First Response to the Crit Rats about AI Risk

Intro

I said on twitter, that I thought that the critical rationalist argument against AI risk, doesn't hold up. Now that I've gotten some more clarity around the shape of that argument,, I want to write up a detailed counterargument to that story.

To start, this is my paraphrase of the basic Crit Rat argument against AI risk.

Paraphrase of the Crit Rat view

The core objection to AI Risk seems to be the following: [read on twitter]

> There are two things that you might call "AI".
>
> The first is non-general AI, which is a program that follows some pre-set algorithm to solve a pre-set problem. This includes modern Machine Learning.
>
> There might be some risk from powerful, but non-general, non-creative AI systems, and it seems fine to think about that some. But non-creative systems are not extremely dangerous: they won't be strategic, creative adversaries.
>
> The second kind of thing called AI, what we would properly call AGI, is different.
>
> An AGI is a program that is CREATIVE, which means that it can generate new explanatory theories as well as create and criticize ideas.
>
> A creative entity is "universal" in the sense of "universal computation": It can comprehend any idea that can be represented by a Turing machine.
>
> Creative entities are PEOPLE. Indeed, The only things that we know of (so far) that is creative are individual humans.
>
> The "goals" of a creative entity are ideas, in its mind, like other ideas. They're not some stable permanent thing that is hard-coded into the mind.
>
> Creative entities consider and criticize different possible goals, and through that process, tend to change (indeed, improve) their goals over time.

That's NOT to say that the starting conditions of an agent's goals don't matter. But they aren't constraining. While a human has some innate predispositions from evolution (things like fear of heights or a propensity to violence) we can overcome those predispositions. (An example that @iamFilos has given several times is one of learning to find sky-diving exhilarating instead of terrifying.)

Similarly, AGIs would presumably have some innate predispositions that result from the parochial details of their design process, but they would likewise be able to overcome those predispositions and change their goals.

Therefore, if a person tries "aligning" an AGI, that must mean inhibiting the ability of the AGI to criticize and change some subset of it's ideas, namely its values. This seems like it must have the outcome of either...

1) breaking the basic general creativity of the AGI mind

...or...

2) Creating some kind of slave-monstrosity that has the values that you chose at the time frozen in.

And in fact, we already know how to live safely in a society with creative entities (people). We raise them (lovingly), and educate them. We give them moral arguments, and have them reason for themselves about what is good to do.

We don't try to inhibit their ability to change their goals.

Soon the AIs will be participating in the great process of figuring out which explanations are best, and helping us to generate new moral knowledge!

There IS some risk that an AI system will get "hung up" on some set of ideas for some amount of time, in the same way that humans can get hung up on a set of ideas (this manifests as our neuroses).

If AIs are particularly powerful,getting hung up like this seems like it could be quite bad, because it might do a lot of damage before its epistemic process got unstuck. But, that seems like a far cry from "the default outcome is doom."

So far, it seems that everyone agrees that this is a pretty good summary of the Crit Rat position, and is not leaving anything out. [links]

What I agree with

Let me start by just stating everything about this story that seems true and correct to me, to minimize miscommunication about what I am claiming.

- First of all, yes, when I am talking about AI risk, I am mostly thinking about general, or creative, AI systems, not narrow systems.
    - I do think that we have related problems with losing control of our civilization, as we hand over responsibility for more and more key functions to extremely powerful, but nevertheless non-general AI systems, but we can limit *this* discussion to risk from AGI.

- I do acknowledge the theory of universal computation and that every turing machine can emulate every other turing machine. I agree that in principle, both humans and AGIs can represent any idea that is computable, modulo space constraints.
    - However, I'll note that you might have to rely on some very awkward work-arounds. If you have specialized hardware designed to do some specific task, running an arbitrary program on that hardware might involve the computational work of emulating a virtual machine to run the program + running the program. This might be horrendously efficient in time or memory compared to running the program on hardware that isn't specialized for something else.
    - In fact, I think this is happening to the human brain. Writing a program that does multiplication on floating point numbers is very computationally simple, but I think that the brain is using this enormously complicated general workspace / conscious mind machinery (which, as Dennett posits, is analogous to a virtual machine), as a way to do multiplication as a specialized kind of neural net. Humans can learn to do multiplication, but only by using orders of magnitude more computation than the python multiplication function.

- I do think that the goals of an AGI system will be "ideas". That is, they won't, or at least won't necessarily have to be, any fundamentally different type signature than other information represented in memory. The goals aren't stored as some separate mystical thing. It's all just implicit or explicit knowledge, stored in memory.

- I *do* expect the goals of an AGI system to change overtime. In fact, I think there is a *stronger* reason to imagine that AGIs will have mutable goals than there is for humans. Humans, while they can critique and modify the ideas in their minds, can't, practically speaking (yet), modify themselves by means of neurosurgery. In contrast "tweaking the neurons directly" is much more practical for an AGI system. So an AGI will be able to not just improve its ideas, but also modify its own source-code.

- Furthermore, I also think that it would be abhorrent to lock in my current values into an AGI. Not only would that be bad for the AGI it would be bad for me, because I expect that my current values, while they are my best guess about the Good, are flawed in

many ways. Let me say this again: AI ALIGNMENT IS NOT ABOUT LOCKING IN YOUR CURRENT VALUES.

- Furthermore, I agree that the ideal outcome is that we build AGI systems that will participate in helping us figure out “moral knowledge”, by which I mean, “what kind of universe we, on reflection, want to live in” and “how to use our massive powers over the cosmos wisely.”

- I don’t think that humans are “safe”, and I agree that humans pose an existential risk to humanity. I agree that we should be trying to solve that problem, and in the case of humans this is mostly done by creating sane, nourishing cultures, that largely avoid traumatizing their participants.

But, nevertheless, believing all of the above, I still think that the argument given doesn’t follow, and the default outcome for AGI is human extinction.

(I encourage you to take a moment, and reflect, asking yourself, *if* it is true that the default AGI path leads to extinction, do you want to know that? If not, you can stop engaging now. If so, let’s go!)

Side note disagreement

One thing that I *don’t* think, which I think is a sideline to the main point of this essay, but still seems good to note explicitly, is that there is a crisp distinction between General Intelligences and narrow AI.

In the same way that there are some important parallels (and also some important differences) between dogs and a human-beings, there are some important parallels (as well as some important differences) between, say, clearly non-general reinforcement learners, and AGIs. In both cases, you can learn something about the more “advanced” thing, from studying the less advanced thing, but you have to be careful of overgeneralizing and assuming too much commonality.

How to engage with this document

I wrote up this document in only a couple of days. It’s pretty rough, and it may not be clear in places. Please be patient with me as I find ways to explain these concepts, effectively efficiently. If needed I will rewrite this so that future iterations of this conversation can go more smoothly. If any section is confusing, in that you literally don’t know what it means, please let me know.

I also apologize if any of this is obvious or overwrought. I’m trying to draw out the argument at a high level of granularity to avoid “hand-waving”, and that might entail spending pages on claims that everyone already agreed with, and that I could have just taken for granted.

My hope is that rather than responding on twitter, people can read, and make comments on this document. That will give us a kind of multi-threading, without many of the frustrations of twitter.

There are two particular ways to engage that would be helpful for me in making progress on this conversation, spelling out the alternative story, and checking for crux-y-ness.

Spelling out the alternative story

In the main body of this essay, I'm going to attempt to spell out, in moderate detail, how I think the process of an AI (or really any agent) changing its goals must go, by necessity. If you dispute my account, the thing that would be most helpful for me, is if you could spell out in a similar level of detail, how you think it *does* work.

Then I could look over your account and try to wrap my head around it.

Checking for cruxes

The second thing that would be helpful for me is if you could check whether my claims are cruxes for you.

That is, if you disagree with some particular claim that I make here, it would be helpful if you could not just note your disagreement, but flag how crucial the disagreement is to your overall view.

It is much more helpful for me to read,

> "I think that the claim that bunnies are secretly extremely advanced aliens in disguise is false, but if I found out that actually they are advanced aliens, that would make me think that maybe they do pose an existential risk. This is a crux for me."

rather than just,

> "But bunnies aren't actually extremely advanced aliens".

The first one flags for me that it is useful to go into more detail about why I think bunnies are technologically advanced aliens, while the later one doesn't help me figure out what to prioritize.

Be wary of an optimism bias

Also, I don't know if I have any business suggesting how you go about running your own cognition, but I'll offer a suggestion, which you are welcome to do whatever you like with.

I find that this topic is often rife with a kind of optimism bias, where people search for reasons why everything will be ok, rather than asking themselves what is most likely to go wrong.

In particular, there's a pattern where people will say "AI risk isn't a problem because of [argument X]." Then after a long discussion they come to see why [argument X] is flawed, and there might actually be some risk from [problem Y]. Noticing this, they immediately jump to stating how we can trivially solve [problem Y], with some solution that they just thought up.

I think in this case it is often more useful to pause arguing for the conclusion that there isn't AI-risk and examine if there any flaws in their proposed solution to [problem Y].

Trying to find solutions to problems is great. *Asserting* that a given solution solves a problem so thoroughly that we don't need to worry about it anymore, before we have carefully critiqued the solution, is not so helpful.

The counterargument

If you want the tldr; this post is a pretty good summary of the core of my counterargument, but I think it will be basically useless to read that post, because it is starting from different assumptions than Crit rats usually do. The rest of this document is an attempt to unpack the details in a way that is (mostly)compatible with the critical rationalist frame.

Please note, that while this document is pretty long, there are lots of places where I am simplifying somewhat or glossing over complexity. It would be nice if you didn't run off and say "Baysians believe exactly X!", without giving context. I would have emphasized different things in a different conversation.

In particular, I'm talking here about the classic Bostromian sharp takeoff, mono-polar scenario. I think this scenario is possible and probable enough to be concerning, but I don't think that it is the all-things-considered most likely way for things to go. I just think that it is the easiest one to explain given my understanding of the Crit Rat perspective.

Some other scenarios are outlined here. (I don't know if that post will be accessible from the Crit Rat frame. It does, sort of, assume a continuum between current ML and transformative, if not general, AI. But it does have the advantage of being much shorter than this essay.)

Also note that this document skips over the question of whether there could be AIs or groups of AIs that are much more intellectually productive than humans. It seemed to me that most of the people I talked with thought that it was possible to have AIs that are, if not qualitatively different than humans, at least much more intellectually productive (either by way of algorithms that are more creative, or by running on hardware that alow them to think, and accumulate knowledge at much faster rates), and that if such AIs were motivated to wipe out humanity, they could. [links]

So, I'm just going to assume, in the context of this essay, that *if* AGIs were motivated to destroy humanity, they would likely succeed. We can go back and retread that ground if necessary.

The Mechanism of Goal Criticism

Again, I do think that AGIs will evaluate and criticize, and ultimately change their goals. But, I want to zoom in on what that process looks like, in detail.[1]

Goals are always evaluated in terms of other goals

My key claim here is that **when considering a potential goal, an agent must use some kind of assessment schema, to evaluate whether to adopt the goal or not. And that assessment schema, at any given time, is defined (in part), by the agent's current goals.**

For instance, let's start by talking about humans.

Choosing to become a doctor or not

Let's say there's a person who is in college (pre-med) and is currently planning to become a doctor, but is considering choosing some other career instead. In making that decision, he might reflect on a number of considerations.

He might introspect and try to determine if he will be happier as a doctor, or as a philosopher, or as a programmer, or as a hermit living in the wilderness.

Perhaps he's heard some argument that he can actually save more lives by working in finance and donating 10% of their salary to malaria eradication. He might try and figure out if that argument is valid.

As one particular mode of analysis, they might examine their reasons for wanting to be a doctor in the first place and evaluate if those reasons are good, for instance maybe he has the implicit belief that his parents will only really love him if he becomes a doctor like his dad, but on reflection, he realizes that he doesn't want his life to be dictated by his parents in that way.

Or maybe he had thought that doctors make a lot of money, and he thinks that he should check how much they actually make.

Or perhaps he has a desire to "give back to society". He might ask how much being a doctor actually gives back to society.

[In my understanding all of these are different ways that the person is criticizing (in Crit Rat parlance) the goal of being a doctor. Please let me know if some of these are not included in

[1] Note that in this section, I am talking about "goals". I think that there are strong reasons why the actual formulation has to be a utility function, and not a basket of goals, but that distinction isn't relevant for this argument.

criticism, or if there is some major “flavor” of criticism, relevant to this situation, that I’ve left out here.]

In general, he considers this potential goal in comparison to other potential goals, and evaluates them on the basis of how well they fare against various arguments (or, if we prefer, “explanations”, though I think that it is somewhat awkward to think of the arguments as explanations for why one goal is better than other.)

In the end, he’ll either decide to keep the “become a doctor” goal, or abandon it, in favor of some other goal that better stands up to criticism.

Crucially, note that each particular criticism is evaluating the goal of becoming a doctor, *in terms of some other goal or value that the person cares about*.

In the example criticisms above, the criteria of evaluation were...

- His personal happiness
- Saving lives / the number of lives saved
- His parents loving him
- making money
- “giving back to society”

...though of course, for any given person making this decision those factors might or might not be relevant to their decision, and there might be other factors not listed that are relevant.

But, whatever considerations / arguments *are* relevant in this case all of them will have a similar form: **they will be arguments for adopting or not adopting this goal, by the standards of *other* goals / values of our pre-med student.**

Maximizing orange

One could, of course, make an argument that justifies adopting the goal of becoming a doctor (or not) on the basis of some other goal or value that our pre-med student doesn’t share, but obviously wouldn’t be compelling.

For instance, a person made the argument, “you shouldn’t become a doctor. Instead, you should instead dedicate your life to painting things orange, because then there would be a lot more orange things in the world.” Our pre-med would give them a funny look, and ignore them.

The argument is valid, becoming a serial orange-painter really does lead to a world with much more orange-colored things than being a doctor does, but the argument isn’t compelling, because who cares about there being more orange-colored things? I*n order for the argument to*

be motivationally compelling, a person has to already care about increasing the number of orange things in the world.

In general, when criticizing some goal, an agent will *always* evaluate it in terms of other goals and values that the agent holds. One never evaluates goals on the basis of goals that one doesn't hold.

If you can think of a counterexample to this general principle, please share it. Depending on the nature of the counter example, I expect it might very well be a crux for me.

Recursion

Now of course, this applies recursively. For every goal X, which is used in evaluating some possible goal Y, the agent could similarly criticize and evaluate X.

If you're deciding whether to become a doctor or not, and one consideration is how much money you'll make, you might recurse, criticizing the goal of "make lots of money", asking how valuable that actually is.

But the same basic situation obtains, just one level down. All of the considerations for or against maintaining "make lots of money" as a goal, are in terms of other goals or other values that you hold (perhaps things like, "security", "freedom from worry", "quality of life", "cool vacations", or "charitable work".

Or to take another example, you might imagine Bob trying to persuade Alice to buy a prius instead of a Hummer, because of the impact on the climate, and the consequences for will have if on future generations. If Alice says "well, I don't care about future generations", Bob might recurse, shifting to trying to persuade Alice that she *should* care about future generations.

But to do this, Bob has to make moral arguments that appeal to *something* that Alice actually cares about, which might mean helping bob to introspect and notice the part of himself that doesn't want people to suffer, or appealing to his sense of fairness or whatever.

Nevertheless, the only way for an argument to be motivationally compelling is if it recruits some other goal that is already motivationally compelling.

(As a corollary of this necessary principle, one can trace back the causal history of an agent's goals, at any given timestep to the "seed goals", that agent was born or built with. It may or may not have abandoned some of those seed goals (I'm not sure if there is any consistent setup which will cause it to abandon ALL of its seed goals?), but regardless, all of the agent's current goals must be justified by other goals that are or were themselves justified, ad inantitem, until we reach the original seed goals.)

Some preementive clarifications

Justification

Note that I when I use the word “justified” in the previous section, I don’t mean *epistemically* justified, necessarily, which is important since (if I understand correctly), Critical Rationalism holds that no knowledge is EVER justified.

I mean merely “motivationally justified”, as in “the reason why I *care* about goal x is *because* of goal Y.”

Ethical schemas

As I’ve been construing this so far, I’ve been imagining that the assessment of a possible goal is a consequentialist calculation. That is, you ask “would I get more of [x thing that I value], if I adopted [goal A]?”

But that doesn’t necessarily need to be the case. A virtue ethicist might ask not “what good things will happen if I adopt the goal of always being honest?” and instead ask “is it virtuous to adopt the goal of always being honest?”. But the point still stands; the virtue ethicist is assessing a potential goal by the criteria of another goal that they already hold.

What’s at the bottom?

If all of our goals can only be justified in terms of our other goals, then we might ask “what happens when we get to the ‘bottom of the stack’?”

It must be the case that we either have terminal goals, or some of our goals are mutually motivationally-justifying.

Terminal goals

At “the bottom” there are some goal(s) that that we value, for their own sake, rather than on the basis of any particular argument that we ought to care about. Some things that we prefer without having a reason for preferring them.

These are typically called “terminal goals” in contrast to instrumental goals (goals that we only care about insofar as they help us achieve some other goal.

Note that talking about “terminal goals” doesn’t imply that these goals are of some fundamentally different type than the other goals. They’re still ideas, or if you prefer, representations in memory. They’re just not motivationally-justified by anything other than the fiat of the mind in question.

Cycles

If there aren’t terminal goals, there have to be cycles in the justification graph, where X is valuable because Y is valuable because Z is valuable because X is valuable.

(Now that I think about it, if there are both goals that are part of cycles, and goals that are not part of cycles, we can draw a box around the cycles in the graph, and treat each of those boxes as isomorphic to a singular terminal goal. It's the self-justifying / section of the goal structure, from which other goals derive their justification.)

(Note that if things bottom out in goal cycles, I expect that the set of goals would settle into some self-reinforcing equilibrium. Assuming that no goals are hard-coded and that all goals can be abandoned, if you have some goals that contradict each other, in the limit, we expect the criticism process to knock out the contradictory goals until they are compatible with each other. More on this is in later sections.)

[Now is a good place to stop and object if any of this seems mistaken, because the next two sections will assume everything up until this point.

In particular, if you think this is not how criticism of potential goals works, the thing that would be most helpful for me, is if you could write up, in specific detail, how it ***does*** *work.*]

Self-Stabilizing Goal Structures

The Gandhi folk theorem

As a consequence of the fact that agents evaluate possible goals according to the criteria of their other goals, there are many goals that are stably self-reinforcing, and there furthermore, there are multiple stable equilibria in the space of goal structures.

Around MIRI we'll sometimes talk about the Gandhi folk theorem: which is the loose generalization that agents will usually not modify their own fundamental/ terminal goals.

Choosing asexuality or sexuality

As a real life, empirical example of the Gandhi folk theorem: consider asexuality, and a hypothetical pill that can, safely and permanently, alter one's sex drive.

I run workshops and training sessions, in which I have, many times, handed out a worksheet that involves eliciting people's intuitions about sexuality and bisexuality, as part of a broader exercise of eliciting intuitions for disagreement.[2]

In general there is a clear pattern:

Asexual people, if offered a pill that would increase their sex drive, they would not take it. From their perspective, this looks developing a new psychological need, which would be a hassle to

[2] The specfic question on the worksheets is "If you have to pick one or the other, would you press a button that made everyone in the world asexual, or a button that made everyone in the wrold bisexual."

meet, and furthermore, seems particularly reality distorting. Furthermore, they often have a reaction of the form “sex is gross and weird, and I don’t want to like it.”

In contrast, when folks with a high-sex drive contemplate taking a pill that would make them asexual, they are similarly vehemently opposed. From their perspective, sex is a core part of what makes life worth living, and it seems actively bad to them to no longer have an interest in it, because then they would have much less sex!

It seems like there are two stable equilibria: there are people who have a low sex drive and who generally prefer having a low sex-drive, and there are people who have a high sex drive and prefer having a high sex drive.

Both these equilibria are consistent. And they are self-stabilizing. So long as they are free to choose to take a sex-drive alteration pill or not, the people with high-sex drives will retain their high sex-drives, or stated differently, continue to hold the goal of having sex, and the people with low sex drives will continue not to have that goal. Their goal in this domain will be stable, *even though they are free to criticize and evaluate that goal*.

The sex-goal is stable, not because it is fundementally engrained, or hard coded, but because whichever state one is in, he or she choses not to change it.

Choosing sociopathy or empathy

We can consider other psychological variables that might illustrate this property.

For instance, consider psychopaths and non-psychopaths. Most of us, if offered a pill that would permanently remove our empathy and care for other human beings would be horrified. Our empathy and care is one of the things we cherish most!

But I speculate that if you asked many high-functioning sociopaths if they would like to take a pill that would “cure” their sociopathy, they would likewise refuse. (I think this because I vaguely remember some comments from *Confessions of a Sociopath*, along the lines of “they are faintly amused by and glad to take advantage of, the way that most people are willing to trade a ‘real’ thing, like money, for a ‘fake’ thing, like some sense of being important.)

Again (supposing for the sake of argument, that my speculation is correct), we have agents maintaining stable goals even in the face of criticism.

Generalizing the Gandhi folk theorem to a hypothetical paperclipper

Changing a goal for other reasons

But this general tendency to preserve one's goals in some circumstances is not sufficient. Presumably, whether one should be sexual or asexual, a psychopath or a non-psychopath (when given the choice to choose one or the other), is *not* mearly a matter of preference. Surely reason must have some bearing on the issue, right?

Perhaps an asexual person has some native revulsion to sex, but if reason (or our current best explainations?) leads us to think that actually, it is better to have a high-sex drive, then that revulsion is just another predispostion to be overcome. And vis versa, if it seem that on balance it is actually better to be asexual.

Even if one's preference and meta-preference are in a stable equilibrium, that doesn't mean that there might be other considerations that dominate that mata-preference.

But remember, in order for those considerations to be compelling, they must be justified in terms of other goals the person has. (Perhaps, in this case, personal happiness, the social good, the propagation of the human race, reduction of violence, etc.)

One can make arguments that it is better to be sexual or asexual, but those arguments have to be grounded in some other goal that you the agent in question cares about.

[Now is another good place to stop and object to anything that seems wrong, before going forward.]

Paperclipers

Let's consider the hypothetical of a paperclip maximizer.

A paperclip maximizer, or paper clipper, is a kind of AGI. It is fully universal, intelligent, and creative, in that it can reason, and develop creative plans, and do good science. But where humans have a basket of goals (things like being a good parent, helping the world, having fun, having good friendships, making cool stuff, and so on), a paperclip maximizer has exactly one goal: to maximize the number of paperclips in the universe.

It has that monomaniacal goal by dint of its initial conditions. Maybe it was built by a malicious actor, or maybe it just happened to start out with this goal by some extremely improbable accident. The details don't really matter, for this hypothetical.

Crucially, being a fully general, universal agent, it regularly generates other possible goals that it might pursue. But every time that happens, it evaluates that possible goal according to its only goal: it asks "Would adopting this goal help me to maximize paper clips?" If the answer is yes, it adopts that goal. If the answer is no it doesn't.

So it might end up with a bunch of (instrumental) goals, compatible with each other, but all of them were chosen on the basis of paperclip maximization. Everything that it does is directed towards, however distantly, maximizing paperclips.[3]

[note here to respond to a potential argument about agents confusing instrumental and terminal goals, and accidently getting value drift, if it comes up.]

Now, of course, you might think that "maximize the number of paperclips in the universe" is a stupid goal. And that's fine. But the question is: what argument could you give it that would compel it to change its mind and decide to pursue something other than maximizing paperclips?

It seems to me that once you have a paperclipper on your hands, persuasion is futile. Not because the paperclipper can't, in principle, respond to arguments, or change its goals, but because in practice, its goal is self-stabilizing.

If you find yourself faced with a powerful AGI with self-stabilizing goals, it seems like you've fucked up. You want to figure out a way to not end up in that situation in the first place.

(I *think* that many of you think that a paperclipper is impossible, though I'm not sure how you can think that, since a cornerstone of your view is that computation is universal. Surely this is a possible mind that could exist, however improbable it is. If not, how is it that this sort of AI couldn't be represented with a Turing machine?)

Evolutionary Micro-structure of Goals

Evolutionary heritage

The thing that makes a papercliper incorrigible is not that it only has *one* goal. The thing that makes it incorrigible is that the paperclipper set of goals has almost no overlap with the set of human goals.

My key claim in all of this is that **human values and human morality rely on a foundation of evolved instincts, urges, and predispositions, which are specific to our own evolutionary trajectory**.

Most humans, preculture, are born with the following instincts / urges / affective reactions / preferences.

- dislike of pain
- a sense of loneliness

[3] You might think that by this method that it would accumulate a number of goals that help to maximize paperclips, but that those goals would take on "a life of their own", and would be a foundation for the AGI to criticize its paperclip maximiring goals. I think this is a sideline, so I won't go into it in detail, but I'll note that this doesn't particularly save us, because those instrumental goals are no

- hunger
- a preference for symmetry
- a desire for attention (especially when young)
- An innate sense of “fairness”
- empathy
- a desire to be liked
- curiosity / a very specific sense of “interestingness”
- a desire to be thought well of by members of a group
- tribal instincts
- sex drive
- sexual attractiveness assessment heuristics
- social emotions like embarrassment, guilt, shame
- an innate desire for autonomy
- competitiveness
- an interest in the social
- a sense of effort and tiredness, and the urge to rest

All of these are the result of specific evolutionary influences on the human mind over the course of animal and mammalian evolution.[4]

They are parochial. Not universal.

I think there are a number of ways that we can tell that they are parochial, starting from the fact that we can see that other animals (which are not general reasoners or in Crit Rat parlance “universal”) have a wide variety of innate predispositions. If a selective breeding project were to select for increasing intelligence in some other species, until it reached generality / universality, we would see that they have a quite different set of urges and predispositions. [Anyone dispute that?]

And the comparison to other animals undershoots how much variation there is in the space of mind designs. We all evolved on earth and are part of the same phylogenetic tree. In the grand

[4] Some caveats:

- Note that I am NOT claiming that everyone has exactly this same compliment of traits. There is obviously variation between humans. For instance, some of us are missing some social emotions that most of us have. And people have different sexual orientations, which is to say that they find very different things sexually attractive.
- I am NOT claiming that these predispositions are permanent and unchangeable. I think that with work, you can substantially modify or even eliminate most of these (note what school does to curiosity, for most people).
- I am NOT saying that the above urges / inclinations / preferences / predispositions ARE human values. Rather, I claim that our values emerge from an interplay between our cultural engagement, and these innate predispositions.

scheme of things, as different as we are, I think that earthly animals have more in common with each other than they have differences between them.

Mindspace is deep and wide, and the variation we've seen in the animal kingdom only barely scratches the surface.

I have close to no idea what sort of predispositions an AI would start with as a result of its design and training process. It depends heavily on the details of the environment it and it's predecessors was trained / formed in.

But it might include things like

- an urge to consume and coopt computational resources?
- a satisfaction with efficient allocation of memory?
- Some kind of sense of something-like-interestingness based on whatever problems it has evolved to solve in its training environment?

(I think those are actually not very plausible suggestions. But I really don't know what kind of urges and predispositions are likely from an AI design / training process.)

The difference between these two sets of initial conditions are crucial because they constrain (not determine, but constrain) the trajectory of goal formation going forward, and limit the set of stable goal-structure equilibria the entity might evolve towards in a process of goal-criticism.

Starting from those different predispositions, a human and AGI are both going to engage in a process of goal generation and goal criticism, guided by reason and argument,

As an illustration, let's talk about my own personal values.

The microstructure of (some) of Eli's goal structure

Much of my life choices are driven by a sense of the grand and glorious project of advancing civilizations: our iterative, compounding steps to understand the universe, and advanced moral progress, and make things better. I am motivated by the vision of a glorious future that could be: brilliant, free, happy beings working and loving together, pursuing discovery and creation, freed from death and trauma and the horrors of our world.

(I imagine that many of you share a comparable ethos and a similar set of high-level goals.)

Asking what the "root" of this set of values is somewhat misguided. I don't think that there is some single root, that it all chains back to. Rather, this feels compelling to me, because these ideas evoke strong affective reactions in me.

I feel inspired by this vision, and have a sense of pride, to be contributing to it.

I feel excitement thinking about what is possible.

I especially feel proud to be standing with people who are putting the improvement of the world, and epistemic integrity above their own ego, and personal-prosperity. When I see someone openly admit that they were mistaken, and change their mind, or refute a bad argument that supports their side, I feel a jolt of pride and inspiration.

I have a sense of defiance, at the cold dark universe that doesn't care about us, and the molochian pressures that push us toward 0-sum competition. A sense of "fuck that, just watch us build something awesome despite you."

I have deep compassion for all the people alive today, who are trapped in hellish circumstances, both physical and psychological, with little power over their own fate. I think of the people dying who don't need to die; the children, traumatized in schools for no reason; the Alzheimer's patients who watch their very self decay. I feel sorrow, and defiance, and fear that that could be me.

I care about "being a good person". I want to be able to, in the end, after all the rationalizations are stripped away, to ask myself "did I do the best that I could

I want to be part of a community of incredible people building the future together. That feels exciting, and fascinating, and

I feel a sense of duty to all the anonymous people that lived, labored, and died, to push civilization incrementally forward. I owe it to them to carry the project forward, to "pass it on" to future generations.

In particular, I think of the men and women who bravely and courageously resisted the NAZIs and fought American slavery, at great risk to themselves. They were called upon to do what was needed in their point in space time, I owe it to them to do what I can in my point in spacetime.

I feel a desire to personally be awesome.

When I think of people being free to choose, I feel a sense of "fuck yeah" and inspiration. I both hate the feeling of being constrained myself, and have some transcendent appreciation at the thought of entities that are free to do what they choose. I feel some strident abohorance of slavery.

My reasons for being compelled by this vision are complex and varied. And they certainly involve *beliefs* about the world (for instance, that unflinching seeking the truth is the best way we know to get to a better world).

But importantly, those reasons are *compelling*, because they recruit innate emotional reaction patterns and emotions. Things like...

- excitement
- moral pride
- defiance
- spite
- a sense of grandeur
- compassion
- a sense of interestingness
- sorrow
- fear (of specific bad things)
- empathy
- a sense of fairness
- transcendent appreciation
- duty and debt
- strident abhorrence

All of these are specific evolutionary adaptations. They are parochial solutions, found by evolution to specific problems of survival and coordination, that humans inherit as part of their evolutionary heritage.[5]

You don't get them for free.

And if a creative entity lacked all or most of those affective reactions, I think it wouldn't be feasible to convince that entity to come to value what I value.

Instrumental moral arguments

Now, aside from my personal parochial preferences for these things, there are some good, more general arguments for many of them.

For instance, I personally get an emotional, inspirational kick out of the idea of a free society, where individuals have protected rights of autonomy and self-determination, and free speech. But there are *also* good arguments for why I should prefer that. Free speech allows for the rapid generation and recombination of new ideas, and for a social process that enables the identification and correction of mistaken beliefs and arguments. And a free society in which people can organize in whatever way they choose allows for hugely more innovation and and productive wealth creation than any other set up we've tried yet.

[5] Note, again I am not saying that you can't overcome your evolutionary heritage. Part of my evolutionary heritage is anger, and I think that I can, in fact train myself out of having an anger response, and it is plausible to me that this is desirable. But my value structure here is self-stable, in that I would not choose to edit or remove these affective responses if I could.

But note that again, we are justifying free speech in terms of other goals that I care about. If I don't care about the error correction or wealth generation of society at large, then these arguments will lack force to move me.

And an agent with a very different goal-structure, especially a powerful agent, is unlikely to care about those things.

Vividly imagining another mind that actually doesn't want what you want

There's a thing that is inherently difficult about reasoning about creatures that have very different goals that we do, because we tend to project our own patterns of thought on to such agents, anthropomorphizing them.

In a conversation on these topics, one of you made the following argument:

> Suppose we had some AGI or some community of AGIs that had the exclusive fundamental goal of curiosity: the only thing they want is to understand the universe.
>
> The AGIs will need to figure out how to coordinate and cooperate among themselves. That requires knowledge, and in particular, the kind of knowledge that we call politics. That knowledge is embodied not just in books but in the actual interactions and social structures of humans.
>
> So the AGIs would have a vested interest in keeping human civilization around,

So first of all, I think we have a problem if the only reason why the AGI community is keeping us around is for our instrumental utility as an object of study. Not only is humanity now effectively in a zoo, overseen by a godlike civilization, but as soon as studying us no longer seems useful, we're done for.

It seems bad to be at the mercy of beings that only have an interest in you so long as you are useful.

But, I don't think that that is what would happen.

Human beings, all things considered are not very good at effectively aligning incentives and creating robust, effective coordination. I'm sure that there is a vast space of innovations in governance and organization much better than our current setups, that we can't even experiment with at scale because of various barriers to societal experimentation.

I'm very skeptical that a community of AI minds running (as per the hypothesis) at 1000x, making compounding intellectual progress, unconstrained to run their own experiments to iterate on better methods, and ultimately, by the nature of their situation facing very different

constraints on cooperation than humans do (things are very different if you can read each other's source code, or make copies of yourself), would see any need at all to study human politics to discover how to cooperate.

But even assuming that for some incredible reason, they do feel that they should study humans to learn politics, I don't think they would just let human civilization run. If I were in their situation, I would start by reading all the human books, to get that knowledge. I get some historical snapshots of the economy and governance mechanisms in practice, so that I could model how they work. Then I would permanently seize control of the world, so that I arrange specific experiments in sociological science: placing various sorts of groups of human in various kinds of constrained challenges in which they need to work together (with various incentive schemes), to explore all of the relevant dynamics

If possible, I would destructively upload all living humans, and save that data, so that I can run experiments with "fresh" subjects, at exactly my specifications, every time, and so that I don't need to keep any humans fed and or healthy perpetually. They're all saved to disk, and I'll boot them up as needed.

If that isn't an option, I would definitely not feel that the need for studying human social structures justfies, maintaining a population of 8 billion people, all consuming resources that I could use for other purposes. It certainly doesn't justify leaving their civilization in control of the planet.

The outcome of "the AGIs need us for our knowledge of politics is not a happy world where we all live in peace. It's a dystopia where the only reason humans are allowed to exist is so that they can function as lab rats in experiments run by a super-civilization that has no moral-concern for us whatsoever.

This issue here is that it isn't just sufficient to find an argument for why it would be good for the AGI's goals to keep us around. We need to find an argument that keeping us around is the literal best way, of all the available options, for the AGIs to accomplish their goals.

If you run the query "why might it be good for the AIs to keep us around?" you'll be able to generate answers.

But the more pertinent question is "what would a completely a-moral (from our perspective), sociopathic genius who only cares about learning more about how the universe works, and nothing else, do if it had sufficient power to do whatever it wanted to with earth and the solar system?"

But it is tricky for humans to think like this, because we're used to generating plans that are the sort of thing that we would consider good. It's tricky to learn to look at the world from through the

eyes of a being that has radically different goals than you do, that really doesn't care about the things that you care about.

This article gives another example of this mistake

> Juergen Schmidhuber of IDSIA, during the 2009 Singularity Summit, gave a talk proposing that the best and most moral utility function for an AI was the gain in compression of sensory data over time. Schmidhuber gave examples of valuable behaviors he thought this would motivate, like doing science and understanding the universe, or the construction of art and highly aesthetic objects.
>
> Yudkowsky in Q&A suggested that this utility function would instead motivate the construction of external objects that would internally generate random cryptographic secrets, encrypt highly regular streams of 1s and 0s, and then reveal the cryptographic secrets to the AI.

Again, we have a person who values art and science, asking "what sort of simple motivational structure would give rise to art and science? But If you take the proposed motivational structure seriously, you don't end up with an agent nicely doing art and science, you end up with an agent doing some much purer form of the specified utility function, usually to the detriment of human civilization.

It's a little like the paper clipper rationalizing that humans would also maximize paperclips, because humans need building materials and they could build things out of all the paperclips they make.

It is true that we could use paperclips as a building material, but it isn't even close to the best way to build structures to meet our goals.

From our perspective, trying to solve the problem of how to build reliable structures, we wouldn't even generate the idea of using paperclips, because we don't have any reason to privilege that hypothesis. It is only from within the goal-frame of the paper clipper that "make paperclips, as a building material, even seems like something worth considering."

We have to be careful to avoid making a similar mistake. That is, we need to be careful not to rationalize a conclusion that *we* find appealing, but rather work out what we expect that the AI would actually do.

It's easy to come up with arguments that sound compelling *to us*, for goals that we like. But what's necessary are arguments that are compelling to a totally different kind of mind.

Instrumental Convergence

The argument so far outlines why “the AI will criticize its goals” is not a sufficient condition for the AI landing on what we consider “good” or “sensible” goals. Because the evolution of one’s goal system is path-dependent, and constrained by one’s initial predispositions, different agents will end up with radically different stable goal-equilibria.

But this on it’s own is not a strong argument for doom. Merely having radically alien goals doesn’t on the face of it, imply a desire to destroy human civilization.

This is an argument of its own, and I’m not going to go into it in detail right here, but the short answer to this objection is instrumental convergence: any goal that doesn’t explicitly value human live and flourishing, can be better and more surely secured with with more resources (computational resources, and material resources) than less. So an agent, executing that goal, is better off seizing all of the resources on earth (including the atoms making up your body), than not.

Links for further reading on instrumental convergence.

https://en.wikipedia.org/wiki/Instrumental_convergence

https://www.nickbostrom.com/superintelligentwill.pdf

https://arbital.com/p/instrumental_convergence/

Conclusion

[This document really needs a conclusion to tie up the points that I’ve attempted to cover, and how they flow together. I’ll write one if this document continues to be relevant to the discussion.

In the meantime, if the spirit moves, you feel free to summarize the point I’m making here. If your summary is good, I won’t have to write one! :)]

Appendix: Babyeaters

As a sort of intuition pump, for thinking about engaging with minds with a very different culture

Some years ago, Yudkowsky wrote a story called *Three Worlds Collide*, in which humans encounter an alien race. That race not only eats its own children, while they are conscious and screaming, but eating children is the central, most meaningful, most important thing in their culture.

> "It's a truism in evolutionary biology that group selection can't work among non-relatives. The exception is if there are enforcement mechanisms, punishment for defectors - then there's no individual advantage to cheating, because you get slapped down. That's what happened with the Babyeaters. They didn't restrain their *individual* reproduction because the more children they put in the tribal pen, the more children of theirs were likely to survive. But the total production of offspring from the tribal pen was greater, if the children were winnowed down, and the survivors got more individual resources and attention afterward. That was how their species began to shift toward a *k*-strategy, an individual survival strategy. That was the beginning of their culture.
>
> "And anyone who tried to cheat, to hide away a child, or even go easier on their own children during the winnowing - well, the Babyeaters treated the merciful parents the same way that human tribes treat their traitors.
>
> "They developed psychological adaptations for enforcing that, their first great group norm. And those psychological adaptations, those emotions, were reused over the course of their evolution, as the Babyeaters began to adapt to their more complex societies. *Honor, friendship, the good of our tribe* - the Babyeaters acquired many of the same moral adaptations as humans, but their brains reused the emotional circuitry of infanticide to do it.
>
> "The Babyeater word for *good* means, literally, to eat children."

Now, I think, as you guys are used to thinking about it, that such aliens must be possible, specifically because of the universality of computation. There's nothing that prohibits such being existing.

(Or maybe you think that that "eating babies" is obviously an error, and a rational mind wouldn't never hold that as a goal for very long. I've attempted to make the argument here that a goal is only reasonable or unreasonable, within the frame of some set of goals, and an agent has a completely different moral frame, can converge to a totally different conclusion while maintaining a rational process.

So it is worth considering, what sorts of moral arguments would you give the baby eaters, to convince them that eating babies was evil?

The whole story is worth reading.