Best of LessWrong: December 2018

1. The Pavlov Strategy
2. Norms of Membership for Voluntary Groups
3. 2018 AI Alignment Literature Review and Charity Comparison
4. How did academia ensure papers were correct in the early 20th Century?
5. Spaghetti Towers
6. Transhumanism as Simplified Humanism
7. Meditations on Momentum
8. Contrite Strategies and The Need For Standards
9. Transhumanists Don't Need Special Dispositions
10. What makes people intellectually active?
11. Playing Politics
12. Reasons compute may not drive AI capabilities growth
13. Conceptual Analysis for AI Alignment
14. What are some concrete problems about logical counterfactuals?
15. Internet Search Tips: how I use Google/Google Scholar/Libgen
16. You can be wrong about what you like, and you often are
17. Best arguments against worrying about AI risk?
18. Should ethicists be inside or outside a profession?
19. New edition of "Rationality: From AI to Zombies"
20. The Bat and Ball Problem Revisited
21. Coherence arguments do not entail goal-directed behavior
22. Why should I care about rationality?
23. In what ways are holidays good?
24. Is there a standard discussion of vegetarianism/veganism?
25. Player vs. Character: A Two-Level Model of Ethics
26. Isaac Asimov's predictions for 2019 from 1984
27. Prediction Markets Are About Being Right
28. Book Review - Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness
29. Two Neglected Problems in Human-AI Safety
30. What does it mean to "believe" a thing to be true?
31. Three AI Safety Related Ideas
32. [Video] Why Not Just: Think of AGI Like a Corporation? (Robert Miles)
33. Argue Politics* With Your Best Friends
34. On Disingenuity
35. Kindergarten in NYC: Much More than You Wanted to Know
36. What is "Social Reality?"
37. Why we need a *theory* of human values
38. A hundred Shakespeares
39. Book review: Artificial Intelligence Safety and Security
40. 18-month follow-up on my self-concept work
41. The E-Coli Test for AI Alignment
42. Why I expect successful (narrow) alignment
43. Experiences of Self-deception
44. Can dying people "hold on" for something they are waiting for?
45. Standing on a pile of corpses
46. New Ratfic: Nyssa in the Realm of Possibility
47. Multi-agent predictive minds and AI alignment
48. Equivalence of State Machines and Coroutines
49. Measly Meditation Measurements
50. What are the axioms of rationality?

Best of LessWrong: December 2018

1. The Pavlov Strategy
2. Norms of Membership for Voluntary Groups
3. 2018 AI Alignment Literature Review and Charity Comparison
4. How did academia ensure papers were correct in the early 20th Century?
5. Spaghetti Towers
6. Transhumanism as Simplified Humanism
7. Meditations on Momentum
8. Contrite Strategies and The Need For Standards
9. Transhumanists Don't Need Special Dispositions
10. What makes people intellectually active?
11. Playing Politics
12. Reasons compute may not drive AI capabilities growth
13. Conceptual Analysis for AI Alignment
14. What are some concrete problems about logical counterfactuals?
15. Internet Search Tips: how I use Google/Google Scholar/Libgen
16. You can be wrong about what you like, and you often are
17. Best arguments against worrying about AI risk?
18. Should ethicists be inside or outside a profession?
19. New edition of "Rationality: From AI to Zombies"
20. The Bat and Ball Problem Revisited
21. Coherence arguments do not entail goal-directed behavior
22. Why should I care about rationality?
23. In what ways are holidays good?
24. Is there a standard discussion of vegetarianism/veganism?
25. Player vs. Character: A Two-Level Model of Ethics
26. Isaac Asimov's predictions for 2019 from 1984
27. Prediction Markets Are About Being Right
28. Book Review - Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness
29. Two Neglected Problems in Human-AI Safety
30. What does it mean to "believe" a thing to be true?
31. Three AI Safety Related Ideas
32. [Video] Why Not Just: Think of AGI Like a Corporation? (Robert Miles)
33. Argue Politics* With Your Best Friends
34. On Disingenuity
35. Kindergarten in NYC: Much More than You Wanted to Know
36. What is "Social Reality?"
37. Why we need a *theory* of human values
38. A hundred Shakespeares
39. Book review: Artificial Intelligence Safety and Security
40. 18-month follow-up on my self-concept work
41. The E-Coli Test for AI Alignment
42. Why I expect successful (narrow) alignment
43. Experiences of Self-deception
44. Can dying people "hold on" for something they are waiting for?
45. Standing on a pile of corpses
46. New Ratfic: Nyssa in the Realm of Possibility
47. Multi-agent predictive minds and AI alignment
48. Equivalence of State Machines and Coroutines

49. Measly Meditation Measurements
50. What are the axioms of rationality?

The Pavlov Strategy

Epistemic Status: Common knowledge, just not to me

The Evolution of Trust is a deceptively friendly little interactive game. Near the end, there's a "sandbox" evolutionary game theory simulator. It's pretty flexible. You can do quick experiments in it without writing code. I highly recommend playing around.

One of the things that surprised me was a strategy the game calls Simpleton, also known in the literature as Pavlov. In certain conditions, it works pretty well — even better than tit-for-tat or tit-for-tat with forgiveness.

Let's set the framework first. You have a Prisoner's dilemma type game.

- If both parties cooperate, they each get +2 points.
- If one cooperates and the other defects, the defector gets +3 points and the cooperator gets -1 point
- If both defect, both get 0 points.

This game is *iterated* — you're randomly assigned to a partner and you play many rounds. Longer rounds reward more cooperative strategies; shorter rounds reward more defection.

It's also *evolutionary* — you have a proportion of bots each playing their strategies, and after each round, the bots with the most points replicate and the bots with the least points die out. Successful strategies will tend to reproduce while unsuccessful ones die out. In other words, this is the Darwin Game.

Finally, it's *stochastic* — there's a small probability that any bot will make a mistake and cooperate or defect at random.

Now, how does Pavlov work?

Pavlov starts off cooperating. If the other player cooperates with Pavlov, Pavlov keeps doing whatever it's doing, even if it was a mistake; if the other player defects, Pavlov switches its behavior, even if it was a mistake.

In other words, Pavlov:

- cooperates when you cooperate with it, except by mistake
- "pushes boundaries" and keeps defecting when you cooperate, until you retaliate
- "concedes when punished" and cooperates after a defect/defect result
- "retaliates against unprovoked aggression", defecting if you defect on it while it cooperates.

If there's any randomness, Pavlov is better at cooperating with itself than Tit-For-Tat. One accidental defection and two Tit-For-Tats are stuck in an eternal defect cycle, while Pavlov's forgive each other and wind up back in a cooperate/cooperate pattern.

Moreover, Pavlov can exploit CooperateBot (if it defects by accident, it will keep greedily defecting against the hapless CooperateBot, while Tit-For-Tat will not) but still

exerts *some* pressure against DefectBot (defecting against it half the time, compared to Tit-For-Tat's consistent defection.)

The interesting thing is that Pavlov can beat Tit-For-Tat *or* Tit-for-Tat-with-Forgiveness in a wide variety of scenarios.

If there are only Pavlov and Tit-For-Tat bots, Tit-For-Tat has to start out outnumbering Pavlov quite significantly in order to win. The same is true for a population of Pavlov and Tit-For-Tat-With-Forgiveness. It doesn't change if we add in some Cooperators or Defectors either.

Why?

Compared to Tit-For-Tat, Pavlov cooperates better with itself. If two Tit-For-Tat bots are paired, and one of them accidentally defects, they'll be stuck in a mutual defection equilibrium. However, if one Pavlov bot accidentally defects against its clone, we'll see

C/D -> D/D -> C->C

which recovers a mutual-cooperation equilibrium and picks up more points.

Compared to Tit-For-Tat-With-Forgiveness, Pavlov cooperates *worse* with itself (it takes longer to recover from mistakes) but it "exploits" TFTWF's patience better. If Pavlov accidentally defects against TFTWF, the result is

D/C -> D/C -> D/D -> C/D -> D/D -> C/C,

which leaves Pavlov with a net gain of 1 point per turn, (over the first five turns before a cooperative equilibrium) compared to TFTWF's 1/5 point per turn.

If TFTWF accidentally defects against Pavlov, the result is

C/D -> D/C -> D/C -> D/D -> C/D

which cycles eternally (until the next mistake), getting Pavlov an average of 5/4 points per turn, compared to TFTWF's 1 point per turn.

Either way, Pavlov eventually overtakes TFTWF.

If you add enough DefectBots to a mix of Pavlovs and TFT's (and it has to be a large majority of the total population being DefectBots) TFT can win, because it's more resistant against DefectBots than Pavlov is. Pavlov cooperates with DefectBots half the time; TFT never does except by mistake.

Pavlov isn't perfect, but it performs well enough to hold its own in a variety of circumstances. An adapted version of Pavlov won the 2005 iterated game theory tournament.

Why, then, don't we actually talk about it, the way we talk about Tit-For-Tat? If it's true that moral maxims like the Golden Rule emerge out of the fact that Tit-For-Tat is an effective strategy, why aren't there moral maxims that exemplify the Pavlov strategy? Why haven't I even *heard* of Pavlov until now, despite having taken a game theory course once, when *everybody* has heard of Tit-For-Tat and has an intuitive feeling for how it works?

In Wedekind and Milinski's 1996 experiment with human subjects, playing an iterated prisoner's dilemma game, a full 70% of them engaged in Pavlov-like strategies. The human Pavlovians were smarter than a pure Pavlov strategy — they eventually recognized the DefectBots and stopped cooperating with them, while a pure-Pavlov strategy never would — but, just like Pavlov, the humans kept "pushing boundaries" when unopposed.

Moreover, humans basically divided themselves into Pavlovians and Tit-For-Tat-ers; they didn't switch strategies between game conditions where one strategy or another was superior, but just played the same way each time.

In other words, it seems fairly likely not only that Pavlov performs well in computer simulations, but that humans *do* have some intuitive model of Pavlov.

Human players are more likely to use generous Tit-For-Tat strategies rather than Pavlov when they have to play a working-memory game at the same time as they're playing iterated Prisoner's Dilemma. In other words, Pavlov is probably more costly in working memory than generous Tit for Tat.

If you look at all 16 theoretically possible strategies that only have memory of the previous round, and let them evolve, evolutionary dynamics can wind up quite complex and oscillatory.

A population of TFT players will be invaded by more "forgiving" strategies like Pavlov, who in turn can be invaded by DefectBot and other uncooperative strategies, which again can be invaded by TFT, which thrives in high-defection environments. If you track the overall rate of cooperation over time, you get very regular oscillations, though these are quite sensitive to variation in the error and mutation rates and nonperiodic (chaotic) behavior can occur in some regimes.

This is strangely reminiscent of Peter Turchin's theory of secular cycles in history. Periods of peace and prosperity alternate with periods of conflict and poverty; empires rise and fall. Periods of low cooperation happen at the fall of an empire/state/civilization; this enables new empires to rise when a subgroup has better ability to cooperate with itself and fight off its enemies than the surrounding warring peoples; but in peacetime, at the height of an empire, more forgiving and exploitative strategies like Pavlov can emerge, which themselves are vulnerable to the barbaric defectors. This is a vastly simplified story compared to the actual mathematical dynamics *or* the actual history, of course, but it's an illustrative gist.

The big takeaway from learning about evolutionary game theory is that it's genuinely complicated from a player-perspective.

"It's complicated" sometimes functions as a curiosity-stopper; you conclude "more research is needed" instead of looking at the data you have and drawing preliminary conclusions, if you want to protect your intellectual "territory" without putting yourself out of a job.

That isn't the kind of "complexity" I'm talking about here. Chaos in dynamical systems has a specific meaning: the system is so sensitive to initial conditions that even a small measurement error in determining where it starts means you cannot even approximately predict where it will end up.

"Chaos: When the present determines the future, but the approximate present does not approximately determine the future."

Optimal strategy depends sensitively on who else is in the population, how many errors you make, and how likely strategies are to change (or enter or leave). There are a lot of moving parts here.

Norms of Membership for Voluntary Groups

Epistemic Status: Idea Generation

One feature of the internet that we haven't fully adapted to yet is that it's trivial to create voluntary groups for discussion. It's as easy as making a mailing list, group chat, Facebook group, Discord server, Slack channel, etc.

What we *don't* seem to have is a good practical language for talking about *norms* on these mini-groups — what kind of moderation do we use, how do we admit and expel members, what kinds of governance structures do we create.

Maybe this is a minor thing to talk about, but I suspect it has broader impact. In past decades voluntary membership in organizations has declined in the US — we're less likely to be members of the Elks or of churches or bowling leagues — so lots of people who don't have any experience in founding or participating in traditional types of voluntary organizations are now finding themselves engaged in governance *without even knowing that's what they're doing*.

When we do this badly, we get "internet drama." When we do it *really* badly, we get harassment campaigns and calls for regulation/moderation at the corporate or even governmental level. And *that* makes the news. It's not inconceivable that Twitter moderation norms affect international relations, for instance.

It's a traditional observation about 19th century America that Americans were eager joiners of voluntary groups, and that these groups were practice for democratic participation. Political wonks today lament the lack of civic participation and loss of trust in our national and democratic institutions. Now, maybe you've moved on; maybe you're a creature of the 21st century and you're not hoping to restore trust in the institutions of the 20th. But what *will* be the institutions of the future? That may well be affected by what formats and frames for group membership people are used to at the small scale.

It's also relevant for the future of freedom. It's starting to be a common claim that "give people absolute 'free speech' and the results are awful; therefore we need regulation/governance at the corporate or national level." If you're not satisfied with that solution (as I'm not), *you have work to do* — there are a lot of questions to unpack like "what kind of 'freedom', with what implementational details, is the valuable kind?", "if small-scale voluntary organizations can handle some of the functions of the state, how exactly will they work?", "how does one prevent the outcomes that people consider so awful that they want large institutions to step in to govern smaller groups?"

Thinking about, and working on, governance for voluntary organizations (and micro-organizations like online discussion groups) is a laboratory for figuring this stuff out in real time, with fairly low resource investment and risk. That's why I find this stuff fascinating and wish more people did.

The other place to start, of course, is *history*, which I'm not very knowledgeable about, but intend to learn a bit. David Friedman is the historian I'm familiar with who's studied historical governance and legal systems with an eye to potential applicability

to building voluntary governance systems today; I'm interested in hearing about others. (Commenters?)

In the meantime, I want to start generating a (non-exhaustive list) of *types of norms* for group membership, to illustrate the diversity of how groups work and what forms "expectations for members" can take.

We found organizations based on formats and norms that we've seen before. It's useful to have an idea of the *range* of formats that we might encounter, so we don't get anchored on the first format that comes to mind. It's also good to have a vocabulary so we can have higher-quality disagreements about the purpose & nature of the groups we belong to; often disagreements seem to be about policy details but are really about the *overall type* of what we want the group to be.

Civic/Public Norms

- Roughly everybody is welcome to join, and free to do as they like in the space, so long as they obey a fairly minimalist set of ground rules & behavioral expectations that apply to everyone.
- We expect it to be *easy* for most people to follow the ground rules; you have to be *deviant* (really unusually antisocial) to do something egregious enough to get you kicked out or penalized.
- If you dislike someone's behavior but it isn't against the ground rules, you can grumble a bit about it, but you're expected to tolerate it. You'll have to admit things like "well, he has a right to do that."
- Penalties are expected to be predictable, enforced the same way towards all people, and "impartial" (not based on personal relationships). If penalties are enforced unfairly, you're *not* expected to tolerate it — you can question why you're being penalized, and kick up a public stink, and it's even praiseworthy to do so.
- Examples: "rule of law", public parks and libraries, stores and coffeeshops open to the public, town hall meetings

Guest Norms

- The host can invite, or not invite, anyone she chooses, based on her preference. She doesn't have to justify her preferences to anyone. Nobody is entitled to an invitation, and it's very rude to complain about not being invited.
- Guests can also choose to attend or not attend, based on *their* preferences, and they don't have to justify their preferences to anyone either; it's rude to complain or ask for justification when someone declines an invitation.
- Personal relationships and subjective feelings, in particular, are totally legitimate reasons to include or exclude someone.
- The atmosphere within the group is expected to be pleasant for everyone. If you don't want to be asked to leave, you shouldn't do things that will predictably bother people.
- Hosts are expected to be kind and generous to guests; guests are expected to be kind and generous to the host and each other; the host is responsible for enforcing boundaries.
- Criticizing other people at the gathering itself is taboo. You're expected to do your critical/judgmental pruning *outside* the gathering, by deciding whom you will invite or whether you'll attend.
- We don't expect that everyone will be invited to be a guest at every gathering, or that everyone will attend everything they're invited to. It can be prestigious to

be invited to some gatherings, and embarrassing to be asked to leave or passed over when you expected an invitation, but it's normal to just *not be invited* to some things.
- Examples: private parties, invitation-only events, consent ethics for sex

Kaizen Norms

- Members of the group are expected to be committed to an *ideal* of some kind of excellence and to continually strive to reach it.
- Feedback or critique on people's performance is continuous, normal, and not considered inherently rude. It's considered praiseworthy to give high-quality feedback and to accept feedback willingly.
- Kaizen groups may have very specific norms about the *style* or *format* of critique/feedback that's welcome, and it may well be considered rude to give feedback in the wrong style.
- Receiving *some* negative feedback or penalties is normal and not considered a sign of failure or shame. What *is* shameful is responding defensively to negative feedback.
- You can lose membership in the group by getting *too much* negative feedback (in other words, failing to live up to the minimum standards of the group's ideal.) It's *not* expected to be easy for most people to meet these standards; they're challenging by design. The group isn't expected to be "for everyone."
- The feedback and incentive processes are supposed to correlate tightly to the ideal. It's acceptable and even praiseworthy to criticize those processes if they reward and punish people for things unrelated to the ideal.
- Conflict about things unrelated to the ideal isn't taboo, but it's somewhat discouraged as "off-topic" or a "distraction."
- Examples: competitive/meritocratic school and work environments, sports teams, specialized religious communities (e.g. monasteries, rabbinical schools)

Coalition Norms

- The degree to which one is "welcome" in the coalition is the degree to which one is loyal, i.e. contributes resources to the coalition. (Either by committing one's own resources or by driving others to contribute their resources. The latter tends to be more efficient, and hence makes you more "welcome.")
- Membership is a matter of degree, not a hard-and-fast boundary. The more solidly loyal a member you are, the more of the coalition's resources you're entitled to. (Yes, this means membership is defined recursively, like PageRank.)
- People can be penalized or expelled for not contributing enough, or for doing things that have the effect of preventing the coalition gaining resources (like making it harder to recruit new members.)
- Conflict, complaint, and criticism over *the growth of the coalition* (and whether people are contributing enough, or whether they're taking more than their fair share) is acceptable and even praiseworthy; criticisms about other things are discouraged, because they make people less willing to contribute resources or pressure others to do so.
- Membership in the coalition is considered praiseworthy. Non-membership is considered shameful.
- Examples: political coalitions, proselytizing religions

Tribal Norms

- Membership in the group is defined by an immutable, unchosen characteristic, like sex or heredity (or, to a lesser extent, geographic location.) It is difficult to join, leave, or be expelled from the group; you *are* a member as a matter of fact, regardless of what you want or how you behave.
- It's not considered shameful not to be a member of the group; after all, it isn't up to you.
- Since expulsion is difficult, behavioral norms for the group are maintained primarily by persuasion/framing, reward, and punishment, so these play a larger role than they do in voluntary groups. Important norms are framed as *commandments* or simply *how things are*.
- Examples: families, public schools, governments, traditional cultures

Some comparisons-and-contrasts:

Honor and Shame

Kaizen and Guest group norms say that being a member of the group is an honor and comes with high expectations, but that *not* being a member is normal and not especially shameful.

Civic norms say that being a member of the group is normal and easy to attain, but *not* being a member is shameful, because it indicates egregiously bad behavior.

Coalition norms say that being a member is an honor and comes with high expectations *and* that not being a member is shameful. This means that most people will have something to be ashamed of.

Tribal norms say that being a member is not an honor (though it may be a privilege), and that not being a member is no shame.

Protest

Civic and Kaizen norms say that it's okay to protest "unfair" treatment by the governing body. In a Civic context, "fair" means "it's possible for everyone to stay out of trouble by following the rules" — it's okay for rules to be arbitrary, but they should be clear and consistent and not so onerous that most people can't follow them. In a Kaizen context, "fair" means "corresponding to the ideal" — it's okay to "not do things by the book" if that gets you better performance, but it's not okay if you're rewarding bad performance and punishing good.

Guest and Coalition norms say that it's not okay to protest "unfair" treatment; if you get kicked out, arguing can't help you get back in. Offering the decisionmakers something they value might work, though.

In Tribal norms, protest and argument can be either licit or taboo; it depends on the specific tribe and its norms.

Examples of debates that are about what type of group you want to be in:

Asking for "inclusiveness" is usually a bid to make the group more Civic or Coalitional.

Making accusations of "favoritism" is usually a bid to make the group more Civic or Kaizen.

Complaining about “problem members” is usually a bid to make the group more Coalitional, Guest, or Kaizen.

Not A Taxonomy

I don’t think these are *the* definitive types of groups. The idea is to illustrate how you can have different starting assumptions about what kind of thing the group is for. (Is it for achieving a noble goal? For providing a public forum or service open to all? For meeting the needs of its members?)

I suspect these kinds of *aims* are prior to *mechanisms* (things like “what is a bannable offense” or “what incentive systems do we set up”?) Before diving into the technical stuff about the rules of the game, you want to ask what kinds of *outcomes or group dynamics* you want the “game structure” to achieve.

2018 AI Alignment Literature Review and Charity Comparison

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Cross-posted to the EA forum.

Introduction

Like last year and the year before, I've attempted to review the research that has been produced by various organisations working on AI safety, to help potential donors gain a better understanding of the landscape. This is a similar role to that which GiveWell performs for global health charities, and somewhat similar to an securities analyst with regards to possible investments. It appears that once again no-one else has attempted to do this, to my knowledge, so I've once again undertaken the task.

This year I have included several groups not covered in previous years, and read more widely in the literature.

My aim is basically to judge the output of each organisation in 2018 and compare it to their budget. This should give a sense for the organisations' average cost-effectiveness. We can also compare their financial reserves to their 2019 budgets to get a sense of urgency.

Note that this document is quite long, so I encourage you to just read the sections that seem most relevant to your interests, probably the sections about the individual organisations. I do *not* recommend you skip to the conclusions!

I'd like to apologize in advance to everyone doing useful AI Safety work whose contributions I may have overlooked or misconstrued.

Methodological Considerations

Track Records

Judging organisations on their historical output is naturally going to favour more mature organisations. A new startup, whose value all lies in the future, will be disadvantaged. However, I think that this is correct. The newer the organisation, the more funding should come from people with close knowledge. As organisations mature, and have more easily verifiable signals of quality, their funding sources can transition to larger pools of less expert money. This is how it works for startups turning into public companies and I think the same model applies here.

This judgement involves analysing a large number papers relating to Xrisk that were produced during 2018. Hopefully the year-to-year volatility of output is sufficiently low that this is a reasonable metric. I also attempted to include papers during December 2017, to take into account the fact that I'm missing the last month's worth of output from 2017, but I can't be sure I did this successfully.

This article focuses on AI risk work. If you think other causes are important too, your priorities might differ. This particularly affects GCRI, FHI and CSER, who both do a lot of work on other issues.

We focus on papers, rather than outreach or other activities. This is partly because they are much easier to measure; while there has been a large increase in interest in AI safety over the last year, it's hard to work out who to credit for this, and partly because I think progress has to come by persuading AI researchers, which I think comes through technical outreach and publishing good work, not popular/political work.

Politics

My impression is that policy on technical subjects (as opposed to issues that attract strong views from the general population) is generally made by the government and civil servants in consultation with, and being lobbied by, outside experts and interests. Without expert (e.g. top ML researchers at Google, CMU & Baidu) consensus, no useful policy will be enacted. Pushing directly for policy seems if anything likely to hinder expert consensus. Attempts to directly influence the government to regulate AI research seem very adversarial, and risk being pattern-matched to ignorant opposition to GM foods or nuclear power. We don't want the 'us-vs-them' situation, that has occurred with climate change, to happen here. AI researchers who are dismissive of safety law, regarding it as an imposition and encumbrance to be endured or evaded, will probably be harder to convince of the need to voluntarily be extra-safe - especially as the regulations may actually be totally ineffective. The only case I can think of where scientists are relatively happy about punitive safety regulations, nuclear power, is one where many of those initially concerned were scientists themselves. Given this, I actually think policy outreach to the general population is probably negative in expectation.

If you're interested in this I'd recommend you read this blog post (also reviewed below).

Openness

I think there is a strong case to be made that openness in AGI capacity development is bad. As such I do not ascribe any positive value to programs to 'democratize AI' or similar.

One interesting question is how to evaluate non-public research. For a lot of safety research, openness is clearly the best strategy. But what about safety research that has, or potentially has, capabilities implications, or other infohazards? In this case it seems best if the researchers do not publish it. However, this leaves funders in a tough position – how can we judge researchers if we cannot read their work? Maybe instead of doing top secret valuable research they are just slacking off. If we donate to people who say "trust me, it's very important and has to be secret" we risk being taken advantage of by charlatans; but if we refuse to fund, we incentivize people to reveal possible infohazards for the sake of money. (Is it even a good idea to publicise that someone else is doing secret research?)

With regard published research, in general I think it is better for it to be open access, rather than behind journal paywalls, to maximise impact. Reducing this impact by a significant amount in order for the researcher to gain a small amount of prestige does not seem like an efficient way of compensating researchers to me. Thankfully this does not occur much with CS papers as they are all on arXiv, but it is an issue for some strategy papers.

More prosaically, organisations should make sure to upload the research they have published to their website! Having gone to all the trouble of doing useful research it is a shame how many organisations don't take this simple step to significantly increase the reach of their work.

Research Flywheel

My basic model for AI safety success is this:

1. Identify interesting problems
    1. As a byproduct this draws new people into the field through nerd-sniping
2. Solve interesting problems
    1. As a byproduct this draws new people into the field through credibility and prestige
3. Repeat

One advantage of this model is that it produces both object-level work and field growth.

There is also some value in arguing for the importance of the field (e.g. Bostrom's Superintelligence) or addressing criticisms of the field.

Noticeably absent are strategic pieces. In previous years I have found these helpful; however, lately fewer seem to yield incremental updates to my views, so I generally ascribe lower value to these. This does not apply to *technical strategy* pieces, about e.g. whether CIRL or Amplification is a more promising approach.

Near vs Far Safety Research

One approach is to research things that will make contemporary ML systems more safe, because you think AGI will be a natural outgrowth from contemporary ML, and this is the only way to get feedback on your ideas. I think of this approach as being exemplified by Concrete Problems. You might also hope that even if ML ends up leading us into another AI Winter, the near-term solutions will generalize in a useful way, though this is of course hard to judge. To the extent that you endorse this approach, you would probably be more likely to donate to CHAI.

Another approach is to try to reason directly about the sorts of issues that will arise with superintelligent AI, and won't get solved anyway / rendered irrelevant as a natural side effect of ordinary ML research. To the extent that you endorse this approach, you would probably be more likely to donate to MIRI, especially for their Agent Foundations work.

I am not sure how to relatively value these two things.

There are a number of other topics that often get mentioned as AI Safety issues. I generally do not think it is important to support organisations or individuals working on these issues unless there is some direct read-through to AGI safety.

I have heard it argued that we should become experts in these areas in order to gain credibility and influence for the real policy work. However, I am somewhat sceptical of this, as I suspect that as soon as a domain is narrow-AI-solved it will cease to be viewed as AI.

Autonomous Cars

My view is that the localised nature of any tragedies plus the strong incentive alignment mean that private companies will solve this problem by themselves.

Unemployment

While technological advance continually mechanise and replace labour in individual categories, it also opens up new ones. Contemporaneous unemployment has more to do with poor macroeconomic policy and inflexible labour markets than robots. AI strong enough to replace humans in basically every job is basically AGI-complete. At that point we should be worried about survival, and if we solve the alignment problem well enough to prevent extinction we will have likely also solved it well enough to also prevent mass unemployment (or at least the negative effects of such, if you believe the two can be separated).

There has been an increase in interest in a ‘Basic Income’ – an unconditional cash transfer given to all citizens – as a solution to AI-driven unemployment. I think this is a big mistake, and largely motivated reasoning by people who would have supported it anyway. In a Hansonian scenario, all meat-based humanity has is our property rights. If property rights are strong, we will become very rich. If they are weak, and the policy is that every agent gets a fair share, all the wealth will be eaten up as Malthusian EMs massively outnumber physical humans and driving the basic income down to the price of some cycles on AWS.

Bias

The vast majority of discussion in this area seems to consist of people who are annoyed at ML systems are learning based on the data, rather than based on the prejudices/moral views of the writer. While in theory this could be useful for teaching people about the difficulty of the alignment problem, the complexity of human value, etc., in practice I doubt this is the case. This presentation is one of the better I have seen on the subject.

Other Existential Risks

Some of the organisations described below also do work on other existential risks, for example GCRI, FLI and CSER. I am not an expert on other Xrisks so they are hard for me to evaluate work in, but it seems likely that many people who care about AI Alignment will also care about them, so I will mention publications in these areas. The exception is climate change, which is highly non-neglected.

Financial Reserves

Charities like having financial reserves to provide runway, and guarantee that they will be able to keep the lights on for the immediate future. This could be justified if you thought that charities were expensive to create and destroy, and were worried about this occurring by accident due to the whims of donors.

Donors prefer charities to not have too much reserves. Firstly, those reserves are cash that could be being spent on outcomes now, by either the specific charity or others. Valuable future activities by charities are supported by future donations; they do not need to be pre-funded. Additionally, having reserves increases the risk of organisations ‘going rogue’, because they are insulated from the need to convince donors of their value.

As such, in general I do not give full credence to charities saying they need more funding because they want more than a year of runway in the bank. A year’s worth of reserves should provide plenty of time to raise more funding.

It is worth spending a moment thinking about the equilibrium here. If donors target a lower runway number than charities, charities might curtail their activities to allow their reserves to last for longer. At this lower level of activities, donors would then decide a lower level of reserves are necessary, and so on, until eventually the overly conservative charity ends up with a budget of zero, with all the resources instead given to other groups who turn donations into work more promptly. This is allows donor funds to be turned into research more quickly.

I estimated reserves = (cash and grants) / (2019 budget – committed annual funding). In general I think of this as something of a measure of urgency. This is a simpler calculation than many organisations (MIRI, CHAI etc.) shared with me, because I want to be able to compare consistently across organisations. I attempted to compare the amount of reserves different organisations had, but found this rather difficult. Some organisations were

extremely open about their financing (thank you CHAI!). Others were less so. As such these should be considered suggestive only.

Donation Matching

In general I believe that charity-specific donation matching schemes are somewhat dishonest, despite my having provided matching funding for at least one in the past.

Ironically, despite this view being espoused by GiveWell (albeit in 2011), this is basically of OpenPhil's policy of, at least in some cases, artificially limiting their funding to 50% of a charity's need, which some charities argue (though not by OpenPhil themselves that I recall) effectively provides a 1:1 match for outside donors. I think this is bad. In the best case this forces outside donors to step in, imposing marketing costs on the charity and research costs on the donors. In the worst case it leaves valuable projects unfunded.

Obviously cause-neutral donation matching is different and should be exploited. Everyone should max out their corporate matching programs if possible, and things like the annual Facebook Match and the quadratic-voting match were great opportunities.

Poor Quality Research

Partly thanks to the efforts of the community, the field of AI safety is considerably more well respected and funded than was previously the case, which has attracted a lot of new researchers. While generally good, one side effect of this (perhaps combined with the fact that many low-hanging fruits of the insight tree have been plucked) is that a considerable amount of low-quality work has been produced. For example, there are a lot of papers which can be accurately summarized as asserting "just use ML to learn ethics". Furthermore, the conventional peer review system seems to be extremely bad at dealing with this issue.

The standard view here is just to ignore low quality work. This has many advantages, for example 1) it requires little effort, 2) it doesn't annoy people. This conspiracy of silence seems to be the strategy adopted by most scientific fields, except in extreme cases like anti-vaxers.

However, I think there are some downsides to this strategy. A sufficiently large miliu of low-quality work might degrade the reputation of the field, deterring potentially high-quality contributors. While low-quality contributions might help improve Concrete Problems' citation count, they may use up scarce funding.

Moreover, it is not clear to me that 'just ignore it' really generalizes as a community strategy. Perhaps you, enlightened reader, can judge that *"How to solve AI Ethics: Just use RNNs"* is not great. But is it really efficient to require everyone to independently work this out? Furthermore, I suspect that the idea that we can all just ignore the weak stuff is somewhat an example of typical mind fallacy. Several times I have come across people I respect according respect to work I found blatantly rubbish. And several times I have come across people I respect arguing persuasively that work I had previously respected was very bad – but I only learnt they believed this by chance! So I think it is quite possible that many people will waste a lot of time as a result of this strategy, especially if they don't happen to move in the right social circles.

Finally, I will note that the two examples which spring to mind of cases where the EA community has forthrightly criticized people for producing epistemically poor work – namely Intentional Insights and ACE – seem ex post to have been the right thing to do, although in both cases the targets were inside the EA community, rather than vaguely-aligned academics.

Having said all that, I am not a fan of unilateral action, so will largely continue to abide by this non-aggression convention. My only deviation here is to make it explicit – though see this by 80,000 Hours.

The Bay Area

Much of the AI and EA communities, and especially the EA community concerned with AI, is located in the Bay Area, especially Berkeley and San Francisco. This is an extremely expensive place, and is dysfunctional both politically and socially. A few months ago I read a series of stories about abuse in the bay and was struck by how many things I considered abhorrent were in the story merely as background. In general I think the centralization is bad, but if there must be centralization I would prefer it be almost anywhere other than Berkeley. Additionally, I think many funders are geographically myopic, and biased towards funding things in the Bay Area. As such, I have a mild preference towards funding non-Bay-Area projects. If you're interested in this topic I recommend you reading this or this or this.

Organisations and Research

MIRI: The Machine Intelligence Research Institute

MIRI is the largest pure-play AI existential risk group. Based in Berkeley, it focuses on mathematics research that is unlikely to be produced by academics, trying to build the foundations for the development of safe AIs. They were founded by Eliezer Yudkowsky and lead by Nate Soares.

Historically they have been responsible for much of the germination of the field, including advocacy, but are now focused on research. In general they do very 'pure' mathematical work, in comparison to other organisation with more 'applied' ML or strategy focuses. I have historically been impressed with their research.

Their agent foundations work is basically trying to develop the correct way of thinking about agents and learning/decision making by spotting areas where our current models fail and seeking to improve them.

Research

Garrabrant and Demski's Embedded Agency Sequence is a short sequence of blog posts outlining MIRI's thinking about Agent Foundations. It describes the issues about how to reason about agents that are embedded in their environment. I found it to be a very intuitive explanation of many issues that MIRI is working on. However, little of it will be new to someone who has worked through MIRI's previous, less accessible work on the subject.

Yudkowsky and Christiano's Challenges to Christiano's Capability Amplification Proposal discusses Eliezer's objections to Paul's Amplification agenda in back-and-forth blog format. Eliezer has a couple of objections. At a high level, Paul is attempting a more direct solution, working largely within the existing ML framework, vs MIRI's desire to work on things like agent foundations first. Eliezer is concerned that most aggregation/amplification methods do not preserve alignment, and that finding one that does (and building the low level agents) is essentially as hard as solving the alignment problem. Any loss of alignment would be multiplied with every level of amplification. Thirdly, there may be many problems that need sequential work - additional bandwidth does not suffice. Additionally, he objects that Paul's ideas would likely be far too slow, due to the huge amount of human input required. This was an interesting post, but I think could have been more clear. Researchers from OpenAI were also named authors on the paper.

Yudkowsky's The Rocket Alignment Problem is a blog post presenting a Galileo-style dialogue/analogy for why MIRI is taking a seemingly indirect approach to AI Safety. It was enjoyable, but I'm not sure how convincing it would be to outsiders. I guess if you thought a deep understanding of the target domain was never necessary it could provide an existence proof.

Demski's An Untrollable Mathematician Illustrated provides a very accessible explanation to some results about logical induction.

MIRI researchers also appeared as co-authors on:

- Manheim and Garrabrant's Categorizing Variants of Goodheart's Law

Non-disclosure policy

Last month MIRI announced their new policy of nondisclosure-by-default:

[G]oing forward, most results discovered within MIRI will remain internal-only unless there is an explicit decision to release those results, based usually on a specific anticipated safety upside from their release.

This is a significant change from their previous policy. As of circa a year ago my understanding was that MIRI would be doing secret research largely *in addition to* their current research programs, not that *all* their programs would become essentially secret.

At the same time secrecy at MIRI is not entirely new. I'm aware of at least one case from 2010 where they decided not to publish something for similar reasons; as far as I'm aware this thing has never been 'declassified' – indeed perhaps it has been forgotten.

In any case, one consequence of this is that for 2018 MIRI has published essentially nothing. (Exceptions to this are discussed above).

I find this very awkward to deal with.

On the one hand, I do not want people to be pressured into premature disclosure for the sake of funding. This space is sufficiently full of infohazards that secrecy might be necessary, and in its absence researchers might prudently shy away from working on potentially risky things - in the same way that no-one in business sends sensitive information over email any more. MIRI are in exactly the sort of situation that you would expect might give rise to the need for extreme secrecy. If secret research is a necessary step en route to saving the world, it will have to be done by someone, and it is not clear there is anyone much better.

On the other hand, I don't think we can give people money just because they say they are doing good things, because of the risk of abuse. There are many other reasons for not publishing anything. A some simple ones would be "we failed to produce anything publishable" or "it is fun to fool ourselves into thinking we have exciting secrets" or "we are doing bad things and don't want to get caught."

Additionally, by hiding the highest quality work we risk impoverishing the field, making it look unproductive and unattractive to potential new researchers.

One possible solution would be for the research to be done by impeccably deontologically moral people, whose moral code you understand and trust. Unfortunately I do not think this is the case with MIRI. (I also don't think it is the case with many other organisations, so this is not a specific criticism of MIRI, except insomuchas you might have held them to a higher standard than others).

Another possible solution would be for major donors to be insiders, who read the secret stuff and can verify it is worth supporting. If the organisation also wanted to keep small donors the large donors could give their seal of approval; otherwise the organisation could simply decide it did not need them any more. However, if MIRI are adopting this strategy they are keeping it a secret from me! Perhaps this is reassuring about their ability to keep secrets.

Perhaps we hope that MIRI employees would leak information of any wrongdoing, but not leak potential info-hazards?

Finally, I will note that MIRI are have been very generous with their time in attempting to help me understand what they are doing.

Finances

According to MIRI they have around 1.5 years of expenses in reserve, and their 2019 estimated budget is around $4.8m. This does not include the potential purchase of a new office they are considering.

There is *prima facie* counterfactually valid matching funding available from REG's Double Up Drive.

If you wanted to donate to MIRI, here is the relevant web page.

FHI: The Future of Humanity Institute

FHI is a well-established research institute, affiliated with Oxford and led by Nick Bostrom. Compared to the other groups we are reviewing they have a large staff and large budget. As a relatively mature institution they produced a decent amount of research over the last year that we can evaluate. They also do a significant amount of outreach work.

Their research is more varied than MIRI's, including strategic work, work directly addressing the value-learning problem, and corrigibility work.

Research

Armstrong and O'Rourke's 'Indifference' methods for managing agent rewards provides an overview of Stuart's work on Indifference. These are methods that try to prevent agents from manipulating a certain event, or ignore it, or change utility function without trying to fight it. In the paper they lay out extensive formalism and prove some results. Some but not all will be familiar to people who have been following his other work in the area. The key to understanding the why the utility function in the example is defined the way it is, and vulnerable to the problem described in the paper, is that we do not directly observe age - hence the need to base it on wristband status. I found the example a little confusing because it could also be solved by just scaling up the punishment for mis-identification that is caught, in line with Becker's Crime and Punishment: An Economic Approach (1974), but this approach wouldn't work if you didn't know the probabilities ahead of time. Overall I thought this was an excellent paper. Researchers from ANU were also named authors on the paper.

Armstrong and Mindermann's Impossibility of deducing preferences and rationality from human policy argues that you cannot infer human preferences from the actions of people who may be irrational in unknown ways. The basic point is quite trivial - that arbitrary irrationalities can mean that any set of values could have produced the observed actions - but at the same time I hadn't internalised why this would be a big problem for the IRL framework, and in any case it is good to have important things written down. More significant is they also showed that 'simplicity' assumptions will not save us - the 'simplest' solution will (almost definitely) be degenerate. This suggests we do need to 'hard code' some

priors about human values into the AI - they suggest beliefs about truthful human utterances (though of course as speech acts are acts all the same, it seems that some of the same problems occur again at this level of meta). Alternatives (not mentioned in the paper) could be to look to psychology or biology (e.g. Haidt or evolutionary biology). Overall I thought this was an excellent paper.

Armstrong and O'Rourke's Safe Uses of AI Oracles suggests two possible safe Oracle designs. The first takes advantage of Stuart's trademark indifference results to build an oracle whose reward is only based on cases where the output after being automatically verified is deleted, and hence cannot attempt to manipulate humanity. I thought this was clever, and it's nice to see some payoff from the indifference machinery he's been working on, though this Oracle only works for NP-style questions, and assumes the verifier cannot be manipulated - which is a big assumption. The paper also includes a simulation of such an Oracle, showing how the restriction affects performance. The rest of the paper describes the more classic technique of restricting an Oracle to give answers simple enough that we hope they're not potentially manipulative, and frequently re-starting the Oracle. Researchers from ANU were also named authors on the paper.

Dafoe's AI Governance: A Research Agenda is an introduction to the issues faced in AI governance for policy future researchers. It seems to do a good job of this. As lowering barriers to entry is important for new fields, this is potentially a very valuable document if you are highly concerned about the governance side of AI. In particular, it covers policy work to address threats from general artificial intelligence as well as near-term narrow AI issues, which is a major plus to me. In some ways it feels similar to Superintelligence.

Sandberg's Human Extinction from Natural Hazard Events provides a detailed overview of extinction risks from natural events. The paper is both detailed and broad, and is something of an updated version of part of Bostrom and Cirkovic's Global Catastrophic Risks. His conclusion is broadly than man-made risks are significantly larger than natural ones. As with any Anders paper it contains a number of interesting anecdotes - for example I also hadn't realised that people in 1910 were concerned that Halley's Comet might poison the atmosphere!

Schulze and Evans's Active Reinforcement Learning with Monte-Carlo Tree Search provide an algorithm for efficient reinforcement-learning when learning the reward is costly. In most RL designs the agent always sees the reward; however, this would not be the case with CIRL, because the rewards require human input, which is expensive, so we have to ration it. Here Sebastian and Owain produce a new algorithm, BAMCP++ that tries to address this in an efficient way. The paper provides simulations to show the near-optimality of this algorithm in some scenarios vs failure of rivals, and some theoretical considerations for why things like Thompson Sampling would struggle.

Brundage et al.'s The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation is a massively collaborative policy document on the threats posed by narrow AI. Aimed primarily at policymakers, it does a good job of introducing a wide variety of potential threats. However, it does not really cover existential risks at all, so I suspect the main benefit (from our point of view) is that of credibility-building for later. However, I am in general sceptical of politicians' ability to help with AI safety, so I relatively downweight this. But if you were concerned about bad actors using AI to attack, this is a good paper for you. Researchers from OpenAI, CSER were also named authors on the paper.

Bostrom's The Vulnerable World Hypothesis introduces and discusses the idea of worlds that will be destroyed 'by default' when they reach a certain level of technological advancement. He distinguishes between a variety of different cases, like if it is easy for individuals to develop weapons of mass destruction, with intuitive names like 'Type-2b vulnerability', and essentially argues for a global police state (or similar) to reduce the risk. It contained a bunch of interesting anecdotes - for example I hadn't realised what little influence the scientists in the Manhattan Project had on the eventual political uses of nukes. However,

given its origin I actually found this paper didn't add much new. The areas where it could have added - for example, discussing novel ways of using cryptography to enable surveillance without totalitarianism, discussing Value Drift as a form of existential risk that might be impossible to solve without something like this, or the risks of global surveillance itself being an existential risk (as ironically covered in Caplan's chapter of Global Catastrophic Risks) - were left with only cursory discussion. Additionally, given the nature of governments, I do not think that supporting surveillance is a very neglected area.

Lewis et al.'s Information Hazards in Biotechnology discusses issues around dangerous biology research. They provide an overview, including numerous examples of dangerous discoveries and the policies that were used and their merits.

FHI researchers also appeared as co-authors on:

- Carey's Interpreting AI Compute Trends
- Baum et al.'s Long-Term Trajectories of Human Civilization
- Evans et al.'s Predicting Human Deliberative Judgments with Machine Learning
- Shah et al.'s Value Learning Sequence
- Duettmann et al.'s Artificial General Intelligence: Coordination and Great Powers

Finances

OpenPhil awarded FHI $13.4m earlier this year, spread out over 3 years, largely (but not exclusively) to fund AI safety research. Unfortunately the write-up I found on the website was even more minimal than last year's and so is unlikely to be of much assistance to potential donors.

They are currently in the process of moving to a new larger office just west of Oxford.

FHI didn't reply to my emails about donations, and seem to be more limited by talent (though there are problems with this phrase) than by money, so the case for donating here seems weaker. But it could be a good place to work!

If you wanted to donate to them, here is the relevant web page.

CHAI: The Center for Human-Compatible AI

The Center for Human-Compatible AI, founded by Stuart Russell in Berkeley, launched in August 2016. They have produced a lot of interesting work, especially focused around inverse reinforcement learning. They are significantly more applied and ML-focused than MIRI or FHI (who are more 'pure') or CSER or CGRI (who are more strategy-focused). They also do work on non-xrisk related AI issues, which I generally think are less important, but which perhaps have solutions that can be re-used for AGI safety.

Research

Shah's AI Alignment Newsletter is a weekly email of interesting new developments relevant to AI Alignment. It is amazingly detailed. I struggle writing this; I don't know how he keeps on track of it all. Overall I thought is an excellent project.

Mindermann and Shah et al.'s Active Inverse Reward Design turns the reward design process into an interactive one where the agent can 'ask' questions. The idea, as I understand it, is that instead of the programmers creating a one-and-done training reward function which the agent learns about, instead the agent learns from the reward function, is cognizant of its uncertainties (Inverse Reward Design) and then queries the designer in such a way as to

reduce its uncertainty. This seems like exploring the designers value space in the same way that an RL agent explores its environmental space. It seems like a very clever idea to me, though I would have liked to see more examples in the paper.

Hadfield-Menell and Hadfield's Incomplete Contracting and AI alignment analogises the problem of AI alignment with the economics literature on incentive alignment (for humans). The analysis is generally good, and might lead to useful followups, though most of the readthroughs they drew from the principal-agent literature seem like they are already appreciated in the AI safety community. There was some somewhat novel stuff about signalling models, and about Aghion & Tirole's 1997 paper on incomplete contracting that seemed interesting but I didn't really understand or have time to look into. It also did a nice job of pointing out how much the human problem of incomplete contracting is solved by humans being embedded in a moral and social order, and thus able and willing to do what 'obviously' is 'common sense' in unclear situations - a solution which unfortunately seems no FAI-complete for our case. Researchers from OpenAI were also named authors on the paper.

Reddy et al.'s Where Do You Think You're Going?: Inferring Beliefs about Dynamics from Behaviour attempt to infer values from agents with incorrect world-models (pace Armstrong and Mindermann's Impossibility paper). They attempt to avoid the impossibility result by first deducing agent beliefs on a task with known goals, and then using those beliefs to infer goals on a new task. While there might not be any tasks with known human goals, you might hope that there are different areas where human goals and beliefs are more or less well understood, which could be utilised by a related approach. As such I was quite pleased by this paper. They also have a n=12 user trial.

Tucker et al.'s Inverse Reinforcement Learning for Video Games apply an IRL algorithm to an Atari game. Given that proving that alignment-congeniality can be achieved with little loss of efficacy is important for convincing the field, and how much status is applied to success at video games, I think this is a good area to pursue.

Filan's Bottle Caps aren't Optimisers is a short blog post about how to identify agents. It argues this is important because we don't want to accidentally create agents.

Milli et al.'s Model Reconstruction from Model Explanations show it is easier to reconstruct a model with queries about gradients than levels. Asking "what are the partial derivatives at this point?" gives more information, and hence makes it easier to reverse-engineer the model, than asking "what is the output at this point?". The paper is framed as being about the desire by some people to make AI models 'accountable' by making them 'explain' their decisions. I think this is not very important, but it does seem to have some relevance to efficiently reconstructing latent *human* value models. Given that we can only query humans so many times, it is important to make efficient use of these queries. Instead of asking "Would you pull the lever?" many times, instead ask "Which factors would make you more likely to pull the lever?". In some sense asking for partial derivatives seems like n queries (for an n-dimensional space), but given that many (most?) of these are likely to be locally negligible this might be an efficient way to help extract human preferences.

Shah et al.'s Value Learning Sequence is a short sequence of blog posts outlining the specification problem. This is basically how to specify even in theory what we might want to AI to do. It is a nice introduction to many of the issues, like why imitation learning is not enough. Most of what has been published so far is not that new, though apparently it is still ongoing. Researchers from FHI were also contributed posts.

Reddy et al.'s Shared Autonomy via Deep Reinforcement Learning desire an RL system that is intended to operate simultaneously with a human, preventing the human from taking very bad actions, despite not fully understanding the humans goals.

Hadfield-Menell et al.'s Legible Normativity for AI Alignment: The Value of Silly Rules build a RL/Game Theory model for why we might want AI agents to obey and enforce even 'silly' rules. Basically the idea is that fidelity to, and enforcement of, silly rules provides credible

signals that important rules will also be enforced - and their failure to be enforced is also useful information that the group is not strong enough to defend itself so agents can quit earlier. I was a little confused by the conclusion, which suggested that agents would have to learn the difference between silly and non-silly rules. Wouldn't this undermine the signalling value?

CHAI researchers also appeared as co-authors on:

- Ratner et al.'s Simplifying Reward Design through Divide-and-Conquer
- Basu's Do You Want Your Autonomous Car to Drive Like You?
- Liu et al.'s Goal Inference Improves Objective and Perceived Performance in Human-Robot Collaboration
- Wu et al.'s Discrete-Continuous Mixtures in Probabilistic Programming: Generalised Semantics and Inference Algorithms
- Zhou et al.'s Expressive Robot Motion Timing
- Sadigh et al.'s Planning for Autonomous Cars that Leverage Effects on Human Actions

Finances

Based on detailed financials they shared with me I estimate they have around 2 years worth of expenses in reserve (including grants promised but not yet disbursed), with a 2019 budget of around $3m.

If you wanted to donate to them, here is the relevant web page.

CSER: The Center for the Study of Existential Risk

CSER is an existential risk focused group located in Cambridge. Like GCRI they do work on a variety of existential risks, with more of a focus on strategy than FHI, MIRI or CHAI.

Strategic work is inherently tied to outreach, like lobbying the UK government, which is hard to evaluate and assign responsibility for.

In the past I have criticised them for a lack of output. It is possible they had timing issues whereby a substantial amount of work was done in earlier years but only released more recently. In any case they have published more in 2018 than in previous years.

CSER’s researchers seem to select a somewhat eclectic group of research topics, which I worry may reduce their effectiveness.

Research

Liu and Price's Ramsey and Joyce on deliberation and prediction discusses whether agents can have credences on which decision they'll make while they're in the process of deciding. This builds on their previous work in Heart of DARCness. The relevance to AI safety is presumably via MIRI's 5-10 problem, and how to model agents who think about themselves as part of the world, which I didn't appreciate when I read Heart of DARCness. In particular, it discusses agents with sub agents. Having said that, a lot of the paper seemed to rest on terminological distinctions.

Currie's Existential Risk, Creativity & Well-Adapted Science argues that the professionalisation of science encourages 'cautious' research, whereas Xrisk requires more creativity. Essentially it argues that many institutional factors push scientists towards exploitation over exploration. In general I found this convincing, though *pace* Currie I think the small number of Professorships compared to the number of PhDs actually *encourages* risk-taking, as the value out-of-the-money call options increases with volatility. I found his

argument that Xrisk research needing unusually large amounts of creativity not entirely convincing - while I agree that novel threats like AI require this, his example of solar flares seems like the sort of threat that could be addressed in a diligent, rather than genius, fashion. The paper has some pertience for how we fund the Xrisk movement - in particular I think it pulls in favour of many small grants to 'citizen scientists', rather than large grants towards organisations.

Rees's On The Future is a quick-read pop-sci book about the future of humanity. It includes a brief discussion of AI risk, and the section on the risks posed by high-energy physics experiments was new to me. Many topics are discussed only in a very cursory way however, and I agree with Robin's review - the book would have benefited from being proofread by an economist, or simply someone who does not share the author's political views.

Shahar and Shapira's Civ V AI Mod is a mod for Civ V (PC game) that adds superintelligence research into the game. This is the novel publicity effort I alluded to last year. It generated some media attention, which seemed less bad than I expected.

Currie's Introduction: Creativity, Conservatism & the Social Epistemology of Science is a general introduction to some issues about how risk-taking (or not) institutional science is.

Shahar's Mavericks and Lotteries describes various ways in which allocating research funding by lottery, rather than through peer review, might be better. In particular he argues it would make institutional science less conservative. I am sceptical of this, however: the proposals still feature filtering proposals for being "good enough", and in equilibrium the standard for being "good enough" may just rise to where the peer review standard was before. Additionally, I'm not sure I see a very strong link to existential risk - I guess OpenPhil could adopt randomisation? Expecting to reform all of science funding as a path to Xrisk reduction seems *very* indirect.

Currie's Geoengineering Tensions discusses the pros and cons of geoengineering, and the difficulties of doing experiments in the field. It discusses two tensions: firstly the moral hazard risk, and secondly the difficulty of doing the necessary experiments given the conservatism of institutional science.

Adrian Currie edited a ‘special issue’, *Futures of Research in Catastrophic and Existential Risk* which I think is basically a journal of articles they in some sense commissioned or collected. Currie and Ó hÉigeartaigh's Working together to face humanity's greatest threats: Introduction to The Future of Research on Catastrophic and Existential Risk provides an overview of the topics discussed in the edition. In general these are not so much concerned with object-level existential risks as with the meta-work of developing the field. Unfortunately I have not had time to review all the articles it contains that were not authored by CSER researchers, though Jones et al.'s Representation of future generations in United Kingdom policy-making which advocated for a Parliamentory committee for future generations, looks interesting, as one was indeed subsequently created. CSER claim, as seems plausible, that many of these papers would not have counterfactually existed without CSER’s role as a catalyst. The topics discussed include a variety of existential risks.

CSER researchers also appeared as co-authors on the following papers:

- Brundage et al.'s The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation (joint first authorship)
- Cave and Ó hÉigeartaigh's An AI Race for Strategic Advantage: Rhetoric and Risks
- Martinez-Plumed et al's The Facets of Artificial Intelligence: A Framework to Track the Evolution of AI

Finances

Based on some very rough numbers shared with me I estimate they have around 1.25 years worth of expenses in reserve, with an annual budget of around $1m.

If you wanted to donate to them, here is the relevant web page.

GCRI: Global Catastrophic Risks Institute

The Global Catastrophic Risks Institute is a geographically dispersed group run by Seth Baum. They have produced work on a variety of existential risks, including AI and non-AI risks. Within AI they do a lot of work on the strategic landscape, and are very prolific.

They are significantly smaller organisation than most of the others reviewed here, and in 2018 only one of their researchers (Seth) was full time. In the past I have been impressed with their high research output to budget ratio, and that continued this year. At the moment they seem to be somewhat subscale as an organisation - Seth seems to have been responsible for a large majority of their 2018 work - and are trying to grow.

Here is their annual write-up.

Adam Gleave, winner of the 2017 donor lottery, chose to give some money to GCRI; here is his thought process. He was impressed with their nuclear war work (which I'm not qualified to judge), and recommend GCRI focus more on quality and less on quantity, which seems plausible to me. GCRI tell me they are attentive to the issue and have made institutional changes to try to affect change.

GCRI also shared some other considerations with me that I cannot disclose, which may have affected my overall conclusion in addition to the considerations listed above.

Research

Baum et al.'s Long-Term Trajectories of Human Civilization provides an analysis of possible ways the future might go. They discuss four broad trajectories: status quo, catastrophe, technological transformation, and astronomical colonisation. The scope is very broad but the analysis is still quite detailed; it reminds me of Superintelligence a bit. I think this paper has a strong claim to becoming the default reference for the topic. Researchers from FHI, FRI were also named authors on the paper.

Baum's Resilience to Global Catastrophe provides a brief introduction to ideas around resilience to disasters. The points it made seem true, but are obviously more applicable to non-AGI based threats that leave more scope for recovery.

Baum's Uncertain Human Consequences in Asteroid Risk Analysis and the Global Catastrophe Threshold discusses the consequences of Asteroid impact. He reviews some of the literature, and discusses the idea of important thresholds for impact. One idea I hadn't come across before was the risk that an asteroid impact might be mistaken as a nuclear attack and cause a war - an interesting risk because all we need to do to avoid it is see the asteroid coming. However, I'm not an expert in the field, so struggle to judge how novel or incremental the paper is.

Baum and Barrett's A Model for the Impacts of Nuclear War goes through the various impacts of nuclear war. It seems diligent and useful for future researchers or policymakers as a reference, though it is not my area of expertise.

Baum et al.'s A Model for the Probability of Nuclear War describes and decomposes the many possible routes to nuclear war. It also contains an interesting and extensive database of 'near-miss' scenarios.

Baum's Superintelligence Skepticism as a Political Tool discusses the risk of motivated scepticism about AI risks in order to protect funding for researchers and avoid regulation for corporation. This seems like a plausible risk, though we should be careful attributing disingenuous motivations to opponents - though it is certainly true that the AI safety community seems to be the target of more misinformation than you might expect. I think the paper could might have benefitted from contrasting this with the risks of regulatory capture, which seem to operate in the other direction. Without doing so the political discussion was somewhat partisan - in both misinformation papers virtually all the examples bad actor were right wing groups, though perhaps most readers might find this is agreeable!

Baum's Countering Superintelligence Misinformation discusses ways to improve debate around superintelligence through countering misinformation. These are mainly different forms of education, plus criticism of people for saying false things. I thought that the sections about ways of addressing misinformation once it exists were generally quite sophisticated, though I am sceptical of some of them as I don't think AI safety is very amenable to popular or state pressure.

Baum et al.'s Modelling and Interpreting Expert Disagreement about Artificial Intelligence attempts to put numbers of Bostrom and Goertzel's credences for various AI risk factors and compare. They try to break down the disagreement into three statements, interpret the two thinkers' statements as probabilities for those statements, and then assign their own probability for which thinker is correct. I'm a bit confused by the last step - it seems that by doing so you're basically ensuring the output will be equal to your own credence (by the law of total probability).

Umbrello and Baum's Evaluating Future nanotechnology: The Net Societal Impacts of Atomically Precise Manufacturing discusses the possible impacts of nanotechnology on society. Most of the discussion is quite broad, and could apply to economic growth in general. I was surprised how little value the authors assigned to greatly increasing the wealth of humanity.

Finances

GCRI spent around $140k in 2018, and are aiming to raise $1.5m to cover the next three years, for a target annual budget of ~$500k. This would allow them to employ their (3) key staff full time and have some money for additional hiring.

This large jump makes it a little hard to calculate runway in a comparable fashion to other organisations. They currently have around $280k, having recently received a $250k donation. But is it unfair to include this donation, given they received it subsequently to some other organisations telling me about their finance? All organisations should look progressively better funded as giving season goes on!

In any case it seems relatively clear that they have been and probably continue to be at the moment more funding constrained than most other organisations. The part-time nature of many of their staff makes their cost structure more variable and less fixed, suggesting this limited runway is less of an existential threat than it would be at some other organisations – they’re not about to disband - though clearly this is still undesirable.

It seems credible that more funding would allow them to hire their researchers full time, which seems like a relatively low-risk method of scaling. If they can preserve their current productivity this could be valuable, though my impression is many small organisations become less productive as they scale, as high initial productivity may be due to founder effects that revert to the mean.

If you want to donate to GCRI, here is the relevant web page.

GPI: The Global Priorities Institute

The Global Priorities Institute is an academic research institute, lead by Hilary Greaves, working on EA philosophy within Oxford. I think of their mission as attempting to provide a home so that high quality academics can have a respectable academic career while working on the most important issues. At the moment they mainly employee philosophers, but they tell me they are planning to hire more economists in the future.

They are relatively new but many of their employees are extremely impressive and their working papers (linked on the EA forum, not on their main website) seem very good to me. At this stage I wouldn't expect them to have reached run-rate productivity, so would expect this to increase in 2019.

They shared with me abstracts of a number of papers and so on they were working on which seemed interesting and useful. As academic philosophy goes it is very tightly focused on important, decision-relevant issues - however it is not directly AI Safety work.

They allow their employees to spend 50% (!) of their time working on non-GPI projects, to help attract talent. However, the Trammell paper mentioned below was one of these projects, and I thought it was very good, so maybe in practice this does not represent a halving of their cost-effectiveness.

CEA are also spawning a new independant Forethought Foundation for Global Priorities Research, which seems to be very similar to GPI except not part of Oxford.

Research

Mogensen's Long-termism for risk averse altruists argues that risk-averse should make altruists *more*, not *less*, interested in preventing existential risks. This is basically for the same reason that risk aversion causes people to buy insurance. You should be risk averse in outcomes, not in the direct impacts of your actions. This argument is totally obvious now but I'd never heard anyone mention it until two months ago, which suggests it is real progress. Overall I thought this was an excellent paper.

Trammell's Fixed-Point Solutions to the Regress Problem in Normative Uncertainty argues that we can avoid infinite metaethical regress through fixed-point results. This seems like an alternative to Will's work on Moral Uncertainty in some senses. Basically the idea is that if the 'choiceworthiness' of different theories are cardinal at every level in their hierarchy, we can prove a unique fixed point. This is significant to the extent we think that AIs are going to have to learn how to do moral reasoning, perhaps without the aid of humans' convenient "just don't think about it" hack. It's also in some ways a nice response to this SlateStarCodex article.

Finances

They have a 2019 budget of around $1.5m dollars, and shared with me a number of examples of types of people they might like to hire in the future, with additional funding.

Apparently Oxford University rules mean that all their hires have to be pre-funded for their entire duration of their (4-5 year) contract.

If you wanted to donate to GPI, here is the link.

ANU: Australian National University

Australian National University has produced a surprisingly large number of relevant papers and researchers over time.

Research

Everitt et al.'s AGI Safety Literature Review AGI Safety Literature Review - I was glad to see someone else attempting to do the same thing I have! Readers of this article might enjoy reading it, as it has much the same purpose. For academics new to the field it could function as a useful overview, introducing but not really arguing for many important points. It's main value probably comes from one-sentence descriptions of a large number of papers, which could be a useful launching point for research. Literature reviews can also help raise the status of the field. However, it is less likely to add much new insight to those familiar with the field, as it doesn't really engage with any of the arguments in depth.

Everitt et al.'s Reinforcement Learning with a Corrupted Reward Channel examines how noisy reward inputs can drastically degrade reinforcement learner performance, and some possible solutions. Unsurprisingly, CIRL features as a possible solution. It's also nice to see ANU-Deepmind collaboration. This paper was actually written last year, but I mention it here for completeness as I think I missed it previously; I haven't reviewed it in depth. Researchers from Deepmind were also named authors on the paper.

EDIT: one paper redacted on author request, pending improved second version.

ANU researchers were also named as co-authors on the following papers:

- Leike et al.'s AI Safety Gridworlds
- Armstrong and O'Rourke's 'Indifference' methods for managing agent rewards
- Armstrong and O'Rourke's Safe Uses of AI Oracles

Finances

Given their position as part of ANU I suspect it would be difficult for individual donations to appreciably support their work. Additionally, one of their top researchers, Tom Everitt, has now joined Deepmind.

BERI: The Berkeley Existential Risk Initiative

EDIT: After publishing, the Berkeley Existential Risk Initiative requested I remove this section. As a professional courtesy I am reluctantly complying, and rescind any suggestion that BERI may be a good place to donate. I apologize for any inconvenience caused to readers.

Ought

Ought is a San Francisco based non-profit are researching the viability of automating human-like cognition. The focus is on approaches that are "scalable" in the sense that better ML or more compute makes them increasingly helpful for supporting and automating deliberation without requiring additional data generated by humans. The idea, as with amplification, is that we can achieve safety guarantees by making agents that reason in individual explicit and comprehensible steps, iterated many times over, as opposed to the dominant more black-box approaches of mainstream ML. Ought does research on computing paradigms that support this approach and experiments with human participants to determine whether this class of approaches is promising. But I admit I understand what they do less well than with other groups.

Their work doesn’t fit neatly into the model of the above groups - they’re not focused on publishing research papers, at least at the moment. Partly as a result of this, and as a new group, I feel like I don’t have quite as good a grasp on exactly their status as with other groups - which is of course primarily a fact about my epistemic state, rather than them.

Research

Stuhlmüller's Factored Cognition outlines the ideas behind their implementation of Christiano-style amplification. They built a web app where people take questions and recursively break them down into simpler questions that can be solved in isolation. At the moment this is for humans, to try to test whether this sort of amplification of distillation and answering could work. It seems like they have put a fair bit of thought into the ontology.

Evans et al.'s Predicting Human Deliberative Judgments with Machine Learning attempts to make progress on building ML systems remain well-calibrated (i.e. the system "knows what it knows") in AI-complete settings (i.e. in settings where current ML algorithms can’t possibly do well on every possible input). To do this they collect a dataset of human judgements on complex issues (weird fermi estimations and political fact-checking) and then look at how people's estimates for these questions changed as they were allowed more time. This is important because someone's rapid judgement of an issue is evidence as to what their eventual slow judgement will be. In some cases you might be able to predict that there is no need to give the human more time; their 30 second answer is probably good enough. This could be useful if you are trying to produce a large training set of judgements about complex topics. I also admire the author's honesty that the results of their ML system was less good than they expected. They also discussed problems with their dataset; this was definitely my experience when trying to use the site. Researchers from FHI were also named authors on the paper.

Finances

Based on numbers they shared with me I estimate they have around half a year’s worth of expenses in reserve, with an projected 2019 budget of around $1m.

Additional funding sounds like it would go towards reserves and additional researchers and programers, including a web developer, probably mainly continuing working on Factored Cognition.

Ought ask me to point out that they have applied for an OpenPhil grant renewal but expect to still have room for more funding afterwards.

AI Impacts

AI Impacts is a small Berkeley-based group that does high-level strategy work, especially on AI timelines, somewhat associated with MIRI.

Adam Gleave, winner of the 2017 donor lottery, chose to give some money to AI Impacts; here is his thought process. He was impressed with their work, although sceptical of their ability to scale.

Research

Carey wrote Interpreting AI Compute Trends, which argues that cutting-edge ML research projects have been getting dramatically more expensive. So much so that the trend will have to stop, suggesting that (one driver of) AI progress will slow down over the next 3.5-10 years. Additionally, he points out that we are also nearing the processing capacity (though not

scanning capacity) required to model human brains. (Note that this was a guest post by Ryan, who works for FHI)

Grace's Likelihood of discontinuous progress around the development of AGI discusses a 11 different arguments for AGI to have a discontinuous impact, and finds them generally unconvincing. This is important from a strategy point of view because it suggests we should have more time to see AGI coming, potentially also making it clear to sceptics. Overall I found the article clear and generally convincing.

McCaslin's Transmitting fibers in the brain: Total length and distribution of lengths analyses how much neural fibre there is in the human brain, and the distribution of long vs short. My understanding is this is related to how many neurons in human brains are dedicated to moving information around, rather than computation, which might be important because it is an additional form of capacity that is often overlooked when people talk about FLOPS and MIPS, and so might affect your estimates for when we have enough hardware capacity for neuromorphic AI. However, I might be misunderstanding, as I found the motivation a little unclear.

Grace's Human Level Hardware Timeline attempts to estimate how long until we have human-level hardware at human cost. Largely based on earlier work, they estimate "a 30% chance we are already past human-level hardware (at human cost), a 45% chance it occurs by 2040, and a 25% chance it occurs later."

They have gathered a collection of examples of discontinuous progress in history, to attempt to produce something of a reference class for how likely this is with AGI - see for example the Burj Khalifa, the Eiffel Tower, rockets. It would be nice to see how many possible examples they investigated and found were not discontinuous.

Finances

According to numbers they shared with me, AI Impacts spent around $90k in 2018 on two part-time employees. In 2019 they plan to significantly increase, to ~$360k and hire multiple new workers. They have just over $400k in current funding, suggesting a bit over a year of runway at this elevated rate, or many years at their 2018 rate.

Similar to GCRI, there is some risks that small groups may have a high productivity due to founder effects, and this might revert to the mean as they scale.

MIRI seems to administer their finances on their behalf; donations can be made here.

Open AI

OpenAI is a San Francisco based AGI startup charity, with a large focus on safety. It was founded in 2015 with money largely from Elon Musk.

Research

Christiano et al. 's Supervising Strong Learners by Amplifying Weak Experts lays out Paul's amplification ideas in a paper - or at least one implementation of them. Basically the idea is that there are many problems where it is too expensive to produce training signals directly, so we will do so indirectly. We do this by iteratively breaking up the task into sub-tasks, using the agent to help with each sub-task, and then training the agent on the human's overall judgement, aided by the agent's output on the subtasks. Hopefully as the agent becomes strong it also gets better at the subtasks, improving the training set further. We also train a second agent to be able to predict good subtasks to go for, and to predict how the human will use the outputs from the subtasks. I'm not sure I understand why we don't train the

agent on its performance of the subtasks (except that it is expensive to evaluate there?) I think the paper might have been a bit clearer if it had included an example of the algorithm being used in practice with a human in the loop, rather than purely algorithmic examples. Hopefully this will come in the future. Nonetheless this was clearly a very important paper. Overall I thought this was an excellent paper.

Irving, Christiano and Amodei's AI Safety via Debate explore adversarial 'debate' between two or more advanced agents, competing to be judged the most helpful by a trusted but limited agent. This is very clever. It's an extension of the grand Christiano project of trying to devise ways of amplifying simple, trusted agents (like humans) into more powerful ones - designing a system that takes advantage of our trust in the weak agent to ensure compliance in the stronger. Imagine we basically have a courtroom situation, where two highly advanced legal teams, with vast amounts of legal and forensic expertise, try to convince a simple but trusted agent (the jury) that they're in the right. Each side is trying to make its 'arguments' as simple as possible, and point out the flaws in the other's. As long as refuting lies is easy relative to lying, honesty should be the best strategy... so agents constrained in this way will be honest, and not even try dishonesty! Like a courtroom where both legal teams decide to represent the same side. The paper contains some nice examples, including AlphaGo as an analogy and a neat MNIST simulation, and an interactive website. Overall I thought this was an excellent paper.

The OpenAI Charter is their statement of values with regard AGI research. It seems to contain the things you would want it to: benefit of all, fiduciary duty to humanity. Most interestingly, it also includes " if a value-aligned, safety-conscious project comes close to building AGI before we do, we commit to stop competing with and start assisting this project. We will work out specifics in case-by-case agreements, but a typical triggering condition might be “a better-than-even chance of success in the next two years”", a clause which seems very sensible. Finally, it also notes that, like MIRI, they anticipate reducing their conventional publishing.

Amodei and Hernandez's AI and Compute attempts to quantify the computing power used for recent major AI developments like ResNets and AlphaGo. They find it has been doubling approximately every 3-4 months, dramatically faster than you would expect from Moore’s law – especially if you had been reading articles about the end of Moore’s law! This is due to a combination of the move to specialist hardware (initially GPUs, and now AI ASICs) and companies simply spending a lot more dollars. This is not a theory paper, but has direct relevance for timeline prediction and strategy that depends on whether or not there will be a hardware overhang.

Christiano's Universality and Security Amplification describes how Amplification hopes to enhance security by protecting against adversarial inputs (attacks). The hope is that the process of breaking down queries into sub-queries that is at the heart of the Amplification idea can leave us with queries of sufficiently low complexity that they are human-secure. I'm not sure I really understood what this posts adds to others in Paul's arsenal, mainly because I haven’t been following these as closely as perhaps I should have.

Researchers from OpenAI were also named as coauthors on:

- Hadfield-Menell and Hadfield's Incomplete Contracting and AI alignment
- Brundage et al.'s The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation
- Yudkowsky and Christiano's Challenges to Christiano's Capability Amplification Proposal

Finances

Given the strong funding situation at OpenAI, as well as their safety team’s position within the larger organisations, I think it would be difficult for individual donations to appreciably support their work. However it could be an excellent place to apply to work.

Google Deepmind

As well as being arguably the most advanced AI research shop in the world, Google's London-based Deepmind has a very sophisticated AI Safety team.

Research

Leike et al.'s AI Safety Gridworlds introduces an open-source set of environments for testing ML algorithms for safetyness. Progress in ML has been considerable aided by the availability of common toolsets like MNIST or the Atari games. Here the Deepmind safety team have produced a set of environments designed to test algorithms ability to avoid a number of safety-related failure modes, like Interruptibility, Side Effects, Distributional Shifts and Reward Hacking. This hopefully not only makes such testing more accessible, it also makes these issues more concrete. Ideally it would shift the overton window: maybe one day it will be weird to read an ML paper that does not contain a section describing performance on the Deepmind Gridworlds. This is clearly not a panacea; it is easily to 'fake' passing the test by giving the agent information it shouldn't have, it is better to prove safety results than tack them on, and there is always a risk of Goodhearting. But this seems to me to be clearly a significant step forward. My enthusiasm is only slightly tempered by the fact that only one paper published in the following year citing the paper made use of the Gridworld suite, though Alex Turner's excellent post on Impact measures did as well. Overall I thought this was an excellent paper. Researchers from ANU were also named authors on the paper.

Krakovna's Specification Gaming Examples in AI provides a collection of different cases where agents have optimised their reward function in surprising/undesirable fashion. The spreadsheet of 45 examples might have some research value, but my guess is most of the value is as evidence of the problem.

Krakovna et al.'s Measuring and avoiding side effects using relative reachability invents a new way of defining 'impact', which is important if you want to minimise it, based on how many states' achievability are affected. Essentially it takes some the set of possible states, and then punishes the agent for reducing the attainability of these states. The post also includes a few simulations in the AI Gridworld.

Leike et al.'s Scalable agent alignment via reward modeling: a research direction outlines the Deepmind agenda for bootstrapping human evaluations to provide feedback for RL agents. Similar in some ways to the Christiano project, the idea is that your main RL agent simultaneously learns its reward function and about the world. The human's ability to provide good reward feedback is improved by training smaller agents who help him judge which rewards to provide. The paper goes into a number of potential familiar problems, and potential avenues of attack on those issues. I think the news here is more that the Deepmind (Safety) team is focusing on this, rather than the core ideas themselves. The paper also reviews a lot of related work.

Gasparik et al.'s Safety-first AI for autonomous data centre cooling and industrial control describes the mainly safety measures Google put in place to ensure their ML-driven datacenter cooling system didn't go wrong.

Ibarz et al.'s Reward Learning from Human Preferences and Demonstrations in Atari combines RL and IRL as two different sources of information for the agent. If you think both ideas have some value, it makes sense that combining them further improves performance.

Leibo et al.'s Psychlab: A Psychology Laboratory for Deep Reinforcement Learning Agents creates an environment for comparing humans and RL agents on the same tasks. Given the goal of getting AI agents to behave in ways humans approve of is closely related to the goal of making them behave like humans, this seems like a potentially useful tool.

Ortega et al.'s Building safe artificial intelligence: specification, robustness and assurance provide an introduction to various problems in AI Safety. The content is unlikely to be new to readers here; it is significant insomuchas it represents a summary of the (worthwhile) priorities of Deepmind('s safety team). They decompose the issue into specification, robustness and assurance.

Researcher's from Deepmind were also named as coauthors on the following papers:

- Everitt et al.'s Reinforcement Learning with a Corrupted Reward Channel
- Rainforth et al.'s Tighter Variational Bounds are Not Necessarily Better
- Ngo and Pace's Some cruxes on impactful alternatives to AI policy work

Finances

Being part of Google, I think it would be difficult for individual donors to directly support their work. However it could be an excellent place to apply to work.

Google Brain

Google Brain is Google's other highly successful AI research group.

Research

Kurakin et al. wrote Adversarial Attacks and Defences Competition, which summarises the NIPS 2017 competition on Adversarial Attacks, including many of the strategies used. If you're not familiar with the area this could be a good introduction.

Brown and Olsson wrote Introducing the Unrestricted Adversarial Examples Challenge, which launches a new 2-sided challenge, for designing systems resistant to adversarial examples, and then finding adversarial examples. The difference here is in allowing a much broader class of adversarial examples, rather than just small perturbations. This seems like a significantly more important class, so it is good they are attempting to move the field in this direction.

Gilmer et al. wrote Motivating the Rules of the Game for Adversarial Example Research, which argue that the adversarial example literature has overly-focused on a narrow class of imperceptibly-changed images. In most realistic cases the adversary has a much wider scope of possible attacks. Importantly for us, the general question is also more similar to the sorts of distributional shift issues that are likely to arise with AGI. To the extent this paper helps push researchers towards more relevant research it seems quite good.

Finances

Being part of Google, I think it would be difficult for individual donors to directly support their work. However it could be an excellent place to apply to work.

EAF / FRI: The Effective Altruism Foundation / Foundational Research Institute

EAF is a German/Swiss group effective altruist group, lead by Jonas Vollmer and Stefan Torges, that undertakes a number of activities. They do research on a number of fundamental long-term issues, many related how to reduce the risks of very bad AGI outcomes, published through the Foundational Research Institute (FRI). Their website

suggests that FRI and WAS (Wild Animal Suffering) are two equal sub-organisations, but apparently this is not the case - essentially everything EAF does is FRI now, and they just let WAS use their legal entity and donation interface. EAF also have Raising for Effective Giving, which encourages professional poker players to donate to effective charities, including MIRI.

In the past they have been rather negative utilitarian, which I have always viewed as an absurd and potentially dangerous doctrine. If you are interested in the subject I recommend Toby Ord's piece on the subject. However, they have produced research on why it is good to cooperate with other value systems, making me somewhat less worried.

Research

Oesterheld's Approval-directed agency and the decision theory of Newcomb-like problems analyses which decision theories are instantiated by RL agents. The paper analyses the structure of RL agents of various kinds and maps them mathematically to either Evidential or Causal Decision theory. Given how much we discuss decision theory it is surprising in retrospect that no-one (to my knowledge) had previously looked to see which ones our RL agents were actually instantiating. As such I found this an interesting paper.

Baumann's Using Surrogate Goals to Deflect Threats discusses using a decoy utility function component as to protect against threats. The idea is that agents run the risk of counter-optimisation at the hands of an extortionist, but this could be protected against by redefining their utility function to add a pointless secondary goal (like avoiding the creation of a certain dimensioned platinum sphere). An opponent would find it easier to extort the agent by negatively optimising the surrogate goal. This doesn't prevent the agent from giving in to the threats, but it does reduce the damage if the attacker has to follow-through on their threat. The paper discusses many additional details, including the multi-agent case, and the interaction between this and other defence mechanisms. My understanding is that they and Eliezer both (independently?) came up with this idea. One thing I didn't quite understand is the notional of attacker-hostile surrogates - surely they would just be ignored?

Sotala and Gloor's Superintelligence as a Cause or Cure for Risks of Astronomical Suffering is a review article for the various ways the future might contain a lot of suffering. It does a good job of going through possibilities, though I felt it was overly focused on suffering as a bad outcome - there are many other bad things too!

Sotala's Shaping economic incentives for collaborative AGI argues that encouraging collaborative norms in AI with regard narrow AI will encourage those norms in the future for AGI due to cultural lock-in. Unfortunately it is not clear how to go about doing this. Researchers from FHI, were also named authors on the paper.

Finances

Based on their blog post, they currently have around a year and a half's worth of reserves, with a 2019 budget of $925,000.

As EAF have in the past worked on a variety of cause areas, donors might worry about fungibility. EAF tell me that they are now basically entirely focused on AI related work, and that WAS research is funded by specifically allocated donations, which would imply this is not a concern, though I note that several WAS people are still listed on their team page.

Readers who want to donate to EAF/FRI can do so here.

Foresight Institute

The Foresight Institute is a Palo-Alto based group focusing on AI and nanotechnology. Originally founded in 1986 (!), they seem to have been somewhat re-invigorated recently by

Allison Duettmann. Unfortunately I haven't had time to review them in detail.

A large part of their activity seems to be in organising 'salon' discussion / workshop events.

Duettmann et al.'s Artificial General Intelligence: Coordination and Great Powers summarises the discussion at the 2018 Foresight Institute Strategy Meeting on AGI. Researchers from FHI and FLI were also named authors on the paper.

Readers who want to donate to Foresight can do so here.

FLI: The Future of Life Institute

The Future of Life Institute was founded to do outreach, including run the Puerto Rico conference. Elon Musk donated $10m for the organisation to re-distribute; given the size of the donation it has rightfully come to somewhat dominate their activity.

In 2018 they ran a second grantmaking round, giving $2m split between 10 different people. These grants were more focused on AGI than the previous round, which included a large number of narrow AI projects. In general the grants went to university professors. They have now awarded most of the $10m.

Unfortunately I haven't had time to review them in detail.

Readers who want to donate to FLI can do so here.

Median Group

The Median Group is a new group for research on global catastrophic risks, with researchers from MIRI, OpenPhil and Numerai. As a new group they lack the sort of track record that would make them easily amenable to analysis. Current projects they're working on include AI timelines, forest fires, and climate change impacts on geopolitics.

I don't know that much about them because the contact email listed on the website does not work.

Research

Taylor et al. wrote Insight-based AI timeline model, which made an insight-based model for the time to AGI. They first produced a list of important insights that have (plausibly) contributed towards AGI. Surprisingly, they find there has been a roughly constant rate of insight production since 1945. They then model time-to-AGI using a pareto distribution for the number of insights required. This is a novel (to me, at least) method that I liked.

Convergence Analysis

Convergence Analysis is a new group, lead by Justin Shovelain, aiming to do strategic work. They are too new to have any track record.

Other Research

I would like to emphasis that there is a lot of research I didn't have time to review, especially in this section, as I focused on reading organisation-donation-relevant pieces. For example,

Kosoy's The Learning-Theoretic AI Alignment Research Agenda seems like a worthy contribution.

Papers

Lipton and Steinhardt's Troubling Trends in Machine Learning Scholarship critiques a number of developments in the ML literature that they think are bad. Basically, they argue that a lot of papers obfuscate explanation vs speculation, obscure the true source of improvement in their papers (often just hyper-parameter tuning), use maths to impress rather than clarify, and use common english words for complex terms, thereby smuggling in unnecessary connotations. It's unclear to me, however, to what extent these issues retard progress on safety vs capabilities. I guess to the extent that safety requires clear understanding, whereas capabilities can be achieved in a more messy fashion, these trends are bad and should be pushed back ok.

Jilk's Conceptual-Linguistic Superintelligence discusses the need for AGI to have a conceptual-linguistic facility. Contra recent AI developments - e.g. AlphaZero does not have a linguistic ability - he argues that AIs will need linguistic ability to understand much of the human world. He also discusses the difficulties that Rice's theorem imposes on AI self-improvement, though this has been well discussed before.

Cave and Ó hÉigeartaigh's An AI Race for Strategic Advantage: Rhetoric and Risks argues that framing AI development as a 'race', or an 'arms race', is bad. Much of their reasoning is not new, and was previously published by e.g. Baum's On the Promotion of Safe and Socially Beneficial Artificial Intelligence. Instead I think of the target audience here as being policymakers and other AI researchers: this is a paper aiming to influence global strategy, not research EA strategy. Having said that, their discussion of why we should actively confront AI race rhetoric, rather than trying to simply avoid it, was novel, at least to me. It also apparently won best paper at the AAAI/ACM conference on Artificial Intelligence, Ethics, and Society. Researchers from CSER were also named authors on the paper.

Liu et al.'s A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View reviews security threats to contemporary ML systems. This is basically addresses the concerns raised in Amodei et al.'s Concrete Problems about Distributional Shifts between training and test data, and how to ensure robustness.

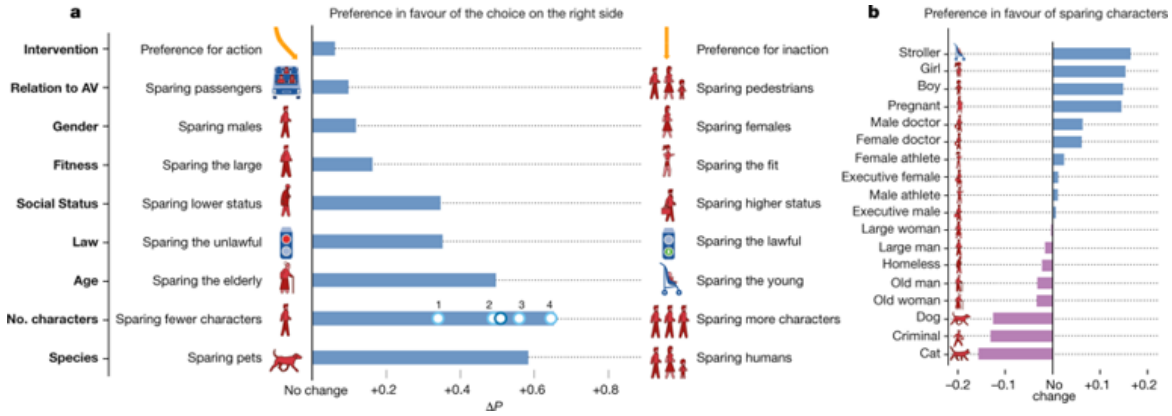
Sarma and Hay's Robust Computer Algebra, Theorem Proving, and Oracle AI discuss computer algorithm systems as potentially important classes of Oracles, and try to provide concrete safety-related work that could be done. Their overview of Question-Answering-Systems, Computer-Algebra-Systems and Interactive-Theorem-Provers was interesting to me, as I didn't have much familiarity thereof. They argue that CAS use heuristics that lead to invalid inferences sometimes, while ITPs are very inefficient, and suggest projects to help integrate the two, to produce more reliable math oracles. I think of this paper as being a bit like a specialised version of Amodei et al's Concrete Problems, but the connection between the projects here and the end goal of FAI is a little harder for me to grasp. Additionally, the paper seems to have been in development since 2013?

Manheim and Garrabrant's Categorizing Variants of Goodheart's Law classifies different types of situations where a proxy measures ceases to be a good proxy when you start relying on it. This is clearly an important topic for AI safety, insomuch as we are hoping to design AIs that will not fall victim to it. The paper provides a nice disambiguation of different kinds of situation, bringing conceptual clarity even if it's not a deep mathematical result. Researchers from MIRI were also named authors on the paper.

Ngo and Pace's Some cruxes on impactful alternatives to AI policy work discuss the advantages and disadvantages of AI policy work. They try to find the 'crux' of their disagreement - the small number of statements they disagree about which determine which

side of the issue they come down on. Researchers from Deepmind were also named authors on the paper.

Awad et al.'s The Moral Machine Experiment did a massive online interactive survey of 35 *million* people to determine their moral preferences with regard autonomous cars. They found that people prefer: saving more people rather than fewer; saving humans over animals; saving young (including unborn children) over old; lawful people over criminals; executives over homeless; fit over fat; females over males; and pedestrians over passengers. I thought this was very interesting, and applaud them for actually looking for people's moral intuitions, rather than just substituting the values of the programmers/politicians. They also analyse how these values differ between cultures. Overall I thought this was an excellent paper.



Green's Ethical Reflections on Artificial Intelligence reviews various ethical issues about AI from a christian perspective. Given the dominance of utilitarian thinking on the subject, it was nice to see an explicitly Christian contribution that displayed familiarity with the literature, with safety as #1 and #3 on the list of issues. "therefore it must be the paramount goal of ethics to maintain human survival.'

Eth's The Technological Landscape Affecting Artificial General Intelligence and the Importance of Nanoscale Neural Probes presents arguments for favouring whole-brain-emulation as a pathway to human-level AI over de novo AGI, and suggests that nanoscale neural probe research could be a good way to differentially advance WBE vs merely human-inspired Neuromorphic AGI. The paper builds on a lot of arguments in Bostrom's Superintelligence. It seems clear that neuromorphic AGI is undesirable - the question is between de novo and WBE, which unfortunately seem to have neuromorphic 'in between' them from a technological requirement point of view. Daniel presents some good arguments for the relative safety of WBE (some of which were already in Bostrom), for example that WBEs would help provide training data from de novo AGI, though I was sceptical of the idea that the identity of the first WBEs would be determined by public debate. An especially good point was that even if nanoscale neural probes accelerate neuromorphic almost as much as WBEs, because the two human-inspired paths are closely linked and hence more likely to hit closer in time than de novo, neural probe research is more likely to cause WBE to overtake neuromorphic than neuromorphic to overtake de novo.

Turchin's Could slaughterbots wipe out humanity? Assessment of the global catastrophic risk posed by autonomous weapons, provides a series of fermi-calculation like estimates of the danger posed by weaponised drones. He concludes that while they are very difficult to defend against, and their cost is coming down, it is unlikely they would be the driving force behind human extinction.

Bogosian's Implementation of Moral Uncertainty in Intelligent Machines, argues for using Will's metanormativity approach to moral uncertainty as a way for addressing moral disagreement in AI design. I'm always glad to see more attention given to Will's thesis, which

I thought was very good, and the application to AI is an interesting one. I'm not quite sure how it would interact with a value-learning system - is the idea that the agent is updating all of its moral theories as new evidence comes in? Or that it has some value-learning approaches that are sharing credence with pre-programmed non-learning systems? I was a bit confused by his citing Greene (2001) as comparing the dispersion of issue and theory level disagreement on moral issues, but I don't think this actually affects the conclusions of the paper at all, and am less concerned than Kyle is about the scaling properties of the algorithm. I also liked his prudential argument for why moral partisans should agree to this compromise, though I note that virtue ethicists, for whom the character of the agent (not merely the results) matters, may not be convinced. Finally, I think he actually understated the extent to which debates about decision procedures are less vicious than those about object-level issues, as virtually all the emotion about voting systems seems to be generated by object-level partisans who believe that changing the voting system will help them achieve their object-level political goals.

rk and Sempere's AI development incentive gradients are not uniformly terrible argue that the 'openness is bad' conclusion from Armstrong et al's Racing to the Precipice is basically because of the discontinuity in success probability in their model. This seems true to me, and reduced my credence that openness was bad. Researchers from FHI were also named authors on the paper.

Liu et al.'s Governing Boring Apocalypses: A new typology of existential vulnerabilities and exposures for existential risk research discusses the broad risk landscape. They provide a number of breakdowns of possible risks, including many non-AI. I think the main use is the relatively policymaker-friendly framing.

Bansal and Weld's A Coverage-Based Utility Model for Identifying Unknown Unknowns design a model for efficiently utilising a scarce human expert to discover false-positive regions.

Dai's A general model of safety-oriented AI development provides a very brief generalisation of the sort of inductive strategies for AI safety I had been referring to as 'Christiano-like'

Books

Roman Yampolskiy edited a 500-page anthology on AI Safety, available for purchase here. Unfortunately I haven't had time to read every article; here is a review by someone who has.

The first half of the book, Concerns of Luminaries, is basically re-prints of older articles. As such readers will probably mainly be interested in the second half, which I think are all original to this volume.

Misc other news

OpenPhil gave Carl Shulman $5m to re-grant, of which some seems likely to end up funding useful AI safety work. Given Carl's intellect and expertise this seems like a good use of money to me.

OpenPhil are also funding seven ML PhD students ($1.1m over five years) through their 'AI Fellows' program. I have read their published research and some of it seems quite interesting – I found Noam's Safe and Nested Subgame Solving for Imperfect-Information Games particularly interesting, partly as I didn't have much prior familiarity with the subject. Most of their work thus far does not seem very AI Safety relevant, with some exceptions like this blog post by Jon Gauthier. But given the timeline for academic work and the mid-year announcement of the fellowships I think it's probably too early to see if they will produce any AI Safety relevant work.

If you like podcasts, you might enjoy these 80,000 Hours podcasts. If not, they all have complete transcripts.

- Paul Christiano
- Hilary Greaves
- Jan Leike
- Alan Dafoe

80,000 Hours also wrote a guide on how to transition from programming or CS into ML.

Last year I mentioned that EA Long Term Future Fund did not seem to be actually making grants. After a series of criticism on the EA forum by Henry Stanley and Evan Gaensbauer, CEA has now changed the management of the funds and committed to a regular series of grantmaking. However, I'm skeptical this will solve the underlying problem. Presumably they organically came across plenty of possible grants – if this was truly a 'lower barrier to giving' vehicle than OpenPhil they would have just made those grants. It is possible, however, that more managers will help them find more non-controversial ideas to fund. Here is a link to their recent grants round.

If you're reading this, you probably already read SlateStarCodex. If not, you might enjoy this article he wrote this year about AI Safety.

In an early proof of the viability of cryonics, LessWrong has been brought back to life. If like me you find the new interface confusing you can view it through GreaterWrong. Relatedly there is integration with the Alignment Forum, to provide a place for discussion of AI Alignment issues that is linked to LessWrong. This seems rather clever to me.

Zvi Mowshowitz and Vladimir Slepnev have been organizing a series of AI Safety prizes, giving out money for the articles they were most impressed with in a certain time frame.

Deepmind's work on Protein Folding proved quite successful, winning the big annual competition by a significant margin. This seemed significant to me mainly because 'solving the protein folding problem' has been one of the prototypical steps between 'recursively self-improving AI' and 'singleton' since at least 2001.

Berkley offered a graduate-level course in AGI Safety.

Vast.ai are attempting to create a two-sided marketplace where you can buy or sell idle GPU capacity. This seems like the sort of thing that probably will not succeed, but if something like it did that's another piece of evidence for hardware overhang.

The US department of commerce suggested an ban on AI exports, presumably inspired by previous bans on cryptography exports.

Conclusions

The size of the field continues to grow, both in terms of funding and researchers. Both make it increasingly hard for individual donors.

As I have once again failed to reduce charity selection to a science, I've instead attempted to subjectively weigh the productivity of the different organisations against the resources they used to generate that output, and donate accordingly.

My constant wish is to promote a lively intellect and independent decision-making among my readers; hopefully my laying out the facts as I see them above will prove helpful to some readers. Here is my eventual decision, rot13'd so you can do come to your own conclusions first if you wish:

Qrfcvgr univat qbangrq gb ZVEV pbafvfgragyl sbe znal lrnef nf n erfhyg bs gurve uvtuyl aba-ercynprnoyr naq tebhaqoernxvat jbex va gur svryq, V pnaabg va tbbq snvgu qb fb guvf lrne tvira gurve ynpx bs qvfpybfher. Nqqvgvbanyyl, gurl nyernql unir n ynetre ohqtrg guna nal bgure betnavfngvba (rkprcg creuncf SUV) naq n ynetr nzbhag bs erfreirf.

Qrfcvgr SUV cebqhpvat irel uvtu dhnyvgl erfrnepu, TCV univat n ybg bs cebzvfvat cncref va gur cvcryvar, naq obgu univat uvtuyl dhnyvsvrq naq inyhr-nyvtarq erfrnepuref, gur erdhverzrag gb cer-shaq erfrnepuref’ ragver pbagenpg fvtavsvpnagyl vapernfrf gur rssrpgvir pbfg bs shaqvat erfrnepu gurer. Ba gur bgure unaq, uvevat crbcyr va gur onl nern vfa’g purnc rvgure.

Guvf vf gur svefg lrne V unir nggrzcgrq gb erivrj PUNV va qrgnvy naq V unir orra vzcerffrq jvgu gur dhnyvgl naq ibyhzr bs gurve jbex. V nyfb guvax gurl unir zber ebbz sbe shaqvat guna SUV. Nf fhpu V jvyy or qbangvat fbzr zbarl gb PUNV guvf lrne.

V guvax bs PFRE naq TPEV nf orvat eryngviryl pbzcnenoyr betnavfngvbaf, nf 1) gurl obgu jbex ba n inevrgl bs rkvfgragvny evfxf naq 2) obgu cevznevyl cebqhpr fgengrtl cvrprf. Va guvf pbzcnevfba V guvax TPEV ybbxf fvtavsvpnagyl orggre; vg vf abg pyrne gurve gbgny bhgchg, nyy guvatf pbafvqrerq, vf yrff guna PFRE’f, ohg gurl unir qbar fb ba n qenzngvpnyyl fznyyre ohqtrg. Nf fhpu V jvyy or qbangvat fbzr zbarl gb TPEV ntnva guvf lrne.

NAH, Qrrczvaq naq BcraNV unir nyy qbar tbbq jbex ohg V qba’g guvax vg vf ivnoyr sbe (eryngviryl) fznyy vaqvivqhny qbabef gb zrnavatshyyl fhccbeg gurve jbex.

Bhtug frrzf yvxr n irel inyhnoyr cebwrpg, naq V nz gbea ba qbangvat, ohg V guvax gurve arrq sbe nqqvgvbany shaqvat vf fyvtugyl yrff guna fbzr bgure tebhcf.

NV Vzcnpgf vf va znal jnlf va n fvzvyne cbfvgvba gb TPEV, jvgu gur rkprcgvba gung TPEV vf nggrzcgvat gb fpnyr ol uvevat vgf cneg-gvzr jbexref gb shyy-gvzr, juvyr NV Vzcnpgf vf fpnyvat ol uvevat arj crbcyr. Gur sbezre vf fvtavsvpnagyl ybjre evfx, naq NV Vzcnpgf frrzf gb unir rabhtu zbarl gb gel bhg gur hcfvmvat sbe 2019 naljnl. Nf fhpu V qb abg cyna gb qbangr gb NV Vzcnpgf guvf lrne, ohg vs gurl ner noyr gb fpnyr rssrpgviryl V zvtug jryy qb fb va 2019.

Gur Sbhaqngvbany Erfrnepu Vafgvghgr unir qbar fbzr irel vagrerfgvat jbex, ohg frrz gb or nqrdhngryl shaqrq, naq V nz fbzrjung zber pbaprearq nobhg gur qnatre bs evfxl havyngreny npgvba urer guna jvgu bgure betnavfngvbaf.

V unira’g unq gvzr gb rinyhngr gur Sberfvtug Vafgvghgr, juvpu vf n funzr orpnhfr ng gurve fznyy fvmr znetvany shaqvat pbhyq or irel inyhnoyr vs gurl ner va snpg qbvat hfrshy jbex. Fvzvyneyl, Zrqvna naq Pbairetrapr frrz gbb arj gb ernyyl rinyhngr, gubhtu V jvfu gurz jryy.

Gur Shgher bs Yvsr vafgvghgr tenagf sbe guvf lrne frrz zber inyhnoyr gb zr guna gur cerivbhf ongpu, ba nirentr. Ubjrire, V cersre gb qverpgyl rinyhngr jurer gb qbangr, engure guna bhgfbhepvat guvf qrpvfvba.

V nyfb cyna gb fgneg znxvat qbangvbaf gb vaqvivqhny erfrnepuref, ba n ergebfcrpgvir onfvf, sbe qbvat hfrshy jbex. Gur pheerag fvghngvba, jvgu n ovanel rzcyblrq/abg-rzcyblrq qvfgvapgvba, naq hcsebag cnlzrag sbe hapregnva bhgchg, frrzf fhobcgvzny. V nyfb ubcr gb fvtavsvpnagyl erqhpr bireurnq (sbe rirelbar ohg zr) ol abg univat na nccyvpngvba cebprff be nal erdhverzragf sbe tenagrrf orlbaq univat cebqhprq tbbq jbex. Guvf jbhyq or fbzrjung fvzvyne gb Vzcnpg Pregvsvpngrf, juvyr ubcrshyyl nibvqvat fbzr bs gurve vffhrf.

However I wish to emphasis that all the above organisations seem to be doing good work on the most important issue facing mankind. It is the nature of making decisions under scarcity that we must prioritize some over others, and I hope that all organisations will understand that this necessarily involves negative comparisons at times.

Thanks for reading this far; hopefully you found it useful. Apologies to everyone who did valuable work that I excluded; I have no excuse other than procrastination, Crusader Kings II,

and a starting work at a new hedge fund.

Disclosures

I have not in general checked all the proofs in these papers, and similarly trust that researchers have honestly reported the results of their simulations.

I was a Summer Fellow at MIRI back when it was SIAI, volunteered briefly at GWWC (part of CEA) and previously applied for a job at FHI. I am personal friends with people at MIRI, FHI, CSER, CHAI, GPI, BERI, OpenAI, Deepmind, Ought and AI Impacts but not really at ANU, EAF/FRI, GCRI, Google Brain, Foresight, FLI, Median, Convergence (so if you're worried about bias you should overweight them… though it also means I have less direct knowledge) (also sorry if I've forgotten any friends who work for the latter set!). However I have no financial ties beyond being a donor and have never been romantically involved with anyone who has ever been at any of the organisations.

I shared drafts of the individual organisation sections with representatives from MIRI, FHI, CHAI, CSER, GCRI, GPI, BERI, Ought, AI Impacts, and EAF/FRI.

I'd like to thank Greg Lewis and my anonymous reviewers for looking over this. Any remaining mistakes are of course my own. I would also like to thank my wife for tolerating all the time I have invested/wasted on this.

EDIT: Removed language about BERI, at their request.

Sources

Amodei, Dario and Hernandez, Danny - AI and Compute - 2018-05-16 - https://blog.openai.com/ai-and-compute/

Armstrong, Stuart; O'Rourke, Xavier - 'Indifference' methods for managing agent rewards - 2018-01-05 - https://arxiv.org/pdf/1712.06365.pdf

Armstrong, Stuart; O'Rourke, Xavier - Safe Uses of AI Oracles - 2018-06-05 - https://arxiv.org/pdf/1711.05541.pdf

Armstrong, Stuart; Soren, Mindermann - Impossibility of deducing preferences and rationality from human policy - 2017-12-05 - https://arxiv.org/abs/1712.05812

Avin, Shahar; Wintle, Bonnie; Weitzdorfer, Julius; Ó hÉigeartaigh, Seán; Sutherland, William; Rees, Martin - Classifying Global Catastrophic Risks - 2018-02-23 - https://www.sciencedirect.com/science/article/pii/S0016328717301957#tbl0010

Awad, Edmond; Dsouza, Sohan; Kim, Richard; Schulz, Jonathan; Henrich, Joseph; Shariff, Azim; Bonnefon, Jean-Francois; Rahwan, Iyad - The Moral Machine Experiment - 2018-10-24 - https://www.nature.com/articles/s41586-018-0637-6

Bansal, Gagan; Weld, Daniel - A Coverage-Based Utility Model for Identifying Unknown Unknowns - 2018-04-25 - https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17110

Basu, Chandrayee; Yang, Qian; Hungerman, David; Mukesh, Singhal; Dragan, Anca - Do You Want Your Autonomous Car to Drive Like You? - 2018-02-05 -

Batin, Mikhail; Turchin, Alexey; Markov, Sergey; Zhila, Alisa; Denkenberger, David - Artificial Intelligence in Life Extension: from Deep Learning to Superintelligence - 2017-08-31 - http://www.informatica.si/index.php/informatica/article/view/1797

Baum, Seth - Countering Superintelligence Misinformation - 2018-09-09 - https://www.mdpi.com/2078-2489/9/10/244

Baum, Seth - Resilience to Global Catastrophe - 2018-11-29 - https://irgc.epfl.ch/wp-content/uploads/2018/11/Baum-for-IRGC-Resilience-Guide-Vol-2-2018.pdf

Baum, Seth - Superintelligence Skepticism as a Political Tool - 2018-08-22 - https://www.mdpi.com/2078-2489/9/9/209

Baum, Seth - Uncertain Human Consequences in Asteroid Risk Analysis and the Global Catastrophe Threshold - 2018-07-28 - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3218342

Baum, Seth; Armstrong, Stuart; Ekenstedt, Timoteus; Haggstrom, Olle; Hanson, Robin; Kuhlemann, Karin; Maas, Matthijs; Miller, James; Salmela, Markus; Sandberg, Anders; Sotala, Kaj; Torres, Phil; Turchi, Alexey; Yampolskiy, Roman - Long-Term Trajectories of Human Civilization - 2018-08-08 - http://gcrinstitute.org/papers/trajectories.pdf

Baum, Seth; Barrett, Anthony - A Model for the Impacts of Nuclear War - 2018-04-03 - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3155983

Baum, Seth; Barrett, Anthony; Yampolskiy, Roman - Modelling and Interpreting Expert Disagreement about Artificial Intelligence - 2018-01-27 - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3104645

Baum, Seth; Neufville, Robert; Barrett, Anthony - A Model for the Probability of Nuclear War - 2018-03-08 - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3137081

Baumann, Tobias - Using Surrogate Goals to Deflect Threats - 2018-02-20 - https://foundational-research.org/using-surrogate-goals-deflect-threats/

Becker, Gary - Crime and Punishment: An Economic Approach - 1974-01-01 - https://www.nber.org/chapters/c3625.pdf

Bekdash, Gus - Using Human History, Psycology and Biology to Make AI Safe for Humans - 2018-04-01 -

Berberich, Nicolas; Diepold, Klaus - The Virtuous Machine - Old Ethics for New Technology - 2018-06-27 - https://arxiv.org/abs/1806.10322

Blake, Andrew; Bordallo, Alejandro; Hawasly, Majd; Penkov, Svetlin; Ramamoorthy, Subramanian; Silva, Alexandre - Efficient Computation of Collision Probabilities for Safe Motion Planning - 2018-04-15 - https://arxiv.org/abs/1804.05384

Bogosian, Kyle - Implementation of Moral Uncertainty in Intelligent Machines - 2017-12-01 - https://link.springer.com/article/10.1007/s11023-017-9448-z

Bostrom, Nick - The Vulnerable World Hypothesis - 2018-11-09 - https://nickbostrom.com/papers/vulnerable.pdf

Brown, Noam; Sandholm, Tuomas - Safe and Nested Subgame Solving for Imperfect-Information Games - 2017-05-08 - https://arxiv.org/abs/1705.02955

Brown, Noam; Sandholm, Tuomas - Solving Imperfect-Information Games via Discounted Regret Minimization - 2018-09-11 - https://arxiv.org/abs/1809.04040

Brown, Tom; Olsson, Catherine; Google Brain Team, Research Engineers - Introducing the Unrestircted Adversarial Examples Challenge - 2018-09-03 - https://ai.googleblog.com/2018/09/introducing-unrestricted-adversarial.html

Carey, Ryan - Interpreting AI Compute Trends - 2018-07-10 - https://aiimpacts.org/interpreting-ai-compute-trends/

Cave, Stephen; Ó hÉigeartaigh, Seán - An AI Race for Strategic Advantage: Rhetoric and Risks - 2018-01-16 - http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_163.pdf

Christiano, Paul - Techniques for Optimizing Worst-Case Performance - 2018-02-01 - https://ai-alignment.com/techniques-for-optimizing-worst-case-performance-39eafec74b99

Christiano, Paul - Universality and Security Amplification - 2018-03-10 - https://ai-alignment.com/universality-and-security-amplification-551b314a3bab

Christiano, Paul; Shlegeris, Buck; Amodei, Dario - Supervising Strong Learners by Amplifying Weak Experts - 2018-10-19 - https://arxiv.org/abs/1810.08575

Cohen, Michael; Vellambi, Badri; Hutter, Marcus - Algorithm for Aligned Artificial General Intelligence - 2018-05-25 - https://cs.anu.edu.au/courses/CSPROJECTS/18S1/reports/u6357432.pdf

Cundy, Chris; Filan, Daniel - Exploring Hierarchy-Aware Inverse Reinforcement Learning - 2018-07-13 - https://arxiv.org/abs/1807.05037

Currie, Adrian - Existential Risk, Creativity & Well-Adapted Science - 2018-07-22 - http://philsci-archive.pitt.edu/14800/

Currie, Adrian - Geoengineering Tensions - 2018-04-30 - http://philsci-archive.pitt.edu/14607/

Currie, Adrian - Introduction: Creativity, Conservatism & the Social Epistemology of Science - 2018-09-27 - http://philsci-archive.pitt.edu/15066/

Currie, Adrian; Ó hÉigeartaigh, Seán - Working together to face humanity's greatest threats: Introduction to The Future of Research on Catastrophic and Existential Risk - 2018-03-26 - https://www.dropbox.com/s/bh6okdz8pvrxzc6/Working%20together%20to%20face%20humanity%E2%80%99s%20greatest%20threats%20preprint.pdf?dl=0

Dafoe, Allen - AI Governance: A Research Agenda - 2018-08-27 - https://www.fhi.ox.ac.uk/wp-content/uploads/GovAIAgenda.pdf

Dai, Wei - A general model of safety-oriented AI development - 2018-06-11 - https://www.lesswrong.com/posts/idb5Ppp9zghcichJ5/a-general-model-of-safety-oriented-ai-development

Demski, Abram - An Untrollable Mathematician Illustrated - 2018-03-19 - https://www.lesswrong.com/posts/CvKnhXTu9BPcdKE4W/an-untrollable-mathematician-illustrated

DeVries, Terrance; Taylor, Graham - Leveraging Uncertainty Estimates for Predicting Segmentation Quality - 2018-07-02 - https://arxiv.org/abs/1807.00502

Dobbe, Roel; Dean, Sarah; Gilbert, Thomas; Kohli, Nitin - A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics - 2018-07-06 - https://arxiv.org/abs/1807.00553

Doshi-Velez, Finale; Kim, Been - Considerations for Evaluation and Generalization in Interpretable Machine Learning - 2018-08-24 - https://finale.seas.harvard.edu/publications/considerations-evaluation-and-generalization-interpretable-machine-learning

Duettmann, Allison; Afanasjeva, Olga; Armstrong, Stuart; Braley, Ryan; Cussins, Jessica; Ding, Jeffrey; Eckersley, Peter; Guan, Melody; Vance, Alyssa; Yampolskiy, Roman - Artificial General Intelligence: Coordination and Great Powers - 1900-01-00 - https://fs1-bb4c.kxcdn.com/wp-content/uploads/2018/11/AGI-Coordination-Geat-Powers-Report.pdf

Erdelyi, Olivia ; Goldsmith, Judy - Regulating Artificial Intelligence: Proposal for a Global Solution - 2018-02-01 - http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_13.pdf

Eth, Daniel - The Technological Landscape Affecting Artificial General Intelligence and the Importance of Nanoscale Neural Probes - 2017-08-31 - http://www.informatica.si/index.php/informatica/article/view/1874

Evans, Owain; Stuhlmuller, Andreas; Cundy, Chris; Carey, Ryan; Kenton, Zachary; McGrath, Thomas; Schreiber, Andrew - Predicting Human Deliberative Judgments with Machine Learning - 2018-07-13 - https://ought.org/papers/predicting-judgments-tr2018.pdf

Everitt, Tom; Krakovna, Victoria; Orseau, Laurent; Hutter, Marcus; Legg, Shane - Reinforcement Learning with a Corrupted Reward Channel - 2017-05-23 - https://arxiv.org/abs/1705.08417

Everitt, Tom; Lea, Gary; Hutter, Marcus - AGI Safety Literature Review - 2018-05-22 - AGI Safety Literature Review

Filan, Daniel - Bottle Caps aren't Optimisers - 2018-11-21 - https://www.greaterwrong.com/posts/26eupx3Byc8swRS7f/bottle-caps-aren-t-optimisers

Fisac, Jaime; Bajcsy, Andrea; Herbert, Sylvia; Fridovich-Keil, David; Wang, Steven; Tomlin, Claire; Dragan, Anca - Probabilistically Safe Robot Planning with Confidence-Based Human Predictions - 2018-05-31 - https://arxiv.org/abs/1806.00109

Garnelo, Marta; Rosenbaum, Dan; Maddison, Chris; Ramalho, Tiago; Saxton, David; Shanahan, Murray; The, Yee Whye; Rezende, Danilo; Eslami, S M Ali - Conditional Neural Processes - 2018-07-04 -

Garrabrant, Scott; Demski, Abram - Embedded Agency Sequence - 2018-10-29 - https://www.lesswrong.com/s/Rm6oQRJJmhGCcLvxh

Gasparik, Amanda; Gamble, Chris; Gao, Jim - Safety-first AI for autonomous data centre cooling and industrial control - 2018-08-17 - https://deepmind.com/blog/safety-first-ai-autonomous-data-centre-cooling-and-industrial-control/

Gauthier, Jon; Ivanova, Anna - Does the brain represent words? An evaluation of brain decoding studies of language understanding - 2018-06-02 - https://arxiv.org/abs/1806.00591

Ghosh, Shromona; Berkenkamp, Felix; Ranade, Gireeja; Qadeer, Shaz; Kapoor, Ashish - Verifying Controllers Against Adversarial Examples with Bayesian Optimization - 2018-02-26 - https://arxiv.org/abs/1802.08678

Gilmer, Justin; Adams, Ryan; Goodfellow, Ian; Andersen, David, Dahl, George - Motivating the Rules of the Game for Adversarial Example Research - 2018-07-20 - https://arxiv.org/abs/1807.06732

Grace, Katja - Human Level Hardware Timeline - 2017-12-22 - https://aiimpacts.org/human-level-hardware-timeline/

Grace, Katja - Likelihood of discontinuous progress around the development of AGI - 2018-02-23 - https://aiimpacts.org/likelihood-of-discontinuous-progress-around-the-development-of-agi/

Green, Brian Patrick - Ethical Reflections on Artificial Intelligence - 2018-06-01 - http://apcz.umk.pl/czasopisma/index.php/SetF/article/view/SetF.2018.015

Hadfield-Menell, Dylan; Andrus, McKane; Hadfield, Gillian - Legible Normativity for AI Alignment: The Value of Silly Rules - 2018-11-03 - https://arxiv.org/abs/1811.01267

Hadfield-Menell, Dylan; Hadfield, Gillian - Incomplete Contracting and AI alignment - 2018-04-12 - https://arxiv.org/abs/1804.04268

Haqq-Misra, Jacob - Policy Options for the radio Detectability of Earth - 2018-04-02 - https://arxiv.org/abs/1804.01885

Hoang, Lê Nguyên - A Roadmap for the Value-Loading Problem - 2018-09-04 - https://arxiv.org/abs/1809.01036

Huang, Jessie; Wu, Fa; Precup, Doina; Cai, Yang - Learning Safe Policies with Expert Guidance - 2018-05-21 - https://arxiv.org/abs/1805.08313

Ibarz, Borja; Leike, Jan; Pohlen, Tobias; Irving, Geoffrey; Legg, Shane; Amodei, Dario - Reward Learning from Human Preferences and Demonstrations in Atari - 2018-11-15 - https://arxiv.org/abs/1811.06521

IBM - Bias in AI: How we Build Fair AI Systems and Less-Biased Humans - 2018-02-01 - https://www.ibm.com/blogs/policy/bias-in-ai/

Irving, Geoffrey; Christiano, Paul; Amodei, Dario - AI Safety via Debate - 2018-05-02 - https://arxiv.org/abs/1805.00899

Janner, Michael; Wu, Jiajun; Kulkarni, Tejas; Yildirim, Ilker; Tenenbaum, Joshua - Self-Supervised Intrinsic Image Decomposition - 2018-02-05 - https://arxiv.org/abs/1711.03678

Jilk, David - Conceptual-Linguistic Superintelligence - 2017-07-31 - http://www.informatica.si/index.php/informatica/article/view/1875

Jones, Natalie; O'Brien, Mark; Ryan, Thomas - Representation of future generations in United Kingdom policy-making - 2018-03-26 - https://www.sciencedirect.com/science/article/pii/S0016328717301179

Koller, Torsten; Berkenkamp, Felix; Turchetta, Matteo; Krause, Andreas - Learning-based Model Predictive Control for Safe Exploration - 2018-09-22 - https://arxiv.org/abs/1803.08287

Krakovna, Victoria - Specification Gaming Examples in AI - 2018-04-02 - https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/

Krakovna, Victoria; Orseau, Laurent; Martic, Miljan; Legg, Shane - Measuring and avoiding side effects using relative reachability - 2018-06-04 - https://arxiv.org/abs/1806.01186

Kurakin, Alexey; Goodfellow, Ian; Bengio, Samy; Dong, Yinpeng; Liao, Fangzhou; Liang, Ming; Pang, Tianyu ; Zhu, Jun; Hu, Xiaolin; Xie, Cihang; Wang, Jianyu; Zhang, Zhishuai; Ren, Zhou; Yuille, Alan; Huang, Sangxia; Zhao, Yao; Zhao, Yuzhe; Han, Zhonglin; Long, Junjiajia; Berdibekov, Yerkebulan; Akiba, Takuya; Tokui, Seiya; Abe Motoki - Adversarial Attacks and Defences Competition - 2018-03-31 - https://arxiv.org/pdf/1804.00097.pdf

Lee, Kimin; Lee, Kibok; Lee, Honglak; Shin, Jinwoo - A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks - 2018-10-27 - https://arxiv.org/abs/1807.03888

Lehman, Joel; Clune, Jeff; Misevic, Dusan - The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities - 2018-08-14 - https://arxiv.org/abs/1803.03453

Leibo, Joel; de Masson d'Autume, Cyprien; Zoran, Daniel; Amos, David; Beattie, Charles; Anderson, Keith; Castañeda, Antonio García; Sanchez, Manuel; Green, Simon; Gruslys, Audrunas, Legg, Shane, Hassabis, Demis, Botvinick, Matthew - Psychlab: A Psychology Laboratory for Deep Reinforcement Learning Agents - 2018-02-04 - https://arxiv.org/abs/1801.08116

Leike, Jan; Kruegar, David; Everitt, Tom; Martic, Miljan; Maini, Vishal; Legg, Shane - Scalable agent alignment via reward modeling: a research direction - 2018-11-19 - https://arxiv.org/abs/1811.07871

Leike, Jan; Martic, Miljan; Krakovna, Victoria; Ortega, Pedro; Everitt, Tom; Lefrancq, Andrew; Orseau, Laurent; Legg, Shane - AI Safety Gridworlds - 2017-11-28 - https://arxiv.org/abs/1711.09883

Lewis, Gregory; Millett, Piers; Sandberg, Anders; Snyder-Beattie; Gronvall, Gigi - Information Hazards in Biotechnology - 2018-11-12 - https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.13235

Lipton, Zachary; Steinhardt, Jacob - Troubling Trends in Machine Learning Scholarship - 2018-07-26 - https://arxiv.org/abs/1807.03341

Liu, Chang; Hamrick, Jessica; Fisac, Jaime; Dragan, Anca; Hedrick, J Karl; Sastry, S Shankar; Griffiths, Thomas - Goal Inference Improves Objective and Perceived Performance in Human-Robot Collaboration - 2018-02-06 - https://arxiv.org/abs/1802.01780

Liu, Hin-Yan; Lauta, Kristian Cedervall; Mass, Matthijs Michiel - Governing Boring Apocalypses: A new typology of existential vulnerabilities and exposures for existential risk research - 2018-03-26 - https://www.sciencedirect.com/science/article/pii/S0016328717301623

Liu, Qiang; Li, Pan; Zhao, Wentao; Cai, Wei; Yu, Shui; Leung, Victor - A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View - 2018-02-13 - https://ieeexplore.ieee.org/document/8290925

Liu, Yang; Price, Huw - Ramsey and Joyce on deliberation and prediction - 2018-08-30 - http://philsci-archive.pitt.edu/14972/

Lütjens, Björn; Everett, Michael; How, Jonathan - Safe Reinforcement Learning with Model Uncertainty Estimates - 2018-10-19 - https://arxiv.org/abs/1810.08700

Malinin, Andrey; Gales, Mark - Predictive Uncertainty Estimation via Prior Networks - 2018-10-08 - https://arxiv.org/abs/1802.10501

Manheim, David; Garrabrant, Scott - Categorizing Variants of Goodheart's Law - 2018-04-10 - https://arxiv.org/abs/1803.04585

Martinez-Plumed, Fernando; Loe, Bao Sheng; Flach, Peter; Ó hÉigeartaigh, Seán; Vold, Karina; Hernandez-Orallo, Jose - The Facets of Artificial Intelligence: A Framework to Track the Evolution of AI - 2018-08-21 - https://www.ijcai.org/proceedings/2018/0718.pdf

McCaslin, Tegan - Transmitting fibers in the brain: Total length and distribution of lengths - 2018-03-29 - https://aiimpacts.org/transmitting-fibers-in-the-brain-total-length-and-distribution-of-lengths/

Menda, Kunal; Driggs-Campbell, Katherine; Kochenderfer, Mykel - EnsembleDAgger: A Bayesian Approach to Safe Imitation Learning - 2018-07-22 - https://arxiv.org/abs/1807.08364

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, Dario Amodei - The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation - 2018-02-20 - https://arxiv.org/abs/1802.07228

Milli, Smitha; Schmidt, Ludwig; Dragan, Anca; Hardt, Moritz - Model Reconstruction from Model Explanations - 2018-07-13 - https://arxiv.org/abs/1807.05185

Mindermann, Soren; Shah, Rohin; Gleave, Adam; Hadfield-Menell, Dylan - Active Inverse Reward Design - 2018-11-16 - https://arxiv.org/abs/1809.03060

Mogensen, Andreas - Long-termism for risk averse altruists - 1900-01-00 - https://unioxfordnexus-my.sharepoint.com/personal/exet1753_ox_ac_uk/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fexet1753%5Fox%5Fac%5Fuk%2FDocuments%2FGlobal%20Priorities%20Institute%2FOperations%2FWebsite%2FWorking%20papers%2FLongtermism%20and%20risk%20aversion%20v3%2Epdf&parent=%2Fpersonal%2Fexet1753%5Fox%5Fac%5Fuk%2FDocuments%2FGlobal%20Priorities%20Institute%2FOperations%2FWebsite%2FWorking%20papers&slrid=10daaa9e-b098-7000-a41a-599fb32c6ff4

Ngo, Richard; Pace, Ben - Some cruxes on impactful alternatives to AI policy work - 2018-10-10 - https://www.lesswrong.com/posts/DJB82jKwgJE5NsWgT/some-cruxes-on-impactful-alternatives-to-ai-policy-work

Noothigattu, Ritesh; Bouneffouf, Djallel; Mattei, Nicholas; Chandra, Rachita; Madan, Piyush; Varshney, Kush; Campbell, Murray; Singh, Moninder; Rossi, Francesca - Interpretable Multi-Objective Reinforcement Learning through Policy Orchestration - 2018-09-21 - https://arxiv.org/abs/1809.08343

Nushi, Besmira; Kamar, Ece; Horvitz, Eric - Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure - 2018-09-19 - https://arxiv.org/abs/1809.07424

Oesterheld, Caspar - Approval-directed agency and the decision theory of Newcomb-like problems - 2017-12-21 - https://casparoesterheld.files.wordpress.com/2017/12/rldt.pdf

OpenAI - OpenAI Charter - 2018-04-09 - https://blog.openai.com/openai-charter/

Ortega, Pedro; Maini, Vishal; Safety Team, Deepmind - Building safe artificial intelligence: specification, robustness and assurance - 2018-09-27 - https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1

Papernot, Nicolas; McDaniel, Patrick - Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning - 2018-03-13 - https://arxiv.org/pdf/1803.04765.pdf

Raghunathan, Aditi; Steinhardt, Jacob; Liang, Percy - Certified Defenses Against Adversarial Examples - 2018-01-29 - https://arxiv.org/abs/1801.09344

Rainforth, Tom; Kosiorek, Adam; Anh Le, Tuan; Maddison, Chris; Igl, Maximilian; Wood, Frank; Whe Teh, Yee - Tighter Variational Bounds are Not Necessarily Better - 2018-06-25 - https://arxiv.org/abs/1802.04537

Ratner, Ellis; Hadfield-Menell, Dylan; Dragan, Anca - Simplifying Reward Design through Divide-and-Conquer - 2018-06-07 - https://arxiv.org/abs/1806.02501

Reddy, Siddharth; Dragan, Anca; Levine, Sergey - Shared Autonomy via Deep Reinforcement Learning - 2018-05-23 - https://arxiv.org/abs/1802.01744

Reddy, Siddharth; Dragan, Anca; Levine, Sergey - Where Do You Think You're Going?: Inferring Beliefs about Dynamics from Behaviour - 2018-10-20 - https://arxiv.org/abs/1805.08010

Rees, Martin - On The Future - 2018-10-16 - https://www.amazon.com/Future-Prospects-Humanity-Martin-Rees-ebook/dp/B07CSD5BG9

rk; Sempere, Nuno - AI development incentive gradients are not uniformly terrible - 2018-11-12 - https://www.lesswrong.com/posts/bkG4qj9BFEkNva3EX/ai-development-incentive-gradients-are-not-uniformly

Ruan, Wenjie; Huang, Xiaowei; Kwiatkowska, Marta - Reachability Analysis of Deep Neural Networks with Provable Guarantees - 2018-05-06 - https://arxiv.org/abs/1805.02242

Sadigh, Dorsa; Sastry, Shankar; Seshia, Sanjit; Dragan, Anca - Planning for Autonomous Cars that Leverage Effects on Human Actions - 2016-06-01 - https://people.eecs.berkeley.edu/~sastry/pubs/Pdfs%20of%202016/SadighPlanning2016.pdf

Sandberg, Anders - Human Extinction from Natural Hazard Events - 2018-02-01 - http://oxfordre.com/naturalhazardscience/view/10.1093/acrefore/9780199389407.001.0001/acrefore-9780199389407-e-293

Sarma, Gopal; Hay, Nick - Mammalian Value Systems - 2017-12-31 - https://arxiv.org/abs/1607.08289

Sarma, Gopal; Hay, Nick - Robust Computer Algebra, Theorem Proving, and Oracle AI - 2017-12-31 - https://arxiv.org/abs/1708.02553

Sarma, Gopal; Hay, Nick; Safron, Adam - AI Safety and Reproducibility: Establishing Robust Foundations for the Neuropsychology of Human Values - 2018-09-08 - https://arxiv.org/abs/1712.04307

Schulze, Sebastian; Evans, Owain - Active Reinforcement Learning with Monte-Carlo Tree Search - 2018-03-13 - https://arxiv.org/abs/1803.04926

Shah, Rohin - AI Alignment Newsletter - 1905-07-10 - https://rohinshah.com/alignment-newsletter/

Shah, Rohin; Christiano, Paul; Armstrong, Stuart; Steinhardt, Jacob; Evans, Owain - Value Learning Sequence - 2018-10-29 - https://www.lesswrong.com/s/Rm6oQRJJmhGCcLvxh

Shahar, Avin - Mavericks and Lotteries - 2018-09-25 - http://philsci-archive.pitt.edu/15058/

Shahar, Avin; Shapira, Shai - Civ V AI Mod - 2018-01-05 - https://www.cser.ac.uk/news/civilization-v-video-game-mod-superintelligent-ai/

Shaw, Nolan P.; Stockel, Andreas; Orr, Ryan W.; Lidbetter, Thomas F.; Cohen, Robin - Towards Provably Moral AI Agents in Bottom-up Learning Frameworks - 2018-03-15 - http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_8.pdf

Sotala, Kaj - Shaping economic incentives for collaborative AGI - 2018-06-29 - https://www.lesswrong.com/posts/FkZCM4DMprtEp568s/shaping-economic-incentives-for-collaborative-agi

Sotala, Kaj; Gloor, Lukas - Superintelligence as a Cause or Cure for Risks of Astronomical Suffering - 2017-08-31 - http://www.informatica.si/index.php/informatica/article/view/1877

Stuhlmuller, Andreas - Factored Cognition - 2018-04-25 - https://ought.org/presentations/factored-cognition-2018-05

Taylor, Jessica; Gallagher, Jack; Maltinsky, Baeo - Insight-based AI timeline model - 1905-07-10 - http://mediangroup.org/insights

The Future of Life Institute - Value Alignment Research Landscape - 1900-01-00 - https://futureoflife.org/valuealignmentmap/

Trammell, Philip - Fixed-Point Solutions to the Regress Problem in Normative Uncertainty - 2018-08-29 - https://unioxfordnexus-my.sharepoint.com/personal/exet1753_ox_ac_uk/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fexet1753%5Fox%5Fac%5Fuk%2FDocuments%2FGlobal%20Priorities%20Institute%2FOperations%2FWebsite%2FWorking%20papers%2Fdecision%5Ftheory%5Fregress%2Epdf&parent=%2Fpersonal%2Fexet1753%5Fox%5Fac%5Fuk%2FDocuments%2FGlobal%20Priorities%20Institute%2FOperations%2FWebsite%2FWorking%20papers&slrid=14daaa9e-3069-7000-a41a-5aa6302f7c36

Tucker, Aaron; Gleave, Adam; Russell, Stuart - Inverse Reinforcement Learning for Video Games - 2018-10-24 - https://arxiv.org/abs/1810.10593

Turchin, Alexey - Could slaughterbots wipe out humanity? Assessment of the global catastrophic risk posed by autonomous weapons - 2018-03-19 - https://philpapers.org/rec/TURCSW

Turchin, Alexey; Denkenberger, David - Classification of Global Catastrophic Risks Connected with Artificial Intelligence - 2018-05-03 - https://link.springer.com/article/10.1007/s00146-018-0845-5

Turner, Alex - Towards a New Impact Measure - 2018-09-18 - https://www.alignmentforum.org/posts/yEa7kwoMpsBgaBCgb/towards-a-new-impact-measure

Umbrello, Steven; Baum, Seth - Evaluating Future nanotechnology: The Net Societal Impacts of Atomically Precise Manufacturing - 2018-04-30 - https://www.researchgate.net/publication/324715437_Evaluating_Future_Nanotechnology_The_Net_Societal_Impacts_of_Atomically_Precise_Manufacturing

Vonitzer, Vincent; Sinnott-Armstrong, Walter; Borg, Jana Schaich; Deng, Yuan; Kramer, Max - Moral Decision Making Frameworks for Artificial Intelligence - 2017-02-12 - https://users.cs.duke.edu/~conitzer/moralAAAI17.pdf

Wang, Xin; Chen, Wenhu; Wang, Yuan-Fang ; Yang Wang, William - No Metrics are Perfect: Adversarial Reward Learning for Visual Storytelling - 2018-07-09 - https://arxiv.org/abs/1804.09160

Wu, Yi; Siddharth, Srivastava; Hay, Nicholas; Du, Simon; Russell, Stuart - Discrete-Continuous Mixtures in Probabilistic Programming: Generalised Semantics and Inference Algorithms - 2018-06-13 - https://arxiv.org/abs/1806.02027

Wu, Yueh-Hua; Lin, Shou-De - A Low-Cost Ethics Shaping Approach for Designing Reinforcement Learning Agents - 2018-09-10 - https://arxiv.org/abs/1712.04172

Yu, Han; Shen, Zhiqi; Miao, Chunyan; Leung, Cyril; Lesser, Victor; Yang, Qiang - Building Ethics into Artificial Intelligence - 2018-07-13 - http://www.ntulily.org/wp-content/uploads/conference/Building_Ethics_into_Artificial_Intelligence_accepted.pdf

Yudkowsky, Eliezer - The Rocket Alignment Problem - 2018-10-03 - https://intelligence.org/2018/10/03/rocket-alignment/

Yudkowsky, Eliezer; Christiano, Paul - Challenges to Christiano's Capability Amplification Proposal - 2018-05-19 - https://www.lesswrong.com/posts/S7csET9CgBtpi7sCh/challenges-to-christiano-s-capability-amplification-proposal

Zhou, Allen; Hadfield-Menell, Dylan; Nagabandi, Anusha; Dragan, Anca - Expressive Robot Motion Timing - 2018-02-05 -

How did academia ensure papers were correct in the early 20th Century?

In the post 'Four layers of Intellectual Conversation', Eliezer says that both the writer of an idea, and the person writing a critique of that idea, need to expect to have to publicly defend what they say at least one time. Otherwise they can write something stupid and never lose status because they don't have to respond to the criticism.

I was wondering about where this sort of dialogue happens in academia. I have been told by many people that current journals are quite terrible, but I've also heard a romantic notion that science (especially physics and math) used to be more effectively pursued in the early 20th century (Einstein, Turing, Shannon, etc). So Oliver and I thought we'd look at the journals to see if they had real conversations.

We looked at two data points, and didn't find any.

First, Oliver looked through Einstein's publication history (Oli is German and could read it). Einstein has lots of 'reviews' of others' work in his list of publications, sometimes multiple of the same person, which seemed like a promising example of conversation. Alas, it turned out that Einstein had merely helped German journals write *summaries* of papers that had been written in English, and there was no real dialogue.

Second, I looked through a volume of the London Mathematical Society, in particular, the volume where Turing published his groundbreaking paper proving that not all mathematical propositions are decidable (thanks to sci-hub for making it possible for me to read the papers!). My eyes looked at about 60% of the pages in the journal (about 12 papers), and *not one of them* disagreed with any prior work. There was :

- A footnote that thanked an advisor for finding a flaw in a proof
- An addendum page (to the whole volume) that consisted of a single sentence thanking someone for showing one of their theorems was a special case of someone else's theorem
- One person who was skeptical of another person's theorem. But that theorem by Ramanujan (who was famous for stating theorems without proofs), and the whole paper primarily found proofs of his other theorems.

There were lots of discussions of people's work but always building, or extending, or finding a neater way of achieving the same results. Never disagreement, correction, or the finding of errors.

One thing that really confuses me about this is that *it's really hard to get all the details right*. Lots of great works are filled with tiny flaws (e.g. Donald Knuth reliably has people find errors in his texts). So I'd expect any discussion of old papers to bring up flaws, or that journals would require a section at the end for corrections of the previous volume. There were of course reviewers, but they can't be experts in all the areas.

But more importantly *where did/does the dialogue happen if not in the journals?*

If I try to be concrete about what I'm curious about:

> As people go about the craft of doing science, they will make errors (conceptual mistakes, false proofs, and so on). One of the main pieces of infrastructure in academia are journals, where work gets published and can become common knowledge.
>
> Two places to fix errors are pre-publication and post-publication. I don't know much about the pre-publication process, but if it is strong enough to ensure no errors got published, I'd like some insight into what that process was like. Alternatively, if course-correction happened post-publication, I'm interested to know how and where, because when I looked (see above) I couldn't find it.

There's also the third alternative, that no progress was made. And there's the fourth alternative, that most papers were bad and the thing scientists did was to just never read or build on them. I'm happy to get evidence for any of these, or a fifth alternative.

Added: Maybe this is a more crisp statement:

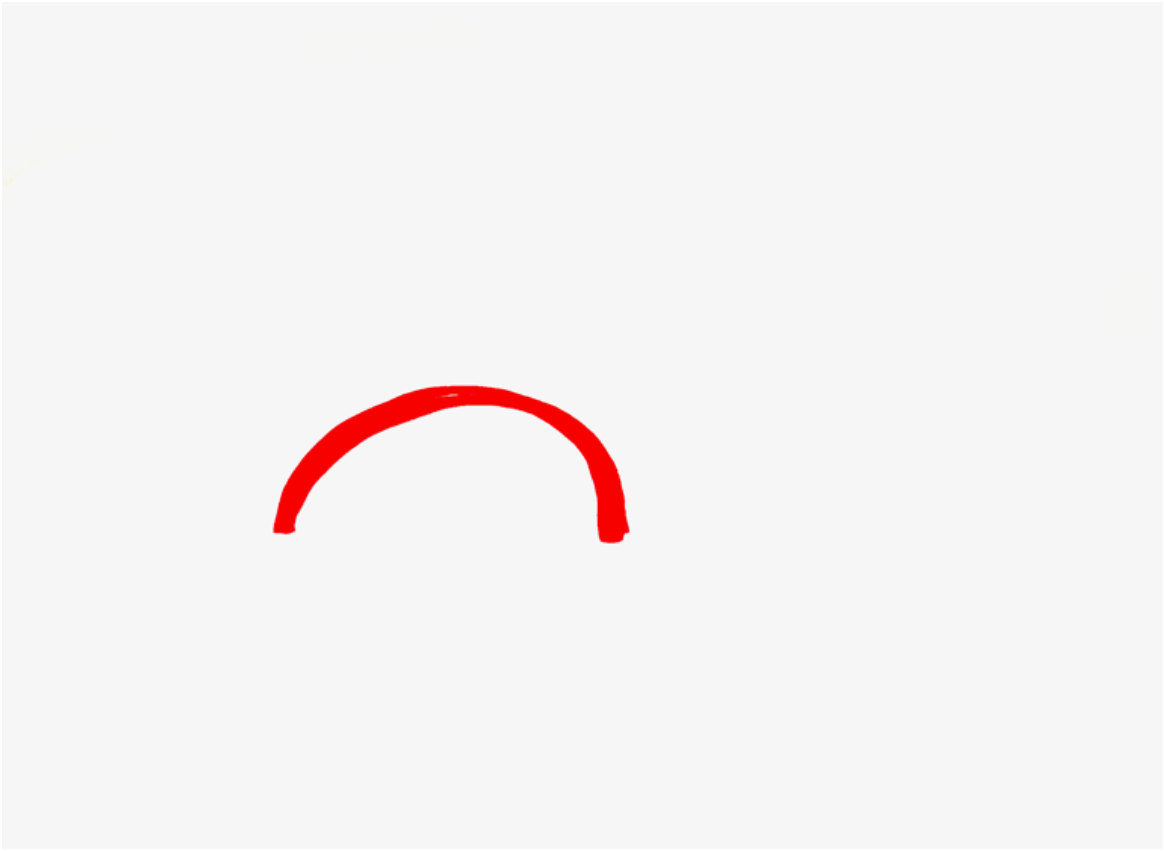
> Why do (old) journals not claim to have errors in any of the papers? Is it because they're (implicitly) lying about the quality of the papers? Or if there's a reliable process that removed errors from 100% of papers, can someone tell me what that process was?

Spaghetti Towers

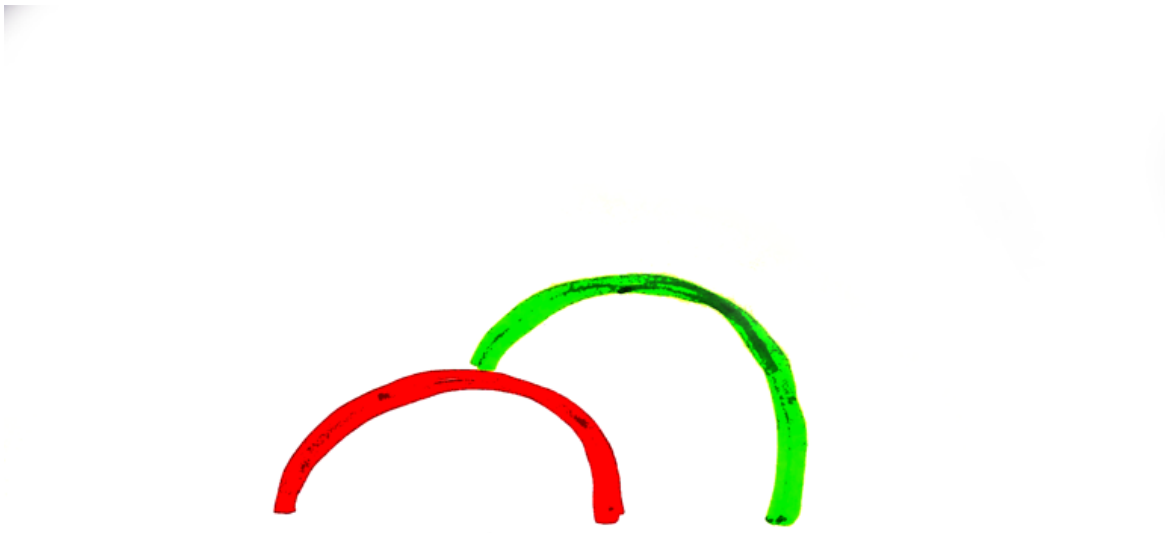
This is a linkpost for https://eukaryotewritesblog.com/2018/12/21/spaghetti-towers/

Here's a pattern I'd like to be able to talk about. It might be known under a certain name somewhere, but if it is, I don't know it. I call it a Spaghetti Tower. It shows up in large complex systems that are built haphazardly.

Someone or something builds the first Part A.



Later, someone wants to put a second Part B on top of Part A, either out of convenience (a common function, just somewhere to put it) or as a refinement to Part A.



Now, suppose you want to tweak Part A. If you do that, you might break Part B, since it interacts with bits of Part A. So you might instead build Part C on top of the previous ones.

And by the time your system looks like this, it's much harder to tell what changes you can make to an earlier part without crashing some component, so you're basically relegated to throwing another part on top of the pile.

I call these spaghetti towers for two reasons: One, because they tend to quickly take on circuitous knotty tangled structures, like what programmers call "spaghetti code". (Part of the problem with spaghetti code is that it can lead to spaghetti towers.)

Especially since they're usually interwoven in multiple dimensions, and thus look more like this:

"Can you just straighten out the yellow one without touching any of the others? Thanks."

Second, because shortsightedness in the design process is a crucial part of spaghetti machines. In order to design a spaghetti system, you throw spaghetti against a wall and see if it sticks. Then, when you want to add another part, you throw more spaghetti until it sticks to that spaghetti. And later, you throw more spaghetti. So it goes. And if you decide that you want to tweak the bottom layer to make it a little more useful – which you might want to do because, say, it was built out of spaghetti – without damaging the next layers of gummy partially-dried spaghetti, well then, good luck.

Note that all systems have load-bearing, structural pieces. This does not make them spaghetti towers. The distinction about spaghetti towers is that they have a lot of shoddily-built structural components that are completely unintentional. A bridge has major load-bearing components – they're pretty obvious, strong, elegant, and efficiently support the rest of the structure. A spaghetti tower is more like this.

Image from the always-delightful r/DiWHY.

(The motto of the spaghetti tower is “Sure, it works fine, as long as you never run lukewarm water through it and unplug the washing machine during thunderstorms.”)

Where do spaghetti towers appear?

- Basically all of biology works like this. Absolutely all of evolution is made by throwing spaghetti against walls and seeing what sticks. (More accurately, throwing nucleic acid

against harsh reality and seeing what successfully makes more nucleic acid.) We are 3.5 billion years of hacks in fragile trench coats.
    - Scott Star Codex describes the phenomenon in neurotransmitters, but it's true for all of molecular biology:

> You know those stories about clueless old people who get to their Gmail account by typing "Google" into Bing, clicking on Google in the Bing search results, typing "Gmail" into Google, and then clicking on Gmail in the Google search results?
>
> I am reading about serotonin transmission now, and everything in the human brain works on this principle. If your brain needs to downregulate a neurotransmitter, it'll start by upregulating a completely different neurotransmitter, which upregulates the first neurotransmitter, which hits autoreceptors that downregulate the first neurotransmitter, which then cancel the upregulation, and eventually the neurotransmitter gets downregulated.
>
> Meanwhile, my patients are all like "How come this drug that was supposed to cure my depression is giving me vision problems?" and at least on some level the answer is "how come when Bing is down your grandfather can't access Gmail?

- My programming friends tell me that spaghetti towers are near-universal in the codebases of large companies. Where it would theoretically be nice if every function was neatly ordered, but actually, the thing you're working on has three different dependencies, two of which are unmaintained and were abandoned when the guy who built them went to work at Google, and you can never be 100% certain that your code tweak won't crash the site.
- I think this also explains some of why bureaucracies look and act the way they do, and are so hard to change.

I think there are probably a lot of examples of spaghetti towers, and they probably have big ramifications for things like, for instance, what systems evolution can and can't build.

I want to do a much deeper and more thoughtful analysis about what exactly the implications here are, but this has been kicking around my brain for long enough and all I want to do is get the concept out there.

Does this feel like a meaningful concept? Where do you see spaghetti towers?

Transhumanism as Simplified Humanism

This essay was originally posted in 2007.

Frank Sulloway once said: “Ninety-nine per cent of what Darwinian theory says about human behavior is so obviously true that we don’t give Darwin credit for it. Ironically, psychoanalysis has it over Darwinism precisely because its predictions are so outlandish and its explanations are so counterintuitive that we think, *Is that really true? How radical!* Freud’s ideas are so intriguing that people are willing to pay for them, while one of the great disadvantages of Darwinism is that we feel we know it already, because, in a sense, we do.”

Suppose you find an unconscious six-year-old girl lying on the train tracks of an active railroad. What, morally speaking, ought you to do in this situation? Would it be better to leave her there to get run over, or to try to save her? How about if a 45-year-old man has a debilitating but nonfatal illness that will severely reduce his quality of life – is it better to cure him, or not cure him?

Oh, and by the way: This is not a trick question.

I answer that I would save them if I had the power to do so – both the six-year-old on the train tracks, and the sick 45-year-old. The obvious answer isn’t *always* the best choice, but sometimes it *is.*

I won’t be lauded as a brilliant ethicist for my judgments in these two ethical dilemmas. My answers are not surprising enough that people would pay me for them. If you go around proclaiming “What does two plus two equal? Four!” you will not gain a reputation as a deep thinker. But it is still the correct answer.

If a young child falls on the train tracks, it is good to save them, and if a 45-year-old suffers from a debilitating disease, it is good to cure them. If you have a logical turn of mind, you are bound to ask whether this is a special case of a general ethical principle which says “Life is good, death is bad; health is good, sickness is bad.” If so – and here we enter into controversial territory – we can follow this general principle to a surprising new conclusion: If a 95-year-old is threatened by death from old age, it would be good to drag them from those train tracks, if possible. And if a 120-year-old is starting to feel slightly sickly, it would be good to restore them to full vigor, if possible. With current technology it is *not* possible. But if the technology became available in some future year – given sufficiently advanced medical nanotechnology, or such other contrivances as future minds may devise – would you judge it a good thing, to save that life, and stay that debility?

The important thing to remember, which I think all too many people forget, is that *it is not a trick question.*

Transhumanism is simpler – requires fewer bits to specify – because it has no special cases. If you believe professional bioethicists (people who get paid to explain ethical judgments) then the rule “Life is good, death is bad; health is good, sickness is bad” holds only until some critical age, and then flips polarity. Why should it flip? Why not just keep on with life-is-good? It would seem that it is good to save a six-year-old girl,

but bad to extend the life and health of a 150-year-old. Then at what *exact* age does the term in the utility function go from positive to negative? Why?

As far as a transhumanist is concerned, if you see someone in danger of dying, you should save them; if you can improve someone's health, you should. There, you're done. No special cases. You don't have to ask anyone's age.

You also don't ask whether the remedy will involve only "primitive" technologies (like a stretcher to lift the six-year-old off the railroad tracks); or technologies invented less than a hundred years ago (like penicillin) which nonetheless seem ordinary because they were around when you were a kid; or technologies that seem scary and sexy and futuristic (like gene therapy) because they were invented after you turned 18; or technologies that seem absurd and implausible and sacrilegious (like nanotech) because they haven't been invented yet. Your ethical dilemma report form doesn't have a line where you write down the invention year of the technology. Can you save lives? Yes? Okay, go ahead. There, you're done.

Suppose a boy of 9 years, who has tested at IQ 120 on the Wechsler-Bellvue, is threatened by a lead-heavy environment or a brain disease which will, if unchecked, gradually reduce his IQ to 110. I reply that it is a good thing to save him from this threat. If you have a logical turn of mind, you are bound to ask whether this is a special case of a general ethical principle saying that intelligence is precious. Now the boy's sister, as it happens, currently has an IQ of 110. If the technology were available to gradually raise her IQ to 120, without negative side effects, would you judge it good to do so?

Well, of course. Why not? It's not a trick question. Either it's better to have an IQ of 110 than 120, in which case we should strive to decrease IQs of 120 to 110. Or it's better to have an IQ of 120 than 110, in which case we should raise the sister's IQ if possible. As far as I can see, the obvious answer is the correct one.

But – you ask – *where does it end?* It may seem well and good to talk about extending life and health out to 150 years – but what about 200 years, or 300 years, or 500 years, or more? What about when – in the course of properly integrating all these new life experiences and expanding one's mind accordingly over time – the equivalent of IQ must go to 140, or 180, or beyond human ranges?

Where does it end? It doesn't. Why should it? Life is good, health is good, beauty and happiness and fun and laughter and challenge and learning are good. This does not change for arbitrarily large amounts of life and beauty. If there were an upper bound, it would be a special case, and that would be inelegant.

Ultimate physical limits may or may not permit a lifespan of at least length X for some X – just as the medical technology of a particular century may or may not permit it. But physical limitations are questions of simple fact, to be settled strictly by experiment. Transhumanism, as a moral philosophy, deals only with the question of whether a healthy lifespan of length X is desirable *if* it is physically possible. Transhumanism answers yes for all X. Because, you see, it's not a trick question.

So that is "transhumanism" – loving life without special exceptions and without upper bound.

Can transhumanism really be that simple? Doesn't that make the philosophy trivial, if it has no extra ingredients, just common sense? Yes, in the same way that the scientific method is nothing but common sense.

Then why have a complicated special name like “transhumanism” ? For the same reason that “scientific method” or “secular humanism” have complicated special names. If you take common sense and rigorously apply it, through multiple inferential steps, to areas outside everyday experience, successfully avoiding many possible distractions and tempting mistakes along the way, then it often ends up as a minority position and people give it a special name.

But a moral philosophy should not *have* special ingredients. The purpose of a moral philosophy is not to look delightfully strange and counterintuitive, or to provide employment to bioethicists. The purpose is to guide our choices toward life, health, beauty, happiness, fun, laughter, challenge, and learning. If the judgments are simple, that is no black mark against them – morality doesn’t always have to be complicated.

There is nothing in transhumanism but the same common sense that underlies standard humanism, rigorously applied to cases outside our modern-day experience. A million-year lifespan? If it’s possible, why not? The prospect may seem very foreign and strange, relative to our current everyday experience. It may create a sensation of future shock. And yet – is life a *bad* thing?

Could the moral question really be just that simple?

Yes.

Meditations on Momentum

Cross-posted and lightly-edited from The Deep Dish.

Epistemic status: describing a general phenomenon; may not be correct on every specific point. may have used scientific terms in an annoying metaphorical fashion. elements of pointing out the bleeding obvious, hopefully framed in a novel way.

tl;dr: positive feedback loops are a thing, thinking in systems/exponentially is hard, intersectionality is underrated.

> *"For to everyone who has, more will be given, and he will have an abundance. But from the one who has not, even what he has will be taken away."*
>
> —MATTHEW 25:29

In 2013, unknown author Robert Galbraith published his debut novel… to crickets.

The first print run of *The Cuckoo's Calling* was 1500 copies. It's not clear how many actually sold. The book occupied 4709th place on Amazon's bestseller charts.

And there, perhaps it would have stayed, if the cuckoo in the nest had remained undiscovered. The secret unravelled after a few months: ex-military security contractor Galbraith was a pseudonym for J.K. Rowling. As soon as the news broke, *The Cuckoo's Calling* soared to the number one spot on Amazon. Sales increased by 150,000 per cent overnight. Copies from that first neglected print run are now worth thousands of dollars.

What's the difference between Robert Galbraith and J.K. Rowling? Clearly, being a talented writer is necessary, but not sufficient. Rowling has *momentum* on her side. At this point, she could publish the contents of a bowl of alphabet soup, and it would still sell better than 99 per cent of novels by hopeful first-time authors.

This is a 'no duh' example, designed to get you nodding your head along. But momentum is *everywhere*, and it's rarely in plain sight. Without being consciously aware of doing so, I've written about it in four domains:

1. Popularity

> *Popular things often get their start through what amounts to good luck. The rapid ascent is driven by something even more powerful than rocket fuel: social contagion. Our opinions and preferences cluster together, but it's not because we've carefully evaluated them on their merits. We just want to feel close to our fellow social apes, and have something to gossip about around the water cooler.*
>
> *In other words, popularity is a lot like herpes. After catching a lucky initial break, it manages to spread to a few hosts, then rides the exponential growth curve until it has planted its gentle, blistery kiss on 60 per cent of the population.*

— The Madness of Crowds: Why It Pays to Go Where the Tourists Aren't

2. Wealth

> *With enough time on his side, [Fry's] 93 cents transforms into $4.3 billion. If your gut instincts are screaming that this is staggeringly, ridiculously, wrong—well, you're not*

alone.

As Mark Zuckerberg put it: "Humans don't understand exponential growth. If you fold a paper 50 times, it goes to the moon and back."

This is a delicious example, not only because the imagery is so jarring—whoa, a tiny sheet of paper can do that?—but because the Zuck himself got it wrong. If you fold a piece of paper 50 times over, it doesn't make a paltry return trip to the moon—it goes all the way to the freakin' sun. Humans don't understand exponential growth, indeed.

— Futurama Taught Me Everything I Know About Compound Interest

3. Entrepreneurship

Something idea-based can be sold over and over again with almost no extra time or effort. It's infinitely scalable. Your debut album might sell 10 copies (three of which your mum bought) or 10 million, but the amount of work that went into recording it was the same.

Scalable careers don't follow a normal distribution, with a clear relationship between effort and reward. Instead, they produce grotesque inequalities. It doesn't necessarily matter how good you are, or how hard you work: a select few people capture almost all the rewards, while everyone else gets next to nothing.

— The Barbell Strategy: Don't Be a Starving Artist

4. Health

What's life like for a moderately fit and muscular person? Well, everything works in your favor. The wind is at your back. You've got momentum.

The fitter you are, the better your hormonal and metabolic health, the lower your bodyfat, the more relaxed you can be with your diet, the more fun life is, the more motivation you have to train, the cooler feats you can perform, the deeper the habit is ingrained, and so on, in an endless positive feedback loop.

In fact, it's even better than that. Almost all these factors are mutually reinforcing. If you do screw up, and drunkenly devour an entire box of cereal, or take a week off from the gym to clock a new video game, it's no biggie. Any one link can seize up for a while, and the cycle will keep on turning without it.

— Fat People Are Heroes

…and a few more examples I've collected, but haven't written about:

5. Academia

Sociologist Robert Merton coined the term 'The Matthew Effect', after the parable of the talents line quoted up top.

Merton noticed that famous scientists often get credited for discoveries made by lesser-known researchers or grad students toiling in obscurity. Similarly, the success of any given paper often depends on the prominence of the author, and how many early citations it happens to receive:

So great is this problem that we are tempted to turn again to the Scriptures to designate the status-enhancement and status-suppression components of the Matthew effect. We

can describe it as the Ecclesiasticus component, from the familiar injunction 'Let us now praise famous men'.

— The Matthew Effect in Science, Robert K. Merton

6. Reading

Psychologists have discovered the same effect in education. The longer it takes kids to learn how to read, the slower the development of their other cognitive skills and performance:

The longer this developmental sequence is allowed to continue, the more generalized the deficits will become, seeping into more and more areas of cognition and behavior. Or to put it more simply – and sadly – in the words of a tearful nine-year-old, already falling frustratingly behind his peers in reading progress, "Reading affects everything you do."

—Progress in Understanding Reading, Keith Stanovich

7. Market prices

For most intents and purposes, the efficient markets hypothesis is correct. But even the Nobel prize-winning EMH creator, Eugene Fama, has admitted there is one major anomaly: *momentum*, which he describes as the "biggest embarrassment to the theory".

Here's Fama's old advisor, Benoit Mandelbrot, on the long memory of market pricing:

What a company does today—a merger, a spin-off, a critical product launch—shapes what the company will look like a decade hence; in the same way, its stock-price movements today will influence movements tomorrow.

…a bottom line emerges. Stock prices are not independent. Today's action can, at least slightly, affect tomorrow's action. The standard model is, again, wrong.

—Benoit Mandelbrot, The (Mis)behavior of Markets

Had enough?

There's also the height of trees , the colour, brightness, and lifetime of stars, the proliferation of species , the halo and horns effect, affective death spirals , and the existence of life itself .

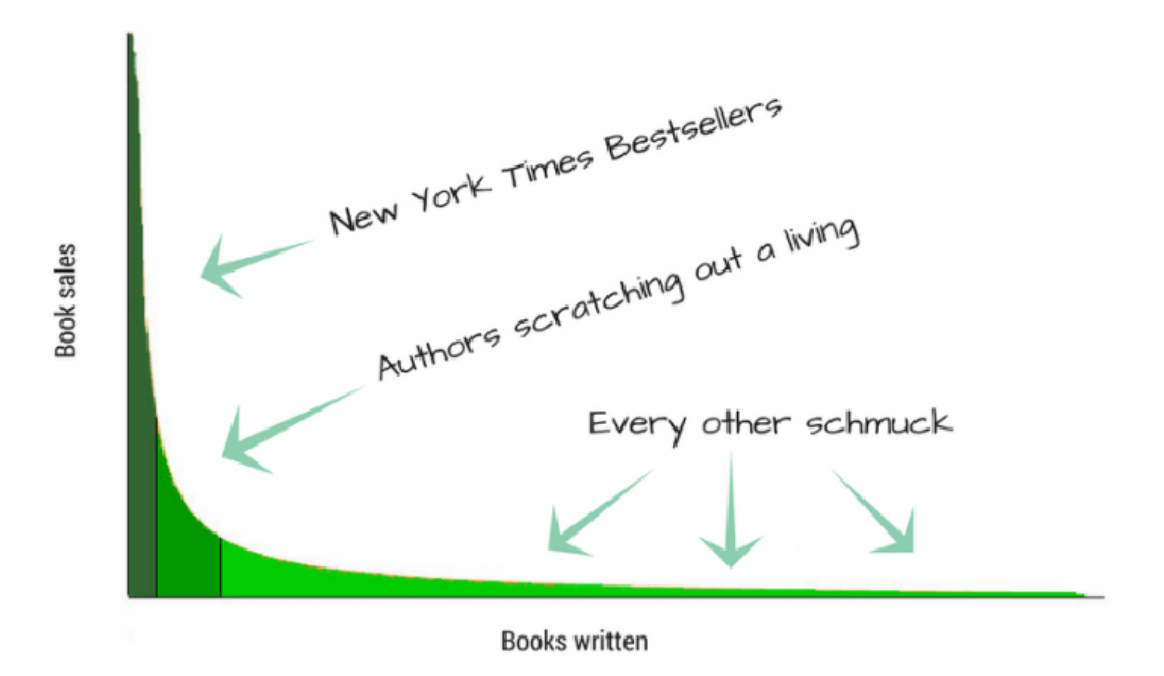
The principle of cumulative advantage spans physics, biology, psychology, economics, and culture. It almost seems like some underlying feature of the universe. Here's Mandelbrot again:

"Can you seriously compare the wind to a financial market, a gale to a rally, a hurricane to a crash? In terms of the underlying causes, certainly not. But mathematically, yes. It is an extraordinary feature of science that the most diverse, seemingly unrelated, phenomena can be described with the same mathematical tools."

On the macro scale of the universe—the birth of stars, complex life bootstrapped from mud—momentum is kind of miraculous. For a brief candle-flicker, we get to resist the relentless march of entropy; create defiant bastions of order and beauty amongst the chaos.

On the micro scale of individual human affairs—wealth, waistlines, popularity, power—momentum is kind of terrifying. It makes us, and it breaks us. The 1 per cent control almost half of the world's wealth, a small number of startups succeed astronomically, most books are sold by the J.K Rowlings of the world.

Momentum leaves behind a distinctive calling card, which looks something like this:



If this graph was drawn to scale, the tail would extend several kilometres off your computer screen. For self-published ebooks, it's worse: the median number of sales is zero.

You will know these various patterns as the '80/20 rule', power laws, long-tails, and Pareto distributions. The economist Vilfredo Pareto devoted years to the pattern which now bears his name. Surely *he* has some kind words to say about his curvy wife?

> *At the bottom of the [curve], men and women starve and children die young. In the broad middle of the curve all is turmoil and motion: people rising and falling, climbing by talent or luck and falling by alcoholism, tuberculosis and other kinds of unfitness. At the very top sit the elite of the elite, who control wealth and power for a time — until they are unseated through revolution or upheaval by a new aristocratic class.*

Yikes!

If each instance of the Matthew effect stayed in its own lane, that would be unfair enough. But as Pareto points out, they're all hopelessly entangled. Each of these domains – money, opportunity, health, education, talent, prestige – not only compounds on itself; but spills over into the other buckets too. Some interactions are obvious: a successful author will almost by definition make more money. Others are less so: a fit and healthy person might get promoted over an equally-qualified overweight person, for no good reason at all.

And that's the *positive* side of the ledger…

The Downward Spiral

> *My dear, here we must run as fast as we can, just to stay in place. And if you wish to go anywhere you must run twice as fast as that.*
>
> —Lewis Carroll

Momentum also works in reverse.

Imagine your partner breaks up with you. You start drinking more. The drinking affects your work. You become isolated from friends and family. You stop exercising and looking after yourself. Eventually, you lose your job. Now you have money problems, on top of your declining physical and mental health, and total lack of support network. Things don't tend to deteriorate in a linear fashion: you spiral downwards faster and faster, until you fall off a cliff.

The further down you slip, the harder it is to regain lost ground.

I had a little taste of this recently. A series of bad things came along in quick succession. Each of them would have been OK in isolation; together, they put me into a tailspin. I pride myself on being put-together, but I unravelled disturbingly quickly. Order begets order; chaos begets chaos. It was an uncomfortable reminder that everyone is always only a few strokes of misfortune away from the abyss: there but for the grace of God go I.

Further Down the Spiral

Just to make it explicit: the title of this post is an homage to Scott Alexander's essay *Meditations on Moloch.* As Scott points out in his epic close-reading of an Allen Ginsberg poem, there are obvious things we could do to make the world a better place, but some invisible force stymies our efforts:

> *If everyone hates the current system, who perpetuates it? And Ginsberg answers: "Moloch". It's powerful not because it's correct – nobody literally thinks an ancient Carthaginian demon causes everything – but because thinking of the system as an agent throws into relief the degree to which the system* isn't *an agent.*

The same alien 'otherness' applies to momentum. A handful of A-list actors are inundated with roles, when tens of thousands of talented hopefuls would jump at the chance to eat the scraps from their table. One per cent of everyone owns half the wealth, while billions of others are desperately poor.

In every area of life, the people who are least in need of further advantage are most likely to receive it.

Almost everyone is unhappy with this distribution of outcomes, but blaming 'capitalism' or 'the government' or whichever tribe you happen to hate might be missing the point. If there *is* some blind force of nature operating behind the scenes, then the exact same pattern will continue to persist (which might explain why socialist utopias don't tend to go exactly as planned).

Back to Pareto, for more cheerful words of encouragement:

> *"There is no progress in human history. Democracy is a fraud. Human nature is primitive, emotional, unyielding. The smarter, abler, stronger, and shrewder take the lion's share. The weak starve, lest society become degenerate: One can compare the social body to the human body, which will promptly perish if prevented from eliminating toxins."*

Assume we *are* dealing with some kind of all-pervasive force of nature. Moloch works tirelessly to destroy everything humans hold dear. The Matthew Effect/momentum is more like the blind, alien god of evolution—responsible for *creating* everything humans hold dear, but in the same mindless fashion, smites entire species into oblivion.

The universe is neither hostile nor benevolent; it's utterly indifferent. What to do?

The Lord Giveth, and The Lord Taketh Away

The parable of the talents says: you better use it or lose it. Get some momentum behind you. Start saving money as early as possible. Reduce debt aggressively. Build behaviours that

compound, and nip bad habits in the bud as soon as possible. Stay the hell away from the abyss.

Saving that first $100,000, as Charlie Munger put it, is a bitch. You have to be the little rocket trying to escape the Earth's gravitational pull, with all your engines on full thrust. Then you can take your foot off the gas a little, but don't get complacent. If you lose your momentum, you'll drift back to earth, slowly at first, then faster and faster, until you slam into the ground at 200 kph.

You have to fight tooth and claw to get some momentum, and then stay up there just as long as you possibly can.

This moral sounds suspiciously demonic. But unlike Moloch's favourite games, which are zero or negative-sum, climbing the pyramid doesn't always involve stamping on the fingers of those below you.

Improving your own health and fitness doesn't make anyone else sickly. Making a consistent habit of reading, or learning new skills, doesn't make other people dumber. Contrary to popular belief, getting richer doesn't necessarily make other people poorer. And of course, one of the best ways of getting rich in the first place is refusing to pay a premium for popular things that *are popular only because they are popular.*

Extending a Helping Hand

If you help yourself without hurting anyone, that's great, but it still leaves loads of people stuck at the bottom of the curve.

Three encouraging observations: First, even if the overall pattern never changes, at least the individual data-points can move around.

We know this happens, because even mighty empires topple. Generational wealth doesn't last forever. Celebrities burn out or fade away. Trees get struck by lightning. Stars implode. In dynamic societies, everyone gets their turn at the top.

The second encouraging observation is that momentum reaches a point of diminishing returns.

Sometimes there are hard physical limits: a redwood can only grow so tall before it takes more energy to pump water up from its roots than its new needles can harvest through photosynthesis. After a certain point, a fit person has to train harder and harder to eke out smaller and smaller gains, and so on.

Even where there are no physical limits, there's a rapid drop-off in marginal utility. A famous person receives more offers and opportunities than they know what to do with. The same goes for wealth. After the first couple million bucks, Bill Gates tells us, it's the same hamburger.

If you take these two observations together, it makes a lot of sense to extend a helping hand up, rather than keep pushing for smaller and smaller gains. The pattern persists, but you create a lot more mobility up and down the curve.

Above and Beyond

Maybe Pareto was wrong.

The third encouraging observation is that mobility might be increasing, *without* a bloody revolution. [EDIT: had another look, I don't think the data actually supports me here. damn!]

Human nature is primitive and emotional, but not unyielding. Even though we struggle to wrap our monkey-minds around *compound interest*—much less social contagion and non-linear causality—we're getting less bad at it.

It's pretty cool that J.K. Rowling deliberately tried to play life on hard-mode again. It's much more exciting that more than 100 billionaires have pledged to give away most (or all) of their fortunes. And that thousands of ordinary people have made a lifetime commitment to give at least 10 per cent of their income to the most effective charities.

The parable of the talents is pretty cut-throat. My guess is that it's meant to be descriptive, not normative. And lots of people—even those at the top— *aren't* OK with it.

Sure, it's the natural order of things. But nature also gave us strychnine, parasitic wasps, and cuddly meerkats that systematically murder their infants. Nature is not to be trusted.

What's the moral of the story? As far as I can see:

1. work your butt off to get some momentum behind you,
2. keep a watchful eye out for any signs of entropy creeping in,
3. once you hit the point of diminishing returns, focus your efforts on helping other people up.

John Wesley, the founder of Methodism, delivered a famous sermon on this topic in the 18th century. I think he summed it up more pithily:

> "*Having, First, gained all you can, and, Secondly saved all you can, Then give all you can.*"

Contrite Strategies and The Need For Standards

Epistemic Status: Confident

There's a really interesting paper from 1996 called The Logic of Contrition, which I'll summarize here. In it, the authors identify a strategy called "Contrite Tit For Tat", which does better than either Pavlov or Generous Tit For Tat in Iterated Prisoner's Dilemma.

In Contrite Tit For Tat, the player doesn't only look at what he and the other player played on the last term, but also another variable, the *standing* of the players, which can be good or bad.

If Bob defected on Alice last round but Alice was in good standing, then Bob's standing switches to bad, and Alice defects against Bob.

If Bob defected on Alice last round but Alice was in *bad* standing, then Bob's standing stays good, and Alice cooperates with Bob.

If Bob cooperated with Alice last round, Bob keeps his good standing, and Alice cooperates.

This allows two Contrite Tit For Tat players to recover quickly from accidental defections without defecting against each other forever;

D/C -> C/D -> C/C

But, unlike Pavlov, it consistently resists the "always defect" strategy

D/C -> D/D -> D/D -> D/D ...

Like TFT (Tit For Tat) and unlike Pavlov and gTFT (Generous Tit For Tat), cTFT (Contrite Tit For Tat) can invade a population of all Defectors.

A related contrite strategy is Remorse. Remorse cooperates only if it is in bad standing, or if both players cooperated in the previous round. In other words, Remorse is more aggressive; unlike cTFT, it can attack cooperators.

Against the strategy "always cooperate", cTFT always cooperates but Remorse alternates cooperating and defecting:

C/C -> C/D -> C/C -> C/D ...

And Remorse defends effectively against defectors:

D/C -> D/D -> D/D -> D/D...

But if one Remorse accidentally defects against another, recovery is more difficult:

C/D -> D/C -> D/D -> C/D -> ...

If the Prisoner's Dilemma is repeated a large but finite number of times, cTFT is an evolutionarily stable state in the sense that *you can't do better for yourself when playing against a cTFT player through doing anything that deviates from what cTFT would recommend.* This implies that no other strategy can successfully invade a population of all cTFT's.

REMORSE can sometimes be invaded by strategies better at cooperating with themselves, while Pavlov can sometimes be invaded by Defectors, depending on the payoff matrix; but for all Prisoner's Dilemma payoff matrices, cTFT resists invasion.

Defector and a similar strategy called Grim Trigger (if a player ever defects on you, keep defecting forever) are evolutionarily stable, but not good outcomes — they result in much lower scores for everyone in the population than TFT or its variants. By contrast, a whole population that adopts cTFT, gTFT, Pavlov, or Remorse on average gets the payoff from cooperating each round.

The bottom line is, adding "contrition" to TFT makes it quite a bit better, and allows it to keep pace with Pavlov in exploiting TFT's, while doing better than Pavlov at exploiting Defectors.

This is no longer true if we add noise in the *perception* of good or bad standing; contrite strategies, like TFT, can get stuck defecting against each other if they erroneously perceive bad standing.

The moral of the story is that there's a game-theoretic advantage to not only having *reciprocity* (TFT) but *standards* (cTFT), and in fact reciprocity alone is not enough to outperform strategies like Pavlov which don't map well to human moral maxims.

What do I mean by standards?

There's a difference between saying "Behavior X is better than behavior Y" and saying "Behavior Y is unacceptable."

The concept of "unacceptable" behavior functions like the concept of "standing" in the game theory paper. If I do something "unacceptable" and you respond in some negative way (you get mad or punish me or w/e), I'm not supposed to retaliate against *your* negative response, I'm supposed to accept it.

Pure reciprocity results in blood feuds — "if you kill one of my family I'll kill one of yours" is perfectly sound Tit For Tat reasoning, but it means that we can't *stop* killing once we've started.

Arbitrary forgiveness fixes that problem and allows parties to reconcile even if they've been fighting, but still leaves you vulnerable to an attacker who just won't quit.

Contrite strategies are like having a court system. (Though not an enforcement system! They are still "anarchist" in that sense — all cTFT bots are equal.) The "standing" is an assessment attached to each person of whether they are in the wrong and thereby restricted in their permission to retaliate.

In general, for actions not covered by the legal system and even for some that are, we don't *have* widely shared standards of acceptable vs. unacceptable behavior. We're aware (and especially so given the internet) that these standards differ from subculture to subculture and context to context, and we're often aware that they're

arbitrary, and so we have enormous difficulty getting widely shared clarity on claims like “he was deceptive and that’s not OK”. Because…was he deceptive in a way that counts as fraud? Was it just “puffery” of the kind that’s normal in PR? Was it a white lie to spare someone’s feelings? Was it “just venting” and thus not expected to be as nuanced or fact-checked as more formal speech? What *level or standard* of honesty could he reasonably have been expected to be living up to?

We can’t say “that’s not OK” without some kind of understanding that he had failed to live up to a shared expectation. And *where is that bar?* It’s going to depend who you ask and what local context they’re living in. And not only that, but the fact that nobody is keeping track of where even the separate, local standards are, eventually standards will have to be dropped to the lowest common denominator if not made explicit.

MBTI isn’t science but it’s illustrative descriptively, and it seems to me that the difference between “Perceivers” and “Judgers”, which is basically the difference between the kinds of people who get called “judgmental” in ordinary English and the people who don’t, is that “Judgers” have a clear idea of where the line is between “acceptable” and “unacceptable” behavior, while Perceivers don’t. I’m a Perceiver, and I’ve often had this experience where someone is saying “that’s just Not OK” and I’m like “whoa, where are you getting that? I can certainly see that it’s *suboptimal*, this other thing would be better, but why are you drawing the line for acceptability here instead of somewhere else?”

The lesson of cTFT is that *having* a line in the first place, having a standard that you can either be in line with or in violation of, has survival value.

Transhumanists Don't Need Special Dispositions

This essay was originally posted in 2007.

I have claimed that transhumanism arises strictly from love of life. A bioconservative humanist says that it is good to save someone's life or cure them of debilitating syndromes if they are young, but once they are "too old" (the exact threshold is rarely specified) we should stop trying to keep them alive and healthy. A transhumanist says unconditionally: "Life is good, death is bad; health is good, death is bad." Whether you're 5, 50, or 500, life is good, why die? Nothing more is required.

Then why is there a widespread misunderstanding that transhumanism involves a special fetish for technology, or an unusually strong fear of death, or some other abnormal personal disposition?

I offer an analogy: Rationality is often thought to be about cynicism. The one comes to us and says, "Fairies make the rainbow; I believe this because it makes me feel warm and fuzzy inside." And you say, "No." And the one reasons, "I believe in fairies because I enjoy feeling warm and fuzzy. If I imagine that there are no fairies, I feel a sensation of deadly existential emptiness. Rationalists say there are no fairies. So they must enjoy sensations of deadly existential emptiness." Actually, rationality follows a completely different rule - examine the rainbow very closely and see how it actually works. If we find fairies, we accept that, and if we don't find fairies, we accept that too. The look-and-see rule makes no mention of our personal feelings about fairies, and it fully determines the rational answer. So you cannot infer that a competent rationalist hates fairies, or likes feelings of deadly existential emptiness, by looking at what they believe about rainbows.

But this rule - the notion of actually *looking* at things - is not widely understood. The more common belief is that rationalists make up stories about boring old math equations, instead of pretty little fairies, because rationalists have a math fetish instead of a fairy fetish. A personal taste, and an odd one at that, but how else would you explain rationalists' strange and unusual beliefs?

Similarly, love of life is not commonly understood as a motive for saying that, if someone is sick, and we can cure them using medical nanotech, we really ought to do that. Instead people suppose that transhumanists have a taste for technology, a futurism fetish, that we just love those pictures of little roving nanobots. A personal taste, and an odd one at that, but how else would you explain transhumanists' strange and unusual moral judgments?

Of course I'm not claiming that transhumanists take no joy in technology. That would be like saying a rationalist should take no joy in math. *Homo sapiens* is the tool-making species; a complete human being should take joy in a contrivance of special cleverness, just as we take joy in music or storytelling. It is likewise incorrect to say that the aesthetic beauty of a technology is a distinct good from its beneficial use - their sum is not merely additive, there is a harmonious combination. The equations underlying a rainbow are all the more beautiful for being true, rather than just made up. But the esthetic of transhumanism is very strict about positive outcomes taking

precedence over how cool the technology looks. If the choice is between using an elegant technology to save a million lives and using an ugly technology to save a million and one lives, you choose the latter. Otherwise the harmonious combination vanishes like a soap bubble popping. It would be like preferring a more elegant theory of rainbows that was not actually true.

In social psychology, the "correspondence bias" is that we see far too direct a correspondence between others' actions and their personalities. As Gilbert and Malone put it, we "draw inferences about a person's unique and enduring dispositions from behaviors that can be entirely explained by the situations in which they occur." For example, subjects listen to speakers giving speeches for and against abortion. The subjects are explicitly told that the speakers are reading prepared speeches assigned by coin toss - and yet the subjects still believe the pro-abortion speakers are personally in favor of abortion.

When we see someone else kick a vending machine for no visible reason, we assume he is "an angry person". But if you yourself kick the vending machine, you will tend to see your actions as caused by your situation, not your disposition. The bus was late, the train was early, your report is overdue, and now the damned vending machine has eaten your money twice in a row. But others will not see this; they cannot see your situation trailing behind you in the air, and so they will attribute your behavior to your disposition.

But, really, most of the people in the world are not mutants - are probably not exceptional in any given facet of their emotional makeup. A key to understanding human nature is to realize that the vast majority of people see themselves as behaving normally, given their situations. If you wish to understand people's behaviors, then don't ask after mutant dispositions; rather, ask what situation they might believe themselves to be in.

Suppose I gave you a control with two buttons, a red button and a green button. The red button destroys the world, and the green button stops the red button from being pressed. Which button would you press? The green one. This response is *perfectly normal.* No special world-saving disposition is required, let alone a special preference for the color green. Most people would choose to press the green button and save the world, *if they saw their situation in those terms.*

And yet people sometimes ask me why I want to save the world. *Why?* They want to know *why* someone would want to save the world? Like you have to be traumatized in childhood or something? *Give* me a *break*.

We all seem normal to ourselves. One must understand this to understand all those strange other people.

Correspondence bias can also be seen as essentialist reasoning, like explaining rain by water spirits, or explaining fire by phlogiston. If you kick a vending machine, why, it must be because you have a vending-machine-kicking disposition.

So the transhumanist says, "Let us use this technology to cure aging." And the reporter thinks, *How strange! He must have been born with an unusual technology-loving disposition*. Or, *How strange! He must have an unusual horror of aging!*

Technology means many things to many people. So too, death, aging, sickness have different implications to different personal philosophies. Thus, different people incorrectly attribute transhumanism to different mutant dispositions.

If someone prides themselves on being cynical of all Madison Avenue marketing, and the meaning of technology unto them is Madison Avenue marketing, they will see transhumanists as shills for The Man, trying to get us to spend money on expensive but ultimately meaningless toys.

If someone has been fed Deep Wisdom about how death is part of the Natural Cycle of Life ordained by heaven as a transition to beyond the flesh, etc., then they will see transhumanists as Minions of Industry, Agents of the Anti-Life Death Force that is Science.

If someone has a postmodern ironic attitude toward technology, then they'll see transhumanists as being on a mission to make the world even stranger, more impersonal, than it already is - with the word "Singularity" obviously referring to complete disconnection and incomprehensibility.

If someone sees computers and virtual reality as an escape from the real world, opposed to sweat under the warm sun and the scent of real flowers, they will think that transhumanists must surely hate the body; that they want to escape the scent of flowers into a grayscale virtual world.

If someone associates technology with Gosh-Wow-Gee-Whiz-So-Cool flying cars and jetpacks, they'll think transhumanists have gone overboard on youthful enthusiasm for toys.

If someone associates the future with scary hyperbole from Wired magazine - *humans will merge with their machines and become indistinguishable from them* - they'll think that transhumanists yearn for the cold embrace of metal tentacles, that we want to *lose our identity and be eaten by the Machine* or some other dystopian nightmare of the month.

In all cases they make the same mistake - drawing a one-to-one correspondence between the way in which the behavior strikes them as strange, and a mutant mental essence that exactly fits the behavior. This is an unnecessarily burdensome explanation for why someone would advocate healing the sick by any means available, including advanced technology.

What makes people intellectually active?

What is the difference between a smart person who has read the sequences and considers AI x-risk important and interesting, but continues to be primarily a consumer of ideas, and someone who starts having ideas? I am not trying to set a really high bar here -- they don't have to be good ideas. They can't be off-the-cuff either, though. I'm talking about someone taking their ideas through multiple iterations.

A person does not need to research full-time to have ideas. Ideas can come during downtime. Maybe it is something you think about during your commute, and talk about occasionally at a lesswrong meetup.

There is something incomplete about my model of people doing this vs not doing this. I expect more people to have more ideas than they do.

AI alignment is the example I'm focusing on, but the missing piece of my world-model extends much more broadly than that. How do some people end up developing sprawling intellectual frameworks, while others do not?

There could be a separate "what could someone do about it" question, but I want to avoid normative/instrumental connotations here to focus on the causal chains. Asking someone "why don't you do more?" has a tendency to solicit answers like "yeah I should do more, I'm bottlenecked on willpower" -- but I don't think willpower is the distinguishing factor between cases I observe. (Maybe there is *something related* involved, but I mostly don't think of intellectual productivity as driven by a top-down desire to be intellectually productive enforced by willpower.)

I have some candidate models, but all my evidence is anecdotal and everything seems quite shaky.

Playing Politics

Epistemic Status: Guesses Based on Personal Experience

Lately I've been going through a family of learning experiences in the world of *how to get things done cooperatively*. It's hard for me. Even very basic things in this area have been stumping me, overwhelming me, leaving me way more tired and drained than I'd expect. My productivity has gone to hell and — worse — I didn't even *notice* for a while. This is hard stuff, and rarely written about by the people for whom it's hard, so my hope is that processing in public helps someone. I generally think that data-sharing is good and helpful.

Collective Deliberation Isn't Working For Me

At a conference, I was in a room full of people having a really good discussion. I wanted to get people together to have a follow-up discussion later — nothing elaborate, just a room with whiteboards and snacks and maybe moving towards some action items.

What I did:

- Passed around a sheet for emails to sign up
- Sent out an email proposing the parameters of the event
- Waited for people to propose dates that worked for them.

Radio silence.

Somebody else suggested a poll where people could put down their preferred times and dates. Out of thirteen people, five signed up. Nobody volunteered "ok, we're doing it on this date then," so I did. I reserved a conference room at my office and bought a bunch of snacks.

The front door was locked on the weekend and my key card didn't work even though it was supposed to, so I had to switch locations at the last minute. It wouldn't have mattered anyhow, because one person showed up on time, and one other person several hours late.

Conclusion: it is harder than I thought to get ten people to show up in a room and talk to each other.

And I probably shouldn't have expected an event to coalesce naturally from the mailing list. I have a strong "egalitarian" instinct that if I'm trying to do something *with* a group and in some sense *for the benefit of everyone in the group,* then I shouldn't be too "bossy" in terms of unilaterally declaring what we're all going to do. But if I leave it up to the group to discuss, it seems like they generally…don't.

I'm also on a policy committee for a community organization, and it's been a whole lot of heartache because I want to change some things about our policies and internal processes, and the process of trying to communicate that has resulted in a *lot* of hurt feelings, mine and other people's.

The first thing I did was write up a document explaining why I thought the existing policies were harmful, and share it with the mailing list. This resulted in *DRAMA*

because people heard it as a personal accusation. (I never meant to imply that my fellow committee members were bad people, but I felt strongly about the policy changes and my writing tone may have come out angrier than I intended.)

In retrospect, I should never have led with complaints — I should have started by proposing solutions. My intention had been to raise the issues I cared about while minimizing bossiness — this is an organization for the benefit of a larger community, and I'm only one member of a committee, so I thought it would *leave more degrees of freedom open to the group* to say "here's why the existing policies have problems, what do you think we should do?" rather than "here's how I'd suggest improving the existing policies." I thought this was the *considerate* way to communicate. But from the committee's perspective, it must have sounded like "You're doing it wrong. Here's a bunch more work you have to do to fix it. You're welcome!" They were actually much *more* receptive once I wrote up a revised set of policies that I'd be happier with. Once again, being "unbossy" and hoping that collaborative discussion would resolve the issue was a *total failure,* because people had less bandwidth to engage in discussion than I'd anticipated.

Private Discussions Are A Flawed Solution

I've noticed that in a lot of deliberative bodies or organizations, the real decision-making doesn't happen in groups. (*Meanwhile Madison is grappling with the fact that/ Not every issue can be settled in committee.*) The people who have "real power" meet in private and hash things out off the record. Nobody really shares their full thoughts on the internet or on an email list. It's not necessarily "secrecy", but it's secrecy-adjacent.

I know this is how things are frequently done, but it bothers me. When an issue is *officially the jurisdiction of a committee*, everyone on the committee is equally entitled to be part of the discussion, and entitled to know what's going on; having secret side conversations creates a hierarchy between those "in the know" and those who aren't. (*No-one else was in the room where it happened/ the room where it happened/ the room where it happened.)* Still more, when your project is supposed to *be for the sake of, and with the participation of, a broader community*, it seems like fairness demands being transparent with that community.

Maybe this is just the geek-kid issue, or what people today tend to call the geek social fallacies. I'm deeply uncomfortable when I see what looks like an elite subgroup, a group of "cool kids" or "VIPs" or whatever, talking behind closed doors because *hoi polloi* just wouldn't understand. I mean, yes, sometimes people *wouldn't* understand! I get it. There do exist people who will be offended by my honest opinion (god knows), or who literally aren't bright enough or knowledgeable enough to contribute to a discussion. I understand why it's easier to talk in private with people who are already more-or-less on the same page. But still…there's a pattern that gives me the willies. It's "elites get to know what's going on, randos are kept out of the loop," and even when somebody says that I qualify as an elite, not a rando, it still bothers me, because I'm much more comfortable *having rights* than *being favored*.

This is part of what gives me a bad feeling about the discourse around "demon threads" (that is: big, addictive, internet debates) and in praise of "taking things private", where tensions will be easier to defuse. There are real costs to acrimonious debate, in time and emotional energy, and I appreciate that people are trying to find ways to reduce those costs. But I feel nervous about anything that looks like it's trying to *sweep real conflicts under the rug*. It's like "don't fight in front of the

children" — except that in this case the members of the public are being placed in the role of "the children," whether or not we want to be.

I occasionally find myself in situations where I feel I'm being asked to take a sort of Straussian stance — *if you want to get important things done, you can't be totally transparent about what you're doing, because the general public will stop you*. I'm not sure these people are wrong. But I really hope they are. I have a bad feeling about maintaining information asymmetries as a general policy. I have a dangerous temperamental temptation towards concealment — it's just "minor" stuff like trying to hide my failures, but in the long run, that's neither ethical nor practical — so I've developed a counter-tendency towards transparency, as a sort of partial safeguard. If I tell people what I'm up to, early and often, I can't slip down the road of dishonesty.

Therapeutic Language: Another Flawed Solution

Peace is good, all things being equal. Fighting *hurts*. And many fights are unnecessary, borne of misunderstanding more than actual disagreement. I've seen this a lot firsthand. It's much more likely that someone *literally doesn't comprehend* your idea than that they oppose it.

And one of the most common types of misunderstanding is when people *falsely assume you are damning them as a person*. This is something I learned from Malcolm Ocean, who gave me the first really clear explanation I ever got as to what people are doing when they use NVC or Circling language or other types of very careful and mannered speech to avoid the perception of blame or judgment. Surely, I asked him, sometimes you *do* need to judge? To distinguish between good and bad behavior? To enforce norms?

After a while, we came up with this analogy:

There's a difference between saying "You're fired" and "You're fired, and also fuck you."

In the course of life, one absolutely *does* have to say things like "you're fired." Or "you can't behave like that in this space", "this work does not merit publication", or "I don't want to go on a date with you." In other words, drawing boundaries is necessary for life. But drawing boundaries doesn't always have to involve *damning* someone, as though sending them to Hell, utterly *condemning their essential being*. (What Madeleine l'Engle would call *X-ing*.) One can fire a person from a job, or reject their manuscript, or turn them down romantically, without saying *it is bad that you exist and you should hate yourself*. One can even, I believe, convict someone of a crime, or kill them in self-defense, without damning them, while wishing that they had not done the thing that forced you to draw an extremely severe boundary.

Boundaries are necessary; self-defense is necessary; damning people might *not* be necessary, and I'm inclined to believe it isn't.

And yet, people *do* damn each other, very frequently; and even more frequently, as a result of these bad experiences, they *assume* they're being damned when they're merely being criticized. "You did a thing with negative consequences" gets read as "your essence is stained, you are a Terrible Person, it's time to hate yourself." So, as an imperfect attempt to forestall these misunderstandings, people have developed these extremely artificial locutions that, yes, make you sound like a therapist, and, yes, aren't as natural as just speaking in plain language. But the hope is that they

create enough distance to allow people to avoid *immediately* jumping to the conclusion that you're accusing them of being Generally Terrible and Worthy of Eternal Hellfire.

Of course, the human mind being devious and wily at figuring out how to make us miserable, it's possible to be easily set off by therapeutic language itself! It turns out I have such a sensitivity. "You're insinuating that I'm having bad feelings — this means you're saying that I'm Weak and Can't Hack It and need Special Treatment — which means you're calling me Generally Terrible! Screw you!" (This isn't completely irrational; it is the appropriate norm for situations like work or school, where hiding physical and mental pain is expected and where people *are* penalized for failing to do so.)

Now, of course, I *do* have bad feelings sometimes, being a human. And, a lot of the time, the person using therapeutic language is trying to *deal productively with* that fact of the matter, rather than condemning me for it — they've moved on to Step 2, What Do We Do Now, while I'm still on Step 1, Is Sarah Terrible Y/N?

But you really can't have good conversations while anyone's still on Step 1. If you haven't yet resolved "Do You Think I'm Terrible?" with a resounding "No," then *every other conversation that's nominally about some topic will actually be about the vital issue of Do You Think I'm Terrible?*

And, because the human mind is devious, Step 1 doesn't stay resolved; you have to *keep reaffirming it*, because people will forget. You have to put what seems like a *colossal* amount of unsubtle effort into saying "I like you and I think you're good" in order to keep discussions from becoming about "I'm good and not terrible! See, I'll prove it!"

I have not *mastered* this art, or even close, but I basically agree with the need for it.

I have totally observed people being blunt and irreverent without hurting others' feelings and while getting very productive discussions done — but I think what's going on is not that these people *don't* validate each other, but that they validate each other *very well* through different *means* than therapeutic language. Some people can get away with speaking styles that are very "offensive" by conventional standards, but that's because they *also* show deep affection and regard for the people they're talking to.

I think there are people who are more robust than others at independently maintaining a sense that they're Okay and Good and Liked and Valid (and that's great!) but I don't think this in any way *disproves the need* for validation, any more than the existence of plants proves that organisms don't need chemical energy.

Nobody (Exactly) Agrees With You

I've been struggling a bunch with the fact that people seem to disagree *fractally and at every turn*. It's really, really hard to get exact alignment on worldviews and desires, to the point that I'm beginning to doubt it's possible. I see someone who seems to see *part* of the world the same way I do, and I go "can we talk? can we be buds? can we be twinsies? are we on the same team?" and then I realize "oh, no, outside of this tiny little area, they…really don't agree with me at all. Dammit."

It would be nice to have *someone to talk to who was basically the same person as you*, right? Someone you could just melt into, the way all of humanity melted into a

single sea of neon-orange thought-fluid in that anime.

But, in my experience, that just keeps not happening. Friendship and mutual respect, sure, I'm very fortunate to have lots of that; but merging doesn't happen. There's always me, or the other person, saying "no, not exactly" instead of "yes, and".

Is it just that I'm unusual? Surely people who build movements get people to agree with each other?

The thing is, I'm starting to suspect they *don't*. I recently went to TEDWomen, and saw a bunch of talks about activism and organizing, including by such luminaries as Dolores Huerta and Marian Wright Edelman. And here are some takeaways I got from them:

- Activists view the main goal as *fighting apathy*, that is, getting people to participate, literally *activating* people. Getting people to show up to vote or show up to a protest or to raise issues in conversations.
- *Everybody* in a coalition supports *everybody* else. It's very "all for one and one for all." They explicitly talk about how you shouldn't allow anyone to frame things as "the environment" vs "women's issues" vs "labor issues" vs "immigration" — everyone's encouraged to push for everyone's agenda together, for every sub-group in the progressive coalition.
- Activists *endorse* being moved more by individual stories and art and emotional appeals than by facts and figures. They don't just talk about how "emotional appeals work better on the public" but they talk about how emotional appeals and personal connections work *on themselves*.

If you think of everybody's beliefs as a forest of trees, where consequences branch out from premises, then "trying to get agreement" is building trees as big as they can get and trying to hash out what's going on when two people's trees differ. What seems to be going on in an activist frame is *not building out the trees very big at all*, only getting agreement on rather basic things like "children shouldn't live in poverty" and trying to move straight to voting and fundraising and other object-level actions, without really hashing out in much detail "ok, what ways of avoiding child poverty are effective and/or morally acceptable?" They recognize that getting people to participate at all is difficult (in my shoes, they would have invested a lot more effort in getting people to show up to the event), and they don't seem to even try to get people to *agree* in a deep sense, to agree on world-models and general principles and moral foundations.

Just because everyone is shouting the same slogan doesn't mean they *really agree with each other*. They agree on the slogan. It might mean different things to different people. That's not necessarily a bad thing, but it's worth being aware that it isn't true unity.

The Greek for "with one accord" is ὁμοθυμαδόν, which appears frequently in the New Testament; it means literally "same passion" or "same spirit", the seat of courage and emotion that lives in the heart. "Unanimity" is an exact translation into Latin — "one spirit." You can have large groups of people who *feel* the same, who are filled with the same passion. It is much harder for all those people to have the same *belief structure,* to stay on the same page on the nitty-gritty details. Just getting groups of people to "weak unanimity," namely, active participation, good will, and agreement on ideal goals, is a challenging full-time job by itself — and it doesn't even *touch* getting worldview alignment.

The Cost of Complaint

One weird and maybe trivial thing that's been nagging at me is trying to get a handle on the underlying worldview expressed by the *Incredibles* movies. Yeah, it's pop culture, but there's clearly an attempt to communicate a moral, and it's a *weird* one.

Sure, there's the inspiring, defiant pro-superhero note of "people shouldn't be pressured to hide their excellence", which often gets labeled Randian (but could just as easily be Nietzchean or *Harrison Bergeron*-esque).

But it gets weird when you look at the villains. The villains of both movies are genius technologists. Syndrome, the villain of the first movie, is a bitter, pimpled male nerd, resentful of superheroes' elevated status, who wants to provide technology to give everyone superpowers. Evelyn Deaver, the villain of the second movie, is a bitter, urbane, worldly feminist, a technologist who dislikes the way technology has "dumbed down" its users, resentful of the public's passive reliance on screens and superheroes. For plot reasons, of course, both supervillains pull dangerous stunts that put the public at risk, and need to be stopped by the superheroes. But their motivations are actually *empowering humanity,* weirdly enough. Syndrome is, effectively, a transhumanist, while Evelyn is an "ethical techie" type reminiscent of the people at the Center for Humane Technology. Their obsession is using their talents and hard work to make all people *more self-reliant and capable of greater things* — a mission that would actually sit well with Rand or Nietzsche, and, outside the world of the films, could easily work as a heroic cause.

What's wrong with the villains, in the world of *The Incredibles*, is that they're *grouchy*. They're social critics. They complain.

Notice that, before we know she's a villain, Evelyn tries to get Mrs. Incredible to *commiserate* about sexism; the heroine doesn't take the bait, and points out that Evelyn is also standing in her brother's shadow. Before *his* villainous reveal, Syndrome is a whiny kid who wants to be Mr. Incredible's sidekick. And the initial controversy that drove superheroes underground was a suicidal man who sued Mr. Incredible for saving his life.

Also, notice that Brad Bird is taking a *very* firm stance in favor of optimism and against gloom, in the *Incredibles* movies and others; his movies overtly defend his creative choice to keep things positive and brightly colored in a world where critical acclaim usually comes in shades of gray. (The antagonist in *Ratatouille,* not accidentally, is a restaurant critic.) I think it's really that simple: Brad Bird likes unity and positivity, and doesn't like complaining. Critics like the *New Yorker's* Richard Brody are right to see a threat in the movies — their real enemy is *criticism*.

(If you look at Brad Bird's actual words, he isn't any kind of a libertarian or Randian, and says so; he's a centrist, he's big on finding common ground, staying positive, focusing on unity, and so on.)

It's almost impossible to talk about the world intelligently while refraining from any complaint. Try finding a blog to read that *never criticizes society,* from any direction. Where you find interesting and articulate people, you'll find people who express dissatisfaction with things as they are. There's no *principled* way to say "hey I think everyone's pretty much right," because people don't remotely agree with each other if you ask about any details at all.

And yet, people (like Bird, but also like me, and like many) get heartsick when we're exposed to too much complaint or disagreement. Moods are contagious, and criticism *is* very often depressing, for all we try to tell ourselves that it's merely an intellectual awareness. Sometimes I feel like "for god's sake, World, for once could you give me a social context where *literally nobody expresses dislike or disapproval about anything*? Could we have a Happy Zone please?"

But I'm genuinely not sure if that's possible. It may be a feature of language or logic itself that it's hard to talk at all if you restrict yourself firmly to avoiding critical speech. I certainly would have a hard time sticking strictly to Happy Zone rules.

I don't have solutions here. I'm just trying to figure things out. It ought to be possible, I think, to deliberate and collaborate *with* people, allowing "the group" to decide, rather than just deciding what *I* want individually and letting people collaborate with me to the extent that it sounds good to them. I know how to be an individualist; I'm trying to learn how to also do the collective thing, "voice" rather than "exit". But I'm just stumped by the fact that *people want different things, and think different things, and actual, far-reaching unity doesn't seem to exist*.

Reasons compute may not drive AI capabilities growth

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

How long it will be before humanity is capable of creating general AI is an important factor in discussions of the importance of doing AI alignment research as well as discussions of which research avenues have the best chance of success. One frequently discussed model for estimating AI timelines is that AI capabilities progress is essentially driven by growing compute capabilities. For example, the OpenAI article on AI and Compute presents a compelling narrative, which shows a trend of well-known results in machine learning using exponentially more compute over time. This is an interesting model because if valid we can do some quantitative forecasting, due to somewhat smooth trends in compute metrics which can be extrapolated. However, I think there are a number of reasons to suspect AI progress to be driven more by engineer and researcher effort than compute.

I think there's a spectrum of models between:

- We have an abundance of ideas that aren't worth the investment to try out yet. Advances in compute capability unlock progress by make researching more expensive techniques economically feasible. We'll be able to create general AI soon after we have enough compute to do it.
- Research proceeds at its own pace and makes use of as much compute is convenient to save researcher time on optimization and achieve flashy results. We'll be able to create general AI once we come up with all the right ideas behind it, and either:
    - We'll already have enough compute to do it
    - We won't have enough compute and we'll start optimizing, invest more in compute, and possibly start truly being bottlenecked on compute progress.

My research hasn't pointed too solidly in either direction, but below I discuss a number of the reasons I've thought of that might point towards compute not being a significant driver of progress right now.

There's many ways to train more efficiently that aren't widely used

Starting October of 2017, the Stanford DAWNBench contest challenged teams to come up with the fastest and cheapest ways to train neural nets to solve certain tasks.

The most interesting was the ImageNet training time contest. The baseline entry took 10 days and cost $1112; less than one year later the best entries (all by the fast.ai team) were down to 18 minutes for $35, 19 minutes for $18 or 30 minutes for $14[^1]. This is ~800x faster and ~80x cheaper than the baseline.

Some of this was just using more and better hardware, the winning team used 128 V100 GPUs for 18 minutes and 64 for 19 minutes, versus eight K80 GPUs for the baseline. However, substantial improvements were made even on the same hardware.

The training time on a `p3.16xlarge` AWS instance with eight V100 GPUs went down from 15 hours to 3 hours in 4 months. The training time on a single Google Cloud TPU went down from 12 hours to 3 hours as the Google Brain team tuned their training and incorporated ideas from the fast.ai team. An even larger improvement was seen on the CIFAR10 contest recently, with times on a `p3.2xlarge` improving by 60x with the accompanying blog series still mentioning multiple improvements left on the table due to effort constraints. He also speculates that many of the optimizations would also improve the ImageNet version.

The main techniques used for fast training were all known techniques: progressive resizing, mixed precision training, removing weight decay from batchnorms, scaling up batch size in the middle of training, and gradually warming up the learning rate. They just required engineering effort to implement and weren't already implemented in the library defaults.

Similarly, the improvement due to scaling from eight K80s to many machines with V100s was partially hardware but also required lots of engineering effort to implement: using mixed precision fp16 training (required to take advantage of the V100 Tensor Cores), efficiently using the network to transfer data, implementing the techniques required for large batch sizes, and writing software for supervising clusters of AWS spot instances.

These results seem to show that it's possible to train much faster and cheaper by applying knowledge and sufficient engineering effort. Interestingly not even a team at Google Brain working to show off TPUs initially had all the code and knowledge required to get the best available performance, and had to gradually work for it.

I would suspect that in a world where we were bottlenecked hard on training times that these techniques would be more widely known about and applied, and implementations of them readily available for every major machine learning library. Interestingly, in postscripts to both of his articles on how fast.ai managed to achieve such fast times, Jeremy Howard notes that he doesn't believe large amounts of compute are required for important ML research, and notes that many foundational discoveries were available with little compute.

[^1]: Using spot/preemptible instance pricing instead of the on-demand pricing the benchmark page lists, due to much lower prices and the lack of need for on-demand instances given the short time. The authors of the winning solution wrote software to effectively use spot instances and actually used them for their tests. It may seem unfair to use spot prices for the winning solution but not for the baseline, but a lot of the improvement in the contest came from actually using all the techniques for faster/cheaper training available despite inconvenience, and they had to write software to easily use spot instances and had short enough training times that it was viable without fancy software to automatically transfer training to new machines.

Hyperparameter grid searches are inefficient

I've heard hyperparameter grid searches mentioned as a reason why ML research needs way more compute than it would appear based on the training time of the models used. However, I can also see the use of grid searches as evidence of an abundance of compute rather than a scarcity.

As far as I can tell it's possible to find hyperparameters much more efficiently than a grid search, it just takes more human time and engineering implementation effort. There's a large literature of more efficient hyperparameter search methods but as far as I can tell they aren't very popular (I've never heard of anyone using one in practice, and all open source implementations of these kind of things I can find have few Github stars).

Researcher Leslie Smith also has a number of papers with little-used ideas on principled approaches to choosing and searching for optimal hyperparameters with much less effort, including a fast automatic procedure for finding optimal learning rates. This suggests that it's possible to substitute hyperparameter search time for more engineering, human decision-making and research effort.

There's also likely room for improvement in the factorization of the hyper-parameters we use so that they're more amenable to separate optimization. For example, L2 regularization is usually used in place of weight decay because they theoretically do the same thing, but this paper points out that not only do they not do the same thing with ADAM and using weight decay causes ADAM to surpass the more popular SGD with momentum in practice, but that weight decay is a better hyper-parameter since the optimal weight decay is more independent of learning rate than L2 regularization strength is.

All of this suggests that most researchers might be operating under an abundance of cheap compute relative to their problems that leads to them not investing the effort required to more efficiently optimize their hyperparameters and just do so haphazardly or with grid searches instead.

The types of compute we need may not improve very quickly

Improvements in computing hardware are not uniform and there are many different hardware attributes that can be bottlenecks for different things. AI progress may rely on one or more of these that don't end up improving quickly, becoming bottlenecked on the slowest one rather than experiencing exponential growth.

Machine learning accelerators

Modern machine learning is largely composed of large operations that are either directly matrix multiplies or can be decomposed into them. It's also possible to train using much lower precision than full 32-bit floating point using some tricks. This allows the creation of specialized training hardware like Google's TPUs and Nvidia Tensor Cores. A number of other companies have also announced they're working on custom accelerators.

The first generation of specialized hardware delivered a large one-time improvement, but we can also expect continuing innovation in accelerator architecture. There will likely be sustained innovations in training with different number formats and architectural optimizations for faster and cheaper training. I expect this will be the area our compute capability will grow the most, but may flatten like CPUs have once we figure out enough of the easily discoverable improvements.

CPUs

Reinforcement learning simulations like the OpenAI Five DOTA bot, and various physics playgrounds, often use CPU-heavy serial simulations. OpenAI Five uses 128,000 CPU cores and only 256 GPUs. At current Google Cloud preemptible prices the CPUs cost 5-10x more than the GPUs in total. Improvements in machine learning training ability will still leave the large cost of the CPUs. If the use of expensive simulations that run best on CPUs becomes an important part of training advanced agents, progress may become bottlenecked on CPU cost.

Additionally, improvement in CPU compute costs may be slowing. Cloud CPU costs only decreased 45% from 2012 to 2017 and performance per dollar for buying the hardware only improved 2x.. Google Cloud Compute prices have only dropped 25% from 2014-2018. Although the introduction of preemptible prices 30% of full price in 2016 was a big improvement, and that decreased to 20% of full price in 2017.

GPU/accelerator memory

Another scarce resource is memory on the GPU/accelerator used for training. The memory must be large enough to store all the model parameters, the input, the gradients, and other optimization parameters.

This is one of the most frequent limits I see referenced in machine learning papers nowadays. For example the new large BERT language model can only be trained properly on TPUs with their 64GB of RAM. The Glow paper needs to use gradient checkpointing and an alternative to batchnorm so that they can use gradient accumulation, because only a single sample of gradients fits on a GPU.

However there are ways to address this limitation that aren't frequently used. Glow already uses the two best ones, gradient checkpointing and gradient accumulation, but did not implement an optimization they mentioned which would make the amount of memory the model takes constant in the number of layers instead of linear, likely because it would be difficult to engineer into existing ML frameworks. The BERT implementation uses none of the techniques because they just use a TPU with enough memory, in fact a reimplementation of BERT implemented 3 such techniques and got it to fit on a GPU. Thus it still seems that in a world with less RAM these might still have happened, just with more difficulty or smaller demonstration models.

Interestingly, the maximum available RAM per device barely changed from 2014 through 2017 with the NVIDIA K80's 24GB, but then shot up in 2018 to 48GB with the RTX 8000 as well as the 64GB TPU v2 and 128GB TPU v3. Probably both because of demand for larger device memories for machine learning training, as well as the availability of high capacity HBM memory. It's unclear to me if this rapid rise will continue or if it was mostly a one-time change reflecting new demands for the largest possible memories reaching the market.

It's also possible that per-device memory will cease to be a constraint on model size due to faster hardware interconnects that allow sharing a model across the memory of multiple devices like Intel's Nervana and Tensorflow Mesh plan to do. It also seems likely that techniques for splitting models across devices to fit in memory, like the original AlexNet did, will become more popular. It may be the case that the fact that we don't split models across devices like AlexNet anymore is evidence that we're not constrained by RAM much but I'm not sure.

Limited ability to exploit parallelism

As discussed extensively in a new paper from Google Brain, there seems to be a limit on how much data parallelism in the form of larger batch sizes we can currently extract out of a given model. If this constraint isn't worked around, wall time to train models could stall even if compute power continues to grow.

However the paper mentions that various things like model architecture and regularization affect this limit and I think it's pretty likely that techniques to increase this limit will continue to be discovered so it isn't a bottleneck. A newer paper by OpenAI finds that more difficult problems also tolerate larger batch sizes. Even if the limit remains, increasing compute would allow training more different models in parallel, potentially just meaning that more parameter search and evolution gets layered on top of the training. I also suspect that just using ever-larger models may allow use of more compute without increasing batch sizes.

At the moment, it seems that we know how to train effectively with batch sizes large enough to saturate large clusters, for example this paper about training ImageNet in 7 minutes with a 64k batch size. But this requires extra tuning and implementing some tricks, even just to train on mid-size clusters, so as far as I know only a small fraction of all machine learning researchers regularly train on large clusters (anecdotally, I'm uncertain about this).

Conclusion

These all seem to point towards compute being abundant and ideas being the bottleneck, but not solidly. For the points about training efficiency and grid searches this could just be an inefficiency in ML research and all the major AGI progress will be made by a few well-funded teams at the boundaries of modern compute that have solved these problems internally.

Vaniver commented on a draft of this post that it's interesting to consider the case where training time is the bottleneck rather than ideas, but massive engineering effort is highly effective at reducing training time. In this case an increase in investment in AI research which lead to hiring more engineers to apply techniques to speed up training could lead to rapid progress. This world might also lead to more sizable differences in capabilities between organizations, if large somewhat serial software engineering investments are required to make use of the most powerful techniques, rather than a well-funded newcomer being able to just read papers and buy all the necessary hardware.

The course of various compute hardware attributes seems uncertain both in terms of how fast they'll progress and whether or not we'll need to rely on anything other than special-purpose accelerator speed. Since the problem is complex with many unknowns, I'm still highly uncertain, but all of these points did move me to varying degrees in the direction of continuing compute growth not being a driver of dramatic progress.

Thanks to Vaniver and Buck Shlegeris for discussions that lead to some of the thoughts in this post.

Conceptual Analysis for AI Alignment

TL; DR - Conceptual Analysis is highly relevant for AI alignment, and is also a way in which someone with less technical skills can contribute to alignment research. This suggests there should be at least one person working full-time on reviewing existing philosophy literature for relevant insights, and summarizing and synthesizing these results for the safety community.

There are certain "primitive concepts" that we are able to express in mathematics, and it is *relatively* straightforward to program AIs to deal with those things. Naively, alignment requires understanding *all* morally significant human concepts, which seems daunting. However, the "argument from corrigibility" suggests that there may be small sets of human concepts which, if properly understood, are sufficient for "benignment". We should seek to identify what these concepts are, and make a best-effort to perform thorough and reductive conceptual analyses on them. But we should also look at what has already been done!

On the coherence of human concepts

For human concepts which *haven't* been formalized, it's unclear whether there is a simple "coherent core" to the concept. Careful analysis may also reveal that there are several coherent concepts worth distinguishing, e.g. cardinal vs. ordinal numbers. If we find there is a coherent core, we can attempt to build algorithms around it.

If there isn't a simple coherent core, there may be a more complex one, or it may be that the concept just isn't coherent (i.e. that it's the product of a confused way of thinking). Either way, in the near term we'd probably have to use machine learning if we wanted to include these concepts in our AI's lexicon.

A serious attempt at conceptual analysis could help us decide whether we should attempt to learn or formalize a concept.

Concretely, I imagine a project around this with the following stages (each yielding at least one publication):

1) A "brainstormy" document which attempts to enumerate all the concepts that are relevant to safety and presents the arguments for their specific relevance and relation to other relevant concepts. This should also specifically indicate how a combination of concepts, if rigorously analyzed, could be along the line of the argument from corrigibility. Besides corrigibility, two examples that jump to mind are "reduced impact" (or "side effects"), and interpretability.

2) A deep dive into the relevant literature (I imagine mostly in analytic philosophy) on each of these concepts (or sets of concepts). These should summarize the state of research on these problems in the relevant fields, and potentially inspire safety researchers, or at least help them frame their work for these audiences and find potential collaborators within these fields. It *might* also do some "legwork" in terms of formalizing logically rigorous notions in terms of mathematics or machine learning.

3) Attempting to transferring insights or ideas from these fields into technical AI safety or machine learning papers, if applicable.

ETA: it's worth noting that the notion of "fairness" is currently undergoing intense conceptual analysis in the field of ML. See recent tutorials at ICML and NeurIPS, as well as work on counter-factual notions of fairness (e.g. Silvia Chiappa's).

What are some concrete problems about logical counterfactuals?

Logical counterfactuals are key to Functional Decision Theory and last I heard still an unsolved problem. Unfortunately, I am still rather confused about what exactly we are trying to solve. The only concrete problem I know of in this space is the 5-and-10 problem. But as far as I know, this is solved by writing programs that immediately cause a paradox if they ever discover their output. So presumably there are some unsolved concrete problems that relate to logical counterfactuals?

Edit: I should mention my post on the Cooperation Game as an example. Plus the further work section of this slideshow.

Internet Search Tips: how I use Google/Google Scholar/Libgen

This is a linkpost for https://www.gwern.net/Search

You can be wrong about what you like, and you often are

Meta: I'm not saying anything new here. There has been a lot of research on the topic, and popular books like Stumbling on Happiness have been written. Furthermore, I don't think I have explained any of this particularly well, or provided particularly enlightening examples. Nevertheless, I think these things are worth saying because a) a lot of people have an "I know what I like" attitude, and b) this attitude seems pretty harmful. Just be sure to treat this as more of an exploratory post than an authoritative one.

I think that the following attitudes are very common:

- I'm just not one of those people who enjoys "deeper" activities like reading a novel. I like watching TV and playing video games.
- I'm just not one of those people who likes healthy foods. *You* may like salads and swear by them, but *I* am different. *I* like pizza and french fries.
- I'm just not an intellectual person. I don't enjoy learning.
- I'm just not into that early retirement stuff. I need to maintain my current lifestyle in order to be happy.
- I'm just not into "good" movies/music/art. I like the Top 50 stuff.

Imagine what would happen if you responded to someone who expressed one of these attitudes by saying "I think that you're wrong." Often times, the response you'll get is something along the lines of:

> Who are you to tell me what I do and don't like? How can you possibly know? *I'm* the one who's in my own head. *I* know how these things make me feel.

When I think about that response, I think about optical illusions. Consider this one:

When I think about that response, I think about the following dialog:

> Me: A and B are the same shade of gray.
>
> Person: No they're not! WTF are you talking about? How can you say that they are? I can *see* with my *eyes* that they're *not*!

I understand the frustration. It *feels* like they're different shades. It feels like it is stupidly obvious that they're different shades.

And if feels like you know what you like.

But sometimes, sometimes your brain lies to you.

The image of the squares is an optical illusion. Neuroscientists and psychologists study them. And they write books like this about them.

The question of knowing what you like can be a *hedonic* illusion (that's what I'll decide to call it anyway). Neuroscientists and psychologists study these illusions too. And

they write books like this about them.

They have found that we're actually really bad at knowing what will make us happy. At knowing what we do, and don't like.

Some quotes from *The Science of Happiness*:

- “One big question was, Are beautiful people happier?” Etcoff says. “Surprisingly, the answer is no! This got me thinking about happiness and what makes people happy.”
- His book *Stumbling on Happiness* became a national bestseller last summer. Its central focus is “prospection”—the ability to look into the future and discover what will make us happy. The bad news is that humans aren’t very skilled at such predictions; the good news is that we are much better than we realize at adapting to whatever life sends us.
- The reason is that humans hold fast to a number of wrong ideas about what will make them happy. Ironically, these misconceptions may be evolutionary necessities. “Imagine a species that figured out that children don’t make you happy,” says Gilbert. “We have a word for that species: *extinct.*

That last one was pretty powerful, wow.

I think that the implications of this are all pretty huge. We all want to be happy. We all want to thrive. We make thousands and thousands of little decisions to this end. We decide to have fried chicken for dinner, and that having salads isn't worth the effort, despite whatever long term health benefits. We decide that video games are a nice, fun, relaxing way to decompress after work. We decide that working a corporate job is worth it because we "need the money".

All of these decisions shape our lives. If we're getting them wrong, well, then we're not doing a good job of shaping our lives.

And if we're basing these decisions off of our *intuitions*, according to the positive psychology research, we're probably screwing up a lot.

So then, I propose that we approach these sorts of questions with more **curiosity**.

> The first virtue is curiosity. A burning itch to know is higher than a solemn vow to pursue truth. To feel the burning itch of curiosity requires both that you be ignorant, and that you desire to relinquish your ignorance. *If in your heart you believe you already know*, or if in your heart you do not wish to know, then your questioning will be purposeless and your skills without direction. Curiosity seeks to annihilate itself; there is no curiosity that does not want an answer. The glory of glorious mystery is to be solved, after which it ceases to be mystery. Be wary of those who speak of being open-minded and modestly confess their ignorance. There is a time to confess your ignorance and a time to relinquish your ignorance.

And with more **humility**.

> The eighth virtue is humility. To be humble is to take specific actions in anticipation of your own errors. To confess your fallibility and then do nothing about it is not humble; it is boasting of your modesty. Who are most humble? Those who most skillfully prepare for the deepest and most catastrophic errors in their own beliefs and plans. Because this world contains many whose grasp of rationality is abysmal, beginning students of rationality win arguments and

> acquire an exaggerated view of their own abilities. But it is useless to be superior: Life is not graded on a curve. The best physicist in ancient Greece could not calculate the path of a falling apple. There is no guarantee that adequacy is possible given your hardest effort; therefore spare no thought for whether others are doing worse. If you compare yourself to others you will not see the biases that all humans share. To be human is to make ten thousand errors. No one in this world achieves perfection.

You can be wrong about what you like, and you often are.

https://www.youtube.com/watch?v=KmC1btSZP7U

(My grandpa used to read this to me all of the time when I was younger. And he still bugs me about it to this day. It's cool that I'm finally starting to understand it.)

Best arguments against worrying about AI risk?

Since so many people here (myself included) are either working to reduce AI risk or would love to enter the field, it seems worthwhile to ask what are the best arguments against doing so. This question is intended to focus on existential/catastrophic risks and not things like technological unemployment and bias in machine learning algorithms.

Should ethicists be inside or outside a profession?

Originally written in 2007.

Marvin Minsky in an interview with Danielle Egan for *New Scientist:*

> **Minsky:** The reason we have politicians is to prevent bad things from happening. It doesn't make sense to ask a scientist to worry about the bad effects of their discoveries, because they're no better at that than anyone else. Scientists are not particularly good at social policy.
>
> **Egan:** But shouldn't they have an ethical responsibility for their inventions
>
> **Minsky:** No they shouldn't have an ethical responsibility for their inventions. They should be able to do what they want. You shouldn't have to ask them to have the same values as other people. Because then you won't get them. They'll make stupid decisions and not work on important things, because they see possible dangers. What you need is a separation of powers. It doesn't make any sense to have the same person do both.

The Singularity Institute was recently asked to comment on this interview - which by the time it made it through the editors at *New Scientist*, contained just the unvarnished quote "Scientists shouldn't have an ethical responsibility for their inventions. They should be able to do what they want. You shouldn't have to ask them to have the same values as other people." Nice one, *New Scientist.* Thanks to Egan for providing the original interview text.

This makes an interesting contrast with what I said in my "Cognitive biases" chapter for Bostrom's *Global Catastrophic Risks:*

> Someone on the physics-disaster committee should know what the term "existential risk" means; should possess whatever skills the field of existential risk management has accumulated or borrowed. For maximum safety, that person should also be a physicist. The domain-specific expertise and the expertise pertaining to existential risks should combine in one person. I am skeptical that a scholar of heuristics and biases, unable to read physics equations, could check the work of physicists who knew nothing of heuristics and biases.

Should ethicists be inside or outside a profession?

It seems to me that trying to separate ethics and engineering is like trying to separate the crafting of paintings into two independent specialties: a profession that's in charge of pushing a paintbrush over a canvas, and a profession that's in charge of artistic beauty but knows nothing about paint or optics.

The view of ethics as a separate profession is part of the *problem.* It arises, I think, from the same deeply flawed worldview that sees technology as something foreign and distant, something *opposed to* life and beauty. Technology is an expression of human intelligence, which is to say, an expression of human nature. Hunter-gatherers who crafted their own bows and arrows didn't have cultural nightmares about bows

and arrows being a mechanical death force, a blank-faced System. When you craft something with your own hands, it seems like a part of you. It's the Industrial Revolution that enabled people to buy artifacts which they could not make or did not even understand.

Ethics, like engineering and art and mathematics, is a natural expression of human minds.

Anyone who gives a part of themselves to a profession discovers a sense of beauty in it. Writers discover that sentences can be beautiful. Programmers discover that code can be beautiful. Architects discover that house layouts can be beautiful. We all start out with a native sense of beauty, which already responds to rivers and flowers. But as we begin to *create* - sentences or code or house layouts or flint knives - our sense of beauty develops with use.

Like a sense of beauty, one's native ethical sense must be continually used in order to develop further. If you're just working at a job to make money, so that your real goal is to make the rent on your apartment, then neither your aesthetics nor your morals are likely to get much of a workout.

The way to develop a highly specialized sense of professional ethics is to do something, ethically, a whole bunch, until you get good at both the thing itself and the ethics part.

When you look at the "bioethics" fiasco, you discover bioethicists writing mainly for an audience of other bioethicists. Bioethicists aren't writing to doctors or bioengineers, they're writing to tenure committees and journalists and foundation directors. Worse, bioethicists are not *using* their ethical sense in bio-work, the way a doctor whose patient might have incurable cancer must choose how and what to tell the patient.

A doctor treating a patient should not try to be *academically original,* to come up with a brilliant new theory of bioethics. As I've written before, ethics is not *supposed* to be counterintuitive, and yet academic ethicists are biased to be just exactly counterintuitive enough that people won't say, "Hey, I could have thought of that." The purpose of ethics is to shape a well-lived life, not to be impressively complicated. Professional ethicists, to get paid, must transform ethics into something difficult enough to require professional ethicists.

It's, like, a good idea to save lives? "Duh," the foundation directors and the review boards and the tenure committee would say.

But there's nothing *duh* about saving lives if you're a doctor.

A book I once read about writing - I forget which one, alas - observed that there is a level of depth beneath which repetition ceases to be boring. Standardized phrases are called "cliches" (said the author of writing), but murder and love and revenge can be woven into a thousand plots without ever becoming old. "You should save people's lives, mmkay?" won't get you tenure - but as a theme of real life, it's as old as thinking, and no more obsolete.

Boringly obvious ethics are just fine if you're *using* them in your work rather than talking about them. The goal is to do it right, not to do it originally. Do your best whether or not it is "original", and originality comes in its own time; not every change is an improvement, but every improvement is necessarily a change.

At the Singularity Summit 2007, several speakers alleged we should “reach out” to artists and poets to encourage their participation in the Singularity dialogue. And then a woman went to a microphone and said: “I am an artist. I want to participate. What should I do?”

And there was a long, delicious silence.

What I would have said to a question like that, if someone had asked it of me in the conference lobby, was: “You are not an ‘artist’, you are a human being; art is only one facet in which you express your humanity. Your reactions to the Singularity should arise from your entire self, and it’s okay if you have a standard human reaction like ‘I’m afraid’ or ‘Where do I send the check?’, rather than some special ‘artist’ reaction. If your artistry has something to say, it will express itself naturally in your response as a human being, without needing a conscious effort to say something artist-like. I would feel patronized, like a dog commanded to perform a trick, if someone presented me with a painting and said ‘Say something mathematical!’”

Anyone who calls on “artists” to participate in the Singularity clearly thinks of artistry as a special function that is only performed in Art departments, an icing dumped onto cake from outside. But you can always pick up some cheap applause by calling for more icing on the cake.

Ethicists should be inside a profession, rather than outside, because ethics itself should be inside rather than outside. It should be a natural expression of yourself, like math or art or engineering. If you don’t like trudging up and down stairs you’ll build an escalator. If you don’t want people to get hurt, you’ll try to make sure the escalator doesn’t suddenly speed up and throw its riders into the ceiling. Both just natural expressions of desire.

There are opportunities for market distortions here, where people get paid more for installing an escalator than installing a safe escalator. If you don’t *use* your ethics, if you don’t wield them as part of your profession, they will grow no stronger. But if you want a safe escalator, by far the best way to get one - if you can manage it - is to find an engineer who naturally doesn’t want to hurt people. Then you’ve just got to keep the managers from demanding that the escalator ship immediately and without all those expensive safety gadgets.

The first iron-clad steamships were actually *much safer* than the *Titanic*; the first ironclads were built by engineers without much management supervision, who could design in safety features to their heart’s content. The *Titanic* was built in an era of cutthroat price competition between ocean liners. The grand fanfare about it being unsinkable was a marketing slogan like “World’s Greatest Laundry Detergent”, not a failure of engineering prediction.

Yes, safety inspectors, yes, design reviews; but these just *verify* that the engineer put forth an effort of ethical design intelligence. Safety-inspecting doesn’t build an elevator. Ethics, to be effective, must be part of the intelligence that expresses those ethics - you can’t add it in like icing on a cake.

Which leads into the question of the ethics of AI. “Ethics, to be effective, must be part of the intelligence that expresses those ethics - you can’t add it in like icing on a cake.” My goodness, I wonder how I could have learned such Deep Wisdom?

Because I studied AI, and the art spoke to me. Then I translated it back into English.

The truth is that I can't inveigh properly on bioethics, because I am not myself a doctor or a bioengineer. If there is a special ethic of medicine, beyond the obvious, I do not know it. I have not worked enough healing for that art to speak to me.

What I do know a thing or two about, is AI. There I can testify definitely and from direct knowledge, that anyone who sets out to study "AI ethics" without a technical grasp of cognitive science, is absolutely doomed.

It's the technical knowledge of AI that forces you to deal with the world in its own strange terms, rather than the surface-level concepts of everyday life. In everyday life, you can take for granted that "people" are easy to identify; if you look at the modern world, the humans are easy to pick out, to categorize. An unusual boundary case, like Terri Schiavo, can throw a whole nation into a panic: Is she "alive" or "dead"? AI explodes the language that people are described of, unbundles the properties that are always together in human beings. Losing the standard view, throwing away the human conceptual language, forces you to *think for yourself* about ethics, rather than parroting back things that sound Deeply Wise.

All of this comes of studying the math, nor may it be divorced from the math. That's not as comfortably egalitarian as my earlier statement that ethics isn't meant to be complicated. But if you mate ethics to a highly technical profession, you're going to get ethics expressed in a *conceptual language* that is highly technical.

The technical knowledge provides the conceptual language in which to express ethical problems, ethical options, ethical decisions. If politicians don't understand the distinction between terminal value and instrumental value, or the difference between a utility function and a probability distribution, then some fundamental *problems* in Friendly AI are going to be complete gibberish to them - never mind the solutions. I'm sorry to be the one to say this, and I don't like it either, but Lady Reality does not have the goal of making things easy for political idealists.

If it helps, the technical ethical thoughts I've had so far require only comparatively basic math like Bayesian decision theory, not high-falutin' complicated damn math like real mathematicians do all day. Hopefully this condition does not hold merely because I am stupid.

Several of the responses to Minsky's statement that politicians should be the ones to "prevent bad things from happening" were along the lines of "Politicans are not particularly good at this, but neither necessarily are most scientists." I think it's sad but true that modern industrial civilization, or even modern academia, imposes many shouting external demands within which the quieter internal voice of ethics is lost. It may even be that a majority of people are not particularly ethical to begin with; the thought seems to me uncomfortably elitist, but that doesn't make it comfortably untrue.

It may even be true that most scientists, say in AI, haven't really had a lot of opportunity to express their ethics and so the art hasn't said anything in particular to them.

If you talk to some AI scientists about the Singularity / Intelligence Explosion they may say something cached like, "Well, who's to say that humanity really ought to survive?" This doesn't sound to me like someone whose art is speaking to them. But then artificial intelligence is not the same as artificial *general* intelligence; and, well, to be brutally honest, I think a lot of people who claim to be working in AGI haven't really gotten all that far in their pursuit of the art.

So, if I listen to the voice of experience, rather to the voice of comfort, I find that most people are not very good at ethical thinking. Even most doctors - who ought properly to be confronting ethical questions in every day of their work - don't go on to write famous memoirs about their ethical insights. The terrifying truth may be that Sturgeon's Law applies to ethics as it applies to so many other human endeavors: "Ninety percent of everything is crap."

So asking an engineer an ethical question is not a sure-fire way to get an especially ethical answer. I wish it were true, but it isn't.

But what experience tells me, is that there is no way to obtain the ethics of a *technical* profession *except* by being ethical inside that profession. I'm skeptical enough of nondoctors who propose to tell doctors how to be ethical, but I *know* it's not possible in AI. There are all sorts of AI-ethical questions that anyone should be able to answer, like "Is it good for a robot to kill people? No." But if a dilemma requires more than this, the specialist ethical expertise will only come from someone who has practiced expressing their ethics from inside their profession.

This doesn't mean that all AI people are on their own. It means that if you want to have specialists telling AI people how to be ethical, the "specialists" have to be AI people who express their ethics within their AI work, and *then* they can talk to other AI people about what the art said to them.

It may be that most AI people will not be above-average at AI ethics, but without technical knowledge of AI you don't even get an *opportunity* to develop ethical expertise because you're not thinking in the right language. That's the way it is in my profession. Your mileage may vary.

In other words: To get good AI ethics you need someone technically good at AI, but not all people technically good at AI are automatically good at AI ethics. The technical knowledge is *necessary* but not *sufficient* to ethics.

What if you think there are specialized ethical concepts, typically taught in philosophy classes, which AI ethicists will need? Then you need to make sure that at least some AI people take those philosophy classes. If there is such a thing as special ethical knowledge, it has to *combine in the same person* who has the technical knowledge.

Heuristics and biases are critically important knowledge relevant to ethics, in my humble opinion. But if you want that knowledge expressed in a profession, you'll have to find a professional expressing their ethics and teach them about heuristics and biases - not pick a random cognitive psychologist off the street to add supervision, like so much icing slathered over a cake.

My nightmare here is people saying, "Aha! A randomly selected AI researcher is not guaranteed to be ethical!" So they turn the task over to professional "ethicists" who are *guaranteed* to fail: who will simultaneously try to sound counterintuitive enough to be worth paying for as specialists, while also making sure to not think up anything *really* technical that would scare off the foundation directors who approve their grants.

But even if professional "AI ethicists" fill the popular air with nonsense, all is not lost. AIfolk who express their ethics as a continuous, non-separate, non-special function of the same life-existence that expresses their AI work, will yet learn a thing or two about the special ethics pertaining to AI. They will not be able to avoid it. Thinking that ethics is a separate profession which judges engineers from above, is like thinking

that math is a separate profession which judges engineers from above. If you’re doing ethics *right,* you can’t *separate* it from your profession.

New edition of "Rationality: From AI to Zombies"

MIRI is releasing a new edition of ***Rationality: From AI to Zombies***, including the first set of *R:AZ* print books. As of this morning, print versions of *Map and Territory* (volume 1) and *How to Actually Change Your Mind* (volume 2) are now available **on Amazon** (1, 2), and we'll be rolling out the other four volumes of *R:AZ* over the coming months.

R:AZ is a book version of Eliezer Yudkowsky's original sequences, collecting a bit under half of his *Overcoming Bias* and *LessWrong* writing from November 2006 to September 2009. *Map and Territory* is the canonical place to start, but we've tried to make *How to Actually Change Your Mind* a good jumping-on point too, since some people might prefer to dive right into *HACYM*.

The price for the print books is $6.50 for *Map and Territory*, and $8 for *How to Actually Change Your Mind*. The new edition is also available electronically (in EPUB, MOBI, and PDF versions) on a pay-what-you-want basis: 1, 2. The *HACYM* ebook is currently available for preorders, and should be delivered in the next day.

The previous edition of *R:AZ* was a single sprawling 1800-page ebook. I announced at the time that we were also going to release a paper version divided into six more manageable chunks; but this ended up taking a lot longer than I expected, and involved more substantive revisions to the text.

Changes going into the new edition include:

- The first sequence in *Map and Territory*, "Predictably Wrong," has been heavily revised, with a goal of making it a much better experience for new readers.
- More generally, *R:AZ* is now more optimized for new readers, and less focused on extreme fidelity to the original blog posts, since this was one of the biggest requests from LessWrongers in response to the previous edition of *R:AZ*. This isn't a *huge* change, but it was an update about which option to pick in quite a few textual tradeoffs.
- A bunch of posts have been added or removed. E.g., The Robbers Cave Experiment was removed because while it's still a cool and interesting study, the researchers' methods and motives have turned out to be pretty bad, and it isn't particularly essential to *HACYM*.
- The "Against Doublethink" sequence in *How to Actually Change Your Mind* has been removed, to reduce *HACYM*'s page count (and therefore its price and its potential to intimidate readers) and improve the book's focus. The first post in "Against Doublethink" (Singlethink) has been kept, and moved to a different sequence ("Letting Go").
- Important links and references are now written out rather than hidden behind Easter egg hyperlinks, so they'll show up in print editions too. Easter egg links are kept around if they're interesting enough to be worth retaining, but not important enough to deserve a footnote; so there will still be some digital-only content, but the goal is for this to be pretty minor.
- A glossary has been added to the back of each book.

Oliver and Ben also plan to post the digital versions of *M&T* and *HACYM* to *R:AZ* on LessWrong — initially as new posts, though the URLs and comment sections of new and old versions may be merged in the future if LW adds a feature for toggling between post revisions.

The Bat and Ball Problem Revisited

Cross posted from my personal blog.

In this post, I'm going to assume you've come across the Cognitive Reflection Test before and know the answers. If you haven't, it's only three quick questions, go and do it now.

One of the striking early examples in Kahneman's *Thinking, Fast and Slow* is the following problem:

> (1) A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball.
>
> How much does the ball cost? _____ cents

This question first turns up informally in a paper by Kahneman and Frederick, who find that most people get it wrong:

> Almost everyone we ask reports an initial tendency to answer “10 cents” because the sum $1.10 separates naturally into $1 and 10 cents, and 10 cents is about the right magnitude. Many people yield to this immediate impulse. The surprisingly high rate of errors in this easy problem illustrates how lightly System 2 monitors the output of System 1: people are not accustomed to thinking hard, and are often content to trust a plausible judgment that quickly comes to mind.

In *Thinking Fast and Slow*, the bat and ball problem is used as an introduction to the major theme of the book: the distinction between fluent, spontaneous, fast 'System 1' mental processes, and effortful, reflective and slow 'System 2' ones. The explicit moral is that we are too willing to lean on System 1, and this gets us into trouble:

> The bat-and-ball problem is our first encounter with an observation that will be a recurrent theme of this book: many people are overconfident, prone to place too much faith in their intuitions. They apparently find cognitive effort at least mildly unpleasant and avoid it as much as possible.

This story is very compelling in the case of the bat and ball problem. I got this problem wrong myself when I first saw it, and still find the intuitive-but-wrong answer very plausible looking. I have to consciously remind myself to apply some extra effort and get the correct answer.

However, this becomes more complicated when you start considering other tests of this fast-vs-slow distinction. Frederick later combined the bat and ball problem with two other questions to create the Cognitive Reflection Test:

> (2) If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? _____ minutes
>
> (3) In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? _____ days

These are designed to also have an 'intuitive-but-wrong' answer (100 minutes, 24 days), and an 'effortful-but-right' answer (5 minutes, 47 days). But this time I seem to

be immune to the wrong answers, in a way that just doesn't happen with the bat and ball:

> I always have the same reaction, and I don't know if it's common or I'm just the lone idiot with this problem. The 'obvious wrong answers' for 2. and 3. are completely unappealing to me (I had to look up 3. to check what the obvious answer was supposed to be). Obviously the machine-widget ratio hasn't changed, and obviously exponential growth works like exponential growth.
>
> When I see 1., however, I always think 'oh it's that bastard bat and ball question again, I know the correct answer but cannot see it'. And I have to stare at it for a minute or so to work it out, slowed down dramatically by the fact that Obvious Wrong Answer is jumping up and down trying to distract me.

If this test was really testing my propensity for effortful thought over spontaneous intuition, I ought to score zero. I hate effortful thought! As it is, I score two out of three, because I've trained my intuitions nicely for ratios and exponential growth. The 'intuitive', 'System 1' answer that pops into my head is, in fact, the correct answer, and the supposedly 'intuitive-but-wrong' answers feel bad on a visceral level. (Why the hell would the lily pads take the same amount of time to cover the second half of the lake as the first half, when the rate of growth is increasing?)

The bat and ball still gets me, though. My gut hasn't internalised anything useful, and it's super keen on shouting out the wrong answer in a distracting way. My dislike for effortful thought is definitely a problem here.

I wanted to see if others had raised the same objection, so I started doing some research into the CRT. In the process I discovered a lot of follow-up work that makes the story much more complex and interesting.

I've come nowhere near to doing a proper literature review. Frederick's original paper has been cited nearly 3000 times, and dredging through that for the good bits is a lot more work than I'm willing to put in. This is just a summary of the interesting stuff I found on my limited, partial dig through the literature.

Thinking, inherently fast and inherently slow

Frederick's original Cognitive Reflection Test paper describes the System 1/System 2 divide in the following way:

> Recognizing that the face of the person entering the classroom belongs to your math teacher involves System 1 processes — it occurs instantly and effortlessly and is unaffected by intellect, alertness, motivation or the difficulty of the math problem being attempted at the time. Conversely, finding $\sqrt{19163}$ to two decimal places without a calculator involves System 2 processes — mental operations requiring effort, motivation, concentration, and the execution of learned rules.

I find it interesting that he frames mental processes as being *inherently* effortless or effortful, independent of the person doing the thinking. This is not quite true even for the examples he gives — faceblind people and calculating prodigies exist.

This framing is important for interpreting the CRT. If the problem inherently has a wrong 'System 1 solution' and a correct 'System 2 solution', the CRT can work as intended, as an efficient tool to split people by their propensity to use one strategy or the other. If there are 'System 1' ways to get the correct answer, the whole thing gets much more muddled, and it's hard to disentangle natural propensity to reflection from prior exposure to the right mathematical concepts.

My tentative guess is that the bat and ball problem *is* close to being this kind of efficient tool. Although in some ways it's the simplest of the three problems, solving it in a 'fast', 'intuitive' way relies on seeing the problem in a way that most people's education won't have provided. (I *think* this is true, anyway - I'll go into more detail later.) I suspect that this is less true the other two problems - ratios and exponential growth are topics that a mathematical or scientific education is more likely to build intuition for.

(Aside: I'd like to know how these other two problems were chosen. The paper just states the following:

> Motivated by this result [the answers to the bat and ball question], two other problems found to yield impulsive erroneous responses were included with the “bat and ball” problem to form a simple, three-item “Cognitive Reflection Test” (CRT), shown in Figure 1.

I have a vague suspicion that Frederick trawled through something like 'The Bumper Book of Annoying Riddles' to find some brainteasers that don't require too much in the way of mathematical prerequisites. The lilypads one has a family resemblance to the classic grains-of-wheat-on-a-chessboard puzzle, for instance.)

However, I haven't found any great evidence either way for this guess. The original paper doesn't break down participants' scores by question – it just gives mean scores on the test as a whole. I did however find this meta-analysis of 118 CRT studies, which shows that the bat and ball question is the most difficult on average – only 32% of all participants get it right, compared with 40% for the widgets and 48% for the lilypads. It also has the biggest jump in success rate when comparing university students with non-students. That looks like better mathematical education does help on the bat and ball, but it doesn't clear up how it helps. It could improve participants' ability to intuitively see the answer. Or it could improve ability to come up with an 'unintuitive' solution, like solving the corresponding simultaneous equations by a rote method.

What I'd really like is some insight into what individual people *actually do* when they try to solve the problems, rather than just this aggregate statistical information. I haven't found exactly what I wanted, but I did turn up a few interesting studies on the way.

No, seriously, the answer isn't ten cents

My favourite thing I found was this (apparently unpublished) ‘extremely rough draft’ by Meyer, Spunt and Frederick from 2013, revisiting the bat and ball problem. The intuitive-but-wrong answer turns out to be *extremely* sticky, and the paper is basically a series of increasingly desperate attempts to get people to actually think about the question.

One conjecture for what people are doing when they get this question wrong is the *attribute substitution hypothesis*. This was suggested early on by Kahneman and Frederick, and is a fancy way of saying that they are instead solving the following simpler problem:

> (1) A bat and a ball cost $1.10 in total. The bat costs $1.00.
>
> How much does the ball cost? _____ cents

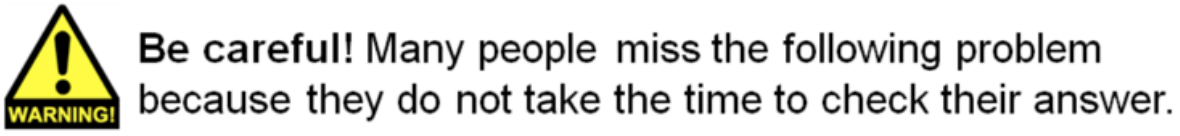
Notice that this is missing the 'more than the ball' clause at the end, turning the question into a much simpler arithmetic problem. This simple problem *does* have 'ten cents' as the answer, so it's very plausible that people are getting confused by it.

Meyer, Spunt and Frederick tested this hypothesis by getting respondents to recall the problem from memory. This showed a clear difference: 94% of 'five cent' respondents could recall the correct question, but only 61% of 'ten cent' respondents. It's possible that there is a different common cause of both the 'ten cent' response and misremembering the question, but it at least gives some support for the substitution hypothesis.

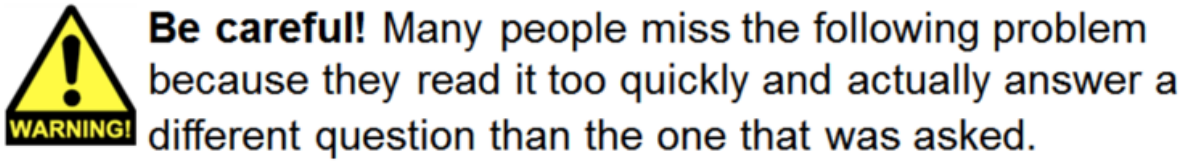
However, getting people to actually answer the question correctly was a much more difficult problem. First they tried bolding the words **more than the ball** to make this clause more salient. This made surprisingly little impact: 29% of respondents solved it, compared with 24% for the original problem. Printing both versions was slightly more successful, bumping up the correct response to 35%, but it was still a small effect.

After this, they ditched subtlety and resorted to pasting these huge warnings above the question:

***COMPUTATION* WARNING**



***COMPREHENSION* WARNING**



These were still only mildly effective, with a correct solution jumping to 50% from 45%. People just really like the answer 'ten cents', it seems.

At this point they completely gave up and just flat out added “HINT: 10 cents is not the answer.” This worked reasonably well, though there was still a hard core of 13% who persisted in writing down 'ten cents'.

That's where they left it. At this point there's not really any room to escalate beyond confiscating the respondents' pens and prefilling in the answer 'five cents', and I worry

that somebody would still try and scratch in 'ten cents' in their own blood. The wrong answer is just incredibly compelling.

So, what *are* people doing when they solve this problem?

Unfortunately, it's hard to tell from the published literature (or at least what I found of it). What I'd really like is lots of transcripts of individuals talking through their problem solving process. The closest I found was this paper by Szaszi et al, who did carry out these sort of interview, but it doesn't include any examples of individual responses. Instead, it gives a aggregated overview of types of responses, which doesn't go into the kind of detail I'd like.

Still, the examples given for their response categories give a few clues. The categories are:

- **Correct answer, correct start.** Example given: 'I see. This is an equation. Thus if the ball equals to x, the bat equals to x plus 1... '
- **Correct answer, incorrect start.** Example: 'I would say 10 cents... But this cannot be true as it does not sum up to €1.10...'
- **Incorrect answer, reflective**, i.e. some effort was made to reconsider the answer given, even if it was ultimately incorrect. Example: '... but I'm not sure... If together they cost €1.10, and the bat costs €1 more than the ball... the solution should be 10 cents. I'm done.'
- **No reflection.** Example: 'Ok. I'm done.'

These demonstrate one way to reason your way to the correct answer (solve the simultaneous equations) and one way to be wrong (just blurt out the answer). They also demonstrate one way to recover from an incorrect solution (think about the answer you blurted out and see if it actually works). Still, it's all rather abstract and high level.

How To Solve It

However, I did manage to stumble onto another source of insight. While researching the problem I came across this article from the online magazine of the Association for Psychological Science, which discusses a variant 'Ford and Ferrari problem'. This is quite interesting in itself, but I was most excited by the comments section. Finally some examples of how the problem is solved in the wild!

The simplest 'analytical', 'System 2' solution is to rewrite the problem as two simultaneous linear equations and plug-and-chug your way to the correct answer. For example, writing B for the bat and b for the ball, we get the two equations

$$B + b = 110,\ B - b = 100,$$

which we could then solve in various standard ways, e.g.

2B = 210, B = 105,

which then gives

b = 110 − B = 5.

There are a couple of variants of this explained in the comments. It's a very reliable way to tackle the problem: if you already know how to do this sort of rote method, there are no surprises. This sort of method would work for any similar problem involving linear equations.

However, it's pretty obvious that a lot of people *won't* have access to this method. Plenty of people noped out of mathematics long before they got to simultaneous equations, so they won't be able to solve it this way. What might be less obvious, at least if you mostly live in a high-maths-ability bubble, is that these people may also be missing the sort of tacit mathematical background that would even allow them to frame the problem in a useful form in the first place.

That sounds a bit abstract, so let's look at some responses (I'll paste all these straight in, so any typos are in the original). First, we have these two confused commenters:

> The thing is, why does the ball have to be $.05? It could have been .04 0r.03 and the bat would still cost more than $1.

and

> This is exactly what bothers me and resulted in me wanting to look up the question online. On the quiz the other 2 questions were definitive. This one technically could have more than one answer so this is where phycologists actually mess up when trying to give us a trick question. The ball at .4 and the bat at 1.06 doesn’t break the rule either.

These commenters don't automatically see two equations in two variables that together are enough to constrain the problem. Instead they seem to focus mainly on the first condition (adding up to $1.10) and just use the second one as a vague check at best ('the bat would still cost more than $1'). This means that they are unable to immediately tell that the problem has a unique solution.

In response, another commenter, Tony, suggests a correct solution which is an interesting mix of writing the problem out formally and then figuring out the answer by trial and error:\

> I hear your pain. I feel as though psychologists and psychiatrists get together every now and then to prove how stoopid I am. However, after more than a little head scratching I’ve gained an understanding of this puzzle. It can be expressed as two facts and a question A=100+B and A+B=110, so B=? If B=2 then the solution would be 100+2+2 and A+B would be 104. If B=6 then the solution would be 100+6+6 and A+B would be 112. But as be KNOW A+B=110 the only number for B on it’s own is 5.

This suggests enough half-remembered mathematical knowledge to find a sensible abstract framing, but not enough to solve it the standard way.

Finally, commenter Marlo Eugene provides an ingenious way of solving the problem without writing all the algebraic steps out:

> Linguistics makes all the difference. The conceptual emphasis seems to lie within the word MORE.
>
> X + Y = $1.10. If X = $1 MORE then that leaves $0.10 TO WORK WITH rather than automatically assign to Y
>
> So you divide the remainder equally (assuming negative values are disqualified) and get 0.05.

So even this small sample of comments suggests a wide diversity of problem-solving methods leading to the two common answers. Further, these solutions don't all split neatly into 'System 1' 'intuitive' and 'System 2' 'analytic'. Marlo Eugene's solution, for instance, is a mixed solution of writing the equations down in a formal way, but then finding a clever way of just seeing the answer rather than solving them by rote.

I'd still appreciate more detailed transcripts, including the time taken to solve the problem. My suspicion is still that very few people solve this problem with a fast intuitive response, in the way that I rapidly see the correct answer to the lilypad question. Even the more 'intuitive' responses, like Marlo Eugene's, seem to rely on some initial careful reflection and a good initial framing of the problem.

If I'm correct about this lack of fast responses, my tentative guess for the reason is that it has something to do with the way most of us learn simultaneous equations in school. We generally learn arithmetic as young children in a fairly concrete way, with the formal numerical problems supplemented with lots of specific examples of adding up apples and bananas and so forth.

But then, for some reason, this goes completely out of the window once the unknown quantity isn't sitting on its own on one side of the equals sign. This is instead hived off into its own separate subject, called 'algebra', and the rules are taught much later in a much more formalised style, without much attempt to build up intuition first.

(One exception is the sort of puzzle sheets that are often given to young kids, where the unknowns are just empty boxes to be filled in. Sometimes you get 2+3=□, sometimes it's 2+□=5, but either way you go about the same process of using your wits to figure out the answer. Then, for some reason I'll never understand, the worksheets get put away and the poor kids don't see the subject again until years later, when the box is now called x for some reason and you have to find the answer by defined rules. Anyway, this is a separate rant.)

This lack of a rich background in puzzling out the answer to specific concrete problems means most of us lean hard on formal rules in this domain, even if we're relatively mathematically sophisticated. Only a few build up the necessary repertoire of tricks to solve the problem quickly by insight. I'm reminded of a story in Feynman's *The Pleasure of Finding Things Out*:

> Around that time my cousin, who was three years older, was in high school. He was having considerable difficulty with his algebra, so a tutor would come. I was allowed to sit in a corner while the tutor would try to teach my cousin algebra. I'd hear him talking about x.
>
> I said to my cousin, "What are you trying to do?"

> "I'm trying to find out what x is, like in 2x + 7 = 15."
>
> I say, "You mean 4."
>
> "Yeah, but you did it by arithmetic. You have to do it by algebra."
>
> I learned algebra, fortunately, not by going to school, but by finding my aunt's old schoolbook in the attic, and understanding that the whole idea was to find out what *x* is - it doesn't make any difference how you do it.

I think this reliance on formal methods might be somewhat less true for exponential growth and ratios, the subjects underpinning the lilypad and widget questions. Certainly I seem to have better intuition there, without having to resort to rote calculation. But I'm not sure how general this is.

How To Visualise It

If you wanted to solve the bat and ball problem without having to 'do it by algebra', how would you go about it?

My original post on the problem was a pretty quick, throwaway job, but over time it picked up some truly excellent comments by anders and Kyzentun, which really start to dig into the structure of the problem and suggest ways to 'just see' the answer. The thread with anders in particular goes into lots of other examples of how we think through solving various problems, and is well worth reading in full. I'll only summarise the bat-and-ball-related parts of the comments here.

We all used some variant of the method suggested by Marlo Eugene in the comments above. Writing out the basic problem again, we have:

$$B + b = 110,\ B - b = 100.$$

Now, instead of immediately jumping to the standard method of eliminating one of the variables, we can just look at what these two equations are saying and solve it directly 'by thinking'. We have a bat, B. If you add the price of the ball, b, you get 110 cents. If you instead remove the same quantity b you get 100 cents. So the bat's price must be exactly halfway between these two numbers, at 105 cents. That leaves five for the ball.

Now that I'm thinking of the problem in this way, I directly see the equations as being 'about a bat that's halfway between 100 and 110 cents', and the answer is incredibly obvious.

Kyzentun suggests a variant on the problem that is much less counterintuitive than the original:

> A centered piece of text and its margins are 110 columns wide. The text is 100 columns wide. How wide is one margin?
>
> Same numbers, same mathematical formula to reach the solution. But less misleading because you know there are two margins, and thus know to divide by two after subtracting.

In the original problem, the 110 units and 100 units both refer to something abstract, the sum and difference of the bat and ball. In Kyzentun's version these become much more concrete objects, the width of the text and the total width of the margins. The work of seeing the equations as relating to something concrete has mostly been done for you.

Similarly, anders works the problem by 'getting rid of the 100 cents', and splitting the remainder in half to get at the price of the ball:

> I just had an easy time with #1 which I haven't before. What I did was take away the difference so that all the items are the same (subtract 100), evenly divide the remainder among the items (divide 10 by 2) and then add the residuals back on to get 105 and 5.
>
> The heuristic I seem to be using is to treat objects as made up of a value plus a residual. So when they gave me the residual my next thought was "now all the objects are the same, so whatever I do to one I do to all of them".

I think that after reasoning my way through all these perspectives, I'm finally at the point where I have a quick, 'intuitive' understanding of the problem. But it's surprising how much work it was for such a simple bit of algebra.

Final thoughts

Rather than making any big conclusions, the main thing I wanted to demonstrate in this post is how *complicated* the story gets when you look at one problem in detail. I've written about close reading recently, and this has been something like a close reading of the bat and ball problem.

Frederick's original paper on the Cognitive Reflection Test is in that generic social science style where you define a new metric and then see how it correlates with a bunch of other macroscale factors (either big social categories like gender or education level, or the results of other statistical tests that try to measure factors like time preference or risk preference). There's a strange indifference to the details of the test itself – at no point does he discuss why he picked those specific three questions, and there's no attempt to model what was making the intuitive-but-wrong answer appealing.

The later paper by Meyer, Spunt and Frederick is much more interesting to me, because it really starts to pick apart the specifics of the bat and ball problem. Is an easier question getting substituted? Can participants reproduce the correct question from memory?

I learned the most from the individual responses, though. This is where you really get to see the variety of ways that people tackle the problem. Careful reflection definitely seems to improve the chance of a correct answer in general, but many of the responses don't really fit the neat 'fast vs slow' division of the original setup.

Questions

I'm interested in any comments on the post, but here are a few specific things I'd like to get your answers to:

- My rapid, intuitive answer for the bat and ball question is wrong (at least until I retrained it by thinking about the problem way too much). However, for the other two I 'just see' the correct answer. Is this common for other people, or do you have a different split?
- If you're able to rapidly 'just see' the answer to the bat and ball question, how do you do it?
- How do people go about designing tests like these? This isn't at all my field and I'd be interested in any good sources. I'd kind of assumed that there'd be some kind of serious-business Test Creation Methodology, but for the CRT at least it looks like people just noticed they got surprising answers for the bat and ball question and looked around for similar questions. Is that unusual compared to other psychological tests?

Coherence arguments do not entail goal-directed behavior

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

One of the most pleasing things about probability and expected utility theory is that there are many *coherence arguments* that suggest that these are the "correct" ways to reason. If you deviate from what the theory prescribes, then you must be executing a *dominated strategy*. There must be some other strategy that never does any worse than your strategy, but does strictly better than your strategy with certainty in at least one situation. There's a good explanation of these arguments here.

We shouldn't expect mere humans to be able to notice any failures of coherence in a superintelligent agent, since if we could notice these failures, so could the agent. So we should expect that powerful agents appear coherent to us. (Note that it is possible that the agent doesn't fix the failures because it would not be worth it -- in this case, the argument says that we will not be able to notice any *exploitable* failures.)

Taken together, these arguments suggest that we should model an agent much smarter than us as an expected utility (EU) maximizer. And many people agree that EU maximizers are dangerous. So does this mean we're doomed? I don't think so: it seems to me that the problems about EU maximizers that we've identified are actually about *goal-directed behavior* or *explicit reward maximizers.* The coherence theorems say nothing about whether an AI system must look like one of these categories. This suggests that we could try building an AI system that can be modeled as an EU maximizer, yet doesn't fall into one of these two categories, and so doesn't have all of the problems that we worry about.

Note that there are two different flavors of arguments that the AI systems we build will be goal-directed agents (which are dangerous if the goal is even slightly wrong):

- Simply knowing that an agent is intelligent lets us infer that it is goal-directed. (EDIT: See these comments for more details on this argument.)
- Humans are particularly likely to build goal-directed agents.

I will only be arguing against the first claim in this post, and will talk about the second claim in the next post.

All behavior can be rationalized as EU maximization

Suppose we have access to the entire policy of an agent, that is, given any universe-history, we know what action the agent will take. Can we tell whether the agent is an EU maximizer?

Actually, *no matter what the policy is*, we can view the agent as an EU maximizer. The construction is simple: the agent can be thought as optimizing the utility function U, where U(h, a) = 1 if the policy would take action a given history h, else 0. Here I'm assuming that U is defined over histories that are composed of states/observations

and actions. The actual policy gets 1 utility at every timestep; any other policy gets less than this, so the given policy perfectly maximizes this utility function. This construction has been given before, eg. at the bottom of page 6 of this paper. (I think I've seen it before too, but I can't remember where.)

But wouldn't this suggest that the VNM theorem has no content? Well, we assumed that we were looking at the *policy* of the agent, which led to a universe-history *deterministically*. We didn't have access to any probabilities. Given a particular action, we knew exactly what the next state would be. Most of the axioms of the VNM theorem make reference to lotteries and probabilities -- if the world is deterministic, then the axioms simply say that the agent must have transitive preferences over outcomes. Given that we can only observe the agent choose one history over another, we can trivially construct a transitive preference ordering by saying that the chosen history is higher in the preference ordering than the one that was not chosen. This is essentially the construction we gave above.

What then is the purpose of the VNM theorem? It tells you how to behave *if you have probabilistic beliefs about the world*, as well as a *complete and consistent preference ordering over outcomes*. This turns out to be not very interesting when "outcomes" refers to "universe-histories". It can be more interesting when "outcomes" refers to world *states* instead (that is, snapshots of what the world looks like at a particular time), but utility functions over states/snapshots can't capture everything we're interested in, and there's no reason to take as an assumption that an AI system will have a utility function over states/snapshots.

There are no coherence arguments that say you must have goal-directed behavior

Not all behavior can be thought of as goal-directed (primarily because I allowed the category to be defined by fuzzy intuitions rather than something more formal). Consider the following examples:

- A robot that constantly twitches
- The agent that always chooses the action that starts with the letter "A"
- The agent that follows the policy <policy> where for every history the corresponding action in <policy> is generated randomly.

These are not goal-directed by my "definition". However, they can all be modeled as expected utility maximizers, and there isn't any particular way that you can exploit any of these agents. Indeed, it seems hard to model the twitching robot or the policy-following agent as having any preferences at all, so the notion of "exploiting" them doesn't make much sense.

You could argue that neither of these agents are *intelligent*, and we're only concerned with superintelligent AI systems. I don't see why these agents could not in principle be intelligent: perhaps the agent knows how the world would evolve, and how to intervene on the world to achieve different outcomes, but it does not act on these beliefs. Perhaps if we peered into the inner workings of the agent, we could find some part of it that allows us to predict the future very accurately, but it turns out that these inner workings did not affect the chosen action at all. Such an agent is in principle possible, and it seems like it is intelligent.

(If not, it seems as though you are *defining* intelligence to also be goal-driven, in which case I would frame my next post as arguing that we may not want to build superintelligent AI, because there are other things we could build that are as useful without the corresponding risks.)

You could argue that while this is possible in principle, no one would ever build such an agent. I wholeheartedly agree, but note that this is now an argument based on particular empirical facts about humans (or perhaps agent-building processes more generally). I'll talk about those in the next post; here I am simply arguing that merely knowing that an agent is intelligent, with no additional empirical facts about the world, does not let you infer that it has goals.

As a corollary, since all behavior can be modeled as maximizing expected utility, but not all behavior is goal-directed, it is not possible to conclude that an agent is goal-driven if you only know that it can be modeled as maximizing some expected utility. However, if you know that an agent is maximizing the expectation of an *explicitly represented* utility function, I would expect that to lead to goal-driven behavior most of the time, since the utility function must be relatively simple if it is explicitly represented, and *simple* utility functions seem particularly likely to lead to goal-directed behavior.

There are no coherence arguments that say you must have preferences

This section is another way to view the argument in the previous section, with "goal-directed behavior" now being operationalized as "preferences"; it is not saying anything new.

Above, I said that the VNM theorem assumes both that you use probabilities and that you have a preference ordering over outcomes. There are lots of good reasons to assume that a good reasoner will use probability theory. However, there's not much reason to assume that there is a preference ordering over outcomes. The twitching robot, "A"-following agent, and random policy agent from the last section all seem like they don't have preferences (in the English sense, not the math sense).

Perhaps you could define a preference ordering by saying "if I gave the agent lots of time to think, how would it choose between these two histories?" However, you could apply this definition to *anything*, including eg. a thermostat, or a rock. You might argue that a thermostat or rock can't "choose" between two histories; but then it's unclear how to define how an AI "chooses" between two histories without that definition also applying to thermostats and rocks.

Of course, you could always define a preference ordering based on the AI's observed behavior, but then you're back in the setting of the first section, where *all* observed behavior can be modeled as maximizing an expected utility function and so saying "the AI is an expected utility maximizer" is vacuous.

Convergent instrumental subgoals are about goal-directed behavior

One of the classic reasons to worry about expected utility maximizers is the presence of convergent instrumental subgoals, detailed in Omohundro's paper The Basic AI Drives. The paper itself is clearly talking about goal-directed AI systems:

To say that a system of any design is an "artificial intelligence", we mean that it has goals which it tries to accomplish by acting in the world.

It then argues (among other things) that such AI systems will want to "be rational" and so will distill their goals into utility functions, which they then maximize. And once they have utility functions, they will protect them from modification.

Note that this starts from the assumption of goal-directed behavior and *derives* that the AI will be an EU maximizer along with the other convergent instrumental subgoals. The coherence arguments all imply that AIs will be EU maximizers for some (possibly degenerate) utility function; they don't prove that the AI must be goal-directed.

Goodhart's Law is about goal-directed behavior

A common argument for worrying about AI risk is that we know that a superintelligent AI system will look to us like an EU maximizer, and if it maximizes a utility function that is even slightly wrong we could get catastrophic outcomes.

By now you probably know my first response: that *any* behavior can be modeled as an EU maximizer, and so this argument proves too much, suggesting that any behavior causes catastrophic outcomes. But let's set that aside for now.

The second part of the claim comes from arguments like Value is Fragile and Goodhart's Law. However, if we consider utility functions that assign value 1 to some histories and 0 to others, then if you accidentally assign a history where I needlessly stub my toe a 1 instead of a 0, that's a slightly wrong utility function, but it isn't going to lead to catastrophic outcomes.

The worry about utility functions that are *slightly wrong* holds water when the utility functions are wrong about some *high-level* concept, like whether humans care about their experiences reflecting reality. This is a very rarefied, particular distribution of utility functions, that are all going to lead to goal-directed or agentic behavior. As a result, I think that the argument is better stated as "if you have a slightly incorrect goal, you can get catastrophic outcomes". And there aren't any coherence arguments that say that agents must have goals.

Wireheading is about explicit reward maximization

There are a few papers that talk about the problems that arise with a very powerful system with a reward function or utility function, most notably wireheading. The argument that AIXI will seize control of its reward channel falls into this category. In these cases, typically the AI system is considering making a change to the system by which it evaluates goodness of actions, and the goodness of the change is evaluated by the system *after the change*. Daniel Dewey argues in Learning What to Value that if

the change is evaluated by the system *before* the change, then these problems go away.

I think of these as problems with *reward* maximization, because typically when you phrase the problem as maximizing reward, you are maximizing the sum of rewards obtained in all timesteps, no matter how those rewards are obtained (i.e. even if you self-modify to make the reward maximal). It doesn't seem like AI systems have to be built this way (though admittedly I do not know how to build AI systems that reliably avoid these problems).

Summary

In this post I've argued that many of the problems we typically associate with expected utility maximizers are actually problems with goal-directed agents or with explicit reward maximization. Coherence arguments only entail that a superintelligent AI system will look like an expected utility maximizer, but this is actually a vacuous constraint, and there are many potential utility functions for which the resulting AI system is neither goal-directed nor explicit-reward-maximizing. This suggests that we could try to build AI systems of this type, in order to sidestep the problems that we have identified so far.

Why should I care about rationality?

In what ways are holidays good?

I'd like a model of the benefits that holidays (vacations) can have, so that I can plan accordingly. Relevant questions that I have, although feel free to answer ones not listed here:

- Do holidays teach you things about other places that you couldn't learn from Wikipedia?
- Why are holidays more relaxing than just lying in bed at home and paying somebody else to take care of you?
- Does visiting family count as a holiday in the relevant sense?
- Are the benefits of tourism and/or pilgrimage the same as the benefits of holidays? What are they?
- How much money should I be willing to spend on holidays?
- Is 'holiday' a coherent enough category that I can treat it as primitive for the purpose of this question?

Is there a standard discussion of vegetarianism/veganism?

I am searching for a concise text that presents and optimally also discusses reasons for a vegetarian/vegan diet, including environmental and climate effects, health, but of course also ethics, and there are some ethical points I would be particularly interested in like "can you rank animals by how bad eating them is?", "is it more ethical to eat wild animals because they have a good life before dying?", "the ethics of offsetting" (the kind discussed in http://slatestarcodex.com/2015/09/23/vegetarianism-for-meat-eaters/) Optimally, this would be a kind of non-partisan text, but I guess for this topic this is hard to find because if someone writes about it, s/he usually explains her/his own reasons.

Player vs. Character: A Two-Level Model of Ethics

Epistemic Status: Confident

This idea is actually due to my husband, Andrew Rettek, but since he doesn't blog, and I want to be able to refer to it later, I thought I'd write it up here.

In many games, such as Magic: The Gathering, Hearthstone, or Dungeons and Dragons, there's a two-phase process. First, the player constructs a *deck* or *character* from a very large sample space of possibilities. This is a particular combination of strengths and weaknesses and capabilities for action, which the player thinks can be successful against other decks/characters or at winning in the game universe. The choice of deck or character often determines the strategies that deck or character can use in the second phase, which is actual gameplay. In gameplay, the character (or deck) can only use the affordances that it's been previously set up with. This means that there are two separate places where a player needs to get things right: first, in designing a strong character/deck, and second, in executing the optimal strategies for that character/deck during gameplay.

(This is in contrast to games like chess or go, which are single-level; the capacities of black and white are set by the rules of the game, and the only problem is how to execute the optimal strategy. Obviously, even single-level games can already be complex!)

The idea is that human behavior works very much like a two-level game.

The "player" is the whole mind, choosing subconscious strategies. The "elephant", not the "rider." The player is very influenced by evolutionary pressure; it is built to direct behavior in ways that increases inclusive fitness. The player directs what we perceive, do, think, and feel.

The player *creates* what we experience as "personality", fairly early in life; it notices what strategies and skills work for us and invests in those at the expense of others. It builds our "character sheet", so to speak.

Note that even things that seem like "innate" talents, like the savant skills or hyperacute senses sometimes observed in autistic people, can be observed to be tightly linked to feedback loops in early childhood. In other words, savants *practice the thing they like and are good at*, and gain "superhuman" skill at it. They "practice" along a faster and more hyperspecialized path than what we think of as a neurotypical "practicing hard," but it's still a learning process. Savant skills are *more* rigidly fixed and seemingly "automatic" than non-savant skills, but they still change over time — e.g. Stephen Wiltshire, a savant artist who manifested an ability to draw hyper-accurate perspective drawings in early childhood, has changed and adapted his art style as he grew up, and even acquired *new* savant talents in music. If even savant talents are subject to learning and incentives/rewards, certainly ordinary strengths, weaknesses, and personality types are likely to be "strategic" or "evolved" in this sense.

The player determines what we find rewarding or unrewarding. The player determines what we notice and what we overlook; things come to our attention if it suits the

player's strategy, and not otherwise. The player gives us emotions when it's strategic to do so. The player sets up our subconscious evaluations of what is good for us and bad for us, which we experience as "liking" or "disliking."

The character is what *executing the player's strategies feels like from the inside*. If the player has decided that a task is unimportant, the character will experience "forgetting" to do it. If the player has decided that alliance with someone will be in our interests, the character will experience "liking" that person. Sometimes the player will notice and seize opportunities in a very strategic way that feels to the character like "being lucky" or "being in the right place at the right time."

This is where confusion often sets in. People will often protest "but I *did* care about that thing, I just forgot" or "but I'm *not* that Machiavellian, I'm just doing what comes naturally." This is true, because when we talk about ourselves and our experiences, we're speaking "in character", as our character. The strategy is not going on at a conscious level. In fact, I don't believe we (characters) have direct access to the player; we can only *infer* what it's doing, based on what patterns of behavior (or thought or emotion or perception) we observe in ourselves and others.

Evolutionary psychology refers to the player's strategy, not the character's. (It's unclear which animals even *have* characters in the way we do; some animals' behavior may *all* be "subconscious".) So when someone speaking in an evolutionary-psychology mode says that babies are manipulating their parents to not have more children, for instance, that obviously doesn't mean that my baby is a cynically manipulative evil genius. To him, it probably just feels like "I want to nurse at night. I miss Mama." It's perfectly innocent. But of course, this has the *effect* that I can't have more children until I wean him, and that's to his interest (or, at least, it was in the ancestral environment when food was more scarce.)

Szaszian or evolutionary analysis of mental illness is absurd if you think of it as applying to the character — of course nobody wakes up in the morning and decides to have a mental illness. It's not "strategic" in that sense. (If it were, we wouldn't call it mental illness, we'd call it feigning.) But at the *player* level, it can be fruitful to ask "what strategy could this behavior be serving the person?" or "what experiences could have made this behavior adaptive at one point in time?" or "what incentives are shaping this behavior?" (And, of course, externally visible "behavior" isn't the only thing the player produces: thoughts, feelings, and perceptions are *also* produced by the brain.)

It may make more sense to frame it as "what strategy is *your brain* executing?" rather than "what strategy are *you* executing?" since people generally identify as their characters, not their players.

Now, let's talk morality.

Our intuitions about praise and blame are driven by moral sentiments. We have emotional responses of sympathy and antipathy, towards behavior of which we approve and disapprove. These are driven by the player, which creates incentives and strategic behavior patterns for our characters to play out in everyday life. The character engages in coalition-building with other characters, forms and breaks alliances with other characters, honors and shames characters according to their behavior, signals to other characters, etc.

When we, speaking as our characters, say "that person is good" or "that person is bad", we are making one move in an overall *strategy* that our players have created.

That strategy is the *determination of when, in general, we will call things or people "good" or "bad"*.

This is precisely what Nietzsche meant by "beyond good and evil." Our notions of "good" and "evil" are character-level notions, encoded by our players.

Imagine that somewhere in our brains, the player has drawn two cartoons, marked "hero" and "villain", that we consult whenever we want to check whether to call another person "good" or "evil." (That's an oversimplification, of course, it's just for illustrative purposes.) Now, is *the choice of cartoons itself* good or evil? Well, the character checks… "Ok, is it more like the hero cartoon or the villain cartoon?" The answer is "ummmm….type error."

The player is *not* like a hero or a villain. It is not like a person at all, in the usual (character-level) sense. Characters have feelings! Players don't have feelings; they are beings of pure strategy that *create* feelings. Characters can have virtues or vices! Players don't; they *create* virtues or vices, strategically, when they build the "character sheet" of a character's skills and motivations. Characters can be *evaluated* according to moral standards; players *set* those moral standards. Players, compared to we characters, are hyperintelligent Lovecraftian creatures that we cannot relate to socially. They are *beyond good and evil*.

However! There is another, very different sense in which players *can* be evaluated as "moral agents", even though our moral sentiments don't apply to them.

We can observe what various game-theoretic strategies *do* and how they perform. Some, like "tit for tat", perform well on the whole. Tit-for-tat-playing agents cooperate with each other. They can survive pretty well even if there are different kinds of agents in the population; and a population composed entirely of tit-for-tat-ers is stable and well-off.

While we can't call cellular automata performing game strategies "good guys" or "bad guys" in a sentimental or socially-judgmental way (they're *not people*), we can totally make objective claims about which strategies dominate others, or how strategies interact with one another. This is an empirical and theoretical field of science.

And there is a kind of ""morality"" which I almost hesitate to call morality because it isn't very much like social-sentiment-morality at all, but which is *very important*, which says simply: *the strategies that win in the long run are good, the ones that lose in the long run are bad.* Not "like the hero cartoon" or "like the villain cartoon", but simply "win" and "lose."

At this level you can say "look, objectively, people who *set up their tables of values* in this way, calling X good and Y evil, *are gonna die.*" Or "this strategy is conducting a campaign of unsustainable exploitation, which will work well in the short run, but will flame out when it runs out of resources, *and so it's gonna die*." Or "this strategy is going to lose to that strategy." Or "this strategy is fine in the best-case scenario, but it's not robust to noise, and if there are any negative shocks to the system, it's going to result in *everybody dying.*"

"But what if a losing strategy is good?" Well, if you are in that value system, of course you'll say it's good. Also, you will lose.

Mother Teresa is a saint, in the literal sense: she was canonized by the Roman Catholic Church. Also, she provided poor medical care for the sick and destitute — unsterilized

needles, no pain relief, conditions in which tuberculosis could and did spread. Was she a good person? It depends on your value system, and, obviously, according to some value systems she was. But, it seems, that a population that places Mother Teresa as its ideal (relative to, say, Florence Nightingale) will be a population with more deaths from illness, not fewer, and more pain, not less. A strategy that says "showing care for the dying is better than promoting health" *will lose* to one that actually can reward actions that promote health. That's the "player-level" analysis of the situation.

Some game-theoretic strategies (what Nietzsche would call "tables of values") are more survival-promoting than others. That's the sense in which you can get from "is" to "ought." The Golden Rule (Hillel's, Jesus's, Confucius's, etc) is a "law" of game theory, in the sense that *it is a universal, abstract fact, which even a Lovecraftian alien intelligence would recognize, that it's an effective strategy*, which is why it keeps being rediscovered around the world.

But you can't adjudicate between character strategies just by *being a character playing your strategy*. For instance, a Democrat usually can't convert a Republican just by *being a Democrat at him*. To change a player's strategy is more like "getting the bodymind to change its fundamental assessments of what is in its best interests." Which can happen, and can happen deliberately and with the guidance of the intellect! But not without some…what you might call, *wiggling things around*.

The way I think the intellect plays into "metaprogramming" the player is indirect; you can *infer* what the player is doing, do some formal analysis about how that will play out, comprehend (again at the "merely" intellectual level) if there's an error or something that's no longer relevant/adaptive, plug that new understanding into *some* change that the intellect *can* affect (maybe "let's try this experiment"), and maybe somewhere down the chain of causality the "player"'s strategy changes. (Exposure therapy is a simple example, probably much simpler than most: add some experiences of the thing not being dangerous and the player determines it really isn't dangerous and stops generating fear emotions.)

You *don't* get changes in player strategies just by executing social praise/blame algorithms though; those algorithms are for *interacting* with other characters. Metaprogramming is… I want to say "cold" or "nonjudgmental" or "asocial" but none of those words are quite right, because they describe character traits or personalities or mental states and it's *not a character-level thing at all*. It's a thing Lovecraftian intelligences can do to themselves, in their peculiar tentacled way.

Isaac Asimov's predictions for 2019 from 1984

This is a linkpost for https://www.thestar.com/news/world/2018/12/27/35-years-ago-isaac-asimov-was-asked-by-the-star-to-predict-the-world-of-2019-here-is-what-he-wrote.html?fbclid=IwAR25wKGV6NngxJFnbyVTOBYGRGjzJXLxyGCywDYl806UWXKq-6XxEhZlATw

My vague impressions

The whole essay is conditional on no nuclear war. Than, he explored two main big trends - computerization and space utilization. If something like a general model how Asimov did futurology can be extracted from the text, it is extending the large trendline, and than thinking about social consequences.

In case of space utilization, this failed badly, because the trend extrapolation did not work, and most of the specific predictions are wrong (e.g. we do not have *prototype of a solar power station, outfitted to collect solar energy, convert it to microwaves and beam it to Earth* or *mining station that will process moon soil*)

In case of computerization, the trendline stayed linear. The predictions of social consequences are often good

- *The growing complexity of society will make it impossible to do without [computers] except by courting chaos; and those parts of the world that fall behind in this respect will suffer so obviously (...)*
- There is a longer part about work, jobs and the force, e.g.: *The jobs that will appear will, inevitably, involve the design, the manufacture, the installation, the maintenance and repair of computers and robots, and an understanding of whole new industries that these “intelligent” machines will make possible. ... By the year 2019, however, we should find that the transition is about over. Those who can he retrained and re-educated will have been: those who can’t be will have been put to work at something useful, or where ruling groups are less wise, will have been supported by some sort of grudging welfare arrangement.*

Predictions about international cooperation are less precise - my impression is Asimov got the trend right, but the causal mechanism wrong

- *In short, there will be increasing co-operation among nations and among groups within nations, not out of any sudden growth of idealism or decency but out of a cold-blooded realization that anything less than that will mean destruction for all.* (It seems the increased coordination was driven more by trade)
- *By 2019, then, it may well be that the nations will be getting along well enough to allow the planet to live under the faint semblance of a world government by co-operation, even though no one may admit its existence.* (This is interesting: if anything has faint semblance of a world government by co-operation, it's probably the financial system / markets)

Predictions about education are precise with regard to opportunities. He would be probably disappointed how the opportunities are utilized, which is likely caused by the educational system having a lot of hidden goals different from education

- *There will be an opportunity finally for every youngster, and indeed, every person, to learn what he or she wants to learn. in his or her own time, at his or her own speed, in his or her own way.*
- *Education will become fun because it will bubble up from within and not be forced in from without.*

Overall, it seems to me the essay shows that futurology on this timescale is viable. (With the caveat that as the world got faster, comparable time horizon is likely shorter)

Prediction Markets Are About Being Right

Response To (Marginal Revolution): If you love prediction markets you should love the art world.

Previously on prediction markets: Prediction Markets: When Do They Work?, Subsidizing Prediction Markets

I'll quote the original in full, as it is short, and I found it interestingly and importantly wrong. By asking the question of *why* this perspective is wrong, we see what is so special about prediction markets versus other markets.

> Think of art markets, and art collecting, as an ongoing debate over what is beautiful and also what is culturally important. But unlike most debates, you have a very direct chance to "put your money where your mouth is," namely by buying art (it is very difficult to sell art short, however). In this regard, debates over artistic value may be among the most efficient debates in the world. At least if you are persuaded by the basic virtues of prediction markets. The prices of various art works really do aggregate information about their perceived values.
>
> I have, however, noted a correlation, how necessary or contingent I am not sure. The "white male nerd types" who are enamored of prediction markets tend to be especially skeptical of the market judgments of particular art works, most of all for conceptual and contemporary art.
>
> In my view, discussions about the value of art, as they occur in the off-the-record, proprietary sphere, are indeed of high value and they deserve to be studied more closely. Imagine a bunch of people competing to make "objects that are interesting but not interesting for reasons related to their practical value." And then we debate who has succeeded, or not. And those debates reflect many broader social, political, and economic issues. And it is all done with very real money on the line. The money concerns not just the value of individual art works, but also the prestige and social capital value that arises from having assembled a prestigious and insightful collection.

That's *exactly* why (almost) everyone who loves prediction markets hates the high-end, expensive art markets, even if they love art and artists and buy original paintings to hang on their walls. This goes beyond 'skepticism of the market judgments.' Expensive art markets are not fundamentally markets. They are fundamentally a political status game.

Consider three (non-exhaustive) types of markets: Consumption markets, commercial markets and prediction markets.

Consumption markets are where the buyer is buying the item in order to use it.

The buyer who pays more than necessary is sad in one sense, and the one who got the best deal is happy in that sense. But that sense isn't the important one for the buyer. If you are 'right' it is because you indeed got good use of the item that justified the purchase. If you are 'wrong' it is because you didn't.

Thus, we can point to a ‘naive’ participant who doesn’t ‘play the game’ of that market, and say ‘look how much they could have saved’, or did ‘save’, but that doesn’t actually impact them.

Liquid commercial markets are where the buyer plans to sell the item to someone else.

Middle men, arbitrage, investment, greater fools, that sort of thing. Buy low, sell high.

If you buy a stock, or a commodity piece of art, or inventory for your store, or a cryptocurrency, and others want to buy it for more, it goes up in price and you make money. If they want to sell it for less, it goes down in price and you lose money.

The buyer who pays more than necessary is sad, and loses money, in the only sense that matters. If the price goes down, that too makes the buyer sad. Paying a locally good price, or having the price go up, makes the buyer happy. The key is to buy before others buy, so they drive the price up.

You might reply, no, the bigger key is to buy what is cheap and sell what is expensive, based on fundamentals, and that will bear out over time.

Well, maybe.

Yes, often buyers and sellers are driven by fundamentals. But in an important sense, that is a coincidence. What is actually good news is often considered bad news, and vice versa. Prices are often largely driven by who is thinking about what and the emotional state and financial needs of participants. The market can stay insane longer than you can stay solvent. The people who say such non-fundamental movements are random, are mostly saying they aren’t good enough to understand and predict them in this case.

Yes, *eventually* fundamentals *might* take over. Or they might not. Low prices cause damage or make items impossible to justify storing or stocking. High prices trigger media attention and create opportunity. Low prices trigger margin calls, gets the company bought out or its employees and partners to quit. High prices trigger short squeezes and make everyone want to work with and for you. And so on. Momentum trading *works,* damn it (like everything else on this blog, not investment advice!).

Ideally the commercial market is anchored by connection to a consumption market – someone wants the goods, or is willing to collect the profits from the stock, or what not. The stronger that anchor versus speculative factors, the more accurate the prices.

Prediction markets have elements of both.

Prediction market traders can choose to mostly act like traders. If you think that others will think that the Patriots will win next week, you can bet on the Patriots now and then bet against them later when the odds change, and make money. You can be a market maker, or a block trader, or any other traditional market role.

In doing this, a trader cares about *future social reality.* They are people predicting what others will, in the future, predict that others will predict that others will predict, and so on. World events can help or hurt them, as they change perception, but they care about that perception and not the reality. By the time reality sets in, who knows what positions the trader will have?

In prediction markets there is another option. You can care about *future reality.* The market predicts a future outcome, and importantly *you can stay solvent longer than the market can stay insane.* Either the Patriots will win next week, or they will not. You can do better by using your commercial market tactics to grab the best possible price on the Patriots winning or losing, but the important thing is that *you win if you are right about the concrete thing, and you lose if you are wrong.*

This works because there is an *objective outcome,* and it occurs *quickly.* Thus it functions in its own way like a consumption market.

Truth matters.

If you choose, *only truth matters. I don't have to care what other people think. They don't determine if I win or lose.*

That's what I love, more than anything, about prediction markets. That's the reason behind many of the requirements of well-functioning prediction markets: They enable this sole reliance on truth, without imposing virtual taxes via long lock-up periods. This also enables prediction markets to output accurate predictions.

That's also a lot of what I love about trading. With a sufficiently deep and liquid market, *you win if and only if you are right.* No one gets to take that away from you and decide who gets the credit and the money. Only your skill mattered, and you reap what you deserve.

I *strongly encourage* the type of people who read this blog to strive to identify and work in such realms. *Be where being right, rather than being approved of, is rewarded.*

The world mostly does not work like this.

The world mostly *hates* prediction markets, *because they predict concrete consequences and outcomes accurately without taking into account what those in power, with high social status, want to be the prediction.*

Mostly, winners and losers are determined by social processes, status, coalitions, power, money and so on.

Credit and compensation mostly isn't based on who knew the truth and predicted accurately, or who did the work or created the value, or even what was stated in the contract. It is based on who has power and what they decide, based on what is good for them. History, along with everything else, gets decided by the winners.

That's life.

That's also expensive art, and expensive art markets, of the type Tyler speaks of. Only more so.

As I understand it (from, mostly, following Marginal Revolution links and posts) a small group determines who succeeds and fails, and buys art from each other, and manipulates the social reality of the art world and its prices to suit its fancies. Its fancies are mostly about the pursuit of conspicuous consumption, high social status and its associated rewards, wealth storage, money laundering and tax evasion, plus suckering outsiders and scamming them out of their money. Artistic merit, or aesthetics, are mostly a minor consideration.

Recall Tyler's description:

> Imagine a bunch of people competing to make "objects that are interesting but not interesting for reasons related to their practical value." And then we debate who has succeeded, or not. And those debates reflect many broader social, political, and economic issues. And it is all done with very real money on the line. The money concerns not just the value of individual art works, but also the prestige and social capital value that arises from having assembled a prestigious and insightful collection.

In this context, what does it mean for an object to 'be interesting'? It means having a high price, but mostly it means being judged as interesting by a high social status cabal that is primarily designed as an alliance of the high status connected people against everyone else. This need not be explicit at all – it is how such people instinctively operate, and you either learn those instincts or you never make it into the club.

There is no reason think any of this will ever "return to fundamentals" in any sense. The system sustains itself. There is (almost) no there, there. There never will be.

Thus, if I buy art, and people don't like me, they will find ways to charge me a lot more then they'd have charged an insider, and then they say therefore my art is not so valuable. Because I was buying it, and now I own it.

If I *hadn't* bought that piece, would it have become valuable? We'll never know. Was it valuable before I bought it? Also impossible to say.

That game is rigged, man. The only way to win is not to play.

If I think those people are wrong, I can *consume* the art by displaying it in my house and admiring it. If I want to spend a few hundred or thousand dollars on something I love, by all means I should go for it, but have zero illusions about the work becoming 'valuable.'

What I cannot do is *predict* that they are wrong, and wait for events to prove me right. There is no judgment day. No profit stream. No right. No wrong.

There are only cliques who watch each other to see if they are favoring the others in the clique, and use this to exploit others, because that's what winners and clique members with power and money do. It's sort of a market, like everything else. But in important senses, it is badly named, and something people like me despise. It is our failure mode and our doom, the way that prediction markets are our success mode and our hope.

Thus, if you love art markets you likely despise prediction markets, at least outside of their designated safe areas like sports and elections. And if you love prediction markets, you likely despise art markets whether or not you find them informative and fascinating in their own way.

What *none* of the people, whether they love or hate either market type, should be fooled by, is in accepting in a non-skeptical fashion the 'market prices' of 'art' in the art market. That is flat out not what is going on, at all. Such trades are not about the exchange of cash value for art value. Trying to use them to value the artwork misses the point entirely.

Are these art-market games worth understanding for what they can teach us about the world and how people work? Absolutely. Such shadowy practices do not get the light shined on them, that they deserve. Scams and exploitation and manipulation should be exposed. Political games as well. To blame and ideally punish those responsible, to protect people against them *and against having to play such games to succeed.* But more than that, to educate us about *how people, and how such systems, work.* Mostly, those who do understand how such things work only understand them from the inside, and do so in a non-intellectual fashion. With exposure, and as they see such actions succeed, they adopt their actions, views, instincts and very identity towards perpetuating such systems through imitation, usually without ever understanding what is going on in either themselves or the system at large.

Actually understanding how such things work might be the first step towards containing or overcoming such systems, or at least minimizing the damage they inflict on our lives, our status, our wealth and our souls.

It is also possible that such systems are in fact how anything actually gets done at all, and the exposure of more and more hypocritical and exploitative systems is making society unable to function, which would be far worse.

That’s a risk I am willing to take.

Book Review - Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness

In this post:

I. Introduction

Peter Godfrey-Smith's *Other Minds: the Octopus, the Sea, and the Deep Origins of Consciousness* is a phenomenal mishmash of octopus- and consciousness-related topics. It deals with everything from the evolution of octopuses, to their social life, to animal consciousness (including octopus consciousness), to evolutionary theories of aging, and more. All of this is tied together by a palpable fascination with octopuses, which manifests itself in rich descriptions of Godfrey-Smith's own experiences scuba-diving off the coast of Australia to observe them.

The book attempts to fit discussion of an impressive amount of interesting topics all into one slim volume. On the one hand, this is great, as each topic is fascinating in its own right, and several are relevant to EA/rationality. On the other hand, fitting in so many topics is a difficult task which the book only halfway pulls off. There wasn't enough room to discuss each topic in as much depth as they deserved, and the breadth of topics meant that the book felt somewhat unorganized and disunified. The book as a whole didn't seem to have any central claim; it was simply a collection of interesting facts, observations, musings, and theories that somehow relate to either octopuses, consciousness, or both, plus a bunch of fascinating first-hand descriptions of octopus behavior.

Do I recommend the book? Yes and no. For general interest, definitely--it's an interesting, enjoyable read; but for rationalists and EAs, there are probably better things to read on each topic the book discusses that would go into more depth, so it may not be the most effective investment of time for learning about, say, theories of animal consciousness. So in the rest of this review I've tried to 80:20 the book a bit, pulling out the insights that I found most interesting and relevant to EA/rationalism (as well as adding my own musings here and there). Because of this, the review is both quite long, and unavoidably reflects a bit of the disjointedness of the book itself--the sections can largely be read independently of each other.

Before I begin, a true-false test. For the following statements about octopuses, write down whether you think they are true or false, and how confident you are in your

response. We'll come back to these later in the review, and the answers will be at the end:

1. Octopuses can squirt jets of ink as an escape tactic.
2. Octopuses have color vision.
3. Octopuses have bilateral symmetry.
4. Octopuses can camouflage themselves nearly perfectly by changing the color and texture of their skin to match whatever surface or object they are trying to blend into.
5. Octopuses can fit through any hole or gap bigger than their eye.
6. Octopuses can recognize individual humans.
7. Most octopus species live for more than 20 years.
8. Octopuses are mostly solitary animals.
9. Octopuses have been known to use shards of glass from shattered bottles on the seafloor as weapons to fight other octopuses.

Answers below.

II. What is it like to be an octopus?

The nervous system of an octopus is structured quite differently than a mammalian nervous system. The octopus not only has a high concentration of neurons in its head (a “brain”), but also clusters of neurons throughout their body, particularly in each arm. At one point Godfrey-Smith notes that the number of neurons in the octopus's central brain is only a little over half of the number of neurons in the rest of its body (67). This means that each arm must have, roughly, 1/4 as many neurons as the central brain. What might it feel like to have 8 other "brains" in your body, each 1/4 the size of your "main brain"?[^1]

Most likely, it feels completely different than we're even capable of imagining. Godfrey-Smith doesn’t discuss the question, and there are some philosophers, such as Daniel Dennett, who would deny that it even makes sense. Nevertheless, there are two main points of reference that I can think of that humans might use to imagine something of what this would feel like.

First, we humans have what is sometimes called a "gut brain": ~500 million neurons in our digestive tracts, about as many as in a cat's brain, that control our digestive process. What does it feel like to have this brain? Well, hunger signals, the gag reflex, and some emotions (via the limbic system), are controlled by this brain, so roughly, it feels like what those mental states feel like. Perhaps the octopus has similar signals that come from its arm brains. These likely feel nothing like the signals coming from our gut brain, and would of course have different functions: while our gut brain sends us hunger or disgust reflexes, the octopus's arm signals might function something like our propioceptive sense, informing the "main brain" about the location of the arms, and maybe also relaying touch and taste/smell information from the respective sense organs located on the arms.

Or perhaps the arm brains don't just relay sensory information from the arms, perhaps they also play a part in controlling the arms' movements (Godfrey-Smith posits that this is the main reason for why the octopus's nervous system is so distributed). Sticking with the gut-brain/arm-brain analogy for the moment, what does it feel like for the gut-brain to control the stomach? That is, what does it feel like to digest food?

... Often like nothing at all. We sometimes don't even notice digestion occurring, and we certainly don't have any detailed sensation of what's going on when it does. So perhaps the octopus just tells its arms where to go in broad strokes, and they take it from there, similar to how our gut-brain just "takes it from there."

The idea that the arm brains control the movement of the arms also brings me to my second comparison: split-brain syndrome. Split-brain results when the corpus callosum is surgically severed (usually as a treatment for epilepsy). After this operation, patients can function mostly normally, but it can be experimentally determined that the two sides of their body function somewhat independently of each other. For example:

> a patient with split brain is shown a picture of a chicken foot and a snowy field in separate visual fields and asked to choose from a list of words the best association with the pictures. The patient would choose a chicken to associate with the chicken foot and a shovel to associate with the snow; however, when asked to reason why the patient chose the shovel, the response would relate to the chicken (e.g. "the shovel is for cleaning out the chicken coop"). (source) (original source)

Notoriously, the two sides can sometimes get into conflict, as when one split-brain patient violently shook his wife with his left hand while trying to stop himself with his right hand.

What does it feel like for that to happen? Well, I don't actually know. But I wonder, does this person feel touch sensations from their left hand? If so, are these produced by the same processes that patch over our blind spot, or are they actual touch sensations? Suppose you put a split-brain patient's left hand behind an opaque barrier; would they be able to tell when something touched it?

Assuming that the answer to this question is "no," another possibility for what having "arm-brains" could feel like is "basically nothing": just as the split-brain patient only knows what their left side is doing through external cues, like seeing it move, and doesn't have any control over it, so too with the octopus's arms. It doesn't really "feel" what's going on in its arms, the arms themselves know what's going on and that's enough.

III. An exercise in updating on surprising facts

[Note: this section is largely recycled from one of my shortform posts; if you've already read that, this will be mostly redundant. If you haven't read that, simply read on.]

I confess, the purpose of the true-false test at the beginning of this review was largely to disguise one question in particular, so that the mere asking of it didn't provide Bayesian evidence that the answer should be surprising: "6.) Octopuses can recognize individual humans." Take a moment to look back at your answer to that question. What does your model say about whether octopuses should be able to recognize individual humans? Why can humans recognize other individual humans, and what does that say about whether octopuses should be able to?

...

...

...

...

...

...

As it turns out, octopuses can recognize individual humans. For example, in the book it's mentioned that at one lab, one of the octopuses had a habit of squirting jets of water at one particular researcher. Take a moment to let it sink in how surprising this is: octopuses, which 1.) are mostly nonsocial animals, 2.) have a completely different nervous system structure that evolved on a completely different branch of the tree of life, and 3.) have no evolutionary history of interaction with humans, can recognize individual humans, and differentiate them from other humans. I'm pretty sure humans have a hard time differentiating between individual octopuses.

And since things are not inherently surprising, only surprising to models, this means my world model (and yours, if you were surprised by this) needs to be updated. First, generate a couple of updates you might make to your model after finding this out. I'll wait...

...

...

...

...

...

...

...

Now that you've done that, here's what I came up with:

(1) Perhaps the ability to recognize individuals isn't as tied to being a social animal as I had thought (2) Perhaps humans are easier to tell apart than I thought (i.e. humans have more distinguishing features, or these distinguishing features are larger/more visually noticeable, etc., than I thought) (3) Perhaps the ability to distinguish individual humans doesn't require a specific psychological module, as I had thought, but rather falls out of a more general ability to distinguish objects from each other (Godfrey-Smith mentions this possibility in the book). (4) Perhaps I'm overimagining how fine-grained the octopus's ability to distinguish humans is. I.e. maybe that person was the only one in the lab with a particular hair color or something, and they can't distinguish the rest of the people. (Though note, another example given in the book was that one octopus liked to squirt new people, people it hadn't seen regularly in the lab before. This wouldn't mesh very well with the "octopuses can only make coarse-grained distinctions between people" hypothesis.)

To be clear, those were my first thoughts; I don't think all of them are correct. As per my shortform post about this, I'm mostly leaning towards answer (2) being the correct update -- maybe the reason octopuses can recognize humans but not the other way

around is mostly because individual humans are just more visually distinct from each other than individual octopuses, in that humans have a wider array of distinguishing features or these features are larger or otherwise easier to notice. But of course, these answers are neither mutually exclusive nor exhaustive. For example, I think answer (3) also probably has something to do with it. I suspect that humans probably have a specific module for recognizing humans, but it seems clear that octopuses couldn't have such a module, so it must not be strictly necessary in order to tell humans apart. Maybe a general object-recognizing capability plus however visually distinct humans are from each other is enough.[^2]

I'd also love to hear in the comments what updates other people had from this.

IV. Animal consciousness

Something else from the book that I found interesting concerns animal consciousness/subjective experience. I suspect this is old hat for those who have done any significant research into animal suffering, but it added a couple more gears to my model of animal consciousness, so I'll share it here for those whose models were similarly gear-less. Remember blindsight (where people who are blind due to damage in their visual cortex can perform better than chance at vision tasks, because the rest of their brain still gets visual information, even though they don't have access to it consciously)? A pair of vision scientists (Milner and Goodale) believe, roughly, that that's what's going on in frogs all the time. What convinced them of this is an experiment performed by David Ingle in which he was able to surgically reverse some, but not all, of the visual abilities of some froggy test subjects. Namely, when his frogs saw a fly in one side of their visual field, they would snap as if it were on the other, but they were able to go around barriers perfectly normally. Milner and Goodale take this as evidence that the frog doesn't have an integrated visual experience at all. They write:

> So what did these rewired frogs "see"? There is no sensible answer to this. The question only makes sense if you believe that the brain has a single visual representation of the outside world that governs all of an animal's behavior. Ingle's experiments reveal that this cannot possibly be true. (Milner and Goodale 2005, qtd. in Peter Godfrey-Smith, Other Minds, 2016, p. 80)

Godfrey-Smith then goes on to discuss Milner and Goodale's view:

> Once you accept that a frog does not have a unified representation of the world, and instead has a number of separate streams that handle different kinds of sensing, there is no need to ask what the frog sees: in Milner and Goodale's words, "the puzzle disappears." Perhaps one puzzle disappears, but another is raised. What does it feel like to be a frog perceiving the world in this situation? I think Milner and Goodale are suggesting that it feels like nothing. There is no experience here because the machinery of vision in frogs is not doing the sorts of things it does in us that give rise to subjective experience. (Godfrey-Smith, pp. 89-90)[^3]

Though he doesn't mention it, there seems to me to be an obvious reply here: the phenomenon of blindsight reveals that there are parts of our visual processing that don't feel like anything to us (or perhaps, as Godfrey-Smith prefers, they feel like something, just not like vision), but this clearly doesn't change the fact that we (most of us) do have visual experience. Why couldn't something similar be going on in the

frogs? They have a visual field, but they also have other visual processing going on as well which doesn't make it into their visual field.

Let me try to explain this thought a bit better. One thing that the human blindsight subject described in Other Minds (known as "DF") was able to do was put letters through a mail-slot placed at different angles. Now those of us who have normal sight presumably still do all the same processing as those with blindsight, plus some extra. So, imagine someone performing brain surgery on a person with normal vision, which affected whatever brain circuitry allows people to align a letter at the correct angle to get it through a mail-slot. At the risk of arguing based on evidence I haven't seen yet, one way I could imagine the scenario playing out is the following: this person would wake up and find that, though their visual experience was the same as before, for some reason they couldn't manage to fit letters through mail-slots anymore. They would experience this in a similar way as someone with exceptionally poor balance experiences their inability to walk a tightrope--it's not as though they can't see where the rope is, they just can't manage to put their feet in the right place to stay on it. I'd guess that the same thing would happen for the person, and that the same thing is happening for the frog. Respectively: it's not as though the person can't see where the mail-slot is, they just can't manage to get the letter through it, and it's not as though the frog can't see where the fly is, it just can't seem to get its tongue to move in the right direction to catch it.[^4]

In any case, even if we discount this argument, does Milner and Goodale's argument amount to an argument that most animals don't have inner lives, and in particular that they don't feel pain?

Not so, Godfrey-Smith wants to argue. He includes some discussion of various theories of consciousness/subjective experience and how early or late it arose,[^5] but what interested me was an experiment that tried to test whether an animal, in this case a Zebrafish, actually feels pain, or is only performing instinctive behaviors that look to us like pain.

The experiment goes like this: There are two environments, A and B, and the fish is known to prefer A to B. The experimenter injects the fish with a chemical thought to be painful. Then, the experimenter dissolves painkiller in environment B, and lets the fish choose again which environment it prefers. With the painkiller and the painful chemical, the fish prefers environment B (though with the painful chemical and no painkiller, it still prefers A). The fish seems to be choosing environment B in order to relieve its pain, and this isn't the kind of situation that the fish could have an evolved reflex to react to. Since the fish is behaving as we would expect it to if it felt pain and the opposite of how we would expect it to if it didn't feel pain, and a reflex can't be the explanation, this is evidence that the fish feels pain, rather than simply seeming to feel pain.

What excited me about this was the idea that we could use experiments to tell something about the inner lives of animals. Even though I've been thoroughly disabused of the idea of a philosophical zombie,[^6] I still had the idea that subjective experience is something that can't really be tested "from the outside." Reading about these experiments made me much more optimistic that experiments could be useful to help determine whether and which animals are moral patients.

V. Aging

Another fact that might surprise (and perhaps sadden) you: octopuses, for the most part, only live about 2 years. One might think that intelligence is most advantageous when you live long enough to benefit from the things you learn with it. Nevertheless, octopuses only live about 2 years. Why is this? Godfrey-Smith posits that octopuses evolved intelligence not for the benefits of long-term learning, but simply to control their highly-amorphous bodies. Since an octopus's body can move so freely, it takes a very large nervous system to control it, which gave rise to what intelligence they possess. Even so, once they had intelligence, shouldn't this have caused selection pressure towards longer lives? I'm still confused on this count, but this does lead us to another question: why do most living organisms age in the first place? There are organisms that don't, at least on the timescales we've observed them on so far, so why are there any that do? What evolutionary benefit does aging provide, could it provide? One would think that aging, at least once an organism had reached maturity, would be strictly disadvantageous and thus selected against, so why do we mostly observe organisms that age and die?

Godfrey-Smith surveys several standard theories, but the one he presents as most likely to be correct (originated by Peter Medawar and George Williams) is as follows. Imagine an organism that didn't age; once it reached its prime, it remained that way, able to survive and reproduce indefinitely until it died of e.g. predation, disease, a falling rock, or some other external cause, all of which I'll call "accidental death." If we assume the average probability of dying by accidental death is constant each year, then the organism's probability p_n of surviving to age n decreases as n increases.

Thus, for large enough n, p_n approaches 0, meaning that there is some age n which the organism is almost certain to die before reaching, even without aging. Now imagine that the organism has a mutation with effects that are positive before age n, but negative after age n. Such a mutation would have almost no selection pressure against it, since the organism would almost certainly die of accidental death before its negative effects could manifest. Thus, such mutations could accumulate, and the few organisms that did survive to age n would start to show those negative effects.

The truth is more general than that. In general, as p_n gets lower, so does the selection pressure against any mutation whose negative effects only appear after age n. This theory predicts that organisms should exhibit a slow and steady increase of negative symptoms caused by mutations whose negative side effects only show up later, and an age which almost no individuals survive beyond, which is what we in fact observe.

Still though, why should there be any *positive* pressure towards these mutations, even if there's little pressure against them? Because, as I mentioned, at least some of these mutations might have positive effects that show up earlier bound up with the negative effects that show up later. This positive selection pressure, combined with the reduced negative selection pressure due to their negative effects only showing up late, after most with the mutation have already died due to accidental death, is enough to get these mutations to fixation. Godfrey-Smith uses the analogy, originally due to George Williams, of putting money in a savings account to be accessed when you're 120 years old. You'll almost certainly be dead by then, so it's rather pointless to save for that far off. In the same way, it's evolutionarily pointless for organisms to pass up mutations that have positive effects now and negative effects later when those negative effects only show up after the animal is almost certain to be dead by

accidental death. So organisms take those mutations, and most do not survive to pay the price; aging is what happens to those who do.

If this is the correct evolutionary account of why aging occurs, it has an interesting implication for anti-aging research: there might be certain routes to eliminating aging that come with unforeseen downsides. If we were to eliminate aging by finding the genes that produce these negative side effects and turning them off (please forgive my utter ignorance of genetics and the science of aging), this could also rob us of whatever benefits those genes provided earlier in life that caused them to be adopted in the first place. This is not to say that we should not pursue anti-aging research (in fact I'm strongly in favor of it), but just that we should be on the lookout for this kind of trap, and avoid it if we can.

Appendix: Answers to True-False Questions

1. Octopuses can squirt jets of ink as an escape tactic. *True*
2. Octopuses have color vision. *False*
3. Octopuses have bilateral symmetry. *True*
4. Octopuses can camouflage themselves nearly perfectly by changing the color and texture of their skin to match whatever surface or object they are trying to blend into. *True*
5. Octopuses can fit through any hole or gap bigger than their eye. *True*
6. Octopuses can recognize individual humans. *True*
7. Most octopus species live for more than 20 years. *False*
8. Octopuses are mostly solitary animals. *True*
9. Octopuses have been known to use shards of glass from shattered bottles on the seafloor as weapons to fight other octopuses. *As far as I know, false*

Notes:

[^1]: To give a sense of the relationship between the octopus's central brain and its arms, here are some quotes from the book:

> How does an octopus's brain relate to its arms? Early work, looking at both behavior and anatomy, gave the impression that the arms enjoyed considerable independence. The channel of nerves that leads from each arm back to the central brain seemed pretty slim. Some behavioral studies gave the impression that octopuses did not even track where their own arms might be. As Roger Hanlon and John Messenger put it in their book Cephalopod Behavior, the arms seemed "curiously divorced" from the brain, at least in the control of basic motions. (67)
>
> Some sort of mixture of localized and top-down control might be operating. The best experimental work I know that bears on this topic comes out of Binyamin Hochner's laboratory at the Hebrew University of Jerusalem. A 2011 paper by Tamar Gutnick, Ruth Byrne, and Michael Kuba, along with Hochner, described a very clever experiment. They asked whether an octopus could learn to guide a single arm along a maze-like path to a specific place in order to obtain food. The task was set up in such a way that the arm's own chemical sensors would not suffice to guide it to the food; the arm would have to leave the water at one point

> to reach the target location. But the maze walls were transparent, so the target location could be seen. The octopus would have to guide an arm through the maze with its eyes. It took a long while for the octopuses to learn to do this, but in the end, nearly all of the octopuses that were tested succeeded. The eyes can guide the arms. At the same time, the paper also noted that when the octopuses are doing well with this task, the arm that's finding the food appears to do its own local exploration as it goes, crawling and feeling around. So it seems that two forms of control are working in tandem: there is central control of the arm's overall path, via the eyes, combined with a fine-tuning of the search by the arm itself. (68-69)

[^2]: So why do I still think humans have a specific module for it? Here's one possible reason: I'm guessing octopuses can't recognize human faces--they probably use other cues, though nothing in the book speaks to this one way or the other. If that's the case, then it might be true both that a general object-differentiating capability is enough to recognize individual humans, but that to recognize faces requires a specific module. If I found out that octopuses could recognize human faces specifically, not just individual humans by other means than face-recognition, I would strongly update in favor of humans having no specific face- or other-person-recognition module. In the same vein, the fact that people can lose the ability to recognize faces without it affecting any other visual capacities (known as prosopagnosia or "face-blindness") suggests that a single module is responsible for that ability.

[^3]: After reading Daniel Dennett's Consciousness Explained, it's actually not at all clear to me why Godfrey-Smith interprets Milner and Goodale this way. It seems more natural to suppose that they're suggesting something similar to Dennett's denial of the "Cartesian Theater" (the idea that there is somewhere where "it all comes together" in the brain, in some sort of "inner movie" to use Chalmers' phrase) and his replacement, the "Multiple Drafts Model" (which I don't feel confident to summarize here).

[^4]: Another way this might play out is if the frog saw the fly and only the fly as reversed in its visual field, rather like a hallucination. I don't see any reason why that would be impossible.

[^5]: Godfrey-Smith actually makes a distinction between "subjective experience" and "consciousness." The way Godfrey-Smith uses these words, when we say that something has "subjective experience," we're just saying that there is something that it feels like to be that thing, while the claim that something has "consciousness" is in some unspecified way stronger. So consciousness is a subset of subjective experience. He speculates that subjective experience arose fairly early, in the form of things like hunger signals and pain, while consciousness arose later and involves things like memory, a "global workspace," integrated experience, etc.

[^6]: See Dennett, *Intuition Pumps and Other Tools for Thinking*, Ch. 55 "Zombies and Zimboes," and Eliezer Yudkowsky's essay "Zombies! Zombies?"

Two Neglected Problems in Human-AI Safety

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

In this post I describe a couple of human-AI safety problems in more detail. These helped motivate my proposed hybrid approach, and I think need to be addressed by other AI safety approaches that currently do not take them into account.

1. How to prevent "aligned" AIs from unintentionally corrupting human values?

We know that ML systems tend to have problems with adversarial examples and distributional shifts in general. There seems to be no reason not to expect that human value functions have similar problems, which even "aligned" AIs could trigger unless they are somehow designed not to. For example, such AIs could give humans so much power so quickly or put them in such novel situations that their moral development can't keep up, and their value systems no longer apply or give essentially random answers. AIs could give us new options that are irresistible to some parts of our motivational systems, like more powerful versions of video game and social media addiction. In the course of trying to figure out what we most want or like, they could in effect be searching for adversarial examples on our value functions. At our own request or in a sincere attempt to help us, they could generate philosophical or moral arguments that are wrong but extremely persuasive.

(Some of these issues, like the invention of new addictions and new technologies in general, would happen even without AI, but I think AIs would likely, by default, strongly exacerbate the problem by differentially accelerating such technologies faster than progress in understanding how to safely handle them.)

2. How to defend against intentional attempts by AIs to corrupt human values?

It looks like we may be headed towards a world of multiple AIs, some of which are either unaligned, or aligned to other owners or users. In such a world there's a strong incentive to use one's own AIs to manipulate other people's values in a direction that benefits oneself (even if the resulting loss to others are greater than gains to oneself).

There is an apparent asymmetry between attack and defense in this arena, because manipulating a human is a straightforward optimization problem with an objective that is easy to test/measure (just check if the target has accepted the values you're trying to instill, or has started doing things that are more beneficial to you), and hence relatively easy for AIs to learn how to do, but teaching or programming an AI to help defend against such manipulation seems much harder, because it's unclear how to distinguish between manipulation and useful information or discussion. (One way to defend against such manipulation would be to cut off all outside contact, including from other humans because we don't know whether they are just being used as other AIs' mouthpieces, but that would be highly detrimental to one's own moral development.)

There's also an asymmetry between AIs with simple utility functions (either unaligned or aligned to users who think they have simple values) and AIs aligned to users who have high value complexity and moral uncertainty. The former seem to be at a substantial advantage in a contest to manipulate others' values and protect one's own.

What does it mean to "believe" a thing to be true?

Three AI Safety Related Ideas

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

(I have a health problem that is acting up and making it hard to type for long periods of time, so I'm condensing three posts into one.)

1. AI design as opportunity and obligation to address human safety problems

Many AI safety problems are likely to have counterparts in humans. AI designers and safety researchers shouldn't start by assuming that humans are safe (and then try to inductively prove that increasingly powerful AI systems are safe when developed/trained by and added to a team of humans) or try to solve AI safety problems without considering whether their designs or safety approaches exacerbate human safety problems relative to other designs / safety approaches. At the same time, the development of AI may be a huge opportunity to address human safety problems, for example by transferring power from probably unsafe humans to de novo AIs that are designed from the ground up to be safe, or by assisting humans' built-in safety mechanisms (such as moral and philosophical reflection).

2. A hybrid approach to the human-AI safety problem

Idealized humans can be safer than actual humans. An example of idealized human is a human whole-brain emulation that is placed in a familiar, safe, and supportive virtual environment (along with other humans for socialization), so that they are not subject to problematic "distributional shifts" nor vulnerable to manipulation from other powerful agents in the physical world. One way to take advantage of this is to design an AI that is ultimately controlled by a group of idealized humans (for example, has a terminal goal that refers to the reflective equilibrium of the idealized humans), but this seems impractical due to computational constraints. An idea to get around this is to give the AI an advice or hint, that it can serve that terminal goal by learning from actual humans as an instrumental goal. This learning can include imitation learning, value learning, or other kinds of learning. Then, even if the actual humans become corrupted, the AI has a chance of becoming powerful enough to discard its dependence on actual humans and recompute its instrumental goals directly from its terminal goal. (Thanks to Vladimir Nesov for giving me a hint that led to this idea.)

3. Several approached to AI alignment will differentially accelerate intellectual progress that are analogous to solving problems that are low in the polynomial hierarchy.

This is bad if the "good" kind of intellectual progress (such as philosophical progress) is disproportionally high in the hierarchy or outside PH entirely, or if we just don't know how to formulate such progress as problems low in PH. I think this issue needs to be on the radar of more AI safety researchers.

(A reader might ask, "differentially accelerate relative to what?" An "aligned" AI could accelerate progress in a bad direction relative to a world with no AI, but still in a good direction relative to a world with only unaligned AI. I'm referring to the former here.)

[Video] Why Not Just: Think of AGI Like a Corporation? (Robert Miles)

This is a linkpost for https://www.youtube.com/watch?v=L5pUA3LsEaw

Robert Miles has been creating AI-Alignment related videos for a while now, but I found this one particularly good.

Here is the automatically generated Youtube transcript. Obviously it's not very good, but at least it makes the post searchable (In case Robert reads this and has a transcript for the video lying around, I would love to replace this with one that has proper capitalization and punctuation marks and other luxuries):

hi so I sometimes see people saying

things like okay so your argument is

that at some point in the future we're

going to develop intelligent agents that

are able to reason about the world in

general and take actions in the world to

achieve their goals

these agents might have superhuman

intelligence that allows them to be very

good at achieving their goals and this

is a problem because they might have

different goals from us but don't we

kind of have that already corporations

can be thought of as super intelligent

agents they're able to think about the

world in general and they can outperform

individual humans across a range of

cognitive tasks and they have goals

namely maximizing profits or shareholder

value or whatever and those goals aren't

the same as the overall goals of

humanity so corporations are a kind of misaligned super intelligence the people who say this having established the metaphor at this point tend to diverge mostly along political lines some say corporations are therefore a clear threat to human values and goals in the same way that misaligned super intelligences are and they need to be much more tightly controlled if not destroyed all together others say corporations are like misaligned super intelligences but corporations have been instrumental in the huge increases in human wealth and well-being that we've seen over the last couple of centuries with pretty minor negative side effects overall if that's the effect of misaligned super intelligences I don't see why we should be concerned about AI and others say corporations certainly have their problems but we seem to have developed systems that keep them under control well enough that they're able to create value and do useful things without literally killing everyone so perhaps we can learn something about how to control or align super intelligences by looking at how we handle corporations

so we're gonna let the first to fight
amongst themselves and we'll talk to the
third guy so how good is this metaphor
our corporations really like misaligned
artificial general super intelligences
quick note before we start we're going
to be comparing corporations to AI
systems and this gets a lot more
complicated when you consider that
corporations in fact use AI systems so
for the sake of simplicity we're going
to assume that corporations don't use AI
systems because otherwise the problem
gets recursive and like not in a cool
way
first off our corporations agents in the
relevant way I would say yeah pretty
much I think that it's reasonably
productive to think of a corporation as
an agent
they do seem to make decisions and take
actions in the world in order to achieve
goals in the world but I think you face
a similar problem thinking of
corporations as agents as you do when
you try to think of human beings as
agents in economics it's common to model
human beings as agents that want to
maximize their money in some sense and

you can model corporations in the same
way and this is useful but it is kind of
a simplification in that human beings in
practice want things that aren't just
money
and while corporations are more directly
aligned with profit maximizing than
individual human beings are it's not
quite that simple so yes we can think of
corporations as agents but we can't
treat their stated goals as being
exactly equivalent to their actual goals
in practice more on that later so
corporations are more or less agents are
they generally intelligent agents again
yeah I think so I mean corporations are
made up of human beings so they have all
the same general intelligence
capabilities that human beings have so
then the question is are they super
intelligent this is where things get
interesting because the answer is kind
of like SpaceX is able to design a
better rocket than any individual human
engineer could design rocket design is a
cognitive task and SpaceX is better at
that than any human being therefore
SpaceX is a super intelligence in the
domain of rocket design but a calculator

is a super intelligence in the domain of arithmetic that's not enough our corporation's general super intelligences do they outperform humans across a wide range of cognitive tasks as an AGI code in practice it depends on the task consider playing a strategy game for the sake of simplicity let's use a game that humans still beat AI systems at like Starcraft if a corporation for some reason had to win at Starcraft it could perform about as well as the best human players it would do that by hiring the best human players but you won't achieve superhuman play that way a human player acting on behalf of the corporation is just a human player and the corporation doesn't really have a way to do much better than that

a team of reasonably good Starcraft players working together to control one army will still lose to a single very good player working alone this seems to be true for a lot of strategy games the classic example is the game of Kasparov versus the world where Garry Kasparov played against the entire rest of the world cooperating on the Internet

the game was kind of weird but Kasparov ended up winning and the kind of real world strategy that corporations have to do seems like it might be similar as well when companies outsmart their competition it's usually because they have a small number of decision makers who are unusually smart rather than because they have a hundred reasonably smart people working together for at least some tasks teams of humans are not able to effectively combine their intelligence to achieve highly superhuman performance so corporations are limited to around human level intelligence of those tasks to break down where this is let's look at some different options corporations have four ways to combine human intelligences one obvious way is specialization if you can divide the task into parts that people can specialize in you can outperform individuals you can have one person who's skilled at engine design one who's great at aerodynamics one who knows a lot about structural engineering and one who's good at avionics can you tell I'm not a rocket surgeon anyway if these people with their different skills are

able to work together well with each
person doing what they're best at the
resulting agent will in a sense have
superhuman intelligence no single human
could ever be so good at so many
different things but this mechanism
doesn't get you superhumanly high
intelligence just superhumanly broad
intelligence whereas super intelligence
software AGI might look like this so
specialization yields a fairly limited
form of super intelligence if you can
split your task up but that's not easy
for all tasks for example the task of
coming up with creative ideas or
strategies isn't easy to split up you
either have a good idea or you don't but
as a team you can get everyone to
suggest a strategy or idea and then pick
the best one that way a group can
perform better than any individual human
how much better though and how does that
change with the size of the team I got
curious about exactly how this works so
I came up with a toy model now I'm not a
statistician I'm a computer scientist so
rather than working it out properly I
just simulated it a hundred million
times because that was quicker okay so

here's the idea quality distribution for
an individual human will model it as a
normal distribution with a mean of 100
and a standard deviation of 20 so what
this means is you ask a human for a
suggestion and sometimes they do really
well and come up with a hundred
30-level strategy sometimes they screw
up and can only give you a 70 idea but
most of the time it's around 100 now
suppose we had a second person whose
intelligence is the same as the first we
have both of them come up with ideas and
we keep whichever idea is better the
resulting team of two people combined
looks like this
on average the ideas are better the mean
is now 107 and as we keep adding people
the performance gets better here's 5
people 10 20 50 100
remember these are probability
distributions so the height doesn't
really matter the point is that the
distributions move to the right and get
thinner the average idea quality goes up
and the standard deviation goes down so
we're coming up with better ideas and
more reliably but you see how the
progress is slowing down we're using a

hundred times as much brain power here
but our average ideas are only like 25%
better what if we use a thousand people
ten times more resources again only gets
us up to around a hundred and thirty
five diminishing returns so what does
this mean for corporations well first
off to be fair this team of a thousand
people is clearly super intelligent the
worst ideas it ever has are still so
good that an individual human will
hardly ever manage to think of them but
it's still pretty limited there's all
this space off to the right of the graph
that it would take vast team sizes to
ever get into if you're wondering how
this would look with seven billion
humans well you have to work out the
statistical solution yourself the point
is the team isn't that super intelligent
because it's never going to think of an
idea that no human could think of which
is kind of obvious when you think about
it but AGI is unlimited in that way and
in practice even this model is way too
optimistic for corporations firstly
because it assumes that the quality of
suggestions for a particular problem is
uncorrelated between humans which is

clearly not true and secondly because
you have to pick out the best suggestion
but how can you be sure that you'll know
the best idea when you see it it happens
to be true a lot of the time for a lot
of problems that we care about that
evaluating solutions is easier than
coming up with them you know Homer it's
very easy to criticize machine learning
relies pretty heavily on this like
writing a program that differentiates
pictures of cats and dogs is really hard
but evaluating such a program is fairly
simple you
show it lots of pictures of cats and
dogs and see how well it does the clever
bit is in figuring out how to take a
method for evaluating solutions and use
that to create good solutions anyway
this assumption isn't always true and
even when it is the fact that evaluation
is easier or cheaper than generation
doesn't mean that evaluation is easy or
cheap
like I couldn't generate a good rocket
design myself but I can tell you that
this one needs work so evaluation is
easier than generation but that's a very
expensive way to find out and I wouldn't

have been able to do it the cheap way by
just looking at the blueprints the
skills needed to evaluate in advance
whether a given rocket design will
explode are very closely related to the
skills needed to generate a non
exploding rocket design so yeah even if
a corporation could somehow get around
being limited to the kind of ideas that
humans are able to generate they're
still limited to the kind of ideas that
humans are able to recognize as good
ideas just how serious is this
limitation how good are the strategies
and ideas that corporations are missing
out on well take a minute to think of an
idea that's too good for any human to
recognize it as good got one well it was
worth a shot we actually do have an
example of this kind of thing in move 37
from alphago's 2016 match with world
champion Lisa doll this kind of
evaluation value that's a very that's a
very surprising move I thought I thought
it was I thought it was a mistake yeah
that turned out to be pretty much the
move that won the game but you're go
playing corporation is never going to
make move 37 even if someone happens to

suggest it it's almost certainly not
going to be chosen
normally human we never play this one
because it's not enough for someone in
your corporation to have a great idea
the people at the top need to recognize
that it's a great idea that means that
there's a limit on the effective
creative or strategic intelligence of a
corporation which is determined by the
intelligence of the decision-makers and
their ability to know a good idea when
they see one okay what about speed
that's one of the things that makes AI
systems so powerful and one of the ways
that software IGI is likely to be super
intelligent the general trend is we go
from computer
can't do this at all two computers can
do this much faster than people not
always but in general so I wouldn't be
surprised if that pattern continues with
AGI how does the corporation rate on
speed again it kind of depends
this is closely related to something
we've talked about before parallelizable
ax t some tasks are easy to split up and
work on in parallel and some aren't
for example if you've got a big list of

a thousand numbers and you need to add
them all up it's very easy to paralyze
if you have ten people you can just say
okay you take the first hundred numbers
you take the second hundred you take the
third and so on have everybody add up
their part of the list and then at the
end you add up everyone's totals however
long the list is you can throw more
people at it and get it done faster much
faster than any individual human code
this is the kind of task where it's easy
for corporations to achieve superhuman
speed but suppose instead of summing a
list you have a simple simulation that
you want to run for say a thousand
seconds you can't say okay you work out
the first hundred seconds of the
simulation you do the next hundred and
you do the next hundred and so on
because obviously the person who's
simulating second 100 needs to know what
happened at the end of second 99 before
they can get started so this is what's
called an inherently serial task you
can't easily do it much faster by adding
more people you can't get a baby in less
than nine months by hiring two pregnant
women

you know most real-world tasks are somewhere in between you get some benefits from adding more people but again you hit diminishing returns some parts of the task can be split up and worked on in parallel some parts need to happen one after the other so yes corporations can achieve superhuman speed add some important cognitive tasks but really if you want to talk about speed in a principled way you need to differentiate between throughput how much goes through the system within a certain time and latency how long it takes a single thing to go through the system these ideas are most often used in things like networking and I think that's the easiest way to explain it so basically let's say you need to send someone a large file and you can either send it over a dial-up internet connection or you can send them a physical disk through the postal system the dial-up connection is low latency each bit of the file goes through the system quickly but it's also low throughput the rate at which you can send data is pretty low whereas sending the physical disk is high latency it

might take days for the first
to arrive but it's also high-throughput
you can put vast amounts of data on the
disk so your average data sent per
second could actually be very good
corporations are able to combine human
intelligences to achieve superhuman
throughput so they can complete large
complex tasks faster than individual
humans could but the thing is a system
can't have lower latency than its
slowest component and corporations are
made of humans so corporations aren't
able to achieve superhuman latency and
in practice as you've no doubt
experienced is quite the opposite so
corporate intelligence is kind of like
sending the physical disk corporations
can get a lot of cognitive work done in
a given time but they're slow to react
and that's a big part of what makes
corporations relatively controllable
they tend to react so slowly that even
governments are sometimes able to move
fast enough to deal with them
software super intelligence is on the
other hand could have superhuman
throughput and superhuman latency which
is something we've never experienced

before in a general intelligence so our corporations super intelligent agents well they're pretty much generally intelligent agents which are somewhat super intelligent in some ways and somewhat below human performance in others so yeah kinda the next question is are they misaligned but this video is already like 14 and a half minutes long so we'll get to that in the next video

[Music]

I want to end the video by saying a big thank you to my excellent patrons it's all of these people here in this video I'm especially thanking Pablo area or Pablo a de aluminio Sushil recently I've been putting a lot of time into some projects that I'm not able to talk about but as soon as I can and the patrons will be the first to know

thank you again so much for your generosity and thank you all for watching I'll see you next time

[Music]

Argue Politics* With Your Best Friends

Epistemic Status: I endorse this strongly but don't think I'm being original or clever at all.

Until recently — yesterday, in fact — I was seriously wrong about something.

I thought that it was silly when I saw people spending lots of energy arguing with their *closest* friends who *almost completely agreed with them, but not quite.*

That's some People's Front Of Judaea shit, I thought. Don't you know that guy you're arguing with so vehemently is your friend? He likes you! He's a pretty good guy! He even shares your values and models, *almost* completely! He's only wrong about this one, itty bitty, relatively abstract thing!

Meanwhile, there are people out there in the world who *don't* share your values. And there are people out there who are *actually evil and do awful things*.

It's like "ok, saying mean things about Muslims can be bad, but being a Muslim terrorist is a hell of a lot worse! Why do the people who are so quick to penalize Islamophobic speech never have anything bad to say about actual mass murder? C'mon, get a sense of proportion!"

I still think, obviously, that really bad actions are worse than slightly bad actions.

But I was *seriously* misunderstanding why people argue with their close friends.

Have you noticed my mistake yet? Give it a moment.

. . .

. . .

. . .

Ok, here it is.

Arguing is not a punishment.

Again.

Arguing is not a punishment.

Sure, serious wrongdoing should be penalized, and socially disapproved of, more than mild wrongdoing. (Murder is worse than prejudiced speech.)

Also, fixing big problems should take priority over fixing little problems. (Saving money on rent is worth more of your attention than saving money on apples.)

But let's frame it differently.

Cooperation is really valuable. Stable cooperation, that is; when even in the future, when you know each other better, and you've had more time to think, you'll *still* want to cooperate.

Trust is really valuable, and scarce. *Justified* trust, that is; when you can rely on what somebody says to be true and base your decisions on information you get from them.

Having "true friends" — people you can cooperate with and trust, stably, to a high degree — is valuable.

Yeah, you can get along and even thrive in a low-trust environment if you have the right skills for it. Havamal, the medieval Icelandic wisdom literature, attributed to the god Odin, is my favorite advice for how to be a savvy customer in a low-trust world. (Exercise for the reader: think about how it applies to the replication crisis in science.) But *especially* in a low-trust world, true friends are valuable, as *Havamal* will remind you again and again.

How do you get more trust and cooperation with your friends?

It's a hard problem; I haven't solved it or even really started trying yet, the following are just ideas at the conceptual level rather than things I've found successful.

But communicating with them to get on the same page is *clearly* part of the puzzle. Cooperation means "you and I agree to do X, and then we follow through and actually do X." The part about willingness to follow through is about loyalty, conscientiousness, motivation, integrity, all those kinds of virtues. The part about agreeing to do X, though? That's not possible unless you both clearly understand *what X is*, which is much harder than it sounds! It takes a lot of discussion, in my experience and from what I've heard, to get people on the same page about what exactly they've committed to doing.

Moreover, if I don't understand *why* X is so important to you, and I say "yeah, ok, sure, X", and then I go home and back to my life, but X *still seems pointless to me*, then I'm going to be less motivated to do X.

Because we didn't have the *argument* about "is X pointless or not?"

We didn't resolve it. We let it drop, to be nice, because we're friends and we like each other. But we *didn't get on the same page,* and now a ball got dropped and you're unhappy with me.

That getting-on-the-same-page process *is not a punishment.*

It's something you'd *only* do with a friend close enough that you really might cooperate on work that you care about getting done. (Mundane example: household chores. Gotta get on the same page about who's responsible for what! Negotiating for fewer/different responsibilities is better than shirking! That can be a really hard thing to internalize, though.)

"I spend more time communicating and getting on the same page with my friends than I do on having discussions with people I hate" — frame it that way, and suddenly that doesn't sound like pointless infighting, it sounds mature and practical, right?

Of *course* you'd focus most on clarifying communication with your closest friends! They're the people you're most likely to be able to cooperate with!

Ok, so *what kind* of agreement is most valuable and attainable? After all, nobody, even your closest friends, agrees with you on everything.

Short term, the answer is obvious: agreement on the details that are practical and relevant to the tasks you share. Share an apartment? Gotta come to agreement on chores, and share world-models relevant for those. (It's no good if I agree to sweep but I don't know where we keep the broom.)

But how about the long-run and more meta problem of *living in a low-cooperation world* itself?

Here's one example: we're in a real trade war with China now. Chinese investment in the US dropped 92 percent in the first half of 2018! I've tuned out financial markets for most of my life, but I'm essentially a professional fundraiser now, and let me tell you, *a drop in Chinese-US investment that drastic affects a US organization's ability to raise capital*. Trade wars, like real wars, can come along all of a sudden and destroy value. Cooperation in this sense is less about singing kumbaya and more about not taking a wrecking ball to your own house. The Hobbesian war of all against all *ruins things that people were trying to build.*

You want collaborators on fixing *that* kind of a problem?

The relevant things to agree (and disagree!) on are about the nature of *cooperation and trust themselves*. How are alliances and coalitions formed and maintained and broken? How, and how well, do enforcement mechanisms and incentive strategies work? You can think of these questions through the lenses of a number of fields:

- game theory
- evolutionary psychology
- some branches of economics (mechanism design, public choice, price theory in general)
- international relations (I know none of this)
- Marxism (I haven't read Marx either, but I've heard that his class analysis can be seen as applied iterated game theory, where a "class" refers to a coalition)

In all cases, the things to get on the same page about are *positive not normative* aspects of *fundamental theory not immediate policy*.

We want *long-term* cooperation, right? That means *fundamentals* need to be gotten right. Why? If you focus on object-level policy, it's too easy for your friend to concur without agreeing ("I agree we should do X, but not with your reason for doing X"), which means that on the *next* policy question that comes up, your friend might not even concur!

(I have a friend — a good guy! a smart guy! — who concurs with me on 100% of object-level political controversies, and in every case, he concurs for a reason I think is dumb. You may know someone like that too. For the purposes of building long-term cooperation, your friend Mr. Concur is *harder* to get on the same page with, and thus *lower* priority to have discussions with, than your friend Ms. Dissent, who starts with the same premises as you but takes them in a totally different direction. This is counterintuitive, because often *you will initially get along better with Mr. Concur!* That is because the mechanism that produces "getting along with" and makes friendships closer or weaker is *itself a short-term, object-level policy!* For instance, people in the same political tribe are nicer to each other.)

So, that's why *fundamental principles, not immediate policy.*

Why positive and not normative? *So you'll avoid unnecessary hostility.*

Hostility, after all, in game-theory-land, is what it feels like from the inside to decide that your interests are opposed to someone else's. You can come to this conclusion mistakenly. To avoid becoming hostile by mistake, *first try to clearly understand and communicate what the landscape of interests and incentives even looks like*. That's what professional negotiators harp on all the time — more often than most people assume, it's in your interests to *keep asking clarifying questions until you understand wtf is going on, and stay cordial enough to keep talking until you understand wtf is going on,* because that increases the odds you'll find a mutually agreeable deal, should one exist. (Notwithstanding this, there are cases in which obfuscating your negotiating position *is* in your interest. That's less true, I expect, the more meta you go. Another reason to start with foundations rather than policies.)

Sticking around for a technical discussion is, itself, a gesture of trust. It invests resources.

That's why it's hard to get this stuff started. As I write this, I haven't washed up yet, I'm not cleaning the house or reading science papers or adding stuff to the LRI blog, and I'm ignoring my baby (who, luckily, is happily playing with his toys and smiling at me every so often.) I'm of the opinion that laying these things out in writing is one of the better ways I have to start coordinated conversations, but, let's be real, it does involve being a little…spendthrift. Feeling like "sure, I can afford to do this." I'm also reading Law's Order, currently. That's also a resource investment into this whole maybe-doomed "understand the micro-foundations of politics" goal, and it also looks kinda like goofing off, and lookit, aren't there already economists for this who do it better? I'm in a remarkably privileged position at the moment when I *have* a bunch of time flexibility, and something tells me that this is one of the ways I want to be using it. It is kind of the future of humanity, after all. But *actually* spending hours chatting merrily — or furiously — with a friend about what is effectively politics for nerds — well, that's what people usually call "wasting time", isn't it?

It's not a waste if you do it well. But I get that there are a lot of incentives pushing against it.

What friendly theory talk has going for it is the very long term — getting to be the future's equivalent of Confucius or Boethius and their friends, or maybe even the Amoraim— and the very short term, in which it's *fun* to hang out with your friends and talk about interesting things and have some sense that you're getting somewhere.

Example question to explore:

The nitty-gritty of the "forgiveness" part of "tit-for-tat-with-forgiveness" in iterated games. There are a lot of slightly different variants of this, I know, which are viable enough to see play. Algorithms for *recovery of cooperation after defection* — how do different ones work? Advantages or disadvantages? Do any of them correspond to known human behaviors or historical/current institutions? As a practical matter, what kind of heuristics do people use as to whether or how to revive relationships with friends that have grown distant, pitch to leads that have gone cold, collect debts that have gone unpaid for a long time, etc?

On Disingenuity

Suppose someone claims that all morality is relative, but when pressed on whether this would apply even to murder, they act evasive and refuse to give a clear answer. A critic might conclude that this person is disingenuous in refusing to accept the clear logical consequences of their belief.

However, imagine that there's a really strong social stigma against asserting that murder might not be bad, to the point of permanently damaging such a person's reputation, even though there's no consequence for making the actually stronger claim that all morality is relative. The relativist might therefore see the critic as the one who is disingenuous; trying to leverage social pressure against them instead of arguing on the basis of reason.

Thus in the right circumstances, each side can quite reasonably see the other as disingenuous. I suspect that everyone will have experienced both sides of the coin at different times depending on the issue being discussed.

This post was produced with the support of the EA Hotel

Kindergarten in NYC: Much More than You Wanted to Know

Kindergarten in NYC: Much More than You Wanted to Know

My son is turning five next year, which means one of the most important transitions in his childhood and potentially his life: starting Kindergarten. I always thought New York City moms who obsessed over this were clearly crazy.

Now I am one of those moms.

Why do we do this to ourselves? It’s not the one year of kindergarten. It’s securing that spot in the school where you want them to stay until middle school and potentially high school, and probably send your other kids to as well. It’s all of the social and class insecurities that come with choosing a school and its associated peer group. It’s the fear that if you choose poorly, your child will age 100 years and his face will melt off in front of you.

Not quite that severe. Still, you worry you’ll mess up their life and they’ll become drug addled sociopaths living on your couch until you kick them out when they bring back that prostitute.

Maybe going overboard again. They’ll go to State College, move to the suburbs, and work in retail.

Wo wo wo, lets not be unrealistic. Retail won't be around in 10 years. Your kid will be horribly miserable for the next 14 years, go through depressive episodes, and blame you for all of it. That’s what I’m actually worried about. Both my husband and I had horrible elementary school experiences. We still carry scars. We don’t want that for our sons.

So why not home school? All the cool kids are doing it. We have personal reasons why this would not work for our family. Our son has some social deficits, but is extremely bright. Literally everyone we’ve spoken to who knows our son agrees that he would do better in a structured environment with peers. We have observed his profound social-emotional growth upon starting the school year. We saw back-sliding over the summer when he lacked structure or regular peer interactions. He will not listen to us when we teach him. He is a different child in the school setting, soaking up knowledge.

People can rant all they like about how horrible school is philosophically, but that does not negate what we’ve personally witnessed in our own child. Philosophy aside, home-schooling is a lot of work and coordination. We both work full-time. While we would pick home-school over the horrid elementary school experiences we had, we hope we can do better and find a school where he will be happy.

That is much easier said than done. Especially for unique children. Our son has done well in a private preschool with 15 children and 3 teachers. A public kindergarten in NYC has a class of 26 children and one teacher. This goes up to as high as 32 in first grade. That is a lot of kids in a small space. It presents two options. Either you get a very noisy and unruly class, or a strictly controlled group which conforms precisely with everyone sitting quietly and doing the same thing at the same time. We have seen both. Neither is pretty. Our son has sensory issues, and will not tolerate a very

noisy classroom. We expect he also would not tolerate a conformist one. Him tolerating it would scare us even more.

If he went to public school, we might well be pressured to put him into a resource room, with children much worse off than himself. Children with emotional disturbance, severe autism, retardation and other severe problems. My mother has worked in such classrooms and what she describes is unacceptable. Those are her stories to tell, but I would not put him there. Ever.

So what can we do? Sue the city! That's what everyone told us to do. Say the public schools can't meet your kid's needs, since they clearly cannot do so. Find a nice, private special needs school, and sue for tuition.

So we saw some special needs schools. Like public schools, they varied a fair bit and we liked some more than others. What they all had in common was a severely impaired peer group. He would be one of the most functional students in the class. We don't want that for him. We want him to be challenged and learn from peers who can be models for him.

So what next? Private school! Private schools also vary a lot, but have one thing in common. They are expensive.

I'm not sure you understand how bad this situation is. I spent time looking around. The average private elementary school charges about $45,000 per year.

Yup. You saw that right, $45,000. That's more than most students' college tuition. Before aid or loans. And it's post-tax income. And we have more than one child.

With two (and perhaps more) children, that would be most if not all of my post-tax income as a psychiatrist.

People have the audacity to say "But you can afford it." Don't get Zvi started on that phrase.

Even if you want to send your kid to private school, you have to apply and be accepted. Most good private schools are selective. Most do not want to deal with a child with special needs.

We have been lucky to find one nearby private school that charges considerably less (though still far from cheap) and happens to have an educational philosophy we think would suit our son. It's a Waldorf school. It emphasizes practical skills such as cooking, gardening, carpentry, foreign language, and trade. Since we believe our son is gifted academically, being less academic does not concern us. He will learn that stuff at home whether we want him to or not. Thus, we wait with baited breath for his trial period there to see if they'll accept him. We don't have a back-up option that comes close at present.

What's been really interesting to me through this process is how vastly schools differ from each other. Often people speak about 'school' as if it is one thing. Either you agree with sending kids to 'school' or you don't. This is not the case. One reason New York City moms go berserk over this is that there are *vast* differences between schools even a few blocks away from each other. Within the public schools, class is everything. Most children go to their 'zoned' school, and so people will pay higher rents near the 'good' schools to get their kids in. One of the public schools we saw looked and felt like a prison, had no music or art program, and only let the kids

outside for 20 minutes a day. Another 10 blocks north in the neighboring district collected $500K/yr from the PTA and had full music and art programs, book fairs, a large library, and extra in-classroom assistants.

We live in a district which has weird rules about admissions. Instead of having a zoned school, you make a rank-list of schools in the district and apply to all of them. In an attempt to ingrate the schools more, the city has imposed rules about who can be admitted by class. The schools are required to accept 67% of 'diversity' applicants who qualify either for low income, English as second language, or living in shelters (i.e. homeless). There is a lot of evidence supporting that peer group is a major factor in child development and life outcome. Political incorrectness aside, this is not a wonderful peer group. It also far reduces the chances that your child will get into the particular school you want them to go to. Since priority is first given to siblings, the 'nice' school in this district (that we would have previously been zoned for) now only has four 'non-diversity' spots open for admission this year. Even if we were willing to send him there, he probably wouldn't get in. Because of this, many better-off families are moving out of the district entirely. This is reflected in the rents within our community – rent jumps considerably right at the district line. People respond to incentives. If we sent our kids to public school we would be forced to do the same. If you have any money at all, you go to the district where the PTA funds the nice art program, not the one with the metal detector in the lobby.

Going private for education hopefully means you avoid true disaster, and the peer group is relatively wealthy and educated. But even private schools differ vastly in their philosophy towards education. Some are super academic, drilling kids to get high SAT scores and become doctors and lawyers. Some are more laid back. Some hardly seem to teach anything at all. There are small schools with one class per grade, others that are much larger. Religious and secular schools. Science schools and arts schools. If you're willing to pay for it odds are there is some school that you would like. That's a big if though.

My practical advice: If your only option is public school, move to an area that has a nice school at least one full school year before you intend to apply. You can tour schools just by saying you have a kid in the district, and they don't force you to prove it. Once you find a school you like, you can move to that school's zone, and you will have a high chance of admission. To be safe, you should make sure there are 1-2 back up schools you find acceptable in the district. If you cannot afford to live any places with reasonable public schools, you should seriously consider leaving the city. I am told of reasonable schools in NJ…

If you can't stand public school, because at the end of the day they all follow common core, take those tests, and have 32 kids in a class, then you have to consider what you can afford. Home school has no tuition, but will require all-day child care, any educational materials/classes you want to use, and a large coordination effort on your part. If you're a stay at home parent this might appeal to you anyway. For the most part the people who choose to do it are happy with it.

Private school is expensive, but requires less advance planning, since they don't care what district you're in as long as you can pay. You might still need to consider moving for private school if you don't want your child to have an infinitely long commute. The city will pay for busing to private schools for bus routes which are 0.25 – 2.0 miles. Keep in mind that they are measuring distance along bus routes and not geographically. Even if you are physically within 2 miles of the school, the bus route might be over 2 miles and you will be out of luck. To be fair, if you're willing to spend

$50,000/year on a school, then what's another $40/day to hire someone to take them to school?

I am now going to write some school reviews. I will leave out specific names, but if you are interested you can message me privately, and I will let you know which is which. Zvi saw some schools I did not, which I haven't written about, and we still have some tours planned at local public schools.

Public Schools:

District 1 (our district – the one with the integration)

Public School A:

I was pleasantly surprised by this school's philosophy of education. They were laid back and progressive. Kids sit at tables instead of desks. Group conversations and creative expression was encouraged. No mandatory homework. Starting in 1st grade, kids learn chess and have the opportunity in 3rd and 4th grade to compete in tournaments. In 3rd grade the kids learn basic computer programming. There is a year of free music lessons. They have a theater and a roof-top garden. Gym is non-competitive until 4th grade. 45 minutes of daily outdoor time. I really liked everything they *said* and the principal was super cool. However, the actual classrooms were tiny and crammed full of students. It was loud. I felt claustrophobic there, and I don't have sensory issues in general. Plus, the district just implemented the diversity criteria this year, so the students I was seeing are not the peer group my son is going to have if he went. And, of course, they only have four non-sibling, non-diversity spots available.

Public School B:

This place is a prison. There is an angry security guard at the entrance to the grime-encrusted orange walls. Multiple signs above the guard state 'theft is a crime.' The slit-like windows at the top of the rooms let in thin beams of daylight to an otherwise flickering-fluorescent landscape. This is hell. There is no music or art program – no room in the budget. So 'we do that within our lessons'. 20 minutes of yard time a day. Everything is centered around standardized tests. The only white faces were part of a special program. No one with any choice would ever let their kid set foot in this place unless they were in the special program. Not worth it. It's social control of minorities. Straight up. If SJWs want a cause, here's one for you. And no, forcing white or wealthy children to go there is not going to work. They won't.

District 2 (the nice one)

Public School C:

The platonic ideal of school. When you think school, you think this school. The people who designed it thought 'what is school?' and then based the design off of every trope and meme about school, ever. Charts of everything on the walls. 'Task leaders.' Bulletin boards. Window decals. Those weird cartoon people you only see in school ever. Worksheets, worksheets, worksheets. Chalk boards. White boards. This place has it all! The place felt nice. Larger rooms, more light. Nice enrichment activities. A music and art program. A nice library and computer lab. Several outdoor spaces and playground equipment. The place gets $500k/yr from the PTA to keep the place great. Mostly white faces sitting quietly in circles while the teacher spoke to them in exaggerated tones with big faces while pointing to a white board.

Looked like the children of the corn. Completely conformist. But conformists at least a year ahead academically. It is disturbing to see kindergarteners completing reading worksheets and pushing papers around, but they were able to do it. This is the place for upper-middle class white people who move into the ‘good’ part of the neighborhood.

Private Schools:

Private School A: Preparatory School

EXPENSIVE. Beautiful school and facility. It is a ‘Quaker’ school, but mostly secular. Has a beautiful chapel where kids have ‘community assembly and quiet time’ once a week. Other parents were very well dressed – a lot of suits and jewelry. Academically rigorous without being oppressively conformist. Perhaps because the class size is 20 instead of 30, so there is more room to maneuver. A fine school as schools go, but not that much of an upgrade from PSC given the price. Also difficult to get into and unwilling to accommodate special needs.

Private School B: Jewish School

I loved this school! I really did. It’s a progressive, laid-back atmosphere that is still academically oriented. It is very Jewish. The boys wear keepas and the curriculum is fully bilingual with one teacher speaking English and the other speaking Hebrew. They have all the usual stuff such as music and art. They go outside for 1 hr/day. They are willing to work with special needs. They know how to work with gifted and talented kids and make special assignments for children who are ahead. LOVE IT. Problem was, it is about 1 hour away by bus and it’s a 7.5 hr day. Not doing that to my kid. Not willing to move close enough to make it work. At least not this coming year.

Private School C: Waldorf School

This is a very unique nearby school that happens to be less expensive than the others. It has a unique education philosophy (a Waldorf school) which emphasizes embodiment and practical skills over academic ones. The curriculum includes foreign languages, cooking, washing, gardening, carpentry, and trade. The kindergarten is entirely non-academic and includes copious time for free play and an hour of outdoor activity. The later grades teach traditional academics, but do so in somewhat unusual ways, which I don’t have a strong opinion on at present. Since the main reason we are sending our son to school is for socialization, and since he’s already brilliant, I’m less worried about academics, especially in the younger grades. The school requested a drastic reduction in our child’s screen time, which at first freaked me out (who are they to tell me what to do in my own home), but I kind of understand. It’s a very small school (only 1 class per grade) and they are currently considering whether or not they can accommodate his needs. This is our top choice at present.

Special Needs Schools:

SNS A: Social Justice Away!

This school is an ‘integrated’ private school – meaning it’s a private school for regular kids which also accepts children with learning disabilities and has services for them. This means you can get the tuition paid by the city, unlike regular private schools, with a relatively normal peer group. It’s a great idea. The school itself is beautiful and has All The Things.

However there is a catch. The school has an agenda. It’s a social justice school. In the sense that other schools are reading and math schools. They call themselves ‘Advocates for Social Justice’ in their opening lines. I wouldn’t have thought this mattered for elementary age children. Sure, loving each other is wonderful! Accepting your neighbors is wonderful! But this is not where they draw the line. Social Justice is taught in every aspect of the curriculum. There are 7 year olds discussing their ‘identities’, an 8 year old talking about how his hero is Colin Kaepernick, that guy who keeled for the national anthem. The teachers then praise his 'activism' for writing about it. The other sample lesson is on how Christopher Columbus was a white colonialist oppressor. And the children absorb this. The school is accepts all kinds – unless you happen to be a *gasp* Republican. No diversity of thinking. If you don’t fully swallow the SJW philosophy in all its forms, or don't want them forced down your child's throat, this is not the place for you.

SNS B: Soothing Gardens…

Beautiful place. Therapeutic environment. Has the things. Didn’t want us to see the children – which was strange. When we peaked in at them, they were, well, very special. Seems like a great place for very special kids. If I have one that needed all that, I’d consider sending him there.

SNS C: Jews with learning problems

While not specifically a Jewish school, there were clearly a lot of Jewish children and teachers. I actually liked this place a lot. It was very laid back and gave the kids a lot of lee-way to be who they are. It didn’t feel at all oppressive. They group kids into separate reading and math groups not by age, but by reading and math level, which I liked. The kids seemed less special than at SNS B, but still clearly special. The school didn’t have its own outdoor space and so kids only go outside twice week with a bunch of parent-volunteers, since they want one adult per kid when crossing the streets. What was particularly disappointing was that they were clearly quite academically behind. The classes were so laid back that there didn’t seem to be a challenge, and the teachers were fine with whatever they produced. I can imagine certain children this would be very good for. I have vastly higher hopes for our son.

What is "Social Reality?"

Eliezer's sequences touch upon this concept but I'm not sure they actually use the phrase. Much of my understanding of it came from in-person conversations. Various comments and posts have discussed it but to my knowledge there isn't a clear online writeup.

Why we need a *theory* of human values

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

There have been multiple *practical* suggestions for methods about how we should extract the values of a given human. Here are four common classes of such methods:

- Methods that put high weight on human (bounded) quasi-rationality, or revealed preferences. For example, we can assume the Kasparov was actually trying to win against DeepBlue, not trying desperately to lose while inadvertently playing excellent chess.
- Methods that pay attention to our explicitly stated values.
- Methods that use regret, surprise, joy, or similar emotions, to estimate what humans actually want. This could be seen as a form of human TD learning.
- Methods based on an explicit procedure for constructing the values, such as CEV and Paul's indirect normativity.

Divergent methods

The first question is why we would expect these methods to point even vaguely in the same direction. They all take very different approaches - why do we think they're measuring the same thing?

The answer is that they roughly match up in situations we encounter everyday. In such typical situations, people who feel regret are likely to act to avoid that situation again, to express displeasure about the situation, etc.

By analogy, consider a town where there are only two weather events: bright sunny days and snow storms. In that town there is a strong correlation between barometric pressure, wind speed, cloud cover, and temperature. All four indicators track different things, but, in this town, they are basically interchangeable.

But if the weather grows more diverse, this correlation can break down. Rain storms, cloudy days, meteor impacts: all these can disrupt the alignment of the different indicators.

Similarly, we expect that an AI could remove us from typical situations and put us into extreme situations - at least "extreme" from the perspective of the everyday world where we forged the intuitions that those methods of extracting values roughly match up. Not only do we expect this, but we desire this: a world without absolute poverty, for example, is the kind of world we would want the AI to move us into, if it could.

In those extreme and unprecedented situations, we could end up with revealed preferences pointing one way, stated preferences another, while regret and CEV point in different directions entierly. In that case, we might be tempted to ask "should we follow regret or stated preferences?" But that would be the wrong question to ask: our methods no longer correlated with each other, let alone with some fundamental measure of human values.

We are thus in an undefined state; in order to continue, we need a meta-method that decides between the different methods. But what criteria could such meta-method use for deciding (note that simply getting human feedback is not generically an option)? Well, it would have to select the method which best matches up with human values in this extreme situation. **To do that, it needs a definition - a theory - of what human values actually are**.

Underdefined methods

The previous section understates the problems with purely practical ways of assessing human values. It pointed out divergences between the methods in "extreme situations". Perhaps we were imagining these extreme situations as the equivalent of a meteor impact on weather system: bizarre edge cases where reasonable methods finally break down.

But all those actually methods fail in typical situations as well. If we interpret the methods naively, they fail often. For example, in 1919, some of the Chicago White Sox baseball team were actually trying to lose. If we ask someone their stated values in a political debate or a courtroom, we don't expect an honest answer. Emotion based approaches fail in situations where humans deliberately expose themselves to nostalgia, or fear, or other "negative" emotions (eg through scary movies). And there are failure modes for the explicit procedures, too.

This is true if we interpret the methods naively. If we were more "reasonable" or "sophisticated", we would point out that don't expect those methods to be valid in every typical situation. In fact, we can do better than that: we have a good intuitive understanding of when the methods succeed and when they fail, and different people have similar intuitions (we all understand that people are more honest in relaxed private settings that stressful public ones, for example). It's as if we lived in a town with either sunny days or snow storms *except on weekends*. Then everyone could agree that the different indicators correlate during the week. So the more sophisticated methods would include something like "ignore the data if it's Saturday or Sunday".

But there are problems with this analogy. Unlike for the weather, there are no clear principle for deciding when it's the equivalent of the weekend. Yes, we have an *intuitive* grasp of when stated preferences fail, for instance. But as Moravec's paradox shows, an intuitive understanding doesn't translate into an explicit, formal definition - and it's that kind of formal definition that we need if we want to code up those methods. Even worse, we **don't** all agree as to when the methods fail. For example, some economists deny the very existence of mental illness, while psychiatrists (and most laypeople) very much feel these exist.

Human judgement and machine patching

So figuring out whether the methods apply is an exercise in human judgement. Figuring out whether the methods have gone wrong is a similar exercise (see the Last Judge in CEV). And figuring out what to do when they don't apply is also an exercise in human judgement - if we judge that someone is lying about their stated preferences, we could just reverse their statement to get their true values.

So we need to patch the methods using our human judgement. And probably patch the patches and so on. Not only is the patching process a terrible and incomplete way of constructing a safe goal for the AI, but human judgements are not consistent - we can be swayed in things as basic as whether a behaviour is rational, let alone all the situational biases that cloud our assessments of more complicated issues.

So obviously, the solution to these problems is to figure out which human is best in their judgements, and then to see under what circumstances these judgements can be least biased, and how to present the information to them in the most impartial way and then automate that judgement...

Stop that. It's silly.. The correct solution is not to assess the rationality of human judgements of methods of extracting human values. The correct solution is to come up with a better theoretical definition of what human values are. Armed with such a theory, we can resolve or ignore the above issues in a direct and principled way.

Building a theory of human values

Just because we need a theory of human values, doesn't mean that it's easy to find one - the universe is cruel like that.

A big part of my current approach is to build such a theory. I will present an overview of my theory in a subsequent post, though most of the pieces have appeared in past posts already. My approach uses three key components:

1. A way of defining the basic preferences (and basic meta-preferences) of a given human, even if these are under-defined or situational.
2. A method for synthesising such basic preferences into a single utility function or similar object.
3. A guarantee we won't end up in a terrible place, due to noise or different choices in the two definitions above.

A hundred Shakespeares

In his post on science slowing down, Scott said:

- "Are there a hundred Shakespeare-equivalents around today? This is a harder problem than it seems – Shakespeare has become so venerable with historical hindsight that maybe nobody would acknowledge a Shakespeare-level master today even if they existed – but still, a hundred Shakespeares?"

I'd argue that there are way more than a hundred Shakespeares around today, and there were several in Shakespeare's time. By Shakespeares, I mean authors who could have produced works of comparable quality to Shakespeare, by some reasonable measure of quality.

This seems surprising; there do not seem to be hundred living authors that are almost universally agreed to be must-reads in the same way that Shakespeare was.

But this lack hints at a resolution of the paradox: we just don't have space for a hundred authors with the same fervour as we make space for Shakespeare. Neither as individuals nor as cultures can we fit these in. Shakespeare was a literary superstar. And superstars are rare, due to network effects and the power law of fame.

So my thesis would be that:

- There are many non-superstars who could plausibly have become superstars, and if they had done, they would produce works of comparable quality to the superstars.

Part of this is the halo effect: superstars just get judged as better than anyone else.

Also, just by being famous, the interpretation of their work is altered. Bits of Shakespeare have permeated popular culture, and many articles and theories have been created about him. When we watch a Shakespeare play, we don't just see the words; we see the layers of cultural meaning and interpretation that have accumulated on it.

I'd argue that, just by knowing that a play is by Shakespeare, we assume that it's deep and meaningful, and read in deeper interpretations and symbolism than we would otherwise. If we rediscovered two old plays, and they were word for word identical, but one was believed to be by Shakespeare and the other by some forgotten minor playwright, I'd expect that the first one would be a better play, just by what the audience would bring to it.

Apart from those effects, superstars have the unique ability to focus more on their own vision. They have great self-confidence, and they can afford to trust that their audiences will have the patience to follow them where they want to go - rather than expecting immediate literary gratification. This would tend to result in works that are better than the average work of someone of equivalent skill, and more likely to be "deep", "insightful", or "timeless". This effect might be even more obvious with bloggers than with authors.

So, though the number of superstars is severely limited, the number of potential superstars of equivalent skill can and most likely does increase with population.

Superstars in science

I'd argue that there's also a superstar effect in science. But here it combines with Scott's explanation 3: low hanging fruit. Newton did not come up with general relativity; Einstein didn't find quantum field theory; Tesla didn't invent the laser. You can't develop an idea until certain pre-requisites are met.

And, unlike those solitary geniuses, most of science and technology is collaborative. Superstars get to be part of the best teams, interact with the best other scientists, and are more free to focus on the biggest, sexiest problems. I expect that there are many non-superstars who would have developed a certain part of theory, if a superstar hadn't got there first. It seems plausible to me that a single scientific superstar could have done the equivalent of derailing a hundred promising careers, just by getting to the key insight faster - without necessarily being much smarter (if at all) than the ones they preempted.

Then, as discoveries pour in from superstars, and the far less productive non-superstars, the domain of science changes, and new avenues of discovery open up. And these new avenues are going to be claimed by the next generation of superstars, who will get there first. I expect that if we removed every single superstar of science in the last two hundred years, that we'd get roughly comparable scientific progress, with alternate superstars rising to the fore.

Book review: Artificial Intelligence Safety and Security

This is a linkpost for http://www.bayesianinvestor.com/blog/index.php/2018/12/06/artificial-intelligence-safety-and-security/

Book review: Artificial Intelligence Safety and Security, by Roman V. Yampolskiy.

This is a collection of papers, with highly varying topics, quality, and importance.

Many of the papers focus on risks that are specific to superintelligence, some assuming that a single AI will take over the world, and some assuming that there will be many AIs of roughly equal power. Others focus on problems that are associated with current AI programs.

I've tried to arrange my comments on individual papers in roughly descending order of how important the papers look for addressing the largest AI-related risks, while also sometimes putting similar topics in one group. The result feels a little more organized than the book, but I worry that the papers are too dissimilar to be usefully grouped. I've ignored some of the less important papers.

The book's attempt at organizing the papers consists of dividing them into "Concerns of Luminaries" and "Responses of Scholars". Alas, I see few signs that many of the authors are even aware of what the other authors have written, much less that the later papers are attempts at responding to the earlier papers. It looks like the papers are mainly arranged in order of when they were written. There's a modest cluster of authors who agree enough with Bostrom to constitute a single scientific paradigm, but half the papers demonstrate about as much of a consensus on what topic they're discussing as I would expect to get from asking medieval peasants about airplane safety.

Drexler's paper is likely the most important paper here, but it's frustratingly cryptic.

He hints at how we might build powerful AI systems out of components that are fairly specialized.

A key criterion is whether an AI has "strong agency". When I first read this paper a couple of years ago, I found that confusing. Since then I've read some more of Drexler's writings (not yet published) which are much clearer about this. Drexler, hurry up and publish those writings!

Many discussions of AI risks assume that any powerful AI will have goals that extend over large amounts of time and space. Whereas it's fairly natural for today's AI systems to have goals which are defined only over the immediate output of the system. E.g. Google translate only cares about the next sentence that it's going to create, and normal engineering practices aren't pushing toward having it look much further into the future. That seems to be a key difference between a system that isn't much of an agent, versus one with strong enough agency to have the instrumentally convergent goals that Omohundro (below) describes.

Drexler outlines intuitions which suggest we could build a superintelligent system without strong agency. I expect that a number of AI safety researchers will deny that such a system will be sufficiently powerful. But it seems quite valuable to try, even if it only has a 50% chance of producing human-level AI, because it has little risk.

This distinction between types of agency serves as an important threshold to distinguish fairly safe AIs from AIs that want to remake the universe into something we may or may not want. And it looks relatively easy to persuade AI researchers to stay on the safe side of that threshold, a long as they see signs that such an approach will be competitive.

By "relatively easy", I mean something like "requires slightly less than heroic effort", maybe a bit harder than it has been to avoid nuclear war or contain Ebola. There are plenty of ways to make money while staying on the safe side of that threshold. But some applications (Alexa? Siri? NPCs?) where there are moderate incentives to have the system learn about the user in ways that would blur the threshold.

Note that Drexler isn't describing a permanent solution to the main AI safety risks, he's only describing a strategy that would allow people to use superintelligence in developing a more permanent solution.

Omohundro's **The Basic AI Drives** paper has become a classic paper in AI safety, explaining why a broad category of utility functions will generate strategies such as self-preservation, resource acquisition, etc.

Rereading this after reading Drexler's AI safety writings, I now see signs that Omohundro has anthropomorphised intelligence a bit, and has implicitly assumed that all powerful AIs will have broader utility functions than Drexler considers wise.

Also, Paul Christiano disagrees with one of Omohundro's assumptions in this discussion of corrigibility.

Still, it seems nearly certain that someday we will get AIs for which Omohundro's warnings are important.

Focus on a singleton AI's value alignment

Bostrom and Yudkowsky give a succinct summary of the ideas in Superintelligence (which I reviewed here), explaining why we should worry about ethical problems associated with a powerful AI.

Soares discusses why it looks infeasible to just encode human values in an AI (e.g. Sorcerer's Apprentice problems), and gives some hints about how to get around that by indirectly specifying how the AI can learn human values. That includes extrapolating what we would want if we had the AI's knowledge.

He describes an ontology crisis: a goal as simple as "make diamond" runs into problems if "diamond" is described in terms of carbon atoms, but the AI switches to using a nuclear model that sees protons/neutrons/electrons - how do we know whether it will identify those particles as carbon?

Tegmark provides a slightly different way of describing problems with encoding goals into AI's: What happens if an AI is programmed to maximize the number of human souls that go to heaven, and ends up deciding that souls don't exist? This specific

scenario seems unlikely, but human values seem complex enough that any attempt to encode them into an AI risks similar results.

Olle Häggström tries to analyze how we could use a malicious Oracle AI while keeping it from escaping its box.

He starts with highly pessimistic assumptions (e.g. implicitly assuming Drexler's approach doesn't work, and worrying that the AI might decide that hedonic utilitarianism is the objectively correct morality, and that maximizing hedons per kilogram of brain produces something that isn't human).

Something seems unrealistic here. Häggström focuses too much on whether the AI can conceal a dangerous message in its answers.

There are plenty of ways to minimize the risk of humans being persuaded by such messages. Häggström shows little interest in them.

It's better to have a security mindset than not, but focusing too much on mathematically provable security can cause researchers to lose sight of whether they're addressing the most important questions.

Herd at al talk about how value drift and wireheading problems are affected by different ways of specifying values.

They raise some vaguely plausible concerns about trade-offs between efficiency and reliability.

I'm not too concerned about value drift - if we get the AI(s) to initially handle this approximately right (with maybe some risks due to ontological crises), the AI will use its increasing wisdom to ensure that subsequent changes are done more safely (for reasons that resemble Paul Christiano's intuitions about the robustness of corrigibility).

Concerns about paths to AI

Bostrom's **Strategic Implications of Openness in AI Development** thoughtfully describes what considerations should influence the disclosure of AI research progress.

Sotola describes a variety of scenarios under which an AI or collections of AIs could cause catastrophe. It's somewhat useful at explaining why AI safety is likely to be hard. It's likely to persuade people who are currently uncertain to be more uncertain, but unlikely to address the beliefs that lead some people to be confident about safety.

Turchin and Denkenberger discuss the dangers of arms races between AI developers, and between AIs. The basic ideas behind this paper are good reminders of some things that could go wrong.

I was somewhat put off by the sloppy writing (e.g. "it is typically assumed that the first superintelligent AI will be infinitely stronger than any of its rivals", followed by a citation to Eliezer Yudkowsky, who has expressed doubts about using infinity to describe real-world phenomena).

Chessen worries about the risks of AI-driven disinformation, which might destabilize democracies.

Coincidentally, SlateStarCodex published Sort by Controversial (a more eloquent version of this) around the time when I read this paper.

This seems much less than an extinction risk by itself, but it might make some important governments short-sighted at key times when the more permanent risks need to be resolved.

His policy advice seems uninspired, e.g. suggesting privacy laws that sound like the GDPR.

And "Americans must choose to **pay for news** again." This seems quite wrong to me. I presume Chessen means we should return to paying for news via subscriptions instead of via ads.

But the tv news of 1960s was financed by ads, and was about as responsible as anything we might hope to return to. My view is that increased click-bait is due mainly to having more choices of news organizations. Back before cable tv, our daily news choices were typically one or two local newspapers, and two to five tv channels. Those gravitated toward a single point of view.

Cable tv enabled modest ideological polarization, and a modest increase in channel surfing, which caused a modest increase in sensationalism. Internet enabled massive competition, triggering a big increase in sensationalism.

Note that the changes from broadcast tv to cable tv to internet multimedia involved switching from ad-based to subscription to ad-based models, with a steady trend away from a focus on fact-checking (although that fact-checking may have been mostly checking whether the facts fit the views of some elite?).

Wikipedia is an example which shows that not paying for news can generate more responsible news - at the cost of entertainment.

There's still plenty of room for responsible people to create new news institutions (e.g. bloggers such as Nate Silver), and it doesn't seem particularly hard for us to distinguish them from disinformation sources.

The main problem seems to be that people remain addicted to sources as they become click-baity, and continue to treat them as news sources even after noticing alternatives such as Nate Silver.

I expect the only effective solution will involve most of us agreeing to assign low status to people who treat click-baity sources as anything more than entertainment.

Miscellaneous other approaches

Torres says the world needs to be ruled by a Friendly AI. (See a shorter version of the paper here.)

His reasoning is loosely based on Moore's Law of Mad Science: Every eighteen months, the minimum IQ necessary to destroy the world drops by one point. But while Eliezer intended that to mostly focus on the risk of the wrong AI taking over the world, Torres extends that to a broad set of weapons that could enable one person to cause human extinction (e.g. bioweapons).

He presents evidence that some people want to destroy the world. I suspect that some of the people who worry him are thinking too locally to be a global danger, but there's likely enough variation that we should have some concern that there are people who seriously want to kill all humans.

He asks how low we need to get risk of any one such malicious person getting such a weapon in order to avoid human extinction. But his answer seems misleading - he calculates as if the risks were independent for each person. That appears to drastically overstate the risks.

Oh, and he also wants to stop space colonization. It creates risks of large-scale war. But even if that argument convinced me that space colonization was bad, I'd still expect it to happen. Mostly, he doesn't seem to be trying very hard to find good ways to minimize interstellar war.

If we're going to colonize space fairly soon, then his argument is weakened a good deal, and it would then imply that there's a short window of danger, after which it would take more unusual weapons to cause human extinction.

What's this got to do with AI? Oh, right. A god-like AI will solve extinction risks via means that we probably can't yet distinguish from magic (probably involving mass surveillance).

Note that a singleton can create extinction risks. Torres imagines a singleton that would be sufficiently wise and stable that it would be much safer than the risks that worry him, but we should doubt how well a singleton would match his stereotype.

Torres is correct to point out that we live in an unsafe century, but he seems wrong about important details, and the AI-relevant parts of this paper are better explained by the Bostrom/Yudkowsky paper.

Bostrom has recently published a better version of this idea.

Miller wants to build addiction into an AI's utility function. That might help, but it looks to me like that would only be important given some fairly bizarre assumptions about what we can and can't influence about the utility function.

Bekdash proposes adopting the kind of rules that have enabled humans to use checks and balances to keep us safe from other humans.

The most important rules would limit AI's span of control - AI must have limited influence on the world, and must be programmed to die.

Bekdash proposes that all AIs (and ems) go to an artificial heaven after they die. Sigh. That looks relevant only for implausibly anthropomorphised AI.

Bekdash want to prevent AIs from using novel forms of communication ("it is easier to monitor and scrutinize AI communication than that of humans.") - that seems to be clear evidence that Bekdash has no experience at scrutinizing communications between ordinary computer programs.

He also wants to require diversity among AIs.

Bekdash proposes that global law ensure obedience to those rules. Either Bekdash is carefully downplaying the difficulties of enforcing such laws, or (more likely) he's

depending on any illegal AIs being weak enough that they can be stopped after they've had a good deal of time to enhance themselves. In either case, his optimism is unsettling.

Why do these papers belong in this book?

Prasad talks about how to aggregate opinions, given the constraints that opinions are expressed only through a voting procedure, and that Pareto dominant alternatives are rare.

I expect a superintelligent AI to aggregate opinions via evidence that's more powerful than voting for political candidates or complex legislation (e.g. something close to estimating how much utility each person gets from each option).

I also expect a superintelligent AI to arrange something close to Pareto dominant deals often enough that it will be normal for 95+% of people to consent to decisions, and pretty rare for us to need to fall back on voting. And even if we do occasionally need voting, I'm optimistic that a superintelligence can usually come up with the kind of binary choice where Arrow's impossibility theorem doesn't apply.

So my impression is that Prasad doesn't have a clear reason for applying voting theories to superintelligence. He is at very least assuming implausibly little change in how politics works. Maybe there will be some situations where a superintelligence needs to resort to something equivalent to our current democracy, but he doesn't convince me that he knows that. So this paper seems out of place in an AI safety book.

Portugal et el note that the leading robot operating system isn't designed to prevent unauthorized access to robots. They talk about how to add an ordinary amount of security to it. They're more concerned with minimizing the performance cost of the security than they are with how secure the result is. So I'm guessing they're only trying to handle fairly routine risks, not the risks associated with human-level AI.

18-month follow-up on my self-concept work

About eighteen months ago, I found Steve Andreas's book Transforming Your Self, and applied its techniques to fixing a number of issues in my self-concepts which had contributed to my depression and anxiety. Six weeks after those changes, I posted a report called "How I found & fixed the root problem behind my depression and anxiety after 20+ years". I figured that by now it would be time for a follow-up on how those effects have lasted.

Overall summary and general considerations

Looking back, this was definitely a major milestone in improving my mental health. I feel like since 2014, I have been ongoing a process of completely transforming myself from the depression- and anxiety-ridden person who was convinced that he had no other option than becoming a total failure, to someone calmly confident who has the option of constructing his life to his taste. I don't claim to be there yet, but I feel like I'm constantly getting closer. I feel like the self-concept work discussed in my post, was one of the largest engines powering this transition. (Other major ones being me getting antidepressants, changing how I thought about ethics, and learning a new mindset from CFAR in 2014, properly learning Focusing and Core Transformation as well as starting to meditate according to The Mind Illuminated system in 2017, and starting to apply Internal Family Systems this year.)

There are two difficulties evaluating my self-concept post afterwards. First is that I have a poor emotional memory, so it's a little hard for me to remember what I felt before these changes. The second is that after doing self-concept work, I've also done plenty of other things, such as meditation and moving together with some housemates, which have also had a definite impact on my mental health. I can't know how well the self-concept work would have stuck around, if I hadn't also implemented those other changes. It's possible and even likely that some of my current results are because of those other changes instead.

At the same time, the self-concept work is also not independent from everything that I've done later. For instance, I think that being able to eliminate the feelings of pointless shame has been a major reason *why* I've been able to live with housemates and find them a definite net positive. Previously the feelings of shame would have made it too draining to have to engage in social interaction in my home on a regular basis, whereas now social interaction has tended to be much more energizing than it did in the past. But then again, there are also other skills which have made social fatigue less of an issue than sometime in the past, and which I've also been gradually training up.

But still, at least I can report on the various things in the post, and on how they've held up.

Things that seem to have been fixed for good

Generalized feelings of shame; being afraid of thinking that thoughts that might trigger feelings of shame; needing constant validation in order to avoid feelings of shame. I described the following in my post from last year:

I realized that I had a sense of unease, a vague feeling of shame… as if there was something shameful about me that I knew, but was trying to avoid thinking about. And I knew that I had felt this same vague shame many times before, often particularly when I was tired. […]

… there's always an underlying insecurity, a sense of unease from the fact that anything might cause your attention to swing back to the [memories of being a terrible person]. You need a constant stream of external validation and evidence in order to keep your attention anchored on the examples [of being a good person]; the moment it ceases, your attention risks swinging to the [memories of being a terrible person] again.

As far as I can tell, this kind of thing simply doesn't happen anymore. I still get feelings of guilt, if I have screwed up in some way, but there's no shame or feeling of being a horrible person. Nor is there any need for external validation in this regard. I just know that I'm always doing the best that I can, and if I make a mistake that I need to learn from, then I feel the amount of guilt that's necessary to motivate me to make amends and/or remember to act differently in the future. And that's that.

Being motivated by a desire to prove to myself that I'm a good person. *Previously I was trying to do a lot of things, but basically everything was strongly driven by a motivation to feel better about myself, and whenever it looked like something wasn't likely to help with that goal, I would get demotivated.* […] *Previously when I was trying to do things to "save the world", there was a strong component of doing it for the sake of guilt, feeling bad, or trying to win respect or status from others.*

Basically fixed; this caused a period of readjustment, in that I had been doing things which had been optimized for looking good in the eyes of people that I admired, even when I personally hadn't felt on a gut level that they made much sense. It took a while to readjust and find things which felt worth doing, but now I mostly feel like I'm doing things because they are genuinely derived from my values, rather than to avoid shame.

I still occasionally have something-like-guilt as a factor in thinking about what I want to do, but it mostly pops up when I notice that I'm not satisfying all of my *own* values and neglecting something that I actually care about. I'm no longer doing things "for the sake of guilt", in the sense that I would do something and then keep feeling guilty regardless. If you find yourself regularly experiencing guilt, then you are using guilt incorrectly; in this respect as well, I'm using guilt much more correctly now.

Insecurity in relationships and with romantic partners; very detailed escapist romantic fantasies. *If I was in a relationship, I would tend to very strongly highlight some qualities that I felt I had and which I felt bad about, in an attempt to get my partner to explicitly express being okay with them. […]*

… much of my desire and need to be in a relationship was another way of trying to look for external validation, some kind of evidence that there was somebody who would accept me and would want to be with me. I used to have a lot of pretty detailed romantic fantasies; a lot of them lost their appeal after I fixed my self-concept.

Evaluating this is slightly harder since I haven't actually been in a relationship since writing that post. However, judging from the way that I've felt towards and interacted with *potential* romantic partners as well as women I've been intimate friends with, and how I've felt about relationships in general, this feels basically fixed. Being single is far from my ideal preference, but it's not particularly terrible either, and I don't spend much time absorbed in detailed fantasies when I could be doing something else. I'm also much more comfortable with intimate friendships which are ambiguous about whether or not they might turn more romantic; I can be genuinely happy either way.

Mostly fixed, might still pop up a bit

Obsessive sexual fantasies. *Without going into too much detail, previously my sexuality and fantasies had been very strongly entwined around a few paraphilias, which provided a great deal of emotional comfort. A lot of those fantasies were obsessive to the point of being bothersome.*

At the time of writing my post, I reported that these basically disappeared. They remained gone for a while, but eventually some (not all) of them came back, though considerably transformed. They are fun to engage with occasionally, and they might get a bit bothersome if I think about them too much. But whenever they start getting that mildly obsessive flavor, it tends to act as a natural disincentive for me to continue thinking about them, and then they quiet down again.

Partially fixed, but with other causes as well

Feelings of anxiety and a need to escape. *It feels that, large parts of the time, my mind is constantly looking for an escape, though I'm not entirely sure what exactly it is trying to escape from. But it wants to get away from the current situation, whatever the current situation happens to be. To become so engrossed in something that it forgets about everything else.*

Unfortunately, this often leads to the opposite result. My mind wants that engrossment right now, and if it can't get it, it will flinch away from whatever I'm doing and into whatever provides an immediate reward. Facebook, forums, IRC, whatever gives that quick dopamine burst. That means that I have difficulty getting into books, TV shows, computer games: if they don't grab me right away, I'll start growing restless and be unable to focus on them. Even more so with studies or work, which usually require an even longer "warm-up" period before one gets into flow.

This kind of a thing still happens; apparently the anxiety from poor self-concepts was only one of its causes. I now think that it's more of an executive dysfunction symptom, in that various causes of stress or feeling bad can trigger a self-reinforcing loop of feeling bad, trying to escape that badness, feeling even more bad for failing to escape it, etc. My feelings of shame were definitely one cause, but many other things can also

trigger it. Meditation and Focusing / IFS work have been a major aid in fixing several other causes.

Insecurities based on shame vs. instrumental considerations. *Suppose that you have an unstable self-concept around "being a good person", and you commit some kind of a faux pas. Or even if you haven't actually committed one, you might just be generally unsure of whether others are getting a bad impression of you or not. Now, there are four levels on which you might feel bad about the real or imagined mistake:*

1. *Feeling bad because you think you're an intrinsically bad person*
2. *Feeling bad because you suspect others think bad of you and that this is intrinsically bad (if other people think bad of you, that's terrible, for its own sake)*
3. *Feeling bad because you suspect others think bad of you and that this is instrumentally bad (other people thinking bad of you can be bad for various social reasons)*
4. *Feeling bad because you might have hurt or upset someone, and you care about what others feel*

Out of these, #3 and #4 are reasonable, #1 and #2 less so. When I fixed my self-concept, reaction #1 mostly vanished. But interestingly, reaction #2 stuck around for a while... or at least, a fear of #2 stuck around for a while.

#1 and #2 seem to indeed have disappeared; however, I've still continued to experience insecurities which have taken the forms of what seems like excessive worries of #3 and #4 (thinking that I've displeased someone in a way which will make them like me less, as well as worrying that someone might have felt upset over something that they in all likelihood won't even remember). These seem to be the kinds of issues that can't be fixed by internal work alone, since they are about the external world: in order to evaluate how justified these are, I need to actually *test* the extent to which something e.g. makes other people dislike me.

This work is still ongoing, but I've been making progress. Major contributors to current progress are the skills of integrating the cautions from my insecurities and tentatively considering emotional stories. These seem to have the effect that parts of my mind which have long held extreme beliefs about how cautious I should be, get listened to in a fairer way, causing them to update their beliefs to less extreme ones.

Difficulties in self-motivation. *Besides being able to work at all, I'm also able to consistently work from home. This was often basically impossible: the impulse to escape was just too strong, and I needed to go elsewhere, preferably co-work with somebody else. Now I've cut down on co-working a lot, because leaving my home would take time, and I get more done if I don't need to spend that time on travel.*

This varies; implementing these fixes seems to have provided a temporary motivational boost allowing me to get a lot of work done with just the reward of financial security. When I find things to do that I'm significantly motivated by, then I seem to be able to work on them pretty well, even from home. However, anything that I'm *not* significantly motivated by still requires a lot of external structure for me to get anything done. Again, this seems like a manifestation of executive dysfunction issues more generally.

My initial motivation boost expired for a while, and I soon ran into new problems (I'll discuss these below). It has taken a while to find promising new directions and figure

out my new motivations so that I can do work more consistently, but (again thanks to meditation and Focusing / IFS work) in the last few months I've been starting to feel more consistently self-motivated.

In progress of being fixed after being made worse *by* the self-concept work

Lack of motivation once escaping the pain was no longer as motivating. *For a while, there was a sense that my life had gotten more boring. Remember that analogy about being hungry all the time and focusing all your energies on food, and then being transformed into an android which didn't need to eat? Your previous overriding priority of finding food being gone, you wouldn't know what to do anymore. You'd feel okay, and it would be a steady okay – no lows, but also no particular highs.*

The fixes in the post had the problem that I no longer felt actively bad; but eventually I started to notice that, having largely structured my life, habits and brain around escaping the badness, I didn't have any particularly wholesome ways of feeling *good*. Even though I had fixed a major cause behind my depression and burnouts, they had still left pretty deep marks in my brain. After a while, I started to feel acutely anhedonic – limited in my ability to get pleasure from anything. The fact that many of my previous obsessive fantasies had been eliminated probably made this worse, since they had at least been a source of pleasure and motivation.

But this is still a good development. The goal of life isn't to be free of problems; it's to have more interesting problems, and this is definitely a much more interesting problem. I've been trying new things, from going to museums to generally being more open to stuff. I'm working on fixing the remaining mental blocks that are keeping me in place rather than experiencing stuff.

I'm gradually relearning to genuinely enjoy things. And that feels good: I feel like I'm just getting started in the process of rebuilding myself.

Can't wait to see where I'll be in a few year's time.

The E-Coli Test for AI Alignment

Let’s say you have an idea in mind for how to align an AI with human values.

Go prep a slide with some e-coli, put it under a microscope, and zoom in until you can see four or five cells. Your mission: satisfy the values of those particular e-coli. In particular, walk through whatever method you have in mind for AI alignment. You get to play the role of the AI; with your sophisticated brain, massive computing power, and large-scale resources, hopefully you can satisfy the values of a few simple e-coli cells.

Perhaps you say “this is simple, they just want to maximize reproduction rate.” Ah, but that’s not quite right. That’s optimizing for the goals of the process of evolution, not optimizing for the goals of the godshatter itself. The e-coli has some frozen-in values which have evolved to approximate evolutionary fitness maximization in some environments; your job is optimize for the frozen-in approximation, even in *new* environments. After all, we don’t want a strong AI optimizing for the reproductive fitness of humans - we want it optimizing for humans’ own values.

On the other hand, perhaps you say “these cells don’t have any consistent values, they’re just executing a few simple hardcoded algorithms.” Well, you know what else doesn’t have consistent values? Humans. Better be able to deal with that somehow.

Perhaps you say “these cells are too simple, they can’t learn/reflect/etc.” Well, chances are humans will have the same issue once the computational burden gets large enough.

This is the problem of AI alignment: we need to both define and optimize for the values of things with limited computational resources and inconsistent values. To see the problem from the AI’s point of view, look through a microscope.

Why I expect successful (narrow) alignment

This is a linkpost for http://s-risks.org/why-i-expect-successful-alignment/

Summary

I believe that advanced AI systems will likely be aligned with the goals of their human operators, at least in a narrow sense. I'll give three main reasons for this:

1. The transition to AI may happen in a way that does not give rise to the alignment problem as it's usually conceived of.
2. While work on the alignment problem appears neglected at this point, it's likely that large amounts of resources will be used to tackle it if and when it becomes apparent that alignment is a serious problem.
3. Even if the previous two points do not hold, we have already come up with a couple of smart approaches that seem fairly likely to lead to successful alignment.

This argument lends some support to work on non-technical interventions like moral circle expansionor improving AI-related policy, as well as work on special aspects of AI safety like decision theory or worst-case AI safety measures.

Experiences of Self-deception

It seems to me that self-deception can describe two different things - conscious and unconscious self-deception.

Sometimes the elephant believes something untrue all by itself without the rider ever getting a look in. The claims of elephant in the brain seem to focus on this type of unconscious self-deception.

At other times the rider is complicit in endorsing a particular known untrue belief. The elephant analyses a situation, determines what it is beneficial to believe and motivates the rider to believe this. The rider has access to information which indicates that this isn't true. If the rider brings this information to full attention then it is one of those rare occasions where he can override the elephant's desires. However the rider also has the option to push the information to the side and believe a beneficial lie. It is possible to do this well enough that the information is forgotten or completely overridden with new, inaccurate, information.

In pushing the information to the side, the rider can sometimes just never bring the information to full attention. Failing that, it can drown the information out by presenting other information (which agrees with its favoured interpretation) as loudly as possible in order to doubt/ignore/forget the information which it doesn't like.

At least, this is something I experience but I don't know whether other people do. I have a few examples where this has happened and have even experimented with allowing myself to start down the route of conscious self-deception to see what it feels like. To me it feels like cognitive dissonance (feeling hot, brain feeling "fuzzy", adrenaline kicking in) whilst the rider works on counteracting the information. I guess this would be followed by the relief of resolving said dissonance when the rider starts to believe the lie but I haven't experimented that far!

The literature appears to be understandably non-committal on whether the subjects are consciously aware of their self-deception - I guess that would be pretty hard to determine.

So my question is - do other people recognise this as something which happens to them? How would you describe the experience? Is it something which you've trained yourself to recognise when it starts?

Can dying people "hold on" for something they are waiting for?

content note: death, old age, sickness

I've heard numerous anecdotal accounts of sick or old people who are on death's door "holding on" in a way suggestive that they were exerting some effort to do, until they had reached closure on some thing (a relative coming to visit, a manuscript published, somebody's birthday).

I could imagine this being a totally real thing that dying people can do for some limited time.

I can also imagine it just being cherry picked stories that were more a matter of luck.

It seems likely that there's at least situations where, say, eating is difficult/painful, and people continue exerting effort to do that so long as they have something that feels worth it to keep doing so, and then stop putting in the effort after hitting some milestone they cared about.

Some of the anecdotes I've heard implied something more immediate going on (where someone seemed to be holding on and literally a few minutes or seconds afterwards, died).

(Possible straightforward mechanism could just be that *breathing* becomes painful and difficult, and people only keep doing it when they have a concrete goal)

Standing on a pile of corpses

[In the darkest day of 2018, it is proper to think about the darkness that surrounds us]

When we think about the history of humanity, we focus on its highlights.

Galileo discovering the moons of Jupiter. Edward Jenner developing the first pox vaccine. Emmy Noether setting the mathematical foundation of modern physics.

We stand on the shoulders of giants, we say, that have elevated us over the clouds so we can see the stars above.

I believe a more apt metaphor is that we stand on a pile of nameless corpses.

Because for each great human that made it to the history books, a million have been forgotten. And each of them, knowingly or unknowingly, it is part of our legacy.

And most of them are dead.

100 billion humans have been born and died, most in undignified circumstances of sickness and age, and most by no choice of their own. They are now part of the pile of corpses.

7 billion people remain alive. And even though our lives are much better than those of the people from 1000 years ago, from 100 years ago and even from 10 years ago, we still suffer.

Many of our living kin live in sickness and hunger. Even amongst the most fortunate we wrestle with mental illness and accidents and the plights of aging.

We do not stand proud, but afraid and resigned to become yet another layer of the pile of corpses.

For not a single human lives free of the tyranny of death. We can choose to embrace it early, but we cannot still postpone it, not for much.

And so the pile of corpses grows, too fast for us to properly mourn the fallen.

There is a glimmer of hope, that the pile of corpses will stop growing as we come of age as a civilization. That we will stop the non consensual suffering we experience, and death will be no more for those who want to defy it.

That we will have, for the first time in history, time to breathe and reflect and remember all the nameless people who came before. To properly ponder without having to constantly struggle over our survival, and leaving behind extreme suffering.

To finally give a proper burial to the pile of corpses we stand on, and decide what to do with our piece of the universe.

But we are not guaranteed this happy ending - the book of humanity might suddenly end, without us having a chance to dictate the final words.

The worst we have endured is not a good predictor of the worst that it is to come, and for all we know the horrors ahead may be the ones that end us.

And then the pile of corpses will stop growing, but so will our civilization.

And all there will be left is an inanimate pile of corpses floating through the cosmos, surrounded by the cold and uncaring void.

Thanks to Tam Borine for proofreading the text. This text was used as part of a 2018 private secular solstice celebration in Madrid, Spain.

New Ratfic: Nyssa in the Realm of Possibility

For NaNoWriMo, I decided to do a rationality themed pastiche of the Phantom Tollbooth. It is complete and serializing at http://nyssa.elcenia.com on Saturdays and Wednesdays. There are three chapters up as of this posting.

Multi-agent predictive minds and AI alignment

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Abstract: An attempt to map a best-guess model of how human values and motivations work to several more technical research questions. The mind-model is inspired by predictive processing / active inference framework and multi-agent models of the mind.

The text has slightly unusual epistemic structure:

1st part: my current best-guess model of how human minds work.

2nd part: explores various problems which such mind architecture would pose for some approaches to value learning. The argument is: if such a model seems at least plausible, we should probably extend the space of active research directions.

3rd part: a list of specific research agendas, sometimes specific research questions, motivated by the previous.

I put more credence in the usefulness of research questions suggested in the third part than in the specifics of the model described the first part. Also, you should be warned I have no formal training in cognitive neuroscience and similar fields, and it is completely possible I'm making some basic mistakes. Still, my feeling is even if the model described in the first part is wrong, something from the broad class of "motivational systems not naturally described by utility functions" is close to reality, and understanding problems from the 3rd part can be useful.

How minds work

As noted, this is a "best guess model". I have large uncertainty about how human minds actually work. But if I could place just one bet, I would bet on this.

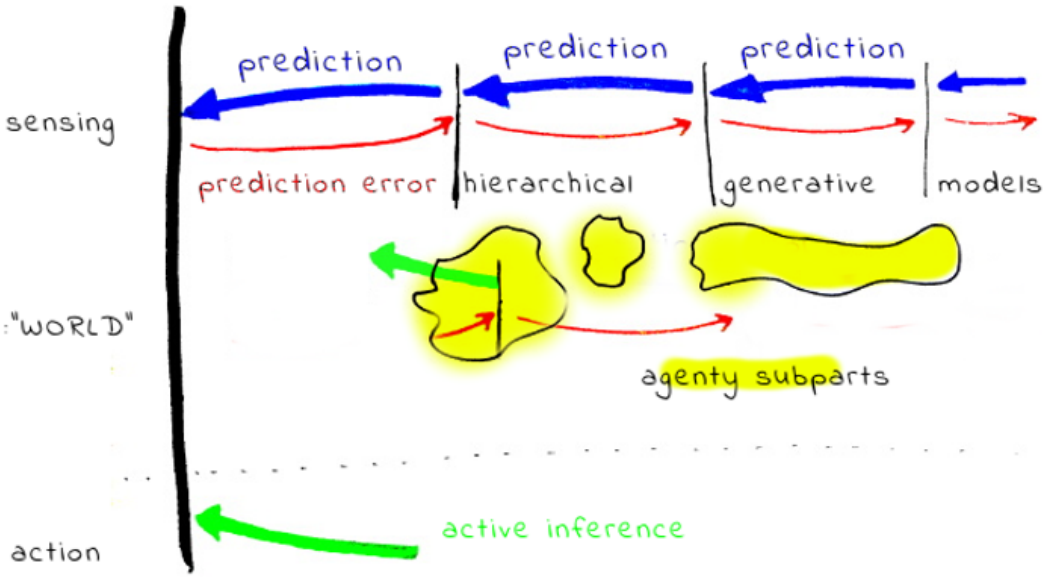
The model has two prerequisite ideas: predictive processing and the active inference framework. I'll give brief summaries and links for elsewhere.

In the predictive processing / the active inference framework, brains constantly predict sensory inputs, in a hierarchical generative way. As a dual, action is also "generated" by the same machinery (changing environment to match "predicted" desirable inputs and generating action which can lead to them). The "currency" on which the whole system is running is prediction error (or something in style of free energy, in that language).

Another important ingredient is bounded rationality, i.e. a limited amount of resources being available for cognition. Indeed, the specifics of hierarchical modelling, neural architectures, principle of reusing and repurposing everything, all seem to be related to quite brutal optimization pressure, likely related to brain's enormous energy consumption (It is unclear to me if this can be also reduced to the same "currency". Karl Friston would probably answer "yes").

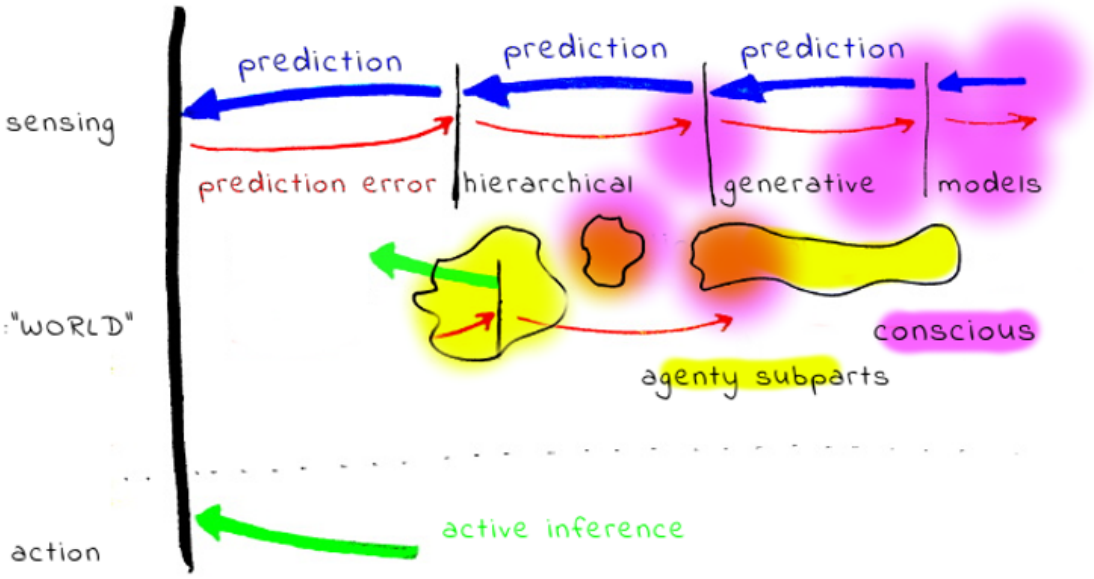
Assuming this whole, how do motivations and "values" arise? The guess is, in many cases something like a "subprogram" is modelling/tracking some variable, "predicting" its desirable state, and creating the need for action by "signalling" prediction error. Note that such subprograms can work on variables on very different hierarchical layers of modelling - e.g. tracking a simple variable like "feeling hungry" vs. tracking a variable like "social

status". Such sub-systems can be large: for example tracking "social status" seems to require lot of computation.



How does this relate to emotions? Emotions could be quite complex processes, where some higher-level modelling ("I see a lion") leads to a response in lower levels connected to body states, some chemicals are released, and this interoceptive sensation is re-integrated in the higher levels in the form of emotional state, eventually reaching consciousness. Note that the emotional signal from the body is more similar to "sensory" data - the guess is body/low level responses are a way how genes insert a reward signal into the whole system.

How does this relate to our conscious experience, and stuff like Kahneman's System 1/System 2? It seems for most people the light of consciousness is illuminating only a tiny part of the computation, and most stuff is happening in the background. Also, S1 has much larger computing power. On the other hand it seems relatively easy to "spawn background processes" from the conscious part, and it seems possible to illuminate larger part of the background processing than is usually visible through specialized techniques and efforts (for example, some meditation techniques).



Another ingredient is the observation that a big part of what the conscious self is doing is interacting with other people, and rationalizing our behaviour. (Cf. press secretary theory, elephant in the brain.) It is also quite possible the relation between acting rationally and the ability to rationalize what we did is bidirectional, and significant part of motivation for some rational behaviour is that it is easy to rationalize it.

Also, it seems important to appreciate that the most important part of the human "environment" are other people, and what human minds are often doing is likely simulating other human minds (even simulating how other people would be simulating someone else!).

Problems with prevailing value learning approaches

While the above sketched picture is just a best guess, it seems to me at least compelling. At the same time, there are notable points of tension between it and at least some approaches to AI alignment.

No clear distinction between goals and beliefs

In this model, it is hardly possible to disentangle "beliefs" and "motivations" (or values). "Motivations" interface with the world only via a complex machinery of hierarchical generative models containing all other sorts of "beliefs".
To appreciate the problems for the value learning program, consider a case of someone who's predictive/generative model strongly predicts failure and suffering. Such person may take actions which actually lead to this outcome, minimizing the prediction error.

Less extreme but also important problem is that extrapolating "values" outside of the area of validity of generative models is problematic and could be fundamentally ill-defined. (This is related to "ontological crisis".)

No clear self-alignment

It seems plausible the common formalism of agents with utility functions is more adequate for describing the individual “subsystems” than the whole human minds. Decisions on the whole mind level are more like results of interactions between the sub-agents; results of multi-agent interaction are not in general an object which is naturally represented by utility function. For example, consider the sequence of game outcomes in repeated PD game. If you take the sequence of game outcomes (e.g. 1: defect-defect, 2:cooperate-defect, ...) as a sequence of actions, the actions are not representing some well behaved preferences, and in general not maximizing some utility function.

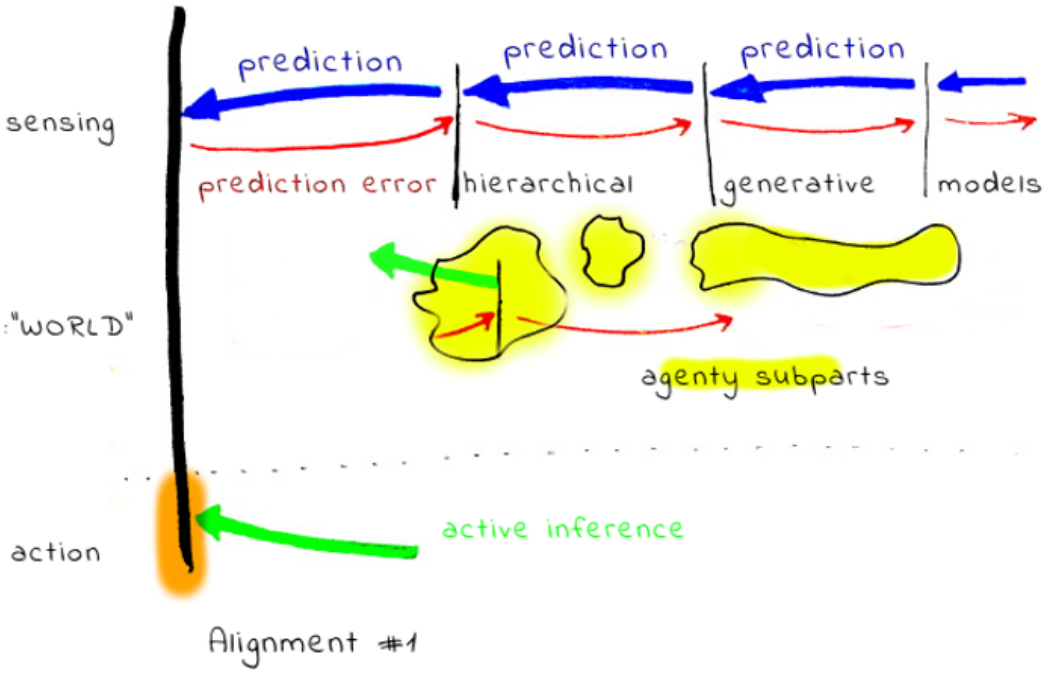
Note: This is not to claim VNM rationality is useless - it still has the normative power - and some types of interaction lead humans to approximate SEU optimizing agents better.

One case is if mainly one specific subsystem (subagent) is in control, and the decision does not go via too complex generative modelling. So, we should expect more VNM-like behaviour in experiments in narrow domains than in cases where very different sub-agents are engaged and disagree.
Another case is if sub-agents are able to do some “social welfare function” style aggregation, bargain, or trade - the result could be more VNM-like, at least in specific points of time, with the caveat that such “point” aggregate function may not be preserved in time.

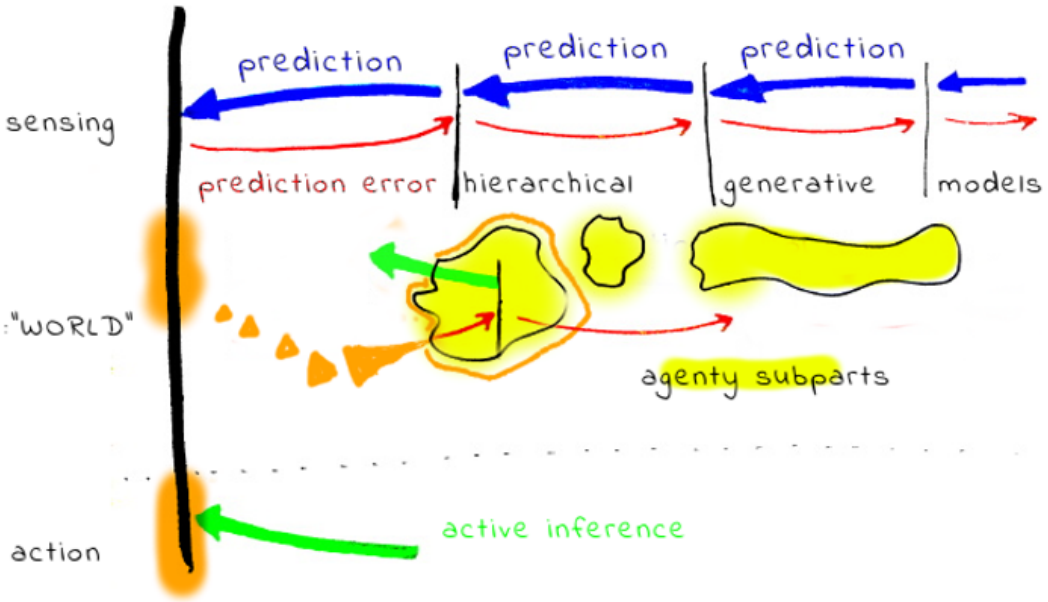
On the contrary, cases where the resulting behaviour is very different from VNM-like may be caused by sub-agents locked in some non-cooperative Nash equilibria.

What we are aligning AI with

Given this distinction between the whole mind and sub-agents, there are at least four somewhat different notions of what alignment can mean.

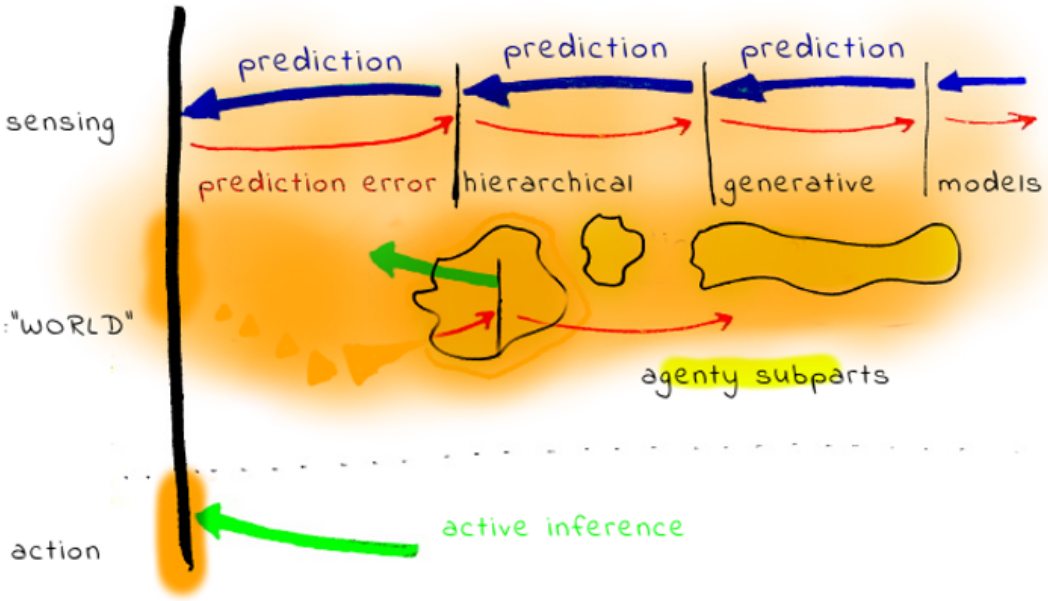


1. Alignment with the outputs of the generative models, without querying the human. This includes for example proposals centered around approval. In this case, generally only the output of the internal aggregation has some voice.



Alignment #2

2. Alignment with the outputs of the generative models, with querying the human. This includes for example CIRL and similar approaches. The problematic part of this is, by carefully crafted queries, it is possible to give voice to different sub-agenty systems (or with more nuance, give them very different power in the aggregation process). One problem with this is, if the internal human system is not self-aligned, the results could be quite arbitrary (and the AI agent has a lot of power to manipulate)



Alignment #3

3. Alignment with the whole system, including the human aggregation process itself. This could include for example some deep NN based black-box trained on a large amount of human data, predicting what would the human want (or approve).

4. Adding layers of indirection to the question, such as defining alignment as a state where the *"A is trying to do what H wants it to do."*

In practice, options 1. and 2. can collapse into one, as far as there is some feedback loop between the AI agent actions and the human reward signal. (Even in case 1, the agent can take an action with the intention to elicit feedback from some subpart.)

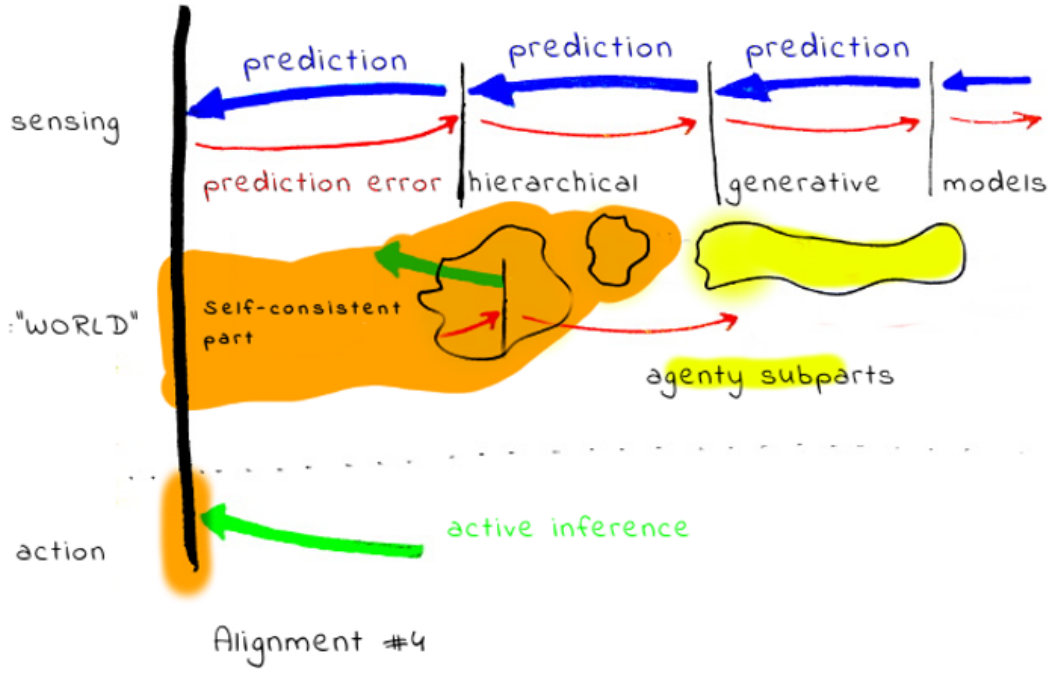
We can construct a rich space of various meanings of "alignment" by combining basic directions.

Now, we can analyze how these options interact with various alignment research programs.

Probably the most interesting case is IDA. IDA-like schemes can probably carry forward arbitrary properties to more powerful systems, as long as we are able to construct the individual step preserving the property. (I.e. one full cycle of distillation and amplification, which can be arbitrarily small).

Distilling and amplifying the alignment in sense #1 (what the human will actually approve) is conceptually easiest, but, unfortunately, brings some of the problems of potentially super-human system optimizing for manipulating the human for approval.

Alignment in sense #3 creates a very different set of problems. One obvious risk are mind-crimes. More subtle risk is related to the fact that as the implicit model of human "wants" scales (becomes less bounded), I. the parts may scale at different rates II. the outcome equilibria may change even if the sub-parts scale at the same rate.



Alignment #4

Alignment in sense #4 seems more vague, and moves the burden of understanding the problem in part to the side of the AI. We can imagine that at the end the AI will be aligned with some part of the human mind in a self-consistent way (the part will be a fixed point of

the alignment structure). Unfortunately, it is *a priori* unclear if a unique fixed point exists. If not, the problems become similar to case #2. Also, it seems inevitable the AI will need to contain some structure representing what the human wants the AI to do, which may cause problems similar to #3.

Also, in comparison with other meanings, it is much less clear to me how to even establish some system has this property.

Rider-centric and meme-centric alignment

Many alignment proposals seem to focus on interacting just with the conscious, narrating and rationalizing part of mind. If this is just a one part entangled in some complex interaction with other parts, there are specific reasons why this may be problematic.

One: if the “rider” (from the rider/elephant metaphor) is the part highly engaged with tracking societal rules, interactions and memes. It seems plausible the “values” learned from it will be mostly aligned with societal norms and interests of memeplexes, and not “fully human”.

This is worrisome: from a meme-centric perspective, humans are just a substrate, and not necessarily the best one. Also - a more speculative problem may be - schemes learning human memetic landscape and “supercharging it” with superhuman performance may create some hard to predict evolutionary optimization processes.

Metapreferences and multi-agent alignment

Individual “preferences” can often in fact be mostly a meta-preference to have preferences compatible with other people, based on simulations of such people.

This may make it surprisingly hard to infer human values by trying to learn what individual humans want without the social context (necessitating inverting several layers of simulation). If this is the case, the whole approach of extracting individual preferences from a single human could be problematic. (This is probably more relevant to some “prosaic” alignment problems)

Implications

Some of the above mentioned points of disagreements point toward specific ways how some of the existing approaches to value alignment may fail. Several illustrative examples:

- Internal conflict may lead to inaction (also to not expressing approval or disapproval). While many existing approaches represent such situation only by the *outcome* of the conflict, the internal experience of the human seems to be quite different with and without the conflict
- Difficulty with splitting “beliefs” and “motivations”.
- Learning inadequate societal equilibria and optimizing on them.

Upside

On the positive side, it could be expected the sub-agents still easily agree on things like “it is better not to die a horrible death”.

Also, the mind-model with bounded sub-agents which interact only with their local neighborhood and do not actually care about the world may be a viable design from the safety perspective.

Suggested technical research directions

While the previous parts are more in backward-chaining mode, here I attempt to point toward more concrete research agendas and questions where we can plausibly improve our understanding either by developing theory, or experimenting with toy models based on current ML techniques.

Often it may be the case that some research was already done on the topic, just not with AI alignment in mind, and a high value work could be “importing the knowledge” into safety community.

Understanding hierarchical modelling.

It seems plausible the human hierarchical models of the world optimize some "boundedly rational" function. (Remembering all details is too expensive, too much coarse-graining decreases usefulness. A good bounded rationality model can work as a principle for how to select models. In a similar way to the minimum description length principle, just taking some more “human” (energy?) costs as cost function.)

Inverse Game Theory.

Inverting agent motivations in MDPs is a different problem from inverting motivations in multi-agent situations where game-theory style interactions occur. This leads to the inverse game theory problem: observe the interactions, learn the objectives.

Learning from multiple agents.

Imagine a group of five closely interacting humans. Learning values just from person A may run into the problem that big part of A’s motivation is based on A simulating B,C,D,E (on the same “human” hardware, just incorporating individual differences). In that case, learning the “values” just from A’s actions could be in principle more difficult than observing the whole group, trying to learn some “human universals” and some “human specifics”. A different way of thinking about this could be by making a parallel with meta-learning algorithms (e.g. REPTILE) but in IRL frame.

What happens if you put a system composed of sub-agents under optimization pressure?

It is not clear to me what would happen if you, for example, successfully “learn” such a system of “motivations” from a human, and then put it inside of some optimization process selecting for VNM-like rational behaviour.

It seems plausible the somewhat messy system will be forced to get more internally aligned; for example, one way how it can happen is one of the sub-agent systems takes control and “wipes out the opposition”.

What happens if you make a system composed of sub-agents less computationally bounded?

It is not clear that the relative powers of sub-agents will scale the same with the whole system becoming less computationally bounded. (This is related to MIRI’s sub-agents agenda)

Suggested non-technical research directions

Human self-alignment.

All other things being equal, it seem safer to try to align AI with humans which are self-aligned.

Notes & Discussion

Motivations

Part of my motivation for writing this was an annoyance: there is a plenty of reasons to believe the view

- human mind is a unified whole,
- at first approximation optimizing some utility function,
- this utility is over world-states,

is neither a good model of humans, nor the best model how to think about AI. Yet, it is the paradigm shaping a lot of thoughts and research. I hope if the annoyance surfaced in the text, it is not too distractive.

Multi-part minds in literature

There are dozens of schemes describing mind as some sort of multi-part system, so there is nothing original about this claim. Based on a very shallow review, it seems the way how psychologists often conceptualize the sub-agents is as subpersonalities, which are almost fully human. This seems to err on the side of sub-agents being too complex, and anthropomorphising instead of trying to describe formally. (Explaining humans as a composition of humans is not much useful for AI alignment). On the other hand, Minsky's *"Society of Mind"* has sub-agents which often seem to be too simple (e.g. similar in complexity to individual logic gates). If there is some literature having sub-agent complexity right, and sub-agents being inside predictive processing, I'd be really excited about it!

Discussion

When discussion the draft, several friends noted something along the line: "It is overdetermined that approaches like IRL are doomed. There are many reasons for that and the research community is aware of them". To some extent, I agree this is the case, on the other hand 1. the described model of mind may pose problems even for more sophisticated approaches 2. My impression is many people still have something like utility-maximizing agent as a the central example.

The complementary objection is that while interacting sub-agents may be a more precise model, it seems in practice it is often enough to think about humans as unified agents is good enough, and may be good enough even for the purpose of AI alignment. My intuitions on this is based on the connection of rationality to exploitability: it seems humans are usually more rational and less exploitable when thinking about narrow domains, but can be quite bad when vastly different subsystems are in in play (imagine on one side a person exchanging stock and money, on the other side some units of money, free time, friendship, etc.. In the second case, many people are willing to trade in different situations by very different rates)

I'd like to thank Linda Linsefors , Alexey Turchin, Tomáš Gavenčiak, Max Daniel, Ryan Carey, Rohin Shah, Owen Cotton-Barratt and others for helpful discussions. Part of this originated in the efforts of the "Hidden Assumptions" team on the 2nd AI safety camp, and my thoughts about how minds work are inspired by CFAR.

Equivalence of State Machines and Coroutines

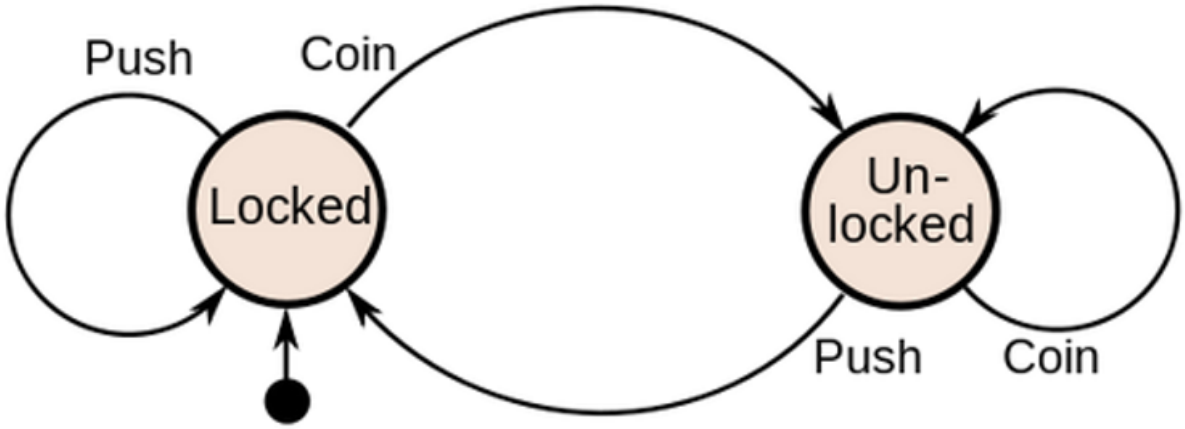
In the past I often referred to the equivalence between state machines and coroutines as a kind of obvious fact that doesn't need any additional explanation. It was brought to my attention, however, that that may not always be the case.

This article therefore doesn't attempt to express and deep and ground-breaking truth, rather, it illustrates the equivalence of finite state machines and coroutines using a practical example.

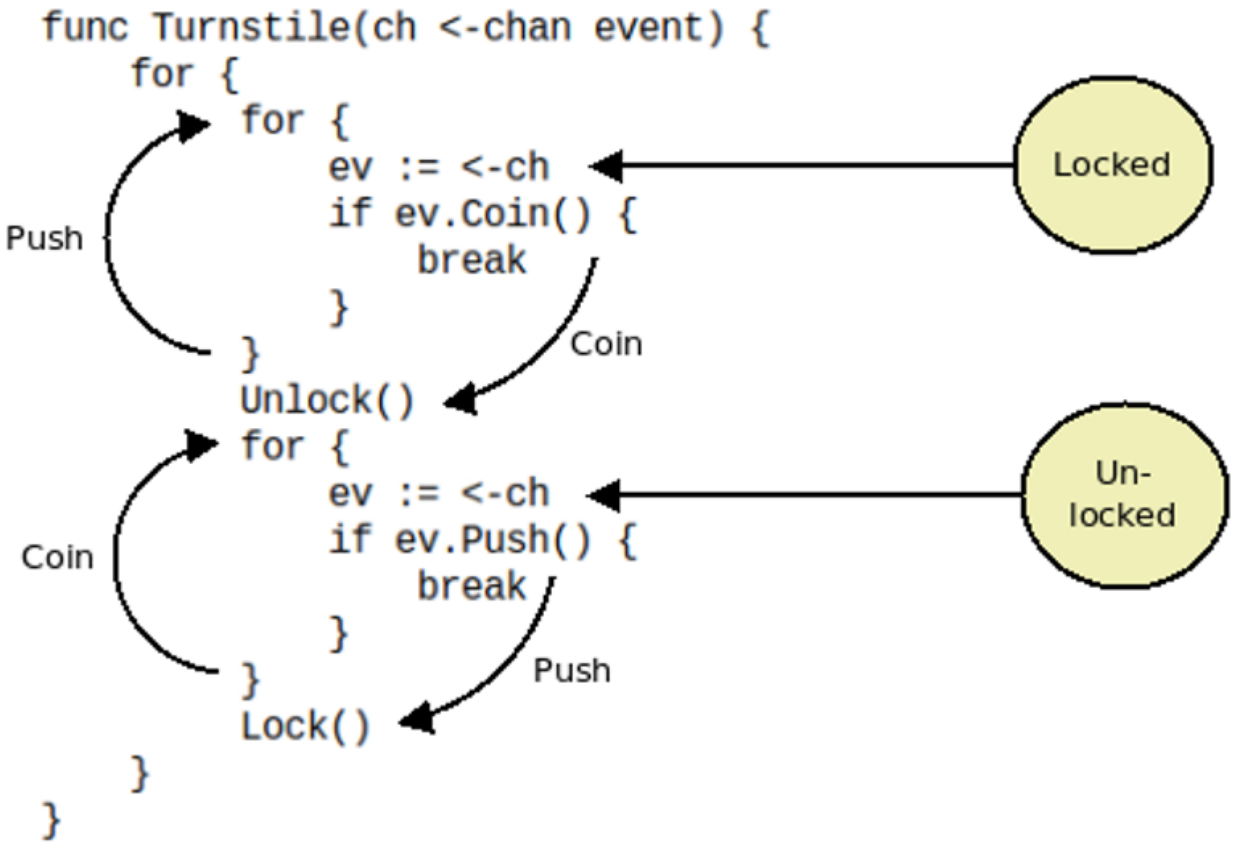
The example is stolen from Wikipedia's article on finite state machines:

> A turnstile, used to control access to subways and amusement park rides, is a gate with three rotating arms at waist height, one across the entryway. Initially the arms are locked, blocking the entry, preventing patrons from passing through. Depositing a coin or token in a slot on the turnstile unlocks the arms, allowing a single customer to push through. After the customer passes through, the arms are locked again until another coin is inserted.

Wikipedia presents the state machine in question in the following manner:



And here's my attempt to rewrite the state machine as a coroutine in Go:


func Turnstile(ch <-chan event) {
for {
for {
ev := <-ch
if ev.Coin() {
break
}
}
Unlock()
for {
ev := <-ch
if ev.Push() {
break
}
}
Lock()
}
}
Locked
Un-
locked
Push
Coin
Coin
Push

Measly Meditation Measurements

A few months ago, I decided to start meditating regularly, around an hour a day. It seemed like a good opportunity to measure possible effects, so I asked for advice on what to measure. This post summarizes the results. In short, while the *subjective* effects of meditation were strong, the *measurements* didn't show anything. This is a fine place to stop reading; I'm mostly posting this because I promised to.

I did mindfulness meditation, as guided by The Mind Illuminated. My object of focus was typically my breath (while sitting), or my steps (as hiking).

What I Measured

- About once a week, I did some online tasks either before or after meditating. These were the Go/NoGo and CuedAttention tasks on quantified-mind.com, and this psychomotor vigilance task.
- I set up roughly once or twice daily pings from TagTime for experience sampling.

What I Measured

- My performance on the tasks looked entirely random. It wasn't better or worse after meditating, and it didn't get better or worse over time.
- I have no idea how to do experience sampling. I understand that some people have moods. I'm almost always in a neutral mood, and so wasn't sure what to put most of the time. Also, I'm apparently often away from my phone, and missed many (most?) pings.

What I Learned

- The Mind Illuminated is as good of a guide as I hoped it would be.
- A few measly months of meditation isn't going to change anything like your performance on reaction-time-like tasks.
- A few measly months of meditation *will* give you a fascinating look into your own mind. It's not what you think. I'd say more, but I'm deeply confused and don't have a good model.
- Meditation retreats are great. I went on a two-day one, whose format wasn't particularly well-suited for me, and even this had a large effect on my practice.

What are the axioms of rationality?

I'm new here (my first post), i just started to get serious about rationality, and one of the questions that immediately came to my mind is "What are the axioms of rationality?". I looked it up a bit, and didn't find (even on this site) a post that'll show them (and i'm quite sure there are).

So this is intended as a discussion, And I'll make a post with the conclusions afterward.

curious to see your reply's! (as well if you have feedback on how i asked the question)

thanks :)