Towards formalizing universality

Paul Christiano Follow
Jan 9 · 22 min read

The scalability of iterated amplification or debate seems to depend on whether large enough teams of humans can carry out arbitrarily complicated reasoning. Are these schemes "universal," or are there kinds of reasoning that work but which humans fundamentally can't understand?

This post defines the concept of "ascription universality," which tries to capture the property that a question-answering system **A** is better-informed than any particular simpler computation **C**.

These parallel posts explain why I believe that the alignment of iterated amplification largely depends on whether HCH is ascription universal. Ultimately I think that the "right" definition will be closely tied to the use we want to make of it, and so we should be refining this definition in parallel with exploring its applications.

I'm using the awkward term "ascription universality" partly to explicitly flag that this is a preliminary definition, and partly to reserve linguistic space for the better definitions that I'm optimistic will follow.

(Thanks to Geoffrey Irving for discussions about many of the ideas in this post.)

I. Definition

We will try to define what it means for a question-answering system **A** to be "ascription universal."

1. Ascribing beliefs to A

Fix a language (e.g. English with arbitrarily big compound terms) in which we can represent questions and answers.

To ascribe beliefs to **A**, we ask it. If **A**("are there infinitely many twin primes?") = "probably, though it's hard to be sure" then we ascribe that belief about twin primes to **A**.

This is not a general way of ascribing "belief." This procedure wouldn't capture the beliefs of a native Spanish speaker, or for someone who wasn't answering questions honestly. But it can give us a sufficient condition, and is particularly useful for someone who wants to use **A** as part of an alignment scheme.

Even in this "straightforward" procedure there is a lot of subtlety. In some cases there are questions that we can't articulate in our language, but which (when combined with **A**'s other beliefs) have consequences that we can articulate. In this case, we can infer something about **A**'s beliefs from its answers to the questions that we can articulate.

2. Ascribing beliefs to arbitrary computations

We are interested in whether **A** "can understand everything that could be understood by someone." To clarify this, we need to be more precise about what we mean by "could be understood by someone."

This will be the most informal step in this post. (Not that any of it is very formal!)

We can imagine various ways of ascribing beliefs to an arbitrary computation **C**. For example:

- We can give **C** questions in a particular encoding and assume its answers reflect its beliefs. We can either use those answers directly to infer **C**'s beliefs (as in the last section), or we can ask what set of beliefs about latent facts would explain **C**'s answers.
- We can view **C** as optimizing something and ask what set of beliefs rationalize that optimization. For example, we can give **C** a chess board as input, see what move it produces, assume it is trying to win, and infer what it must believe. We might conclude that **C** believes a particular line of play will be won by black, or that **C** believes general heuristics like "a pawn is worth 3 tempi," or so on.
- We can reason about how **C**'s behavior depends on facts about the world, and ask what state of the world is determined by its current behavior. For example, we can observe that **C**(113327) = 1 but that **C**(113327) "would have been" 0 if 113327 had been composite, concluding that **C**(11327) "knows" that 113327 is prime. We can extend to probabilistic beliefs, e.g. if **C**(113327) "probably" would have been 0 if 113327 had been composite,

then we might that **C** knows that 113327 is "probably prime." This certainly isn't a precise definition, since it involves considering logical counterfactuals, and I'm not clear whether it can be made precise. (See also ideas along the lines of "knowledge is freedom".)

- If a computation behaves differently under different conditions, then we could use restrict attention to a particular condition. For example, if a question-answering system appears to be bilingual but answers questions differently in Spanish and English, we could ascribe two different sets of beliefs. Similarly, we could ascribe beliefs to any subcomputation. For example, if a part of **C** can be understood as optimizing the way data is laid out in memory, then we can ascribe beliefs to that computation about the way that data will be used.

Note that these aren't intended to be efficient procedures that we could actually apply to a given computation **C**. They are hypothetical procedures that we will use to define what it means for **A** to be universal.

I'm not going to try to ascribe a single set of beliefs to a given computation; instead, I'll consider all of the reasonable ascription procedures. For example, I think different procedures would ascribe different beliefs to a particular human, and don't want to claim there is a unique answer to what a human "really" believes. A universal reasoner needs to have more reasonable beliefs than the beliefs ascribed to that a human using any particular method.

An ascription-universal reasoner needs to compete with any beliefs that can be ascribed to **C**, so I want to be generous with this definition. For example, given a chess-playing algorithm, we might rationalize it as trying to win a game and infer its beliefs about the rules of chess. Or we might rationalize it as trying to look like a human and infer its beliefs about what a human would do. Or something different altogether. Most of these will be kind of crazy ascriptions, but I want to compete with them anyway (competing with crazier beliefs will turn out to just be easier).

It's not totally clear what counts as a "reasonable" ascription procedure, and that's the biggest source of informality. Intuitively, the key property is that the ascription itself isn't doing the "hard work." In practice I'm using an informal extensional definition, guided by examples like those in the bulleted list.

3. Comparing beliefs

What does it mean to say that one agent is "better-informed" than another?

It's natural to try to express this in terms of empirical information about the world, but we are particularly interested in the different inferences that agents are able to draw from the same data. Another natural approach is to compare their "knowledge," but I have no idea how to define knowledge or justified belief. So I'm reduced to working directly with sets of beliefs.

Consider two sets of beliefs, described by the subjective expectations $\mathbb{E}^1$ and $\mathbb{E}^2$. What does it mean to say that $\mathbb{E}^1$ is better-informed than $\mathbb{E}^2$?

This framing makes it tempting to try something simple: "for every quantity, $\mathbb{E}^1$'s belief about that quantity is more accurate." But this is property is totally unachievable. Even if $\mathbb{E}^1$ is obtained by conditioning $\mathbb{E}^2$ on a true fact, it will almost certainly happen to update in the "wrong" direction for some claims.

We will instead use a subjective definition, i.e. we'll define this concept from a particular epistemic position represented by another subjective expectation $\mathbb{E}$.

Then we say that $\mathbb{E}^1$ **dominates** $\mathbb{E}^2$ (w.r.t. $\mathbb{E}$) if, for every bounded quantity X and for every "nice" property Φ:

- $\mathbb{E}[X|\Phi(\mathbb{E}^1, \mathbb{E}^2)] = \mathbb{E}[\mathbb{E}^1[X]|\Phi(\mathbb{E}^1, \mathbb{E}^2)]$

(By "nice" I mean something like: simple to define and open in the product topology, viewing $\mathbb{E}^1$ and $\mathbb{E}^2$ as infinite tables of numbers.)

Intuitively, this means that $\mathbb{E}$ always "trusts" $\mathbb{E}^1$, even if given arbitrary information about $\mathbb{E}^1$ and $\mathbb{E}^2$. For example, if $\mathbb{E}$ was told that $\mathbb{E}^1[X] \approx x$ and
$\mathbb{E}^2[X] \approx y$, then it would expect X to be around x (rather than y). Allowing arbitrary predicates Φ allows us to make stronger inferences, effectively that $\mathbb{E}$ thinks that $\mathbb{E}^1$ captures *everything* useful about $\mathbb{E}^2$.

I'm not sure if this is exactly the right property, and it becomes particularly tricky if the quantity X is itself related to the behavior of $\mathbb{E}^1$ or $\mathbb{E}^2$ (continuity in the product topology is the minimum

plausible condition to avoid a self-referential paradox). But I think it's at least roughly what we want and it may be exactly what we want.

Note that dominance is *subjective*, i.e. it depends on the epistemic vantage point $\mathbb{E}$ used for the outer expectation. This property is a little bit stronger than what we originally asked for, since it also requires $\mathbb{E}$ to trust $\mathbb{E}^1$, but this turns out to be implied anyway by our definition of universality so it's not a big defect.

Note that dominance is a property of the *descriptions* of $\mathbb{E}^1$ and $\mathbb{E}^2$. There could be two different computations that in fact compute the same set of expectations, such that $\mathbb{E}$ trusts one of them but not the other. Perhaps one computation hard-codes a particular result, while the other does a bunch of work to estimate it. Even if the hard-coded result happened to be correct, such that the two computations had the same outputs, $\mathbb{E}$ might trust the hard work but not the wild guess.

4. Complexity and parameterization

There are computations with arbitrarily sophisticated beliefs, so no fixed **A** can hope to dominate everything. To remedy this, rather than comparing to a fixed question-answerer **A**, we'll compare to a parameterized family **A**[**C**].

I'll consider two different kinds of potentially-universal reasoners **A**:

- In the "idealized" case, **A**[**C**] depends only on the complexity of **C**.
  For example, we might hope that an n-round debate dominates any beliefs that could be ascribed to a fast computation with (n-1) rounds of alternation. In particular, this **A**[**C**] is the same for any two computations **C** of the same complexity.
- In the "practical" case, **A**[**C**] depends on the complexity of **C** but also uses the computation **C** as a hint. For example, if **C** is the training process for a neural net, then we might take **A**[**C**] to be a debate in which the debaters are able to share weights and activations with the neural net throughout the entire training process.

I'm generally interested in the case where **A**[**C**] is only slightly more powerful than **C** itself. This mirrors the setting where a universal Turing machine is able to run any other Turing machine with only a modest slowdown.

Putting it all together

We say that a set of beliefs $\mathbb{E}^A$ *epistemically dominates* a computation **C** (w.r.t. some beliefs $\mathbb{E}$ and language L) if the beliefs ascribed to A by the "straightforward" procedure, using L, dominate (w.r.t. $\mathbb{E}$) the beliefs ascribed to **C** by any reasonable ascription procedure.

We say that a family of question-answering systems **A**[·] are *ascription universal* (w.r.t. $\mathbb{E}$ and L) if **A**[**C**] epistemically dominates **C** for every computation **C**.

II. Discussion

Why is (subjective) dominance sufficient?

This universality condition requires that *we* believe that **A**[**C**] is better-informed than **C**. Naively we might have wanted it to *actually* be the case that **A**[**C**] is better-informed than **C**; the stronger condition is clearly unachievable, but why should we be satisfied with the weaker condition?

In applications of this property, the subjective condition is what we need in order for *us* to believe that **A**[**C**] will cope with the challenges posed by **C**. For example, suppose that **C** formulates a plan to "trick" **A**[**C**]. Then the subjective universality condition implies that we don't expect **C** to succeed.

This isn't as good as actually knowing that **C** won't succeed. But I think it should be good enough for us—the reason we are thinking about AI safety is because we are concerned that something bad will happen. If we find a technique that defuses this argument, then we've addressed the motivating problem. It may still be the case that bad things happen (and we should still search for additional reasons that bad things might happen), but we don't particularly expect them to.

Of course if you select over a large number of computations, then you may find one that will succeed in tricking **A**. But if we are concerned about that, then we can instead apply ascription universality to the entire process including the selection.

Why trust opaque computation?

If **C** uses some clever heuristics that I don't understand, then **C**'s "beliefs" might be excellent, but I might not *expect* them to be excellent. In this sense understanding may seem almost vacuous. If there is some heuristic that I trust, wouldn't **A** just use it?

To see why the definition is demanding, consider the special case where **C** performs an extensive search to find a computation that works well empirically. For example, **C** might be the following computation:

- Start with a training set of (image, label) pairs.
- Search over simple programs to find one that makes good predictions.
- Run that simple program on a new image to predict its label.

In this case, we can ascribe beliefs to **C** about the contents of the new image. And because those beliefs are coming from a simple program that works empirically, I expect them to be accurate (in some respects).

For example, a simple classifier **C** may "believe" that the new image contains a particular curve that typically appears in images labeled "dog;" or a really sophisticated classifier may perform complex deductions about the contents of the scene, starting from premises that were empirically validated on the training set.

So it's not OK for **A** to simply ignore whatever heuristics **C** is using—if those heuristics have the kind of empirical support that makes us think they actually work, then A needs to be able to understand everything that those heuristics imply about the domain.

Why be so general?

I've formulated universality as competing with arbitrary computations **C**. It seems totally possible that the form of **C** discussed in the last section—searching for a program that works well in practice and then using it in a new situation—is so central that the definition of universality should focus entirely on it.

One reason to use the broader definition is because sometimes this "selection" process can be embedded in a non-trivial way in a larger computation. For example, if I have a sufficiently large group of humans, I might expect memetic selection to occur and produce systems that could be said to have "beliefs," and I'd like universal systems to dominate those beliefs as well.

The other reason to use this very general definition is because I don't see an easy way to simplify the definition by using the additional

structural assumption about **C**. I do think it's likely there's a nicer statement out there that someone else can find.

Universal from whose perspective?

Unfortunately, achieving universality depends a lot on the epistemic perspective $\mathbb{E}$ from which it is being evaluated. For example, if $\mathbb{E}$ knows any facts, than a universal agent must know all of those facts as well. Thus "a debate judged by Paul" may be universal from Paul's perspective, but "a debate arbitrated by Alice" cannot be universal from my perspective unless I believe that Alice knows everything I know.

This isn't necessarily a big problem. It will limit us to conclusions like: Google engineers believe that the AI they've built serves the user's interests reasonably well. The user might not agree with that assessment, if they have different beliefs from Google engineers. This is what you'd expect in any case where Google engineers build a product, however good their intentions.

(Of course Google engineers' notion of "serving the user's interests" can involve deferring to the user's beliefs in cases where they disagree with Google engineers, just as they could defer to the user's beliefs with other products. That gives us reason to be less concerned about such divergences, but eventually these evaluations do need to bottom out somewhere.)

This property becomes more problematic when we ask questions like: is there a way to seriously limit the inputs and outputs to a human while preserving universality of HCH? This causes trouble because even if limiting the human intuitively preserves universality, it will effectively eliminate some of the human's knowledge and know-how that can only be accessed on large inputs, and hence violate universality.

So when investigating schemes based on this kind of impoverished human, we would need to evaluate universality from some impoverished epistemic perspective. We'd like to say that the impoverished perspective is still "good enough" for us to feel safe, despite not being good enough to capture literally everything we know. But now we risk begging the question: how do we evaluate whether the impoverished perspective is good enough? I think this is probably OK, but it's definitely subtle.

I think that defining universality w.r.t. $\mathbb{E}$ is an artifact of this definition strategy, and I'm optimistic that a better definition wouldn't have this dependence, probably by directly attacking the notion of "justified" belief (which would likely also be useful for actually establishing universality, and may even be necessary). But that's a hard problem. Philosophers have thought about very similar problems extensively without making the kind of progress that seems adequate for our purposes, and I don't see an immediate angle of attack.

III. Which A might be universal?

Two regimes

I'm interested in universality in two distinct regimes:

- Universality of idealized procedures defined in terms of perfect optimization, such as debate under optimal play or max-HCH, where **A**[**C**] depends only the computational complexity of **C**.
- Universality of practical variants, in which the perfect optimization is replaced by a bounded optimization, which might use the computation **C** as a "hint." For example, we might consider amplification or debate where the agents are allowed to inspect a transcript of **C**; or if **C** is a neural net training process, we might train them jointly with **C**.

Idealized models

A[**C**] could be:

- A debate, judged by a human, with perfect debaters, whose length depends on the complexity of **C**. (Hopefully in such a way that **A**[**A**[**C**]] is only a slightly longer debate than **A**[**C**].)
- max-HCH, using perfect optimization rather than a bounded optimizer. Again, we can impose a budget limit that depends on the complexity of **C**, ideally such that **A**[**A**[**C**]] does not involve a much larger computation than **A**[**C**].
- Use HCH to implement two functions, Info and Answer. Define **A**[**C**](Q) to be Answer(Q, a^*), where a^* = argmax Info(a, Q), and the optimization is over advice strings a of length comparable to the description of **C**.

I don't think any of these would be universal if instantiated with an open-ended human policy—providing arbitrary complex inputs to a human is likely to lead to bad behavior. Instead, I imagine restricting the human to some small enough space of inputs that they can behave "reasonably" over the whole space, while still being rich enough to be universal. Even as such, I think additional measures might be needed to achieve security (since a team of humans may behave badly in some exotic corners of the input space, even if individual humans are sheltered from exotic inputs).

Practical models

In practice I'm particularly interested in the case where **C** has the special form described in the "Trusting opaque computations" section: it trains a neural net on some dataset, and then runs that neural network on new inputs.

A[**C**] could be:

- Debates arbitrated by a human where the debaters are trained jointly with **C**, and can make references to individual examples from the dataset.
- max-HCH where the agent is trained jointly with **C**, and the overseer can consult individual entries from the dataset.
- HCH, combined with interpretability techniques that allow it to get information about what **C** is thinking and to relate its behavior to the dataset on which it was trained.

For each of these practical variants, I don't expect it to be universal unless the corresponding idealized model is. So we can think of dividing the universality question into two pieces:

- Would our procedure be universal *if the optimization was performed perfectly*?
- Is weight sharing / interpretability strong enough to solve the optimization problems needed for universality?

I think none of the suggestions in this section are plausibly universal without some additional mechanisms, and so even if HCH or debate is universal, additional ideas are needed in order to translate those to practical universal systems. I'm still inclined to focus on the idealized case, because I want to better understand what is needed for universality without thinking about algorithmic difficulties.

IV. Which C are hard to epistemically dominate?

The definition in this post was somewhat informal, but the discussion in this appendix will be extremely informal. The purpose is to communicate some of my early thinking about universality and my sense for what a universal reasoner might look like.

The first three examples will consider programs **C** with the following form:

- Iterate over programs to find an *f* that successfully predicts the answers to some dataset of historical questions for which we know the answer.
- Use that program to predict the answers to future questions.

The different examples explore features of *f* that might make it difficult to dominate the overall computation.

Deduction

Suppose *f* takes the following form:

- Given a question and context, it performs a short series of deductions from facts in the context, together with some axioms (e.g. "Bob's eyes are more likely to twitch when he is lying," the rules of poker, *etc.*)
- When it arrives at answers that constrain the possible answers to the given question, it outputs an answer consistent with those constraints.

We could attribute the statements produced by this process to **C** as beliefs. We expect this set of beliefs to have useful information because it was selected for making correct predictions.

Fortunately, if HCH is able to see the training dataset then it may be able to arrive at similarly accurate views:

- It can compute that P(observations | axioms) is high by following along with the deductive process used by *f*.
- Bayes implies that the axioms are likely to be true (fully capturing the argument which leads us, from our current perspective, to expect them to be true).

- Once HCH knows the axioms are likely to be true, it can follow along with the deductive process to reproduce all of *f*'s beliefs.

Modeling

Suppose *f* takes the following form:

- Given a question Q and context, set up a relevant physical situation.
- Use a dynamics model M to predict what will happen at some future time.
- Use the resulting state to answer Q.

We could say that **C** believes everything implied by the outcome of this physical simulation. We expect these beliefs to be accurate because *f* has predicted well in the past.

As in the deduction case, hopefully HCH can compute that P(observations|M is accurate) is high, since it can also compute the consequences of M. Then Bayes implies the dynamics model is accurate, and HCH can use that model to compute physical states.

Inferring all the beliefs from a dynamics model is not trivial though. As an extreme example, if *f* is performing an atom-by-atom simulation of a room, and that room contains Alice and Bob, then we could ascribe extensive beliefs about Alice and Bob to the computation **C**.

(Here we run head on into the fuzziness about what counts as a "reasonable" ascription procedure, but for the moment I'll assume that some reasonable procedure ascribes beliefs about Alice and Bob to the computation.)

To compete with these ascriptions, HCH needs to infer those high-level beliefs about Alice and Bob from the low-level computation involving atoms. One way to do this is to search over possible "bridging" hypotheses that relate low-level physical facts to high-level facts about the environment. If such a hypothesis can explain additional high-level facts, then a Bayesian can learn that it is true. Similarly, if the bridging hypothesis relates facts about the model to constraints we know from the high-level interpretation, then the Bayesian can potentially use that as evidence. (This kind of reasoning will be discussed in a bit more detail in the next section.)

We could further hope that searching for a bridging hypothesis isn't much harder than performing the original search over low-level physics, given that the low-level physics needed to explain a bunch of high-level facts and so already must encode some part of that correspondence.

(Note that the "deduction" example in the previous case could also involve alien concepts or models, in which case the same kind of work would be needed.)

Alien reasoning

In the previous section we described two styles of reasoning we already understand. But there are probably many kinds of reasoning that work well in practice but that would be more alien, and those might be more challenging. This section will explore one example in some detail to try to help anchor our reasoning about the general phenomenon. It will also elaborate on some of the reasoning about "bridging" hypotheses mentioned in the last section.

Suppose that our predictions are always of the same form (e.g. what is the probability the stock market will go up today), and *f* works as follows (the details are long but not very important):

- Find the PSD matrix A with maximum log determinant subject to the constraints in the next bullet points, then output the (0, 0) entry.
- There is an implicit correspondence between the rows/columns of A, and some uncertain properties X(0), X(1), X(2), ... (which we'll view as 0–1 variables), where X(0) is the property we want to forecast.
- If the (i, j) entry of A represented the expectation $E[X(i)X(j)]$, then the matrix would necessarily satisfy a bunch of constraints, which we impose A. For example:
- If the context implies that $X(i) = 1$, then $E[X(i)X(j)] = E[X(j)] = E[X(j)^2]$, so $A(i, j) = A(j, j)$.
- If $X(i)$ and $X(j)$ together imply $X(k)$, then we must have $E[X(i)X(j)] \leq E[X(i)X(k)]$ and hence $A(i, j) \leq A(i, k)$.
- For any constants $a, b, \ldots$, $E[(a\,X(1) + b\,X(2) + \ldots)^2] \geq 0$—i.e., the matrix A must be PSD.

The chosen matrix A(opt) corresponds to a set of beliefs about the propositions X(i), and we can ascribe these beliefs to **C**. Because f predicts well, we again expect these beliefs to say something important about the world.

I chose this procedure f in part because we can give a kind of argument for why the matrix A(opt) should tend to encode accurate beliefs. But I don't think that a universal reasoner can make use of that argument:

- Finding the argument that f works is an additional problem, beyond finding f itself, which might be much harder.
- A comprehensible version of that argument may be much larger than the strategy itself, so even in the idealized cases like debate with perfect optimization, we may need to increase the scale.
- I don't expect that all "good" reasoning strategies have clean understandable arguments in their favor (and even in this case, if it the scheme worked well it would be largely an empirical fact rather than a consequence of the simple theorems we could prove). I think this kind of example is useful because we can easily imagine a human debate judge not having the argument while still being apparently universal. This makes it a useful analogy for cases where the argument really doesn't exist.

Instead, I think a universal reasoner needs to be able to infer the efficacy of this reasoning procedure from its empirical success. It's relatively easy for a Bayesian to learn the regularity "f makes good predictions." Recovering the rest of the matrix A, and learning how to interpret and whether to trust them, is the hard part.

This is going to require the same kind of bridging/identification we discussed in the last section. Let's write X(A) for the set of beliefs about the world implied by the "intended" identification. Searching over possible identifications to find X (or something like it) is the only way we can ever relate the rows of A to the quantities X(i). Again, we can hope that it isn't much harder than finding the original reasoning procedure.

I think that a sufficiently sophisticated Bayesian would probably be able to learn to trust X(A):

- If f is performing well enough that we think it's more likely to be right in the future, then the Bayesian is going to end believing

some claim like "the predictions of *f* are good" (since it explains the data so well).

- This is a complicated statement, and without some kind of explanation this claim has a low prior probability (roughly decaying with the complexity of *f*). The Bayesian is motivated to find an explanation with higher prior probability.

- The correspondence X can explain the constraints on the matrix A, in terms of facts that we already know about the world. This explanation may end up being simpler (or at least higher prior) than a direct enumeration of the constraints on A—I hope (and think it's plausible) that this happens iff we'd actually believe on reflection that X(A) captures reality.
  (To the extent that we are uncertain and think A's beliefs have a non-negligible chance of capturing reality, then hopefully we can capture that by the same mechanism by ending up with a non-degenerate posterior.)

- Now the Bayesian is faced with at least two kinds of explanations:
  (a) "If you use the constraints implied by correspondence X(A) + positive semidefiniteness, and then optimize log det, you get a matrix A for which X(A) makes good predictions,"
  (b) "The actual situation in the real world is described by positive semi-definite matrices with higher log determinant (under the correspondence X)."

- Explanation (b) is explaining two things at once: both why the optimization done by *f* respects the constraints on our beliefs, and why that optimization leads to good predictions. Hopefully this is simpler than making two separate bridging claims, one which explains *f* as respecting the constraints implied by X, and one which claims that *f* makes good predictions. Ideally, this 2-for-1 that favors (b) exactly mirrors the underlying reasoning that leads us to actually believe that X(A) is correct, rather than resembling what we know about reality and making good predictions "by coincidence."

This is a pretty speculative discussion—it's not very careful, and it's hard to make it careful in part because I don't have a formalization of Bayesian reasoning that can even really be applied to this setting. But it seems to match my intuitions about what reasonable Bayesian reasoning "should" do, which gives me a lot more optimism that a careful Bayesian would be able to epistemically dominate **C**.

Deliberation and self-improvement

Often we expect the computation **C** to have accurate beliefs because it uses a strategy that appears to work in practice—the last 3 examples have discussed that case. But there are other reasons to trust a computation.

For example, humans often write code and trust it (to some extent) even without extensive empirical testing—instead, we have a reason to think it will work, and need only modest testing to make sure that we haven't made an error in our implementation or reasoning. If I write an automated mathematician that works by finding proofs that pass a proof checker, I don't expect it to be correct because of the empirical record (Empirical data backs up some key assumptions, but isn't being used to directly establishing the correctness of the method.)

Likewise, if we train a powerful agent, that agent might initially use strategies that work well in training, but over time it might use learned reasoning to identify other promising strategies and use those. Reasoning might allow it to totally skip empirical testing, or to adopt the method after much less testing than would have been necessary without the reasoning.

To dominate the beliefs produced by such reasoning, we can't directly appeal to the kind of statistical inference made in the previous section. But in these cases I think we have access to an even more direct strategy.

Concretely, consider the situation where **C** contains a process f that designs a new reasoning process g. Then:

- From the outside, we trust g because we trust f and it trusts g.
- An otherwise-universal reasoner **A** will dominate f's beliefs, and in particular if f is justified in thinking that g will work then **A** will believe that and understand why.
- Once we understand f's beliefs, dominating g is essentially another instance of the original ascription universality problem, but now from a slightly stronger epistemic state that involves both what $\mathbb{E}$ knows and what f knows. So unless our original approach to universality was tightly wedded to details of $\mathbb{E}$, we can probably dominate g.

At the end of the day we'd like to put all of this together into a tight argument for universality, which will need to incorporate both statistical arguments and this kind of dynamic. But I'm tentatively optimistic about achieving universality in light of the prospect of agents designing new agents, and am much more worried about the kind of opaque computations that "just work" described in the last few sections.