Comments on the Open Philanthropy Project's Anonymized Reviews of Three Recent MIRI Papers

Nate Soares
Machine Intelligence Research Institute
nate@intelligence.org

September 2016

As part of an evaluation process for their grant decision, the Open Philanthropy Project had twelve anonymous reviewers (four technical advisers internal to the Open Philanthropy Project, plus eight external reviewers with relevant expertise) assess recent papers by MIRI researchers. The result is a very valuable record of experts' responses to a sampling of our recent papers. I provide comments on these reviews below.

MIRI senior researcher Eliezer Yudkowsky wrote up his thoughts about MIRI's research methodology in advance of the reviews; these are excerpted in Appendix A at the end of this document. I wrote up a series of predictions about how I thought the review process would probably go in an email exchange with Open Philanthropy Project Program Officer Nick Beckstead, collected in Appendix B.

From the external review of "Proof-Producing Reflection for HOL"

> *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.*

(1) *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. [. . .]*

> *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.*

I agree with this overview. The follow-on project the reviewer mentions is being headed by Ramana, with support from the Future of Life Institute (Kumar 2015). It's been slow going, due to a number of rough edges in HOL. Our two general objectives here are:

(a) shoring up our theoretical understanding: formalizing an idea for reflection in code helps vet the idea, and tells us whether there are nontrivial problems that arise in implementation.

(b) encouraging the research community to push in the direction of theorem provers that can handle adaptive systems: there's a vibrant automated theorem-proving community, but we don't think they'll necessarily develop tools that would be useful for verifying important properties of safety-critical components of advanced AI systems by default.

To be explicit about why we're interested in these problems: building practical real-world AI systems out of reflective theorem provers is in no way the goal. Rather, the goal is to nudge the program verification community in the direction of tools that can aid in the design of safety-critical adaptive (and potentially reflective) systems. We think that this sort of research helps with that goal by demonstrating that useful reflection is possible in practice even though Gödel's theorems show that it's impossible in full generality.

Our follow-on project to implement reflection in HOL has hit lots of practical snags requiring that HOL itself receive some development. As a major contributor to HOL, Ramana is in a relatively good position to push on this front. We also have an intern working on similar things in a dependently typed language (and a job opening for a full-time researcher to focus on reflection in type theory).

From the first internal review of "Proof-Producing Reflection for HOL"

(2) *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.*

This paper was selected mainly for Ramana's contributions on the theorem prover end of things. Note that model polymorphism underwent peer review and was presented at an AGI conference (Fallenstein and Soares 2014).

(3) *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.*

This doesn't quite capture the thought process behind this line of research. The systems we're thinking about needn't be "meaningfully analogous" to theorem-provers in any stronger sense than humans are.

The Open Philanthropy Project's technical advisers are already aware of this, but I'll sketch out why I (and some others at MIRI) use tools in logic for problems like this. Although practical advanced AI systems will almost certainly not be logic-based, I expect any artificial general intelligence to face problems where finding a solution requires managing uncertainty in the face of heavy deductive limitations, and I think that developing a theory of probabilistic reasoning under deductive limitations would make it easier to design systems that admit of a theoretically-principled understanding of how they do their reasoning. (I think that ensuring that the system's abstract reasoning is robust and reliable is a

crucial part of AI alignment, and I'm skeptical of humanity's ability to do this without principled understanding.) Our work on reflection is mainly aimed at helping us get a better understanding of deductively limited reasoning.

We spend time on methods for managing uncertainty about *logical sentences* (describing, e.g., claims about long-running computations or mathematical conjectures) mainly because there already exists a large body of theory and machinery for representing and manipulating logical sentences. If we want to talk about an AI system that is uncertain about what different computer programs do, one easy way to do this is to write down claims about computer programs (using some formal language) and then design a method for reasoning probabilistically about those claims.

In short, probabilistic methods for managing uncertainty about logical sentences seem to us to be the simplest formal environment in which to study toy models of deductively limited reasoning. Reflective reasoning is a special case of this where it's easy to see that even highly general formalisms like AIXI aren't applicable.

(4)

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.*

I agree with most of this, and with the reviewer's later comment: "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." One way to justify this intuition is to note that *humans* manage to more-or-less trust themselves most of the time.

One of the goals of MIRI's research is to take intuitions like these and formalize them. We're actively working on developing principled probabilistic reasoning methods that avoid the standard paradoxes of self-reference, and we've recently had some success here; see our forthcoming paper on logical induction (Garrabrant, Taylor, et al. 2016).

I don't think there's any disagreement here about the default probability that humans are going to have a strong principled understanding of the reasoning used by early smarter-than-human AI systems. If I had to guess, I would say that the underlying disagreement is about how hard it is to align smarter-than-human AI systems with human interests in the absence of that strong principled understanding. Another point of disagreement may be that I think it's easier to shift what kind of transformative AI system comes first, by way of basic research. Combined, I expect that these two disagreements explain most of the difference between the research MIRI prioritizes and the research the Open Philanthropy Project's technical advisers prioritize. I'll write more about these two points when I have time to write up a series of blog posts from the conversation notes that I generated during the Open Philanthropy Project review period.

From the second internal review of "Proof-Producing Reflection for HOL"

(5) *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.

I think this is more or less consistent with my views, though I'd put the emphasis elsewhere. The way I see it is this: there are basic gaps in our models of what it means to do good reasoning (especially when it comes to things like long-running computations, and doubly so when those computations are the reasoner's source code). Logic happens to be the main tool mathematics has for representing and manipulating claims like "this program outputs 7". We're happy to use better tools as they become available, but until then, we're forced to use the tools we have on hand. (Our logical uncertainty research is about developing adequate probabilistic tools for this task.)

(6) *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 aren't agents at all.

Although real-world agents will rely on heuristics, if the heuristics are approximating something that wouldn't work, then the approximation probably won't work too well either. (Diagonalization isn't specific to logic-based systems, after all.) If the goal is to design heuristics that can reason about themselves in a sensible way, I expect it's important to know what distinguishes the reliable self-reference from the failure-prone kind. There are plenty of good reasons to think that this problem might not need to be solved right now, but I don't think the fact that we'll rely on heuristics is one of them.

I don't think that whether or not the AI system is used as an "agent" (if by that we mean "a system that autonomously selects and executes actions with minimal oversight") vs. a "tool" bears much on the question of whether this research is useful. If it seems like these problems don't apply to "tool AIs" then I suspect there's been a miscommunication. I plan to write more about this before long.

(7) *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.*

We tried a modular approach, and quickly found that it doesn't help with this problem.

Finitistic approaches help in a sense, but we don't think they get at the underlying problem. We aren't imagining a scenario where we literally have arithmetical agents that we want to run and we're just unwilling to pick a big number and give them a finite amount of self-trust. In most cases, we don't pick particular research directions by backchaining from imagined disaster scenarios at all; we instead imagine trying to design an aligned AI system in theory (e.g., with a Jupiter-sized computer), while paying attention to obstacles that we don't yet know how to overcome even in principle.

In this case, we notice that our current best models of abstract reasoning can't handle self-reference and reflection. The symptoms of this are often easy enough to treat in isolation; we're trying to use the symptom for insight into the underlying ailment. I think that this strategy of tugging on loose threads and seeing where they lead is fairly common across the sciences, and has led to many mathematical and scientific breakthroughs in the past.

I'll write up some blog posts in the future about why we we're tugging on *this* particular loose end. The question of which loose threads to tug on is very much a question of mathematical intuition, which I think is one of the reasons why the conversation between MIRI folks and Open Philanthropy Project folks has required so much depth. Transferring the reasoning behind inchoate mathematical intuitions turns out to be really difficult. Hopefully, it will be easier to pass objective judgement about these mathematical intuitions over time as we develop a track record for actually solving (or failing to solve) these problems.

From the first external review of "Inductive Coherence"

(8) *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.*

We're definitely thinking about these problems differently from most of the people in the field. The idea behind this is that we expect to have a bigger marginal impact if we look at questions that aren't a natural fit for established methods and subfields in CS. The hope is that some deep problems may not turn out to be especially *hard*, just ill-suited to the standard methods and ways of thinking. Indeed, the most interesting things we've found (in my opinion) come from picking low-hanging fruit on strange branches. For example:

- "Inductive Coherence" takes a strange view of probabilistic logic, and the ideas in "Asymptotic Convergence in Online Learning with Unbounded Delays" only barely fit in an online learning framework, but having both in hand led us to a (forthcoming) result about how to reason inductively about mathematical conjectures and logical facts that we've found useful (Garrabrant, Taylor, et al. 2016).
- Löb's theorem and reflective oracles are strange tools to use in game theory, but they gave us answers to the questions "Can realistic non-identical agents unexploitably cooperate in one-shot prisoner's dilemmas?"(Critch 2016) and "Does there exist a large class of policies with the grain of truth property?" (Fallenstein, Taylor, and Christiano 2015; Leike, Taylor, and Fallenstein 2016)
- Representing non-causal dependencies with causal graphs is a strange way to think about philosophical decision theory, but it helped us answer the question "Is there a simple, principled rule that simultaneously gets more wealth than causal decision theory in Newcomb's problem and more wealth than evidential decision theory in medical Newcomb problems?" (Soares and Fallenstein 2015)

Part of the rationale behind having a research nonprofit work on these problems is that they're often an awkward fit for academia.

From the second external review of "Inductive Coherence"

(9) *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.*

I share this view. Neither "Inductive Coherence" nor "Asymptotic Convergence in Online Learning with Unbounded Delays" is particularly compelling on its own; what's interesting from our perspective is the relationship between the two papers. (See my advance predictions in Appendix B.)

"Inductive Coherence" gives us a method for reasoning probabilistically about mathematical claims while respecting logical relationships between claims; e.g., claims known to be mutually exclusive are assigned probabilities that sum to at most 1, even when their truth-value is unknown. "Asymptotic Convergence in Online Learning with Unbounded Delays" gives us a method for reasoning probabilistically about mathematical claims while respecting observed regularities and patterns; e.g., after observing enough digits of π, one's subjective confidence that a certain late digit is 5 should be the same as one's confidence that it's 3, even if one hasn't proven that π is normal. (Assuming they haven't found a better way to compute digits quickly, of course.) Neither algorithm given is particularly interesting in its own right. What's surprising here is that the simple brute force methods for attaining each desideratum separately were (on the surface) incompatible (Soares 2016). Subsequent attempts to combine the two approaches' advantages led to Garrabrant, Critch, et al.'s (2016) much more interesting and general formalism.

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

I agree that the lack of natural examples is a weakness of the paper. The canonical example here is assigning positive probability to the set of complete consistent extensions of Peano arithmetic, so that the distribution can answer questions about claims written in PA. Solomonoff induction fails at this: if you fix an enumeration of theorems and treat the bits as bits revealing the truth-values of the sentences, then you can't condition on the axioms of PA, because that event has probability zero. The same is true for Demski's prior. In both cases, the problem is the infinitely many induction axioms, which are treated independently, meaning that their simultaneous truth is assigned probability zero. This in turn means that Solomonoff inductors can't "reason" about logical sentences. (Of course, there are many other ways to attempt to get a prior over logical sentences out of Solomonoff induction. For an alternative approach, refer to Garrabrant, Benson-Tilsen, et al. (2016).)

(11) *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.*

Our goal is to develop a generalization of ideal Bayesian reasoning that can handle predictions about, e.g., computer programs. Note that we're not looking for practical algorithms. We're starting with things like Solomonoff induction and working inwards, not trying to come up with alternatives to the factor-graph algorithm for belief propagation in Bayesian networks (Pearl and Russell 2002) or Auto-Encoding Variational Bayes (Kingma and Welling 2013). The existing formalisms for ideal induction are uncomputable, and they still don't tell us how to do induction about things like mathematical conjectures and the outputs of computer programs. From my perspective, "doubly exponential" mainly means "computable" in that context, which is progress.

From the internal review of "Inductive Coherence"

(12) *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.*

That's a good point. What I had in mind were specifically probability distributions used to reason about logical sentences, by having variables representing the logical sentences related in the obvious way (coherently). You can of course create a probability distribution where one variable is named "ϕ" and the other is named "not ϕ", and then have those two variables be completely uncorrelated — the variable names are only labels, after all. That distribution won't help you reason about logical claims, though.

To build a distribution that captures reasoning about logical claims, we need to weaken coherence *somehow*, because coherence implies logical omniscience. The question is how exactly we should set up the probabilistic system so that it's actually useful for, e.g., predicting the outputs of certain computations. What's interesting about this result and "Asymptotic Convergence in Online Learning with Unbounded Delays" put together is that the simple brute-force methods for learning statistical patterns and the simple brute-force methods for respecting logical constraints don't mesh very well.

(13) *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.*

("Uniform coherence" is a deprecated name for inductive coherence.)

I'm completely on board with "the perfect is the enemy of the good". I see this research not as "aiming for perfection", but rather as "using simple mathematical models in attempts to fill the most basic gaps in our understanding of good reasoning", more in line with the other reasons the reviewer considers.

(14) *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.*

I agree that that sort of research is a good place to focus efforts. I think the existing ML research community is well-positioned to make inroads on those sorts of problems, so I expect MIRI's impact on the margin to be smaller here than on problems that the existing community is not likely to prioritize.

I'm guessing that the main disagreement here is not about the detailed merits of double Q-learning vs. asymptotic logical uncertainty, but about things like (a) whether dedicated research can cause sizable shifts in the way that AI research is done in broad strokes (e.g., shifts comparable to the shift from expert systems to deep learning); and (b) whether that needs doing (e.g., to shift towards a paradigm where it's significantly easier to design AI systems that admit of a principled understanding of how they do their abstract reasoning). As I mentioned in response to **(4)**, I expect that I'm much more pessimistic about how hard it is to design AI systems that scale up safely without having a principled understanding of how they do their abstract reasoning. (Of course, I think there are also disagreements about whether MIRI's approach could cause the relevant shifts if executed well, and about how well MIRI is executing; much of this comes back to the "inchoate intuition" issue that I mentioned in **(7)**.)

From the first external review of "Asymptotic Convergence in Online Learning with Unbounded Delays"

> *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 $o_{<s_i}$.*
>
> (15)
>
> - *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.*

The reviewer is correct that the proof is a bit lax. Jessica Taylor has uploaded clarifications confirming that H_i and G_i depend only on $o_{\prec s_k}$, and that G_1,H_1,G_2,H_2,... is a Markov chain (Taylor 2016). We've also updated the eprint to add these clarifications.

(16) *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).*

As the final Open Philanthropy Project reviewer points out, there are some subtle differences between what we're doing and standard work in online learning, and in this case we do actually care about the differences. That said, the notation we used was surely more unusual than it needed to be — we aren't natives to this subfield.

(17) *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).

Our goal isn't to predict "several types of computations", but rather to have the system face *all* the computations, and learn to predict the predictable ones. The rough framework we're modeling is this: Fix a universal Turing machine and an enumeration of Turing machines. Run a Levin search, and reveal the results (e.g., "it is time 1019; machine #37 just halted with output 1") as they occur. The goal is to observe this sequence and (by induction) learn to predict machines late in the sequence pretty well, in the same way that Solomonoff induction sees only a sequence of 1s and 0s and learns which environment generated them pretty well. (Yeah, weird setup, I know.)

I'm not sure I understand the reviewer's objection here, but I think they're saying that in this setup it would be extraordinarily hard to predict any meaningful family of observations without some other sort of data, such as contextual information, or a guarantee that every outcome is eventually revealed (which amounts to guaranteeing that all non-halting Turing machines have been filtered out). I agree that prediction in this framework is quite difficult, but we *are* more or less trying to develop an algorithm that learns to predict computations from only the temporal relations. (Though in fact the algorithm gets to see the machine's number, the time at which the machine halts, and the bit it outputs when it halts, which makes the task possible.)

The reviewer is correct that it's still very difficult to do induction in this framework; this is why the paper makes very strong assumptions. (EvOp only attempts to predict subsequences, and we only guarantee convergence when it has access to a Bayes-optimal predictor for that subsequence.)

The thing that makes this paper non-trivial is the fact that EvOp converges to Bayes-optimal predictions on any subsequence on which it has access to a Bayes-optimal predictor, including on environments like $P^{\#}$, constructed in the paper, which are difficult for standard methods to handle. This still isn't a very large result, but I think it's *less* trivial than the reviewer recognized.

(18) *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?*

EWA is a very natural algorithm to try here, but it fails for reasons discussed in section 3. In fact, the environment $P^{\#}$ constructed in section 3 is designed for the purpose of demonstrating this, and you can verify that EWA fails in the example given there. (Let the class of predictors be f^1, f^0, and f^* as in the example. Then whenever one of the fair coins comes up heads, by the end of the run, EWA will be fooled into predicting 1 with very high probability, and whenever one of the fair coins comes up tails, by the end of the run, EWA will be fooled into predicting 0 with very high probability.) The natural patch is to have the weights decay over time; this still fails, and this is the example discussed explicitly in section 3 (at the bottom of p. 5 / top of p. 6).

From the second external review of "Asymptotic Convergence in Online Learning with Unbounded Delays"

(19) *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.*

I agree that the algorithm presented isn't a large addition to the existing base of knowledge, and that it's pretty easy to come up with. See **(9)** for my perspective on the result's significance.

Overall, I think this review is quite good, and I agree with pretty much all of it.

From the internal review of "Asymptotic Convergence in Online Learning with Unbounded Delays"

(20) *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.*

That scenario (self-modifying agents modifying themselves to conform to a good formal theory of bounded rationality) seems relatively plausible, but it's not high up on my list of concerns. A similar idea higher up on my list of concerns is that any sufficiently advanced AI system has instrumental pressures to correct irrationalities in its reasoning, which means that if a safety requirement depends on the system possessing a certain irrationality, we'll need to think hard about how to avoid the default instrumental pressures (Benson-Tilsen and Soares 2016).

My main reasons for working on this sort of thing, though, are more like what the reviewer mentions later in the review: "more theoretically-principled methods are likely to be more safe than less theoretically-principled methods". I think that having a solid understanding of how the AI system does its abstract reasoning is fairly indispensable if we actually want sufficient safety and robustness guarantees for smarter-than-human AI, I think the places where we currently lack even a bare-bones theory are often places where our basic understanding needs shoring up, and I think the problem of coming up with generalizations of probability theory that handle uncertainty about long-running computations, mathematical conjectures, and logical facts is surprisingly tractable. (If we didn't have classical probability theory with which to analyze and understand algorithms for managing empirical uncertainty, we'd be trying to develop those tools too, for analogous reasons.)

(21) *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).*

I agree with this, more or less. I'd phrase it more like: "Physical uncertainty is logical uncertainty (about what the laws of physics entail) plus indexical uncertainty (about where we are in the universe), *and* logical uncertainty can be used to model high-level/coarse/abstract reasoning, so good tools for handling logical uncertainty will likely subsume good tools for handling empirical uncertainty." So, just as classical probability theory is a pretty useful tool when designing modern AI systems and reasoning about their behavior, I expect basic methods for handling logical uncertainty to be useful for designing and understanding systems that carry out abstract reasoning.

(22) *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.*

Again, I agree with this, more or less. This analysis doesn't seem to get at the core of why I care about these problems in particular, though. The reason why I care about logical uncertainty and decision theory problems is something more like this: The whole AI problem can be thought of as a particular logical uncertainty problem, namely, the problem of taking a certain function $f : Q \to R$ and finding an input that makes the output large. To see this, let f be the function that takes the AI agent's next action (encoded in Q) and determines how "good" the universe is if the agent takes that action. The reason we need a principled theory of logical uncertainty is so that we can do function optimization, and

the reason we need a principled decision theory is so we can pick the right version of the "if the AI system takes that action…" function.

The fact that we don't know how to do either of these things right in principle strikes me as worrying. For this reason (and a few others), these are two of the basic research threads that I think are worth tugging on (per **(7)** above).

(23) *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.*

Dropping i.i.d. and attaining a strange convergence guarantee is indeed the goal. It's a non-standard framework. See **(17)** for a quick discussion of the problem we're trying to use this to solve. Basically, we're trying to predict (subsequences of) *all* computations, and the strong assumptions on the convergence guarantee are part of what make that possible at all.

The reviewer's response here is the particular response I was imagining when I wrote in my advance predictions (Appendix B):

> I also expect a high probability that people will say that the paper [fails] to use obvious Online Learning simplifications (which is also true, and an artifact of the fact that we care about the solution for very different reasons from the standard reasons, which means that most of the usual simplifications would break the correspondence between the OL problem and [the] actual problem we're trying to study).

(24) *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.*

In deterministic settings, counterfactual reasoning requires counterpossible reasoning, which involves some sort on assigning probabilities to logical sentences. For example, it seems intuitively correct for two copies of an AI agent playing a non-iterated prisoner's dilemma to note that if the one cooperates, the other does too; whereas if the one defects, then the other defects. But two deterministic algorithms will in fact either cooperate or defect, so formalizing this kind of reasoning in full generality requires some method for assigning probabilities to "the mathematical function I'm computing outputs `cooperate`" and "the mathematical function I'm computing outputs `¬cooperate`", even though at least one of these is logically impossible. As such, logical uncertainty seems like a prerequisite to solid counterfactual reasoning. Soares and Fallenstein (2015) discuss this in more detail.

Our recent progress on logical uncertainty has already spun out some interesting new decision theory ideas, though these are all still in a very early stage.

A Pre-Review Comments by Eliezer Yudkowsky[1]

The point of current AI safety work is to cross, e.g., the gap between [...] saying "Ha ha, I want AIs to have an off switch, but it might be dangerous to be the one holding the off switch!" to, e.g., realizing that utility indifference is an open problem. After this, we cross the gap to solving utility indifference in unbounded form. Much later, we cross the gap to a form of utility indifference that actually works in practice with whatever machine learning techniques are used, come the day.

Progress in modern AI safety *mainly* looks like progress in conceptual clarity — getting past the stage of "Ha ha it might be dangerous to be holding the off switch." Even though Stuart Armstrong's original proposal for utility indifference completely failed to work (as observed at MIRI by myself and Benya), it was still a lot of conceptual progress compared to the "Ha ha that might be dangerous" stage of thinking.

Simple ideas like these would be where I expect the battle for the hearts of future grad students to take place; somebody with exposure to Armstrong's first simple idea knows better than to walk directly into the whirling razor blades without having solved the corresponding problem of fixing Armstrong's solution. A lot of the actual increment of benefit to the world comes from getting more minds past the "walk directly into the whirling razor blades" stage of thinking, which is not complex-math-dependent.

Later, there's a need to have real deployable solutions, which may or may not look like impressive math per se. But actual increments of safety there may be a long time coming. [...]

Any problem whose current MIRI-solution *looks* hard (the kind of proof produced by people competing in an inexploitable market to look impressive, who gravitate to problems where they can produce proofs that look like costly signals of intelligence) is a place where we're flailing around and grasping at complicated results in order to marginally improve our understanding of a confusing subject matter. Techniques you can actually adapt in a safe AI, come the day, will probably have very simple cores — the sort of core concept that takes up three paragraphs, where any reviewer who didn't spend five years struggling on the problem themselves will think, "Oh I could have thought of that." Someday there may be a book full of clever and difficult things to say about the simple core — contrast the simplicity of the core concept of causal models, versus the complexity of proving all the clever things Judea Pearl had to say about causal models. But the planetary benefit is mainly from posing understandable problems crisply enough so that people can see they are open, and then from the simpler abstract properties of a found solution — complicated aspects will not carry over to real AIs later.

1. I've here excerpted an email from a conversation I had with MIRI Senior Research Fellow Eliezer Yudkowsky, which provides some useful background about our research strategy. Excerpts from this email were shared with the Open Philanthropy Project prior to the review process.

B Pre-Review Comments by Nate Soares[2]

Hi Nate/Eliezer,

We're about to send out the MIRI papers for external review. We want this external review process (which will involve obviously qualified CS people with relevant expertise) assessing things we should expect external reviewers to be able assess, such as:

1. *Reasonableness of argumentation*
2. *Novelty of results*
3. *How hard results are to get*
4. *Significance of the results (not for significance like "reduce AI risk" but for significance like "improve understanding of reasoning under deductive limitations")*

Then we are having our technical advisers (Chris, Dario, Jacob, Paul) give input on the significance of the results for:

1. *How much progress they constitute on MIRI's research agenda (holding aside what they think of the merit of MIRI's research agenda)*
2. *How much producing more results like these would be likely to help with AI risk (including what they think about the merit of MIRI's research agenda)*

I'd be curious to get your predictions on what kind of results we're going to get from these two types of reviewers, and what results you think should/shouldn't be seen as meaningful (for someone in our position, where we don't have, and can't easily have, all the context you have). We feel that getting this information before we see results might help us decide what to make of the results.

A reasonable way to give your predictions might be (be needn't exactly be):

A. *Probability distribution over "good," "bad," "meh," and "I don't know what this is" from external reviewers.*
B. *Probability distribution over "good," "bad," "meh," and "I don't know what this is" from internal reviewers (re: progress on MIRI's research agenda)*
C. *Probability distribution over "good," "bad," "meh," and "I don't know what this is" from internal reviewers (re: overall importance of progress for AI risk)*

Would you be up for sending us your thoughts on A-C, plus your thoughts on what kinds of results you think should/shouldn't be seen as informative from our perspective?

I'd also welcome thoughts of the form, "This whole review exercise sounds like a waste of time to me. If you really want to know if MIRI has made progress, you should do XYZ instead."

Best, Nick

P.S. The results that people are reviewing include:

- *"Uniform Coherence"*
- *"Asymptotic Convergence in Online Learning with Unbounded Delays"*
- *"Optimal Predictors: A Bayesian Notion of 'Approximation Algorithms'"*
- *"Proof-Producing Reflection for HOL: With an Application to Model Polymorphism"*
- *Secret result (only Paul and Daniel)*

2. I've here excerpted an email I sent to Nick Beckstead (who is quoted in italics) prior to the review process.

Disclaimers: I haven't thought too hard on these probabilities; they likely aren't stable under reflection.

UC

○ External

- bad: 20%
- meh: 60%
- good 20%

By default I expect the reviews to be similar to the IJCAI reviews: I expect most reviewers to say "it's pretty well written and the math is correct but it's not clear that the math was actually hard and I don't quite understand the problem they're trying to solve or how/[whether] this bears on it and I'm not sure what applications it has." I expect a small subset to think it's quite cool (if they've ever tried to pull of a similar result themselves, for instance); and a small subset to think it's terrible (likely correlated with them fixating on a particular pet peeve such as "didn't cite the right related work" or "didn't motivate applications well enough" etc. etc.).

○ Internal reviewers re: Progress from MIRI's perspective

- bad: 15%
- meh: 40%
- good: 45%

My mainline expectation is that your internal reviewers will be somewhat uncomfortable answering this question as they won't feel like they really understand why we care so much about logical uncertainty (with the exception of Paul), and thus I imagine them feeling unsure of their ability to evaluate how much progress it actually makes towards the hard part of the problem-as-we-see-it. (And I have a lot of uncertainty about whether internal reviewers who feel unable to give a solid evaluation of how the progress appears from our perspective will go with 'meh' or 'good'.) Insofar as the answer is "bad", I weakly predict that the reason will be something like "the math doesn't seem hard" / "the solution seems fairly obvious."

I would be very surprised if one of them said "I feel like I understand precisely why the MIRI people think logical uncertainty is important, and I feel like I understand the result in this paper, and regardless of how difficult the result was to get I think that it simply doesn't bear on the question at hand."

○ Internal re: Importance of progress for AI risk

- bad: 45%
- meh: 35%
- good: 20%

Depends quite a bit on how the question is phrased and framed, and on how the internal reviewers are thinking about the question. (e.g., "does this research have positive sign?" vs "do you think this research direction is tractable?" vs "do you think this is worth the effort of MIRI people?" vs "do you think this would be worth your effort?" etc. etc.

AC

○ External

- bad: 35%
- meh: 57%
- good 8%

This paper is further from traditional Online Learning than the UC paper is from traditional Probability Logic, so I expect a much higher probability that reviewers will be off-put (thanks to a vague gut-level sense that the paper is trying to make a problem fit into a place that it doesn't actually fit, which is true). I also expect a high probability that people will say that the paper [fails] to use obvious Online Learning

simplifications (which is also true, and an artifact of the fact that we care about the solution for very different reasons from the standard reasons, which means that most of the usual simplifications would break the correspondence between the OL problem and [the] actual problem we're trying to study).

I'd be quite surprised if people said the result was incorrect (aside from typos).

I think it would take a fairly peculiar person to find this paper really exciting, as we find it mostly exciting in the context of the UC paper (above). Scott Aaronson or someone of similar caliber would probably get at least a little excited about it if he saw both papers together; I expect him to say things like "this is exciting but very preliminary".

○ Internal re: Progress from MIRI's perspective

Similar to the UC paper, with probabilities adjusted a bit downwards both because (a) this paper is more awkward and (b) this result is a little less interesting on its own. 70% that Paul thinks the two papers together constitute legitimate logical uncertainty progress, though.

○ Internal re: Importance of progress for AI risk

Similar to UC paper.

OP

○ External

- bad: 40%
- meh: 35%
- good 25%

I'm highly uncertain about this one. On the one hand, the prose is worse and the paper is very unfinished. On the other hand, the math appears impressive on this one. I expect most reviewers to bounce off and say things like "I couldn't understand it much at all", but the ones who do understand it I expect to say "I think this is onto something pretty neat". I expect a number of small bugs in the math (it's still a pretty rough draft), but I would be fairly surprised if people seemed to both understand the math and say it was wrong. (I would be *very* surprised if they continued thinking it was wrong after corresponding with Vadim about the apparent problem.)

○ Internal re: Progress from MIRI's perspective

Similar to UC/AC papers (except that I expect a much higher chance that the internal reviewers just say "it was too poorly written / too hard to understand; I didn't want to take the time; so I'm not sure").

○ Internal re: Importance of progress for AI risk

Similar to UC/AC papers.

HiH

○ External

- bad: 25%
- meh: 65%
- good 10%

IIRC, this paper strips out the AI motivation entirely, so I doubt you'll get good responses wrt to "how does this pertain to AI?" I expect a small subset of people will think this result is *very* cool and surprising, but that most will not understand it, and a small subset will say "whatever they're trying to do is impossible, so these results must be silly, because something something Gödel."

◦ Internal re: Progress from MIRI's perspective

- bad: 25%
- meh: 25%
- good 50%

As I said, the paper strips out the AI motivation entirely (IIRC), so I think this comes down in large part to how well the internal reviewers are trying to read between the lines and able to correctly figure out which parts are relevant to the study of reflective reasoning. Insofar as they do that, though, I expect them to say "yeah this is pretty neat from the MIRI perspective." I'd be surprised if you got responses of the form "this is factually inaccurate."

◦ Internal re: Importance of progress for AI risk

- bad: 55%
- meh: 30%
- good 15%

This also depends on how much the internal reviewers are attempting to charitably extrapolate why we care about the result, given that we don't talk about it. (I expect these numbers would be upped a bit, but not a ton, after spending an afternoon chatting; though I think that that afternoon would be better spent chatting about the logical uncertainty stuff instead.)

* * *

Overall, I'd be quite surprised if the reviewers found serious technical flaws, though there are a couple of instances where I expect external reviewers to think they might have (especially in the HiH paper), and I expect some reviewers may argue something along the lines of "they're attempting to do the impossible it will never work."

Beyond that I'm pretty highly uncertain about the range of results, but I don't generally expect results better than the peer reviews that I forwarded.

* * *

Finally, FYI, I was considering AC+UC to be a single result in the top five, rather than two separate ones. The fifth result I listed (the "solution to the grain of truth" paper) is, I think, actually a fair bit more impressive to reviewers than the [HOL in HOL] paper or the Optimal Predictors paper; for that one my estimation of external reviewers returning "good" is in the 40%+ range (unlike any of the above papers).

References

Benson-Tilsen, Tsvi, and Nate Soares. 2016. "Formalizing Convergent Instrumental Goals." In *2nd International Workshop on AI, Ethics and Society at AAAI-2016.* Phoenix, AZ.

Critch, Andrew. 2016. "Parametric Bounded Löb's Theorem and Robust Cooperation of Bounded Agents." arXiv: 1510.03370 [cs.GT].

Fallenstein, Benja, and Ramana Kumar. 2015. "Proof-Producing Reflection for HOL." In *Interactive Theorem Proving: 6th International Conference, ITP 2015,* edited by Christian Urban and Xingyuan Zhang, 9236:170–186. Lecture Notes in Computer Science. Springer International Publishing.

Fallenstein, Benja, and Nate Soares. 2014. "Problems of Self-Reference in Self-Improving Space-Time Embedded Intelligence." In *Artificial General Intelligence: 7th International Conference, AGI 2014, Quebec City, QC, Canada, August 1–4, 2014. Proceedings,* edited by Ben Goertzel, Laurent Orseau, and Javier Snaider, 21–32. Lecture Notes in Artificial Intelligence 8598. New York: Springer.

Fallenstein, Benja, Jessica Taylor, and Paul F. Christiano. 2015. "Reflective Oracles: A Foundation for Game Theory in Artificial Intelligence." In *Logic, Rationality, and Interaction: 5th International Workshop, LORI 2015,* edited by Wiebe van der Hoek, Wesley H. Holliday, and Wen-fang Wang, 9394:411–415. Lecture Notes in Computer Science. Springer Berlin Heidelberg.

Garrabrant, Scott, Tsvi Benson-Tilsen, Siddharth Bhaskar, Abram Demski, Joanna Garrabrant, George Koleszarik, and Evan Lloyd. 2016. "Asymptotic Logical Uncertainty and the Benford Test." In *9th Conference on Artificial General Intelligence (AGI-16),* edited by Bas Steunebrink, Pei Wang, and Ben Goertzel, 9782:202–211. Lecture Notes in Artificial Intelligence. Springer International Publishing.

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

Garrabrant, Scott, Jessica Taylor, Andrew Critch, Tsvi Benson-Tilsen, and Nate Soares. 2016. *Logical Induction.* Working Paper. Berkeley, CA: Machine Intelligence Research Institute.

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

Kingma, Diederik P., and Max Welling. 2013. "Auto-Encoding Variational Bayes." arXiv: 1312.6114 [stat.ML].

Kosoy, Vadim. 2015. *Optimal Predictors: A Bayesian Notion of Approximation Algorithms.* Unpublished draft.

Kumar, Ramana. 2015. *Applying formal verification to reflective reasoning.* Grant Proposal. http://futureoflife.org/first-ai-grant-recipients/#Kumar.

Leike, Jan, Jessica Taylor, and Benya Fallenstein. 2016. "A Formal Solution to the Grain of Truth Problem." In *Uncertainty in Artificial Intelligence: 32nd Conference (UAI 2016),* edited by Alexander Ihler and Dominik Janzing, 427–436. Jersey City, New Jersey, USA.

Pearl, Judea, and Stuart J. Russell. 2002. "Bayesian Networks." In *Handbook of Brain Theory and Neural Networks,* 2nd, edited by Michael A. Arbib. MIT Press.

Soares, Nate. 2016. "New papers dividing logical uncertainty into two subproblems." *Intelligence.org* (blog). https://intelligence.org/2016/04/21/two-new-papers-uniform/.

Soares, Nate, and Benja Fallenstein. 2015. "Toward Idealized Decision Theory." arXiv: 1507.01986 [cs.AI].

Taylor, Jessica. 2016. *Errata for "Asymptotic Convergence in Online Learning with Unbounded Delays."* Brief Technical Note. Berkeley, CA: Machine Intelligence Research Institute. http://intelligence.org/files/AsymptoticConvergenceErrata.pdf.