From: Eliezer Yudkowsky <[redacted]>
To: Jeffrey Epstein <jeevacation@gmail.com>
Subject: What MIRI does, and why
Date: Wed, 19 Oct 2016 21:49:52 +0000

To be sent to a hedgefundish person who asked for more reading. Rob, Nate, any priority objections to this text, especially the financial details at bottom?

===

This is probably the best overall online intro to what we do and why:

https://intelligence.org/2015/07/27/miris-approach/

In my own words:

MIRI is the original organization (since 2005) trying to get from zero to not-zero on the AGI alignment problem. This is not an easy problem on which to do real work as opposed to fake work.

Our subjects of study were and are selected by the criterion: "Can we bite off any chunk of the alignment problem where we can think of anything interesting, non-obvious, and discussable to say about it?" (As opposed to handwaving, "Well, we don't want the AI to kill humans," great, now what more can you say about that which would not be obvious in five seconds to somebody skilled in the art?) We wanted to get started on building a sustained technical conversation about AGI alignment.

There is indeed a failure mode where people go off and do whatever math interests them, and afterwards argue for it being relevant. We honestly did not do this. We stared at a big confusing problem and tried to bite off pieces such that we could do real work on them, pieces that would force us to think deeply and technically, instead of waving our hands and saying "We must sincerely ponder the ethics of blah blah etcetera." If the math we work on is interesting, I would regard this as validation of our having picked real alignment problems that forced us to do real thinking, which is why we eventually arrived at interesting and novel applied math.

Why consider AGI alignment a math/conceptual problem at all, as opposed to running out and building modern deep learning systems that don't kill people?

- The first chess paper in 1950, by Shannon, gave in passing the algorithm for a program that played chess using unbounded computing power. Edgar Allen Poe in 1833, on the other hand, gave an a priori argument that no mechanism, such as the calculating engine of Mr. Babbage, could ever play chess, when arguing that the Mechanical Turk chessplayer must have a concealed human operator. When you don't know how to solve a problem even using unbounded computing power, you are in some sense confused about the problem. We are presently unable to code, even in principle, the sort of superintelligent program that would e.g. synthesize a strawberry on a plate and then stop without destroying the universe, never mind all the stuff humans think of as deep moral dilemmas. (Modern computer science has a much stronger idea of how to give an unbounded formula for a superintelligence without the alignment part, e.g. Marcus Hutter's AIXI.)

- Almost all of what we expect to be the interesting and difficult problems of AGI alignment do not spontaneously materalize at humanity's current level of AI capabilities, and if they did, other people would provide funding to work on them. For example, an AGI doesn't spontaneously resist having its off-switch pressed until the AGI understands enough of the big picture to realize what an off-switch is and that its other goals will not be achieved if there is no AGI trying to achieve them. Some of these problems can be sort of modeled in toy systems, but not all, and they're hard to toy-model in an interesting way that doesn't just produce trivial results

that anyone skilled in the art would easily predict in advance. To build a sustained conversation about what would happen later in more advanced AI systems, we need to say things precisely enough that a critic can say "No, that just kills everybody, because..." and the proposer is forced to agree. Precise talk about unbounded systems can sometimes well-incarnate a stumbling-block, or something we are confused about, which by default we'd also expect to show up in a sufficiently advanced bounded system. Then we can talk precisely about the stumbling-block, and have conversations where somebody says "Wrong" and the original proposer stares off for a second and nods.

===

One of the major classes of interesting problem we tried to bite off has to do with stability of self-modification, or coherent cognitive reflection. For example, we'd want to show that "be concerned for other sentient life" sticks around as an invariant when the AI self-improves or builds other AIs.

There's an obvious informal argument that an agent optimizing a utility function U would by default try to build another agent that optimized U (because that leads to higher expected U than building an agent that optimizes not-U). When we tried to formalize this argument in what seemed like the obvious way, we started running into Godelian problems which seemed to imply issues like: "If the AI reasons in a system T, will it be able to trust a theorem asserted by its own memory, or by an exact copy of itself, given that things usually blow up when a system T asserts that anything T proves can be trusted?"

The point of this wasn't to investigate reflectivity and self-description in mathematical systems for the sake of their pure interestingness. We were trying to get to the point where we could write down any coherent description whatsoever of a self-modifying stable AI. It turned out that we actually did run into the obvious Godelian obstacle. And so far as we could tell, there was no trivial cunning way to dance around the Godelian obstacle without losing what seemed like some critical desideratum of a bounded self-modifying AI. For example, System X can readily reason about System Y if System X is much larger than System Y; but a bounded AI should expect its future self to be larger and more capable than its current self, not the other way around.

On this avenue, we've recently published a major new result of which Scott Aaronson and others have spoken favorably. We can now exhibit a system that can talk sensibly and probabilistically about logical propositions (like the probability that the trillionth decimal digit of pi is a 9, in advance of computing that digit exactly) and get many nice desiderata simultaneously. E.g., the probability distribution learning from experience as it discovers more mathematical facts, without any obvious Godelian limitations on what can be learned. More importantly from our perspective, the probability distribution is able to talk about itself with almost perfect precision (absolute precision still being unattainable for Godelian reasons). We took "math talking about math" out of the realm of logical statements and into the realm of probabilistic statements, much further than had previously been done.

https://intelligence.org/2016/09/12/new-paper-logical-induction/

This in turn takes us a lot closer to being able to write down a description of a not-totally-boundless stably self-modifying AI that reasons about itself or other cognitive systems that assign probabilities to logical propositions. Which in turn gets us closer to being able to talk formally about many other reflective and self-modeling and cognitive-systems-modeling problems in a sufficiently advanced AI that we'd like to be stable.

===

The current academically dominant version of formal decision theory, "causal decision theory", is problematic from our own standpoint in that CDT is not reflectively consistent: if you have a CDT agent, it will immediately run off and try to self-modify into something that's not a CDT agent. The same theory also says that a formally-rational selfish superintelligence should defect in the Prisoner's Dilemma against its own clone, which some might regard as unreasonable.

Trying to resolve this issue led to another behemothic set of results we're still trying to write up, so I can only point you to a relatively informal and unfinished online intro:

https://arbital.com/p/logical_dt/

But we have now shown, for example, how two dissimilar AIs with common knowledge of each other's source code--and no other means of communication, coordination, or enforcement--could end up robustly and reliably cooperating on the Prisoner's Dilemma:

https://arxiv.org/abs/1401.5577

This new decision theory also has economic implications about negotiations a la the Ultimatum Game, about whether it's 'rational' to vote in elections, and a number of other informal problems in which the 'rational' choice was previously said to be something unpalatable (see the first link above).

===

https://intelligence.org/files/Corrigibility.pdf

...is an overview of the "how do you make an AI that wants to keep its suspend button" problem, aka "how do you build an agent that learns over two possible utility functions given one bit of information, without giving the agent an incentive to modify that bit of information, or assigning it a definite value in advance of your trying to press the button or not".

===

Everything above is on the more theory-laden side of our work. We only recently started spinning up work on alignment within the paradigm of modern machine learning algorithms; half our researchers have decided to allocate themselves to that going forward. Here's our current broad technical agenda for that research:

https://intelligence.org/files/AlignmentMachineLearning.pdf
https://intelligence.org/2016/05/04/announcing-a-new-research-program/

An example of a bit of specific work we may try to do there soon is figure out how to translate an autoencoder's representation into something closer to a human-readable representation--a representation that more easily 'exposes' or makes easier to compute, the answers to the sort of questions that humans ask. (An unsupervised system learns a representation; we want to learn a transform from that representation onto another representation, such that the second representation makes it easy to compute the answers to a class of supervised questions.)

===

MIRI has 8 fulltime researchers including me, plus 2 interns and a few more funded to do part-time work. 2 more researchers with signed offer letters are starting later this year / early 2017. We have 6 ops staff, plus 1 more starting in November (handling research workshops, popular writeups, keeping the lights on, etcetera). We're on track to spend $1.6M this year and we hope to spend $2.2M in 2017. We are currently at $257,000 out of a needed $750,000 basic target in our annual public fundraiser, with 12 days remaining:
https://intelligence.org/donate/

===

Hope this conveys some general idea of what MIRI does and why!

--

Eliezer Yudkowsky
Research Fellow, Machine Intelligence Research Institute