Anonymized Reviews of Three Recent Papers from MIRI's *Agent Foundations* Research Agenda

Compiled by the Open Philanthropy Project in 2016

Reviews for Paper 1

Benja Fallenstein and Ramana Kumar. 2015. "Proof-Producing Reflection for HOL: With an Application to Model Polymorphism." In *Interactive Theorem Proving: 6th International Conference*, ITP 2015, Nanjing, China, August 24-27, 2015, Proceedings. Springer.

One external review, two internal reviews.

External Review

Summary of the paper

What are the main questions this paper is addressing?

The paper considers the problem of reflection, which in this case has to do with allowing a formal logical system to reason about itself. Here we face the classic tension between competing goals: we want our proof-checking machinery to be simultaneously flexible, to support encoding of as many useful arguments as possible; and trustworthy, relying on as few axioms as possible and supporting a short proof-checker implementation.

One fairly natural thing to want to do in a formal system is reason about the convincingness of the formal system's own proofs. The paper highlights the potential application area of self-updating systems, where, for instance, some software wants to install a new version of itself, but only after analyzing that new version to make sure it is reasonable in some way. Gdel's incompleteness theorems established that very general logical systems cannot encode proofs of their own consistency, which on the face of it seems to rule out a software program, proved correct in some logic, that is able to check proofs in the same logic that come attached to future updates. However, Gdel's theorems don't rule out a sound formal system that can check proofs in a (perhaps only slightly) different logic.

What are its main conclusions?

This paper presents an implementation of a new approach to internalizing such reasoning patterns within the HOL logic. Consider a self-updating system that should be able to survive n layers of checking future versions, which can then survive $n - 1$ layers of checking future versions, and so on. A general HOL assumption is defined, parameterized on n, asserting the existence of n nested inaccessible cardinals. Such an assumption basically says "hey, there are some pretty big sets out there!" It's a common sort of assumption in the world of metamathematics.

An embedding of HOL into itself is defined, partly drawing on the authors' past work. The embedding is parameterized on a correct implementation of set theory. One new result is that the existence of the right inaccessible cardinal implies the existence of a suitable set-theory implementation. Another new result is that, if n inaccessible cardinals are enough to prove some formula in the reflected world, then that formula is also true "in reality", so long as $n + 1$ inaccessible cardinals exist.

A prototype implementation to has been built. It automatically proves reflection theorems of the kind I just sketched.

Novelty

Do you know of other investigators or groups pursuing similar questions?

Please also comment on other investigators or groups that could be in a good position to pursue similar questions.

Logical systems that can reason about themselves are a venerable topic, having been studied by Turing, among others, as the paper mentions. What is relatively less common is tackling such questions from an engineering perspective, building systems that can be used to check actual theorems automatically. This

paper acknowledges its inheritance from Harrison's work on reflecting HOL in HOL, which I'll cite here for completeness:

Harrison, John. *Towards self-verification of HOL Light.* IJCAR 2006.

The CakeML and Milawa projects cited in this paper have also been leaders in the space of reflected HOL.

The paper mentions computational reflection as a closely related but distinct idea, centered as it is on optimizing the performance of proof procedures, rather than on classic metatheory questions. In this approach, one implements an executable proof procedure in the logic (taking as input reified syntax trees of logical formulas) and proves it sound once and for all. Computational reflection seems to be most associated today with proof assistants based on type theory. All the major platforms of that kind have had recent work on computational reflection. A few examples:

- *Agda*: Kokke, Pepijn and Wouter Swierstra. Auto in Agda. MPC 2015.
- *Coq*: Malecha, Gregory, Adam Chlipala, and Thomas Braibant. Compositional Computational Reflection. ITP 2014.
- *Idris*: Brady, Edwin. First-class Type-safe Reflection in Idris. DTP 2013.

It's a pretty abstruse area, generally avoided even by the committed user communities of these proof assistants. In my opinion, all of the investigators in the area are extremely competent, both as implementors and as planners of higher-level research agendas.

There has also been work with Coq that is closer in spirit to what this paper presents.

Barras, Bruno. *Sets in Coq, Coq in Sets.* Journal of Formalized Reasoning (2010).

If correct, what would the paper's conclusions add to what is already known?

As far as I know, this paper gives the first demonstration of a system that can support an arbitrarily deep tower of reflective reasoning within a general-purpose mechanized logic. That is, we can reason about a program that checks proofs about programs that check proofs about programs that check proofs... up to a depth based on how large of an inaccessible cardinal we assume. It's a new and intuitively appealing capability that is nontrivial to implement.

Technical quality

Did the paper address its main questions in a logically defensible way based on reasonable premises?

The main evidence in favor of the approach is a prototype implementation, which proves families of reflection principles. The implementation itself cannot be verified so straightforwardly, but presumably it has been put through its paces, checking that it generates the expected principles for a variety of different parameters. The logical foundations (e.g., which set-theoretic axioms to assume) seem to be standard for this domain.

Though the paper presents an implementation of the reflection principle, it does not discuss any case study using that principle for a verification result of inherent interest. That is, while scenarios of self-updating programs are used as motivation, it appears that no such scenarios have been explored concretely yet. It could happen that the approach fails to hold up when applied at scale, failing in ways that are hard to

predict today. We can view this omission either as a flaw in the paper or as a clear suggestion of follow-on work, as the authors make in the conclusion.

A subtlety here that might easily be missed by non-experts is that it does not seem that the authors have developed any machinery for checking syntactic proofs embedded within HOL, which seems necessary for the proposed applications in self-updating programs. That is, while a proof system is formalized, it does not seem that an executable checker for it has been implemented within HOL.

Did the paper build on and draw from the most important and relevant prior work?

Yes, I believe that the most relevant prior work is suitably built upon, especially since one of the authors was involved in some of that work.

How hard would it be for someone else to derive results like these?

(We're looking for an answer of the rough form, "If working on these questions, a graduate student in my department could make comparable progress with about a month of work" or "There are only a few people in our field who would be able to make comparable progress on these questions with 6 months of work", or somewhere in between.)

My back-of-the-envelope estimate is that there are about 20 people in the world today with the kind of background to begin a project like this immediately. What is required is a mix of metamathematical knowledge and software-implementation skills; each is rather rare at the level of ability required, and they don't occur together too frequently today. The universities of the world have the capacity to graduate a few PhD students per year with the required skills, but mostly those educational resources are turned toward other, more fashionable topics today.

For those with the right background, the implementation presented here probably wouldn't be a huge task. I would estimate that the authors completed it in less than a year of combined work, if they were sufficiently close to working on it full-time. The conceptual design decisions are the less predictable part.

So, to summarize, (1) there is a relatively small set of people in the world qualified to do this kind of work; (2) some nontrivial conceptual leaps are required to plan out the project; and (3) a member of this class of very-qualified people can probably complete the implementation in well under a year of full-time effort.

Significance

The authors of this paper think that its results shed light on standards for good reasoning under deductive limitations. They say more about that here.

How significant do you feel these results are for that?

I see these results as an important milestone toward formal analysis of systems with some level of self-understanding. It is well-understood "folklore" in the area that, to support sound checking of a system's own proofs within itself, some sort of stratification is necessary, and inaccessible cardinals had already been used that way (for instance, by Barras in Sets in Coq, Coq in Sets). The conceptual jump in the present paper seems to be (1) identifying the potential connection to e.g. self-updating software and (2) implementing the first concrete machinery for towers of reflective self-reasoning in a general-purpose logic (HOL).

As I wrote above, it is important to be clear on how far this paper goes, along the path to reasoning about self-updating software. No such case study has been carried out, and it does not seem that even algorithmic proof-checking within HOL has been implemented yet. It is hard to predict how far these methods can go in a practical application, which definitely puts this project in the high-risk-high-reward quadrant, and it seems that the authors are looking to that sort of experiment as the next follow-on project.

Do the methods and results seem potentially fruitful in the sense that they or related work could shed additional light on these issues in the future?

Yes, the work to date belongs to a short list of projects that have explored these themes with realistic implementations, and I believe that supporting follow-on work is likely to have a significant positive effect on our understanding of the pragmatics of algorithmic systems that reason about the code of very similar (though never quite exactly the same!) systems. I believe that the authors should be encouraged to structure a follow-on project around a particular case study that shows a specific software system in action, applying the metatheoretical results in a way that demonstrates concrete benefit understandable to, say, expert software developers who have not used proof assistants.

Clarity

Is the paper written in a way that will allow others to understand it and build on it?

I found the paper to be exceptionally well-written, and I have no concerns about the ability of others to use it as the basis for follow-on work. I would say that significant background in proof assistants is required to follow along, but that no familiarity is required with the particular implementation approach that the authors build on, coming from papers published in the last 10 years.

Internal Review 1

Summary of paper and its potential significance for AI safety (this can be very brief)

What are the main questions this paper is addressing?

a) Can we implement a natural reflection principle in a concrete theorem prover, and instantiate parametric model polymorphism?

b) Parametric model polymorphism is an idea Benya proposed a while ago, it is not exactly clear whether the idea itself is to be considered a contribution in this paper (given that it hasn't previously appeared in peer-reviewed work). I will not treat it as one.

What are its main conclusions?

They implement a reflection principle and parametric model polymorphism.

How does MIRI think these questions are related to potential risks from advanced AI?

a) One important problem for advanced AI systems is designing new AI systems. If we want to have provable guarantees about such systems, and about the systems they create, and so on ad infinitum, it seems we would need some kind of reflection like this.

b) MIRI believes that the most natural approaches to robust reasoning will be meaningfully analogous to logical reasoning, such that the obstructions to obtaining proofs will also be obstructions for practical robust reasoners.

c) Parametric polymorphism is one of the more satisfying approaches to this problem in the context of logical reasoning, and pretty much any approach is going to involve some reflection principle of the kind that they implement. Actually implementing these systems in a concrete theorem prover is a natural way to understand them better. Maybe you'll learn something surprising that you overlooked in informal arguments (there are a lot of details). For many people, having an implementation is also a prerequisite to getting them to take you seriously, since they will otherwise be skeptical of informal arguments.

What obstacle(s) to developing safe AI does MIRI think work in this direction could help to overcome?

See my previous answer.

Comments on the novelty, technical quality, and difficulty of the results (optional)

I think that the file system safety example is pretty good and it seems like a natural problem. Parametric polymorphism seems like a nice solution to this problem and I'm not aware of alternatives. But I'm not really qualified to comment on technical merit of this work.

Significance (most important)

Consider again the obstacle(s) to developing safe AI that MIRI thinks this work may help overcome.

Do you think this is a real/important obstacle?

a) I am skeptical. I agree with Paul Christiano's take on this question here.

b) I do think that there is room for legitimate disagreement and that anyone who thinks the case is open-and-shut has probably not seriously engaged with the argument in favor. Mathematical proof is in fact powerful machinery that has had a historical influence, there is a legitimate hole in our understanding of how to apply mathematical proof to systems like this one, and that hole may be resolvable. Moreover, we are sufficiently uncertain about what future reliable reasoners will look like that we ought to at least entertain the possibility it will encounter some of the same difficulties as mathematical proof.

c) That said, I think that the issues with reflection in logical systems are unlikely to be serious issues for practical AI systems.

   (a) First, I think that they probably are non-issues for probabilistic reasoners who use the same kinds of evidence that humans use about their own reliability. This is based on my own analysis, but I've considered the problem in significant detail.

   (b) Second, based on the recent history of AI I think that we are unlikely to develop such a strongly principled understanding of the reasoning used by practical AI systems.

Assuming the obstacle is real, how much do the results in this paper address the obstacle?

I think it's a meaningful step forward. This is the first paper that actually implements the kind of reflection principle that is potentially interesting for AI. They implement the novel part of a solution to a nice model problem (an OS that never overwrites some file, but which is able to rewrite its own code). It would be more compelling if they surveyed other possible solutions to that model problem. Is it the case that their technique is actually the only existing approach? It kind of looks like it, but it's a bit hard to tell.

Is the approach in the paper well-suited to overcoming the obstacle compared to other possible approaches?

Can you think of other research directions/approaches/people that might be better positioned to overcome the obstacle?

a) This depends on exactly how you slice things up, and how broadly we define the obstruction.

b) For some ways of slicing it up this is a very promising approach to the obstruction but I don't believe the obstruction is real (because there are ways to route around it)—see above.

c) For other ways of slicing things up, this is a real obstruction (AI systems really will need to help us build new AI systems, and that really does involve some tricky reasoning), but I don't believe that this is a productive step towards resolving that obstruction.

What are the most likely ways that this work could turn out to be relevant to the safety of future ML systems?

a) It's possible that we will in fact prove strong correctness properties of the AI systems we build (e.g. proving that they will never do something "obviously dumb") while at the same giving them a very rich and flexible space of actions. If we could do this it would be quite impressive, and this kind of work would probably be very relevant.

b) It's also possible that logical reasoning will be usefully analogous to whatever kinds of reasoning practical systems do, even if there are no proofs involved. I think there are reasons to be skeptical of this but again it's not open-and-shut.

What are the most likely ways that this work could turn out to be irrelevant to the safety of future ML systems?

a) I think it is most likely that we won't be able to prove strong claims about the AI systems we build, and that principled probabilistic reasoning wouldn't encounter analogous problems with reflection.

b) I think it's likely that we won't be able to rely on highly principled reasoning, and instead will have to build systems that are trained to reason well but whose reliability is ultimately an empirical question. (Either [a] or [b] would probably make this research irrelevant.)

Overall, how promising does this work seem in comparison with other approaches to developing safe AI?

Please weigh both probability of relevance and importance if relevant.

Fairly typical. This work seems about as promising as other research I have seen proposed (e.g., FLI proposals.)

In my view, most research on AI safety is relatively speculative and only likely to be useful under a particular set of assumptions about how AI is developed. I think that the assumptions underlying this work are moderately less plausible than the implicit assumptions underlying some other work on AI safety (though not as implausible as it may at first appear), but the case for relevance (given those assumptions) is significantly stronger.

Internal Review 2

Summary of paper and its potential significance for AI safety (this can be very brief)

What are the main questions this paper is addressing?

How can we implement reflection and model polymorphism inside a theory?

What are its main conclusions?

They implement reflection in HOL (and release source code) in a way that requires more typical axioms than previous implementations. They extend it to model polymorphism.

How does MIRI think these questions are related to potential risks from advanced AI?

This paper does not discuss its connection to safety research, besides some brief allusions. However, the reviewer understands it to be addressing a concern MIRI often raises about the safety of AI systems.

In this concern, MIRI imagines a powerful agent that reasons in a very logical way, proving theorems about the world. Such agents would need to reason about self-modification or creating new agents. This would seem to introduce fundamental difficulties, because the agent now needs to prove theorems about systems that prove theorems, and so on, running into Godelian/Lobian issues.

MIRI believes that studying this problem may shed light on more general issues in other kinds of agents.

What obstacle(s) to developing safe AI does MIRI think work in this direction could help to overcome?

By demonstrating a way to embed HOL within itself, the paper takes a step towards solving this problem.

Comments on the novelty, technical quality, and difficulty of the results (optional)

I'm not well qualified to judge the technical quality of MIRI's work, but this seems impressive, attacking really hard topics in logic with sophisticated tools. Of the MIRI work I'm aware of, it looks like the most impressive.

Significance (most important)

Consider again the obstacle(s) to developing safe AI that MIRI thinks this work may help overcome.

Do you think this is a real/important obstacle?

I think we probably don't live in a world where this particular problem is an issue. This is mostly because I expect agents to be heuristic reasoners, which use logic when useful but aren't fundamentally reasoning by proving theorems about things. I am doubtful that the kind of foundational logic problem this paper is attacking has relevance to such systems.

We might also create tool AIs that arent agents at all.

If we did create a kind of "theorem prover" AI, this would be a bit more of an issue, although I think there would be easier, if less satisfying, solutions (discussed more later).

Assuming the obstacle is real, how much do the results in this paper address the obstacle?

The work of this paper advances a natural approach to solving the problem.

Is the approach in this paper well-suited to overcoming the obstacle compared to other possible approaches?

Can you think of other research directions/approaches/people that might be better positioned to overcome the obstacle?

It seems like there are easier ways out of the Godelian obstacle than MIRI typically pursues. For example, it seems like one should be able to have agents produce new agents which produce new agents...up to any finite depth by repeatedly adding reflection axioms to your system. Alternatively, the agent might be able to self-improve by having a modular structure and axioms that allow it to improve certain subcomponents as long as they continue to maintain relatively weak properties.

MIRI is clearly aware that these are possibilities, but doesn't find them fully satisfying. That seems reasonable: if we were really going to have theorem proving based agents, I'd be inclined to think this was somewhat worthwhile, even if I thought the problem would be solved in other ways. However, combined with how improbable the problem itself seems, I feel really unexcited about this as safety research.

What are the most likely ways that this work could turn out to be relevant to the safety of future ML systems?

If we end up with very different systems than we are presently on trajectory for, and those systems reason in a very "theorem proving" / rules based way.

What are the most likely ways that this work could turn out to be irrelevant to the safety of future ML systems?

- If we end up with primarily heuristic reasoners.
- If we can't wrap around these approaches with heuristics.
- If the easy ways out don't work in some situations.

Overall, how promising does this work seem in comparison with other approaches to developing safe AI

Please weigh both probability of relevance and importance if relevant.

Less promising than usual. This is because:

- I don't think we live in a world where this is an issue.
- If we do live in such a world, we're probably on a much longer timeline and it's less urgent.
- If we do live in such a world, there are probably easier ways out of the problem.

Reviews for Paper 3

Scott Garrabrant, Benya Fallenstein, Abram Demski, and Nate Soares. 2016. "Inductive Coherence." arXiv:1604.05288 [cs:AI]. Previously published as "Uniform Coherence."

Two external reviews, one internal review.

External Review 1

Summary of the paper

What are the main questions this paper is addressing?

The paper is trying to come up with methods to answer questions of the form "How likely is it that this computation will output 3?".

The assumption is that these questions are expressed as formulas in a rich logic (like Peano Arithmetic). By assigning a probability to all formulas in Peano Arithmetic (or some other suitably rich logic), we get, in particular, a probability on statements like "How likely is it that this computation will output 3?".

Unfortunately, the problem of defining a probability distribution on Peano Arithmetic that is coherent (which simply means that it satisfies some minimal desiderata for probability, like the probability of "false" is 0 and the probability of two mutually exclusive formulas f_1 and f_2 is the sum of the individual probabilities of f_1 and f_2) is undecidable; that is, there is no Turing Machine (TM) that, given as input a formula, returns its probability. So the authors back off from that goal and try to find an approximation to a probability distribution. Their notion of an approximation to a probability distribution Pr is a TM M that gets as input both a number n and a formula f such that $M(n, f)$ converges to $Pr(f)$ as n gets large. But the authors do not want an arbitrary approximation scheme; they want one that they call uniformly coherent, which means it satisfies some additional properties.

What are its main conclusions?

The key results of this paper are (1) the definition of uniform coherence, (2) the proof of some (fairly straightforward) properties of uniform coherence, and (3) the proof of the existence of a uniform coherent approximation scheme.

Novelty

Do you know of other investigators or groups pursuing similar questions?

Please also comment on other investigators or groups that could be in a good position to pursue similar questions.

I'm not aware of anyone doing research on anything quite like this. At a very high level, there is some overlap between the type of work considered here and work on making decisions optimally in the face of computational constraints, since making such decision might involve approximating probability. Work on the latter topic goes back to Russell and Horvitz in computer science; it is currently a hot topic in Cognitive Science. (In the interests of full disclosure, it is also a topic that I am actively working on.). However, as I said, the thrust of this paper is quite different (and to my mind, less interesting – more on this below). In spirit, the paper is also close to work going back to the 1960s on language identification in the limit. Work on this topic has been largely superseded by a weaker model, Valiant's notion of PAC (probably approximately correct) learning. (As an aside, if I were a referee of this paper for a journal, I would ask the authors to compare their work to Gold's work on language identification.)

If correct, what would the paper's conclusions add to what is already known?

See "Significance" below.

Technical quality

Did the paper address its main questions in a logically defensible way based on reasonable premises?

Not answered.

Did the paper build on and draw from the most important and relevant prior work?

Not answered.

How hard would it be for someone else to derive results like these?

(We're looking for an answer of the rough form, "If working on these questions, a graduate student in my department could make comparable progress with about a month of work" or "There are only a few people in our field who would be able to make comparable progress on these questions with 6 months of work", or somewhere in between.)

Not answered.

Significance

The authors of this paper think that its results shed light on standards for good reasoning under deductive limitations.

How significant do you feel these results are for that?

As I said, I don't find the results of this paper particularly interesting. The first paragraph suggests that this problem is motivated by the concern of assigning probabilities to computations. This can be viewed as an instance of the more general problems of (a) modeling a resource-bounded decision maker computing probabilities and (b) finding techniques to help a resource-bounded decision maker compute probabilities. I find both of these problems very interesting. But I think that the model here is not that useful for either of these problems. Here are some reasons why:

1. It's not clear why the properties of uniform coherence are the "right" ones to focus on. Uniform coherence does imply that, for any fixed formula, the probability converges to some number, which is certainly a requirement that we would want. This is implied by the second property of uniform coherence. But that property considers not just constant sequences of formulas, but sequence where the n^{th} formula implies the $(n+1)^{\text{st}}$. Why do we care about such sequences? I see no reason that

why the approximations to probability used by actual decision makers should satisfy these properties (important for desideratum (a)) nor why these are the particular properties we should care about to help decision makers in making decisions (desideratum (b)).

2. The TM that computes the uniformly coherent approximation scheme given in Theorem 3.1 runs in time double-exponential in n on input (n, f). My guess is that any approximation scheme would be similarly slow. The issue of computational complexity is not discussed in the paper, but it is clearly highly relevant.

3. Suppose that a decision maker was actually using a uniformly coherent approximation scheme and had to make a decision regarding the probability of formula f after a short time. Presumably he would want to use $M(n*, f)$ for a small value of $n*$. But what assurance does the decision maker have that $M(n*, f)$ is a good approximation to the desired probability? I strongly suspect that the problem of deciding how close $M(n*, f)$ is to the limiting value of $M(n, f)$ is undecidable. (More precisely, given any epsilon, the problem of deciding whether $M(n*, f)$ is within epsilon of the limit is undecidable.) This makes approximation schemes not terribly useful for the problem of approximating real probabilities.

4. Suppose that we're interested in a formula f. Now if f follows from Peano arithmetic (or whatever the underlying theory being used is), then $M(n, f)$ converges to 1, as we would hope. Similarly, if the negation of f follows from Peano arithmetic, then $M(n, f)$ converges to 0. But if f is independent of Peano arithmetic, to what extent does $M(n, f)$ converge to a value that really is a useful probability (useful in terms of corresponding to any notion of uncertainty of interest for the application at hand)?

5. There is no obvious way of incorporating prior beliefs in uniformly coherent approximation schemes.

I see no obvious modification of uniformly coherent schemes that would address these concerns. Even worse, despite the initial motivation, the authors do not seem to be thinking about these motivational issues. But ignoring motivational concerns, the paper is written clearly and the theorems seem correct.

Do the methods and results seem potentially fruitful in the sense that they or related work could shed additional light on these issues in the future?

Not answered.

Clarity

Is the paper written in a way that will allow others to understand it and build on it?

Not answered.

Further comments

Additional comments: I've looked (not very carefully) at 2-3 other MIRI papers, and I had much the same reaction in terms of motivation. These are smart guys, but they have no real computer science sensibilities (although their steering committee certainly has terrific folks with great CS sensibility!). I found myself unexcited by the particular problems they were trying to solve (although this should be taken with a huge grain of salt; I didn't look at the papers carefully). But I am quite enthusiastic about the general space they were working in.

External Review 2

Summary of the paper

What are the main questions this paper is addressing?

Classically, the study of probability has been partitioned into two schools of thought. Frequentists view probability as measuring the frequency that some event will occur under repeated experiments. Bayesians view probability as quantifying an observer's uncertainty about some event (what odds would you place a bet on the event occurring). However, even Bayesians typically consider only uncertainty due to a lack of information and not due to a lack of computation. Thus for example, even though we have no idea what is the 10100-th decimal digit of π, traditional Bayesian analysis would say that the probability that this digit is 7 is equal to either zero or one, since the question is fully specified and it's only a matter of computation.

Intuitively, we would like to have a theory that would assign to the statement "The 10100-th digit of π is 7" a probability of roughly 1/10, but would still satisfy some basic "sanity checks" such as that if we sum the probabilities of these statements over all the digits from 0 to 9 then we get the value 1. This is what this paper attempts to achieve.

What are its main conclusions?

The paper comes up with a variant of a notion of a "nice probability theory" called "uniform coherence" which is a variant on definitions proposed in the past. The idea is that this stronger notion should somehow give stronger assurances on the quality of approximation to the probabilities that we obtain in finite time. Their main results are this definition and an algorithm that achieves it.

Whether these assurances, and the related algorithm, have important significance is a matter for debate. It is to a large extent a subjective question. This reviewer is not extremely impressed but others might feel differently.

What I would have liked to see are concrete natural examples where their algorithm assigns some natural probabilities and prior constructions do not.

Also, there is an inherent issue with algorithms that work by enumerating over all proofs. They run in exponential time and even practically it seems that this enumeration will quickly explode before we see any reasonable probabilities.

Novelty

Do you know of other investigators or groups pursuing similar questions?

Please also comment on other investigators or groups that could be in a good position to pursue similar questions.

There are several people that study similar questions. This reviewer is not an expert on all of the relevant literature, but one line of work that is not cited but is relevant is the work on proof complexity and related convex relaxations that yield probabilities (for some pointers, see this blog post; this paper from the MIRI website also mentions some of those proof systems, though it doesn't include citations).

If correct, what would the paper's conclusions add to what is already known?

Not answered.

Technical quality

Did the paper address its main questions in a logically defensible way based on reasonable premises?

Did the paper build on and draw from the most important and relevant prior work?

How hard would it be for someone else to derive results like these?

Yes I think the paper addresses its questions in a logically defensible way and builds on some prior work. I cannot judge how hard it would have been to do so.

(We're looking for an answer of the rough form, "If working on these questions, a graduate student in my department could make comparable progress with about a month of work" or "There are only a few people in our field who would be able to make comparable progress on these questions with 6 months of work", or somewhere in between.)

Significance

The authors of this paper think that its results shed light on standards for good reasoning under deductive limitations.

How significant do you feel these results are for that?

I don't believe the authors made a strong enough case that they shed light on how to practically assign probabilities in settings where we have uncertainty due to a computational limited budget.

Do the methods and results seem potentially fruitful in the sense that they or related work could shed additional light on these issues in the future?

Not answered.

Clarity

Is the paper written in a way that will allow others to understand it and build on it?

It could be clearer. In particular, the abstract should state clearly and precisely what is the main result.

Internal Review 1

Summary of paper and its potential significance for AI safety (this can be very brief)

What are the main questions this paper is addressing?

This paper introduces uniform coherence, a property one might wish for in systems assigning probabilities to mathematical propositions.

What are its main conclusions?

The paper provides an algorithm that satisfies uniform coherence, demonstrating that it is theoretically possible. They also help us build intuition about uniform coherence by proving a few properties about it.

How does MIRI think these questions are related to potential risks from advanced AI?

No answer given.

What obstacle(s) to developing safe AI does MIRI think work in this direction could help to overcome?

The paper does not address these questions, or even suggest that is addressing safety. Their blog post also doesn't really directly address this.

They do say, regarding this paper and the uniform convergence paper, "for us, the exciting result is that we have teased apart and formalized two separate notions of what counts as "good reasoning" under logical uncertainty, both of which are compelling." But they don't explicitly spell out the how they see this applying to safety and I'd have to extrapolate a fair amount.

(One thing that does seem striking to me is this line by Nate in the blog post: "if you give a classical probability distribution variables for statements that could be deduced in principle, then the axioms of probability theory force you to put probability either 0 or 1 on those statements." I'm not sure this is an accurate presentation of the situation. Rather, "coherent distributions," a kind of distribution over statements, requires this. In fact, their paper seems to cite some other work that doesn't require it.)

Comments on the novelty, technical quality, and difficulty of the results (optional)

This paper was significantly better written than the average paper I review for conferences, and somewhat better written than the average paper I read.

Significance (most important)

Consider again the obstacle(s) to developing safe AI that MIRI thinks this work may help overcome.

Do you think this is a real/important obstacle?

Most powerful AI systems I can think of involve a component of "reasoning" – evaluating a hypothesis not just by immediate pattern recognition, but by considering its logical relationships to to other hypotheses. If an AI system's reasoning was pathological in some way that could be extremely dangerous. I don't think this is a very likely way for a system to become unsafe, but it is possible.

Since we don't know how full reasoning will work, studying it in simplified settings is natural. Assigning probabilities to logical propositions seems like a nice simplified setting for studying this problem.

Assuming the obstacle is real, how much do the results in this paper address the obstacle?

Unfortunately, I don't think the paper does much to address the problem of potentially pathological reasoning. There are a few reasons for this.

1. Perfection vs Safety. Uniform coherence seems to be an attempt at capturing something like "eventually perfect reasoning." It guarantees that the predictor will eventually take advantage of certain kinds of patterns. This seems more like a statement about the strength of the model than addressing some subtlety about safety. A system that doesn't recognize some class of pattern could be perfectly safe as long as it is well calibrated, assigning reasonable uncertainties in the absence of its ability to prove something. Conversely, a system that has uniform coherence might be dangerous when used with a finite computational budget; for example uniform coherence doesn't prohibit the system from saying it is certain something is true when it hasn't proved it, as long as it would fix this with more compute.
2. Asymptotic Nature. Uniform coherence is a guarantee about what a systems will eventually do, given enough time to think. Most of the work goes into making sure it will start to do reasonable things before it reaches its final answer, but those are still eventual properties. It's not clear to me that it sheds any light on a system that just thinks for a finite amount of time and then acts based on what it's concluded at that point.
3. No Deep Transferable Insights. Uniform coherence is not the ultimate notion of good or perfect reasoning. MIRI clearly doesn't think that it is. That being the case, the value of this work would be conceptual insight or the introduction of building blocks that might be used in developing more sophisticated ideas.

   Since the core of this paper seems to be dealing with very abstract asymptotic properties of systems, I'm not optimistic that it will help us with reasoning about real systems. (For a real system, I expect us to be able to say things like "after n steps XYZ" quite easily; see the next section on alternatives.) Beyond that – while I don't think I understand this result deeply enough to judge it with confidence – it just doesn't feel to my aesthetic like the sort of deep result that illuminates lots of other things.

   Perhaps more interesting is that MIRI has another notion of good reasoning and is finding the two difficult to reconcile (this is what their blog post says is important). I'm still not super optimistic, but I could imagine that panning out to be something I'd be more excited about.

My conclusion is that the value of this paper would have to rest on conceptual insight and cognitive tools it gives us for thinking about the reasoning algorithms. That's hard to asses, but I don't feel very optimistic

about it. I wouldn't be surprised if a lot of work in theorem proving or mathematical logic was about equally useful.

Is the approach in the paper well-suited to overcoming the obstacle compared to other possible approaches?

Can you think of other research directions/approaches/people that might be better positioned to overcome the obstacle?

I think stronger results in a similar vein to the work of this paper would easily fall out of studying a real system.

For example, one can consider AlphaGo to do be doing something like theorem proving: assigning probabilities that it can win a game. There is an analogous property to coherence, here: if it has a winning strategy for board state s2, then it should also realize it has a winning strategy for board state s1 which can be turned into s2 by taking an action. One can then ask whether this system will converge to perfect behavior given infinitely large tree searches (it will) or a perfectly flexible function approximator and infinite training (it will). Then one can ask about it's behavior as one scales it up: I'm fairly certain the case that you scale up both will easily give you nice properties about the convergence and increasing coherence of the distribution, much stronger than the analogy of uniform coherence.

The bigger issue, however, is that I don't think the asymptotics of these systems is really the right thing to be worrying about in this space. Instead, I want to know if a fixed system is behaving reasonably or not. There is basic research in reinforcement learning that feels a lot more promising to me in this regard. For example, it is understood that Q-learning, a very popular algorithm for training RL systems, is systematically over optimistic about the value of states. In a reasoning system, this could translate to systematic overconfidence. Recently, an algorithm called double Q-learning was developed that fixes this. It was then applied to modern deep RL and improved performance there. This to me is an example of what I by default expect good research in this area to look like.

What are the most likely ways that this work could turn out to be relevant to the safety of future ML systems?

- I haven't understood this result and have misjudged it. (I think this is very possible.)
- We see radically different AI systems than the natural extrapolation of deep RL.
- We have a very fast take off, such that any forethought is super important, even if it is much easier to reason about the actual system we're working with.
- We operate in a near infinite-compute situation, where the asymptotic properties of algorithms as we scale them is the critical issue.

What are the most likely ways that this work could turn out to be irrelevant to the safety of future ML systems?

- Reasoning isn't where we see most safety issues.
- Safety problems relating to reasoning are subtle pathologies, not the system being "weak" in the sense of failing to recognizing patterns or sub-propositions or outputting nonsense until it proves things.

- Reasoning systems operate at a relatively restricted computational budget and make lots of heuristic decisions, such that their asymptotic properties aren't very important.
- It's very easy to get traction on the asymptotic reasoning properties of a system if that's relevant. (Much more so than reasoning in the abstract.)

Overall, how promising does this work seem in comparison with other approaches to developing safe AI?

Please weigh both probability of relevance and importance if relevant. Somewhere between "No special claim to relevance" and "Less promising than usual". This is because, as described in more depth earlier, I think:

- I don't think reasoning flaws are very likely to be where safety problems arise from.
- If we have reasoning problems, I expect them to be issues of some pathology, rather than our system not exploiting certain patterns or assigning bad probabilities until it completely solves a problem.
- If we do have problems of this form, this paper's contribution would be conceptual insight. I don't expect these contributions to be much more helpful than a lot of other work in mathematical logic or theorem proving.

Reviews for Paper 4

Scott Garrabrant, Nate Soares, and Jessica Taylor. 2016. "Asymptotic Convergence in Online Learning with Unbounded Delays." arXiv:1604.05280 [cs:LG].

Two external reviews, one internal review.

External Review 1

Summary of the paper

What are the main questions this paper is addressing?

Technically, the paper studies full information prediction with expert advice under delayed feedback and a stochastic environment.

Given a set of "experts" (i.e., prediction strategies), in prediction with expert advice the goal is to compete with the best expert in hindsight: in each time step experts make predictions and the learner selects one of them, while the environment chooses an outcome (stochastically, in the case of the present paper). The predictions and the outcome determine the loss of each expert in each time step, and the learner wants to minimize the difference of its loss and that of the best expert in hindsight. An unimportant detail is that the paper allows the experts to abstain from making a prediction.

In the delayed feedback model, in each time step the learner makes a prediction, but learns the outcome for that time step only after some delay (when the outcome arrives, the learner also learns which time step this outcome was generated for). In this paper the delays are also stochastically chosen by the environment, without the knowledge of the learner's choices. The question is whether, despite the delays, a learner is still able to keep its total cumulated loss close to that of the best expert in hindsight.

In the particular setting studied here, the delays are allowed to grow unbounded, while the loss is strongly convex in the predictions. The motivation for studying this particular problem is because it is thought to model an imagined scenario when the goal is to predict the results of computations that become increasingly more expensive to carry out – hence the results arrive after increasingly longer delays. Can then an algorithm still keep up (in terms of cumulative loss) and be almost as good as the best expert in hindsight? If this is not possible, can some weaker goal be satisfied?

What are its main conclusions?

The answer to the first question is negative: No algorithm can keep up with the best expert in hindsight as the delays get unbounded. This is in fact trivial to see: Just let the observation for time step n be delayed by (e.g.) $\exp(n)$. Then by time T, all but $\log T$ observations are missing and no learner can figure out which expert to listen to. After discussing this, the authors settle on a less ambitious goal: First, they make a strong assumption that there is a so-called *Bayes-optimal* expert (an expert such that for any time step t, the expected loss for the expert's prediction at step t, given the observations up to step t, is the smallest among all possible predictions).

Then, they ask the question of whether it is possible to "match" the loss of a Bayes-optimal expert, in an asymptotic sense (precisely: the absolute deviation between the instantaneous loss of the learner and that of the Bayes-optimal expert must converge to zero). They show that this is possible, under an additional strong assumption, namely that the loss function, as a function of the predictions, is strongly convex. Their algorithm is based on the idea that the observed cumulative loss differences for every pair of experts, when matched for identical time steps, behaves like a random walk with a drift, where the drift's sign depends on whether one of the experts is a Bayes-optimal one.

Novelty

Do you know of other investigators or groups pursuing similar questions?

Please also comment on other investigators or groups that could be in a good position to pursue similar questions.

Learning under delayed feedback is studied by a number of authors (a list of relevant work, some cited by the paper and some missed, is given at the end of this answer). The subject has been studied in the control literature from the point of view of distributed computations (see the book *Parallel and Distributed Computation: Numerical Methods*, PrenticeHall, 1989 by John Tsitsiklis and Dimitri Bertsekas; republished in 1997 by Athena Scientific). From the same point of view, several recent papers have studied delayed feedback optimization, in particular stochastic optimization. See Mania et al (2015) and the references therein for the work in the stochastic optimization setting.

In the context of online learning, Weinberger and Ordentlich (2002) studied online sequence prediction under a fixed delay in feedback. Mesterharm (2005, 2007) has studied label prediction for specific adversarial environments under delayed feedback. Joulani, Gyorgy and Szepesvari (2013, 2016) studied general reductions from delayed feedback to non-delayed online learning, for partial as well as full-information feedback, in both stochastic and adversarial environments. Quanrud and Khashabi (2015) together with Taghvaei (2016) studied online learning with convex and strongly convex losses under delayed feedback, providing bounds in terms of the cumulative delay. Riabko (2005) proposes the "weak teacher" model for learning with missing information, which provides another point of view for learning with delayed feedback.

All the researchers mentioned here (and many more) are in the position of working on the aforementioned questions. Papers mentioned above:

- Marcelo Weinberger, Erik Ordentlich, "On delayed prediction of individual sequences", IEEE Transactions on Information Theory, 48.7, 2002.
- Chris J. Mesterharm, "Online learning with delayed label feedback", Algorithmic Learning Theory (ALT), 2005.
- Chris J. Mesterharm, "Improving online learning",. PhD thesis, Department of Computer Science, Rutgers University, New Brunswick, NJ, 2007.
- P Joulani, A Gyrgy, C Szepesvri, "Online learning under delayed feedback", International Conference on Machine Learning (ICML), June 2013.
- P Joulani, A Gyrgy, C Szepesvri, "Delaytolerant online convex optimization: Unified analysis and adaptivegradient algorithms", AAAI Conference on Artificial Intelligence (AAAI), 2016.
- K Quanrud, D Khashabi, "Online Learning with Adversarial Delays", Advances in Neural Information Processing Systems (NIPS), December 2015.
- Daniel Khashabi, Kent Quanrud, Amirhossein Taghvaei, " Adversarial Delays in Online StronglyConvex Optimization", arXiv:1605.06201v1, May 2016.
- Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Ben Recht, Kannan Ramchandran, Michael Jordan, "Perturbed Iterate Analysis for Asynchronous Stochastic Optimization", arXiv:1507.06970v2, July 2015.
- Daniil Riabko, "On the flexibility of theoretical models for pattern recognition", PhD thesis, University of London, April 2005.

If correct, what would the paper's conclusions add to what is already known?

The observation that under unbounded delays no algorithm can compete with the best expert in hindsight, albeit a trivial one, has not appeared in the literature before. The same is true for the particular positive result proven.

Technical quality

Did the paper address its main questions in a logically defensible way based on reasonable premises?

There are several problems in the proofs and the validity of the results cannot be established (though the reviewer suspects that the proof can be fixed with considerable effort). In particular, there are important problems in the proof of Lemma 7, which is a main ingredient of Theorem 5, the main result of the paper. For example:

- The definitions of H_i and G_i depends on $|s|$, the length of the independent sequence returned by 'test_seq'. Since this length is a random quantity itself (e.g., it depends on the values of predictions y_i and on the values of past and future observations o_i), observing the value of G_k for some k gives more information than $o_{<s_k}$: for example, if one observes G_k and finds out that G_k differs from $o_{<\infty}$, one could infer that $|s| > k$.

  Thus, x_{s_k} may not have the same conditional distribution given G_k than it has given $o_{<s_k}$. As such, it is not clear why $\mathbb{E}[r(H_i)|G_i] = 0$, a condition required for applying Lemma 10, should hold (the cleanest way to see this is to write the definitions formally with the help of indicators). In particular, per the discussion above, the above expectation is not the same as $E[r(H_i)|o_{<s_i}]$, because G_i gives more information than $o_{<s_i}$.

- The quantity t_n is random. As such, one cannot go from Eq (2) to Eq (3), since Eq (2) holds only for fixed (non-random) values of M (to see this, consider the case when M is $\sum r(H_i)v(G_i)$, i.e., M is the same random quantity whose probability is being bounded; the event will always be true and have probability 1, while the right hand side will be less than 1). A possible remedy is to write (3) for any fixed t instead of the random t_n, and then observe that the event "{exists n with relscore$_n > l$}" implies the event "{there is t such that (3) holds}".

- On top of the above, an assertion that the authors don't prove is that $G_1, H_1, G_2, H_2, \ldots$ form a Markov chain. Why is this the case? (It is unclear though whether seeing this is really necessary for all the proofs to go through; in particular, Doob's optional skipping processes together with Martingale arguments should give the required results.)

- Similar problems propagate to the results of Lemma 8, since the proof uses the same formalism as Lemma 7, in particular the same G_i and H_i, and depends on the (incorrect) use of the random quantity t_n outside of the probability inequalities.

Did the paper build on and draw from the most important and relevant prior work?

I take this question is whether the authors are knowledgeable of the literature. Indeed, they are. At the same time, their reference of previous work is not completely correct. For example, the work of Joulani et al (2013) does not require bounded delays. In fact, their Theorem 1 holds even in the case of unbounded delays, and results in sub-linear regret for full-information online learning as long as the delays up to time t are sublinear in t. This is also the case for the work of Quanrud and Khashabi (e.g. Theorem 2.1), which the authors had missed. (See also the discussion under Question 8).

Also, using the existing framework of online learning under delayed feedback (developed by Mesterharm, Joulani et al, and others mentioned above) would have been completely sufficient and satisfactory for this paper (rather than reintroducing this framework with unclear notation and definitions).

How hard would it be for someone else to derive results like these?

(We're looking for an answer of the rough form, "If working on these questions, a graduate student in my department could make comparable progress with about a month of work" or "There are only a few people in our field who would be able to make comparable progress on these questions with 6 months of work", or somewhere in between.)

The main idea of both the negative and positive results are fairly simple and straightforward. Yet, there are some technical challenges (e.g., considering infinitely many experts, properly handling the stochastic arguments, which the authors didn't succeed at). Given this, I would think that an expert (junior professor, post-doc) with proper prior knowledge can replicate the results in a couple of weeks, while writing up a formally correct argument takes more time (e.g., a month or two).

Significance

The authors of this paper think that its results shed light on standards for good reasoning under deductive limitations. They say more about that here.

How significant do you feel these results are for that?

The results are vaguely relevant for this problem.

First, the requirement of asymptotic consistency (in the sense of matching the loss of a Bayes-optimal expert) is on one hand very weak (asymptotics is "easy"), while at the same time, the particular assumptions, being able to access a Bayes-optimal expert and the strong convexity of the "loss", are very strong. In any remotely practical setting, we don't expect these to hold and there is not much novelty in achieving asymptotic consistency by comparing losses.

On the website, it is promised that this paper makes a step towards figuring out how to come up with "logically non-omniscient reasoners". In particular, the present paper, supposedly contributes to the following requirement:

"They [=logically non-omniscient reasoners] should be able to notice patterns in sentence classes that are true with a certain frequency."

The authors later state:

"We give an algorithm that eventually assigns the 'right' probabilities to every predictable subsequence of observations, in a specific technical sense."

This surely sounds impressive, but there is the question whether this is a correct interpretation of Theorem 5. In particular, one could imagine two cases: a) we are predicting a single type of computation, and b) we are predicting several types of computations. In case (a), why would the delays matter in asymptotic convergence in the first place? Note that the authors are assuming that every outcome is "eventually revealed", so eventually a non-delayed algorithm will have enough samples to come up with an accurate probability. Why do we need EvOp then?

In case (b), the setting that is studied is not a good abstraction: in this case there should be some "contextual information" available to the learner, otherwise the only way to distinguish between two types of computations will be based on temporal relation, which is a very limiting assumption here (and falls back to case (a) unless one engages in predicting the probability that the next prediction is of a particular type, a concern not addressed by the paper).

Finally, even if this interpretation is correct, it is only made possible by a computationally expensive algorithm under the strong assumption of having access to a Bayes-optimal forecaster, and only for strongly convex losses. Note that while the authors are trying to provide an answer to this question "in principle", computational concerns do matter for this specific question, as discussed next.

As usual, online learning of course offers a powerful set of tools for addressing prediction problems, but online learning can only do so much. In particular, one can push online learning towards philosophy (lacking any practical relevance), e.g., by using infinite expert sets and unbounded computational power, as done here. But it is rather intriguing if not self-defeating if this is done in the context of trying to answer a question that is derived from the lack of sufficient computational resources!

Of course, the paper's results remain true if we only have finitely many experts. But then perhaps much simpler alternatives are also available (see below for some comments of what else is perhaps missing from the paper). And then the next problem is who selects the experts and how? In particular, if we only have finitely many experts, but we want to predict the results of very complex calculations, maybe we can keep up with the best expert, but even the best expert is expected to perform very poorly. Hence, results of these kind might not be too instructive.

More generally, the authors somehow miss the opportunity of studying the simplest algorithm of all, the exponentially weighted average (EWA) forecaster. In particular, it is unclear what's wrong with just using this standard method with whatever information is currently available; such an idea was used, for example, by Mesterharm (2005, 2007) in the stochastic setting, and by Quanrad and Khashabi (2015) and Joulani et. al. (2016) in the adversarial setting (indeed the last of these works might be directly applicable to EWA). If experts are abstaining, some modifications may be necessary, but EWA should handle delays as well as any other algorithm. If one demands uncountably many experts then one can use EWA with a prior. Even if this idea does not work, it should be discussed why this is the case. In face of larger delays, would this simpler approach not give even asymptotic results?

Do the methods and results seem potentially fruitful in the sense that they or related work could shed additional light on these issues in the future?

Much more work would be necessary to figure this out.

Clarity

Is the paper written in a way that will allow others to understand it and build on it?

The paper could be much better written. Sometimes it over-complicates matters: e.g., the presentation of the stochastic setting is messy. The motivation for studying the particular setting is not very well explained (Why do the experts abstain? If this is because many different types of computations need to be predicted, then shouldn't there be some "contextual" information available to the learner? In particular, at the moment, an expert could only know when to abstain based on the temporal relation of interactions, not the current context, which does not sound reasonable. Wouldn't contextual information be important / required for the motivating goal anyways? Why do we need infinitely many experts? Why strong convexity?).

There are other small presentation issues, but these are relatively minor (typos, etc.)

External Review 2

Summary of the paper

What are the main questions this paper is addressing?

This paper considers the problem of making predictions sequentially in an environment where the true outcomes are revealed with delays. These delays may be unbounded, and the goal is to be able to make predictions that are competitive with the predictions made by an optimal forecaster in a given class of forecasters. This paper considers the following questions in this setup:

a) How should performance of a forecaster be measured when there are unbounded delays in feedback?

b) Is it possible to construct a forecaster that is consistent in the sense that its performance approaches the best forecaster in the class?

What are its main conclusions?

The main conclusion of the paper is that if the outcomes and observations are generated stochastically and if the class of forecasters contains a Bayes optimal one (defined in a certain sense in the paper) then it is possible to construct an algorithm (called "EvOp" in this paper) that asymptotically approaches the performance of the Bayes optimal forecaster.

Novelty

Do you know of other investigators or groups pursuing similar questions?

Please also comment on other investigators or groups that could be in a good position to pursue similar questions.

As far as I know, this is the only paper to consider the problem of unbounded delays. John Langford has worked on learning with delays as well, but not unbounded delays.

If correct, what would the paper's conclusions add to what is already known?

This paper's conclusions add very little to what is already known. The problem is that the setup considered in this paper (of unbounded delays) is far too general to be analyzable theoretically or useful practically. The authors themselves recognize both issues: standard performance measures become meaningless in this general setup, and the convergence bounds of the algorithm EvOp are too weak for the algorithm to be useful practically. Furthermore, reading the paper, it is a simple exercise to construct an algorithm that terminates in finite time this is the main result of the paper. The authors are not concerned with efficiency at all, simply decidability, which makes the results of this paper somewhat insignificant.

Technical quality

Did the paper address its main questions in a logically defensible way based on reasonable premises?

The authors do attempt to build up their definitions and analysis in a logically sound manner. However, I am not convinced all the arguments they have presented are rigorously correct. There is a lot of hand-waving/ad-hoc argumentation in the paper that is difficult to verify without a formal proof.

For example, section 3 provides the key motivations for the definitions they have used in the paper and their algorithm. I am not convinced that the argument presented there is correct. Since the definition of regret in stochastic environments such as the one considered in the paper used expectations of the losses, I believe that the Bayes optimal forecaster in the example given in section 3 has vanishing regret in expectation, even though it is true that the actual regret for any realization of the outcomes and observations can be large. Thus, I am not convinced that standard definitions of regret should be discarded in favor of the definitions in this paper.

Similarly, definition 3 is ambiguous since it leaves the subsequence s of length at least n unspecified: there could be multiple such subsequences; which one is used for measuring consistency?

Did the paper build on and draw from the most important and relevant prior work?

Since the setup considered in this paper is pretty novel, it doesn't build on or draw from the prior work.

How hard would it be for someone else to derive results like these?

(We're looking for an answer of the rough form, "If working on these questions, a graduate student in my department could make comparable progress with about a month of work" or "There are only a few people in our field who would be able to make comparable progress on these questions with 6 months of work", or somewhere in between.)

Any competent graduate student with mathematical maturity should be able to derive the main results in this paper in a day or two. This is simply because the algorithm in this paper is very straightforward and only finite time termination is shown with no regard for efficiency. This is very easy to achieve.

Significance

The authors of this paper think that its results shed light on standards for good reasoning under deductive limitations. They say more about that here.

How significant do you feel these results are for that?

I do not think these results any significant light on standards for good reasoning under deductive limitations. The primary reason is that the setup is far too general and too unrealistic, and the algorithm completely impractical. It would be far better to model the problem of delays more realistically and come up with efficient algorithms to deal with delay.

Do the methods and results seem potentially fruitful in the sense that they or related work could shed additional light on these issues in the future?

I do not believe the methods and results in this paper would lead to significant future work, for the same reasons as above.

Clarity

Is the paper written in a way that will allow others to understand it and build on it?

The paper is generally pretty well-written but the proofs definitely need to be tightened and written more rigorously. In general there is a lot of hand-waving/non-rigor that makes for good light reading but only confuses the reader who is scrutinizing the work critically.

Internal Review

Summary of paper and its potential significance for AI safety (this can be very brief)

What are the main questions this paper is addressing?

This paper considers a formalism for making predictions about computations that may take a long time to finish. Formally, at time step t we would like to make a prediction y_t about x_t (and then receive loss $L(x_t, y_t)$). Due to computational limitations, we might not observe x_t immediately, but instead observe it after time step $t+d_t$, for some delay d_t (we think of d_t as the time that it takes the computation to complete). It is also possible that in some cases we never observe the output of the computation (d_t is infinite). The goal is to find a good predictor, i.e. something that given the observations so far (i.e., s such that $s+d_s < t$) outputs a prediction y_t where $L(x_t, y_t)$ is small. The particular notion of goodness is based on regret relative to some class of predictors, and later based on Bayes-optimality.

What are its main conclusions?

First, they show that a naive notion of regret does not admit any non-trivial algorithms — it is possible to construct a sequence (x_t, d_t) such that no algorithm can have average regret converging to zero (the regret basically oscillates back and forth ad infinitum). Then they show that under the assumption that there is a Bayes-optimal predictor on any subsequence of observations, it is possible to asymptotically converge to the Bayes-optimal predictor in the sense that the difference in pointwise loss between the Bayes-optimal predictor and our predictor converges to zero almost-surely.

How does MIRI think these questions are related to potential risks from advanced AI?

This problem is related to logical uncertainty. I understand logical uncertainty to mean the problem of formalizing what it should mean to have subjective uncertainty about logically determined computations (such as the 1000th digit of pi), although it seems to me that MIRI at least sometimes uses logical uncertainty in a somewhat broader sense. In either case, the current paper seems to be focused on the narrower notion stated above.

What obstacle(s) to developing safe AI does MIRI think work in this direction could help to overcome?

I'm not entirely sure what obstacles MIRI has in mind. I know that they have been interested in logical uncertainty for a long time (I think going back to Wei Dai, though perhaps someone else should get the credit). I think that logical uncertainty is a very interesting topic, though it is not obviously safety-related to me except insofar as systems about which we have a better theoretical understanding are likely safer, and logical uncertainty is definitely an obstacle to theoretically principled approaches to machine learning. Note that I think this is already a relatively good reason to work on something for safety purposes.

I'll now speculate on reasons why MIRI cares about logical uncertainty. One reason is that it seems likely to be a necessary component of any good formal theory of bounded rationality. I would expect MIRI to care a good deal about building up such a formal theory, and I expect them to care about *this* because they think that most self-modifying agents will eventually modify themselves to conform to such a theory,

and so it is better to start with agents that have these properties so that we don't have to worry about the self-modification process going wrong; in addition, understanding the properties of such agents might help us to better understand what failure modes might exist.

Two other reasons MIRI might care: first, coarse-grained human concepts might also be seen as invoking a form of logical uncertainty (maybe "atoms" don't really exist or aren't the fundamental building block of reality, but this doesn't stop us from reasoning about them; but if an AI realized that atoms weren't real, and its value function was defined in terms of atoms, perhaps that would lead to issues). In addition, MIRI is interested in counterfactual reasoning / distributional shift, and this delayed computation model brings in aspects of distributional shift; one could imagine that there are connections here to be explored.

Update after discussing with other reviewers: There is another sense in which logical uncertainty ties into distributional shift that I had not originally appreciated. In particular, if one assumes a well-specified model then a large number of issues related to distributional shift go away (e.g. because model error, which is hard to reason about, gets replaced with parameter error, which is easier to reason about). One might obtain a well-specified model by taking a highly expressive model family, e.g. Turing machines of some (large) bounded length, or very large neural networks. In such cases, computational constraints become a large bottleneck and in practice one would currently resort to using heuristics that form an obstruction to most of the guarantees one would hope to get out of a well-specified model. One goal of logical uncertainty, then, is to recover these guarantees even in the presence of (an appropriate family of) heuristic approximations.

A summary of this argument might be: taking a sufficiently expressive model family replaces model error (very hard!) with parameter error ("easy") + computational error (also hard, but maybe less hard). So a good understanding of computational error would be valuable in e.g. outputting principled estimates of uncertainty. One aspect of this argument I find unsatisfying is the assumption that "sufficiently expressive model" implies "well-specified model". I think one could have a very complex model that is nevertheless misspecified in some subtle way; indeed, MIRI is worried about this as well, see "Reflective Variants of Solomonoff Induction and AIXI" and this post by Wei Dai. The place where we differ is that whereas their approach seems to be to try really hard to make sure they have the right model family, I would rather assume that we are likely to get things wrong in some way, and make sure that systems are robust to such mistakes. While this might seem like a minor difference, I think it is relatively important (and probably correlates with several other disagreements) and would lead to a relatively different research agenda. While I am therefore somewhat uneasy about this general approach, I still think it's worthwhile for someone to be exploring it (this is independent of my opinion of the particular way that MIRI is going about it).

Comments on the novelty, technical quality, and difficulty of the results (optional)

We can think about the problem in the following way: if we actually observed the x_t sequentially with no delays, standard online learning algorithms could achieve vanishing regret. The difficulty comes from the delays, which we can think of as inducing a form of covariate shift on the inputs (since we see the inputs in a different order, and in particular the empirical distribution over inputs observed so far might not converge in any meaningful way to the true distribution over inputs; though this is rough intuition since the inputs are not even i.i.d. in this setting).

If we think about the problem in terms of covariate shift, the existence of a Bayes-optimal predictor is saying that the model is well-specified, in which case covariate shift should not create problems — learning over any subset of the observations should yield the Bayes-optimal predictor (modulo issues due to censoring of the inputs). There may be some difficulties due to the fact that some of the delays could be infinite, but the definition of Bayes-optimality in the paper basically assumes this away.

As a result, I do not think that achieving these sorts of results is very difficult — it is well-known that

realizability/well-specification/Bayes-optimality makes one immune to covariate shift, and I think that the only differences in this setting are (1) the x_t are not i.i.d., and (2) we are trying to get a very strong notion of convergence (identical losses to the Bayes-optimal predictor, vs. vanishing average regret relative to the Bayes-optimal predictor). The proof seems a bit complicated relative to the standard approach (I do not understand why they take a countable enumeration of the prediction family F, rather than using standard geometric approaches from stochastic optimization) and I am not sure how much of this comes from insufficient acquaintance with the relevant literature vs. needing a different approach to get this stronger convergence notion. I could imagine that this stronger notion does increase the mathematical difficulty of the results, but I couldn't understand from the paper why this stronger notion is so important, or what makes the stronger notion difficult to achieve.

So overall, I think that novelty/difficulty are not very high from what I can understand. The technical quality is reasonable, in the sense that while there is some degree of non-standard notation/terminology that makes it harder than necessary to understand, I was overall able to understand the results without too much difficulty. On the other hand, see my comment above about potential over-complication of their proof approach relative to more standard approaches.

Significance (most important)

Consider again the obstacle(s) to developing safe AI that MIRI thinks this work may help overcome.

Do you think this is a real/important obstacle?

As I mentioned, I am not personally convinced that logical uncertainty is inordinately safety-relevant, except insofar as more theoretically-principled methods are likely to be more safe than less theoretically-principled methods. I do think that this is a valid consideration. For the three possible connections that I drew (self-modification, human concepts, counterfactual reasoning) I would say:

- Self-modification: I am not very convinced by the self-modification argument for caring about logical uncertainty.
- Human concepts: it is not clear to me either way. I could imagine that some version of logical uncertainty is useful in modeling human concepts.
- Counterfactual reasoning: I think the tie-in is unidirectional — counterfactual reasoning would help with logical uncertainty, but I don't think that the reverse is true.

Assuming the obstacle is real, how much do the results in this paper address the obstacle?

I don't think that the results address the obstacle very much at all. The results basically say to assume (via Bayes-optimality) that the same predictor which performs well on short computations also performs well on long computations, and that we can therefore generalize well even in a computationally-bounded setting. At a high level this is already obvious, and doesn't seem to address the fundamental obstacles posed by logical uncertainty.

Is the approach in this paper well-suited to overcoming the obstacle compared to other possible approaches?

Can you think of other research directions/approaches/people that might be better positioned to overcome the obstacle?

Here are some approaches that I think do more to address the problem of computationally-bounded learning (this is a very biased sampling of the literature, not meant to be representative or exhaustive):

Computational-statistical tradeoffs:

- Paul Christiano's online local learning paper
- Jacob Steinhardt's memory-bounded learning paper
- Venkat Chandrasekaran's convex relaxation paper

Some of Jacob Steinhardt's more concrete work on avoiding pathological behavior on computationally-intractable tasks: reified context models and relaxed supervision.

Boaz Barak's sum-of-squares approach (though I don't know anyone who is actively exploring that direction, and I remain skeptical of that approach as well).

What are the most likely ways that this work could turn out to be relevant to the safety of future ML systems?

If the formalism of online learning with delays ended up being the "right" formalism for understanding computationally-bounded reasoning, and/or it was possible to significantly relax the constraint of Bayes-optimality in this framework.

What are the most likely ways that this work could turn out to be irrelevant to the safety of future ML systems?

I think this would be the default case, since I didn't feel that there was a substantial technical contribution in this paper, relative to what is already known about covariate shift and Bayes-optimality.

Overall, how promising does this work seem in comparison with other approaches to developing safe AI?

Please weigh both probability of relevance and importance if relevant.

I think I have to go with either "Less promising than usual" or "No special claim to relevance", again based on the fact that I did not think there was enough of a technical contribution.

Additional notes

Below are notes I made for myself while reading the paper (though I stopped taking them at some point). I don't know if this is useful or not, but I'm including them since it's low-cost. They may not make much sense to people other than me.

Thoughts upon reading abstract (note that I also thought about a related problem that Holden forwarded me, so this isn't my very first exposure to the problem): okay, I suspect that their model is going to be the following: at time t there is a function f_t and a delay d_t, drawn from some joint distribution over (f_t, d_t) which is the same for all t. However you don't get to observe function t until time $t + d_t$. Our goal is to construct a sequence of inputs $w_1, \ldots, w_T$ such that $(1/T)\sum_t f_t(w_t)$ converges to $\sum_t f_t(w*)$, where $w*$ is the minimizer of $\mathbb{E}[f(w)]$ (and where w_t can only depend on the f_s that have been observed so far, i.e. where $s + d_s < t$).

Apparently, if d_t is unbounded then you can run into trouble, and it can be impossible to construct such a sequence. I assume that what goes wrong is that if $\mathbb{E}[d_t] = \infty$ then the empirical average of the delays can get larger and larger as time goes on, and so the distribution of observed f_t can end up very far from the actual true distribution over f. The problem is essentially that the censoring process means that we aren't actually observing i.i.d. samples. It seems that they plan to instead consider regret or a similar notion with respect to some sparse subsequence. Presumably, if this sequence is sufficiently sparse, then the censoring process has a negligible effect and we get samples that approach i.i.d. in the limit.

It's not immediately obvious to me that this result should even be true, as one might worry that there is a cat-and-mouse type of behavior where however sparsely we sample the sequence, one could pick a distribution that has even heavier tails. I think what might save us is the fact that the samples are i.i.d., so there is only a fixed distribution and perhaps we can make sure to eventually have heavier tails than any fixed distribution does.

Now let's think about how to actually prove this. Maybe I can consider the following trivial strategy: wait until I see f_1, and then update on f_1. Next, wait until I see f_2, and then update on f_2, etc. By doing this I can guarantee that I am basically performing stochastic optimization on the original sequence $f_1, f_2, \ldots$, but I might have to wait a long time. If that is the actual algorithm, then that would be pretty lame, so I hope that's not it (especially since it shouldn't take 16 pages to show that). I'm going to start reading beyond the abstract to check.

Thoughts after reading through intro (and peaking at their model / main algorithm): one thing I forgot is that the impossibility result might be somewhat interesting. Let's see how it might work: since I typically get regret $O(\sqrt{T})$ in regular stochastic optimization settings, if I can make sure that I observe less than $\sqrt{T}$ samples after T steps, then I should get non-vanishing regret. This still seems kind of hard — I should have a non-zero probability of having a delay of zero (or some bounded amount) in which case I see $\Omega(T)$ samples. So I probably have to make these undelayed examples look atypical: maybe there's some correlation between the delay and the value of f. I also note that there probably has to be something about their model other than what I said above — perhaps rather than observing f_t after some delay, you observe some specific value of f_t (maybe $f_t(w_t)$). It's pretty late where I am, so I'm going to sleep for now and will think/read more tomorrow.

Thoughts after reading their model: okay, so it looks like you have a loss function L, and your loss on round t is $L(x_t, y_t)$. At time t you have to guess y_t, and after d_t additional steps you observe x_t. It also seems that it's possible to never be able to observe x_t. I am guessing that this is the reason why my trivial algorithm above doesn't work — you can't actually wait to observe x_t, because your wait might never finish.

It looks like their results are the following: in at least some cases, it is impossible to obtain vanishing regret under this model. However, if there is a Bayes-optimal predictor then we can actually obtain vanishing

regret. I assume the reason for this is that for each x there is a unique y that could be Bayes-optimal, so each observation we make just lets us make strictly more progress. (Although I don't yet know what Bayes-optimal should mean in their setting.)

Okay, apparently Bayes-optimal means that y_n minimizes $\mathbb{E}[L(x_n, y)|\text{observations before time } n]$. I am kind of confused by this (what is the expectation taken over? x_n?) and also it seems to imply that the observations are not i.i.d., which I didn't realize before (it's kind of weird to call it the "stochastic" setting since that typically means i.i.d. in the context of online learning).

Lower bounds: okay, is it possible to get a lower bound even in the i.i.d. setting? (Not sure yet whether their lower bound is for this setting.) How about, with probability $\frac{1}{n(n+1)}$ I have a delay of 2^n, and x is either $+1$ or -1 with independent probability (but the same value, v_n, whenever the delay is 2^n). Then after T steps, I have observed at most the v_n with $n \leq \log(T)$. But this only accounts for $1 - \frac{1}{\log(T)}$ of the total probability mass. Okay but this still gives vanishing regret, it just vanishes at the slow rate of $\frac{1}{\log(T)}$. I would maybe guess that in the stochastic case, you always can get non-vanishing regret, at least if the delays need not be infinite.

Let's see what lower bound they do... okay, it looks like the lower bound is to have non-i.i.d. data in batches of increasing size, where within each batch the data are identical (but observations are delayed until the end of the batch). Clearly we can't do better than chance here, and just need to make sure that at least something in our hypothesis class can do better than chance.

Upper bounds: if we assume that there is a Bayes-optimal strategy, it seems to me that we should be able to get an upper bound. In particular, the existence of a Bayes-optimal strategy is saying that the model is well-specified, in which case covariate shift (which is basically what's happening with the different delays) should not create problems. Therefore, doing online learning over any subset of the observations should do well.

Okay, I don't see how their Theorem 1 could be true, since $x_{\{s_n\}}$ is stochastic... perhaps this is just due to the fact that the Bayes-optimal response is unique for strongly convex functions, so eventually you just have to output the same thing every time?

At this point I'm stopping my note-taking and just reading the rest of the paper.