prior that can be generated from any distribution Φ over all sentences. A "generator" is used to generate complete theories (by drawing sentences at random from Φ), and Demksi's prior assigns probability to a theory T according to the probability that the generator would generate T.

More formally, Demski's generator is given by Algorithm 1. It takes an initial set B of known sentences and a distribution Φ over sentences. It constructs a complete theory T by starting with B and selecting sentences ϕ at random from Φ. It either adds ϕ to T (if ϕ is consistent with T) or adds $\neg\phi$ to T (otherwise).

For example, let the base theory be Peano Arithmetic ($\mathcal{PA}$), and let Φ be a simplicity prior over sentences which assigns each sentence ϕ probability $2^{-|\phi|}$ where $|\phi|$ is the length[5] of ϕ. Clearly, the simplicity prior does not describe a satisfactory logical prior over sentences, as it puts significant probability mass on short contradictory sentences such as "$0 = 1$". Demski's generator, however, only generates consistent theories, and therefore, it places probability 0 on all contradictions. Similarly, because all sentences of $\mathcal{PA}$ are

4. Remember that a probability distribution over complete theories can be treated as a probability distribution over sentences which assigns probability to sentences in accordance with the proportion of theories in which that sentence is true.

5. Where sentences are encoded in binary, preferably using some encoding of length where the length of ϕ is the same as the length of $\neg\phi$, so that the prior is not biased in disfavor of negations.

Algorithm 1: The Demski generator

Data: A probability distribution Φ over sentences
Data: A base theory B of known sentences
Result: A complete theory T
begin
  $T \longleftarrow B$
  loop
    $\phi \longleftarrow$ genrandom(Φ)
    if $T \cup \{\phi\}$ *is consistent* **then**
      $T \longleftarrow T \cup \{\phi\}$
    else
      $T \longleftarrow T \cup \{\neg\phi\}$

included in B, the prior only generates theories T consistent with $\mathcal{PA}$, and so it assigns probability 1 on all sentences implied by $\mathcal{PA}$. Now consider a sentence ϕ which is independent of $\mathcal{PA}$: the probability of this sentence depends upon how often Demski's generator generates a theory T in which ϕ is true. Clearly, this probability is positive, as with probability $2^{-|\phi|}$, ϕ will be the first random sentence added to T. Similarly, because there is a chance that the first random sentence is $\neg\phi$, the probability of ϕ will not be 1. Thus, Demski's prior defines a probability distribution over complete theories extending $\mathcal{PA}$.

While Demski's prior is uncomputable, Demski (2012) has shown that the resulting prior probability distribution is computably *approximable*: There is a computable procedure which will output successive approximations of the probability of a sentence ϕ, converging in the limit to the probability assigned to ϕ by the uncomputable procedure. Even this computable approximation, however, is not a tractable algorithm; recently, Christiano (2014) has proposed an alternative approach to constructing priors which borrows from standard machine learning techniques, making it more likely that the priors developed in this way can be used in realistic algorithms.

These priors, however, have some undesirable properties. For example, starting with B as the empty set, Demski's prior places zero probability on the set of complete theories where $\mathcal{PA}$ holds.[6] Agents approximating Demski's would not be able to learn Peano Arithmetic: Demski's prior, while approximable, is not weak enough.

In order for a reasoner using Demski's prior to believe $\mathcal{PA}$, it must be included in the base theory B. This reveals a related issue: there are two different ways to "update" Demski's prior on a sentence ϕ. The prior can either be completely regenerated from the base theory $B \cup \{\phi\}$, or it can be *conditioned* on ϕ (by removing all theories in which ϕ is false). These two different updates result in two different posterior probability distributions.

Consider the posterior probability of a sentence ψ such that both ψ and $\neg\psi$ are consistent with ϕ. If the prior is regenerated from $B \cup \{\phi\}$, then the resulting posterior still places at least $2^{-|\psi|}$ probability on ψ, because this is the probability that ψ is the first sentence selected at random from Φ. But if the prior is conditioned on ϕ, then it may be the case that the posterior probability of ψ is arbitrarily low. For example, if $\neg\psi \rightarrow \phi$, then all theories with $\neg\psi$ will have ϕ, and if it is also the case that almost all theories with ψ also contain $\neg\phi$, then the posterior probability may place arbitrarily small positive probability on ψ. In other words, conditioning the prior on ϕ favors *explanations* for ϕ, while regenerating the prior does not alter the Φ-based lower bound probability of any sentence that does not directly contradict ϕ.

This double update is strange. An agent reasoning using Demski's prior would treat facts that it "learns" (through observation and conditioning) differently from facts that it "always knew" (sentences from the base theory). This phenomenon is not well understood. Why does the double update occur? Is it undesirable? Can it be avoided? These questions remain open, and it is possible that answers to these questions will lend insight into the generation of satisfactory approximable logical priors.

It is not at all clear what it would mean for a logical prior to be "satisfactory," in the first place: part of the problem is that it is not yet clear what desiderata to demand from a logical prior. Candidate desiderata include:

1. **Coherence:** A prior $\mathbb{P}(\cdot)$ is coherent if it is a probability distribution over complete theories. (This requires $\mathbb{P}(\phi) = 1 - \mathbb{P}(\neg\phi)$, and so on.)

2. **Computable approximability:** A prior $\mathbb{P}(\cdot)$ is computably approximable if there is an algorithm which computes an approximation of the probability which $\mathbb{P}$ assigns to ϕ that converges to $\mathbb{P}(\phi)$ in the limit.

3. **The Occam property:** A prior has the Occam property if there exists a length-based lower bound on the probability of any consistent sentence.

4. **Inductive:** A prior is inductive if its probability for sentences of the form $\forall n.\psi(n)$ goes to 1 as it conditions on more and more (going to all) confirmations of $\psi(\cdot)$.

5. **$\mathcal{PA}$-weakness:** A prior is $\mathcal{PA}$-weak if it places non-zero probability mass on the set of complete extensions of $\mathcal{PA}$.

6. Specifically, Demski's prior places zero probability mass on any theory which is not finitely axiomatizable. The induction schema of $\mathcal{PA}$ consists of infinitely many axioms, all independent of each other. With probability 1, Demski's generator will eventually select the negation of one of these axioms from Φ.

6. **Bounded regret:** It may be desirable to show that a prior has regret (in terms of log loss or some similar measure) at most a constant worse than any other probability distribution over complete theories.

7. **Practicality:** The more tractable the algorithm which approximates a prior, the more practical that prior is.

8. **Reflectivity:** A prior $\mathbb{P}(\cdot)$ is reflective if there is some symbol P in the logical language which can be interpreted as a representation of $\mathbb{P}(\cdot)$, such that $\mathbb{P}(\cdot)$ assigns accurate probabilities to statements about P.

Coherence is an extremely desirable property; while *approximations* of a logical prior must be incoherent, it is prudent to demand coherence in the distribution being approximated. Reflectivity has been shown to be possible (up to infinitesimal error) (Christiano et al. 2013) but difficult to do in a satisfactory manner (Fallenstein 2014).

Hutter's prior is coherent, inductive, and $\mathcal{PA}$-weak. It has the Occam property so long as the probability distribution which it is generated from has the Occam property; but it is not computably approximable and it is far from practical. By contrast, Demski's prior is coherent and computably approximable, and has the Occam property if Φ does, but lacks most other desirable properties.

Inductivity has been suggested in the literature, where it is better known as the *Gaifman condition* (Gaifman 1964). Roughly, the Gaifman condition requires that if $\mathbb{P}(\cdot)$ is a logical prior and ϕ is a true Π_1[7] sentence of the form $\forall n\colon \psi(n)$, then if $\mathbb{P}(\cdot)$ is conditioned on the truth values of the first N instances of $\psi(n)$, then as N goes to infinity, $\mathbb{P}(\phi)$ tends to 1. In other words, the Gaifman condition requires that if a reasoner learns that $\psi(n)$ is true for more and more n, then it eventually become arbitrarily confident that it is true for all n. However, computably approximable probability distributions which satisfy the Gaifman condition must assign probability 0 to some true Π_2 sentences (Sawin and Demski 2013),[8] which seems strongly undesirable: computably approximable priors that satisfy the Gaifman condition are not sufficiently "weak"; they prevent reasoners from deducing true sentences.

Hahn (2013) reports on an investigation of a more involved desirable property: If $\phi(\cdot)$ is a generic predicate symbol (that is, the initial set of axioms makes no claims about $\phi(\cdot)$), and if the prior is conditioned on the statement that $\phi(n)$ is true for exactly 90% of the first 10^{100} natural numbers, then the posterior probability of $\phi(0)$ should be (approximately) 0.9. (This captures some of the intuition that a tractable algorithm will assign probability ≈ 0.1 that the $(10^{100})^{\text{th}}$ digit of π is a 7, a desideratum that is difficult to formalize.) Unfortunately, even this property appears to be very difficult to obtain, and we are not aware of any proposed logical priors that have been shown to possess this property.

In large part, generating logical priors is difficult because it's not yet clear what properties such a prior should possess, nor which properties are possible. Continued investigation into well-behaved logical priors is warranted, as the development of satisfactory computably approximable logical priors promises insight into problems of reasoning under logical uncertainty more generally.

4 Beyond Logical Sentences

A study of logical uncertainty with respect to sentences of first-order logic has proven insightful, but even if a practically approximable prior distribution over complete theories were defined, it would not provide a full theoretical understanding of reasoning under logical uncertainty in practice.

Ultimately, logical sentences are not the right tool for reasoning about the behavior of objects in the real world. It is *possible* to construct a logical sentence which is true if and only if the Rube Goldberg machine deposits the ball in the top chute, but this sentence would be long and awkward. The manipulations that are easy to do to the sentence don't obviously correspond to realistic reasoning shortcuts about Rube Goldberg machines. Realistic reasoning under logical uncertainty will likely require hierarchical and context-filled models of the problem. It is possible that these things could be built *atop* practical methods for reasoning according to a probability distribution over logical sentences, but the ability to reason with uncertainty about the truth-values of logical sentences will not solve these problems directly.

Furthermore, while logical sentences are quite expressive, it is not clear that sentences of first order logic are the "correct" underlying structure for logically uncertain reasoning about the world. Practical logically uncertain reasoning inevitably requires reasoning about states of reality, and while most simple real-world questions can be translated into a sentence of first-order logic, it is by no means clear that uncertainty over logical sentences is the best foundations upon which to build practical reasoning.

By analogy, consider a billiards player who lacks knowledge of physics. This reasoner would do well to learn classical mechanics, not because it behooves the player to start modeling the billiards table in terms of individual atoms, but because various insights from classical mechanics apply at the high level of billiards. But though the billiards player may use knowledge from classical mechanics in their high-level model of

7. ϕ is a Π_1 sentence if it can be written in the form $\forall n\colon \psi(n)$ and there is a primitive recursive function which takes n as input and computes whether $\psi(n)$ is true or false.

8. ϕ is a Π_2 sentence if it can be written in the form $\forall m.\ \exists n.\ \psi(m,n)$, where $\psi(m,n)$ is primitive recursive.

the world, it is not the case that the high-level model is "merely" a computational expedient standing in for the "real" atomic model of reality. The atomic model, too, is simply a model, and one which does not quite explain all the phenomena in the quantum world of the billiards player.

We are like the billiards player: our state of knowledge is one where a study into uncertainty over logical sentences may provide significant insight that we can use to understand logical uncertainty as it pertains to "high level" objects, but this does not mean that practical logically uncertain reasoning will be done in terms of logical sentences, and nor does it mean that practical logically uncertain reasoning *could* always be reduced to uncertainty about logical sentences. It is merely the case that, given our present state of knowledge, a better understanding of logical uncertainty in the context of logical sentences is likely to provide insight into reasoning under logical uncertainty more generally.

5 Discussion

Smarter-than-human artificial systems must do most reasoning under both logical and environmental uncertainty. If high confidence in this reasoning is to be justified, even in a wide array of esoteric situations, then a theoretical understanding of logically uncertain reasoning is necessary: without it, it is difficult to ask the right questions (Soares and Fallenstein 2014a).

The development of reliable methods for reasoning under logical uncertainty is work that must be done in advance of the development of smarter-than-human systems, if those systems are to be safe. While it may be possible to delegate significant AI research to early smarter-than-human systems, the creation of reliable methods for reasoning under logical uncertainty cannot be delegated, because logically uncertain reasoning is precisely what the delegatee must use in order to perform its research! How could a smarter-than-human system be trusted to accurately discover methods for reasoning under logical uncertainty, while using unreliable methods of reasoning under logical uncertainty?

Furthermore, a better theoretical understanding of logical uncertainty is necessary in order to formalize many open problems related to the alignment of smarter-than-human systems. For example, consider the problem of constructing realistic world models: an agent faced with learning about the environment in which it is embedded must reason according to a distribution over environments that contain the agent. This problem must be fully described in order to check whether a practical program implements a solution, but describing the problem requires a better understanding of reasoning under logical uncertainty (Soares 2015).

Or consider the problem of counterfactual reasoning: formalizing the decision problem faced by an agent embedded within its environment requires some way to formalize the problem of agents which may have an accurate description of their program, but uncertainty about which action it will take. Specifying this problem, too, requires a better theory of logical uncertainty. In fact, satisfactory decision theory additionally requires an understanding of *logical counterfactuals*, the ability to reason about what "would" happen if a deterministic program did something that it doesn't. It is likely that a better understanding of logical uncertainty will yield insight in this domain (Soares and Fallenstein 2014b).

Many existing tools for studying reasoning, such as game theory, standard probability theory, and Bayesian networks, all assume that reasoners are logically omniscient. If these tools are to be extended and improved, a better understanding of logical uncertainty is required.

In short, a developed theory of logical uncertainty would go a long way towards putting theoretical foundations under the study of smarter-than-human systems. We are of the opinion that those theoretical foundations are essential to the process of aligning smarter-than-human systems with the interests of humanity.

References

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies.* New York: Oxford University Press.

Christiano, Paul. 2014. *Non-Omniscience, Probabilistic Inference, and Metamathematics.* Technical report 2014–3. Berkeley, CA: Machine Intelligence Research Institute. `http://intelligence.org/files/Non-Omniscience.pdf`.

Christiano, Paul F., Eliezer Yudkowsky, Marcello Herreshoff, and Mihaly Barasz. 2013. *Definability of Truth in Probabilistic Logic.* Working Paper. Machine Intelligence Research Institute, Berkeley, CA, April 2. `https://intelligence.org/files/DefinabilityTruthDraft.pdf`.

Demski, Abram. 2012. "Logical Prior Probability." In *Artificial General Intelligence: 5th International Conference, AGI 2012, Oxford, UK, December 8–11, 2012. Proceedings,* edited by Joscha Bach, Ben Goertzel, and Matthew Iklé, 50–59. Lecture Notes in Artificial Intelligence 7716. New York: Springer. doi:`10.1007/978-3-642-35506-6_6`.

Fallenstein, Benja. 2014. *Procrastination in Probabilistic Logic.* Working Paper. Machine Intelligence Research Institute, Berkeley, CA. `http://intelligence.org/files/ProbabilisticLogicProcrastinates.pdf`.

Gaifman, Haim. 1964. "Concerning Measures in First Order Calculi." *Israel Journal of Mathematics* 2 (1): 1–18. doi:`10.1007/BF02759729`.

———. 2004. "Reasoning with Limited Resources and Assigning Probabilities to Arithmetical Statements." *Synthese* 140 (1–2): 97–119. doi:`10.1023/B:SYNT.0000029944.99888.a7`.

Gödel, Kurt, Stephen Cole Kleene, and John Barkley Rosser. 1934. *On Undecidable Propositions of Formal Mathematical Systems.* Princeton, NJ: Institute for Advanced Study.

Hacking, Ian. 1967. "Slightly More Realistic Personal Probability." *Philosophy of Science* 34 (4): 311–325. http://www.jstor.org/stable/186120.

Hahn, Jeremy. 2013. *Scientific Induction in Probabilistic Mathematics.* Brief Technical Note. Machine Intelligence Research Institute, Berkeley, CA. http://intelligence.org/files/ScientificInduction.pdf.

Halpern, Joseph Y. 2003. *Reasoning about Uncertainty.* Cambridge, MA: MIT Press.

Hutter, Marcus, John W. Lloyd, Kee Siong Ng, and William T. B. Uther. 2013. "Probabilities on Sentences in an Expressive Logic." *Journal of Applied Logic* 11 (4): 386–420. doi:10.1016/j.jal.2013.03.003.

Łoś, Jerzy. 1955. "On the Axiomatic Treatment of Probability." *Colloquium Mathematicae* 3 (2): 125–137. http://eudml.org/doc/209996.

Sawin, Will, and Abram Demski. 2013. *Computable probability distributions which converge on Π_1 will disbelieve true Π_2 sentences.* Machine Intelligence Research Institute, Berkeley, CA, July. http://intelligence.org/files/Pi1Pi2Problem.pdf.

Soares, Nate. 2015. *Formalizing Two Problems of Realistic World-Models.* Technical report 2015–3. Berkeley, CA: Machine Intelligence Research Institute. https://intelligence.org/files/RealisticWorldModels.pdf.

Soares, Nate, and Benja Fallenstein. 2014a. *Aligning Superintelligence with Human Interests: A Technical Research Agenda.* Technical report 2014–8. Berkeley, CA: Machine Intelligence Research Institute. https://intelligence.org/files/TechnicalAgenda.pdf.

———. 2014b. *Toward Idealized Decision Theory.* Technical report 2014–7. Berkeley, CA: Machine Intelligence Research Institute. https://intelligence.org/files/TowardIdealizedDecisionTheory.pdf.

Yudkowsky, Eliezer. 2008. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks,* edited by Nick Bostrom and Milan M. Ćirković, 308–345. New York: Oxford University Press.