Best of LessWrong: July 2018

Best of LessWrong: July 2018

48. A Gym Gridworld Environment for the Treacherous Turn
49. How To Use Bureaucracies
50. June gwern.net newsletter

Are ethical asymmetries from property rights?

These are some intuitions people often have:

- You are not required to save a random person, but you are definitely not allowed to kill one
- You are not required to create a person, but you are definitely not allowed to kill one
- You are not required to create a happy person, but you are definitely not allowed to create a miserable one
- You are not required to help a random person who will be in a dire situation otherwise, but you are definitely not allowed to put someone in a dire situation
- You are not required to save a person in front of a runaway train, but you are definitely not allowed to push someone in front of a train. By extension, you are not required to save five people in front of a runaway train, and if you have to push someone in front of the train to do it, then you are not allowed.

Here are some more:

- You are not strongly required to give me your bread, but you are not allowed to take mine
- You are not strongly required to lend me your car, but you are not allowed to unilaterally borrow mine
- You are not strongly required to send me money, but you are not allowed to take mine

The former are ethical intuitions. The latter are implications of a basic system of property rights. Yet they seem very similar. The ethical intuitions seem to just be property rights as applied to lives and welfare. Your life is your property. I'm not allowed to take it, but I'm not obliged to give it to you if you don't by default have it. Your welfare is your property. I'm not allowed to lessen what you have, but I don't have to give you more of it.

My guess is that these ethical asymmetries—which are confusing, because they defy consequentialism—are part of the mental equipment we have for upholding property rights.

In particular these well-known asymmetries seem to be explained well by property rights:

- **The act-omission distinction** naturally arises where an act would involve taking someone else's property (broadly construed—e.g. their life, their welfare), while an omission would merely fail to give them additional property (e.g. life that they are not by default going to have, additional welfare).
- **'The asymmetry'** between creating happy and miserable people is because to create a miserable person is to give that person something negative, which is to take away what they have, while creating a happy person is giving that person something extra.
- **Person-affecting views** arise because birth gives someone a thing they don't have, whereas death takes a thing from them.

Further evidence that these intuitive asymmetries are based on upholding property rights: we also have moral-feeling intuitions about more straightforward property rights. Stealing is wrong.

If I am right that we have these asymmetrical ethical intuitions as part of a scheme to uphold property rights, what would that imply?

It might imply something about when we want to uphold them, or consider them part of ethics, beyond their instrumental value. Property rights at least appear to be a system for people with diverse goals to coordinate use of scarce resources—which is to say, to somehow use the resources with low levels of conflict and destruction. They do not appear to be a system for people to achieve specific goals, e.g. whatever is actually good. Unless what is good is exactly the smooth sharing of resources.

I'm not actually sure what to make of that—should we write off some moral intuitions as clearly evolved for not-actually-moral reasons and just reason about the consequentialist value of upholding property rights? If we have the moral intuition, does that make the thing of moral value, regardless of its origins? Is pragmatic rules for social cohesion all that ethics is anyway? Questions for another time perhaps (when we are sorting out meta-ethics anyway).

A more straightforward implication is for how we try to explain these ethical asymmetries. If we have an intuition about an asymmetry which stems from upholding property rights, it would seem to be a mistake to treat it as evidence about an asymmetry in consequences, e.g. in value accruing to a person. For instance, perhaps I feel that I am not obliged to create a life, by having a child. Then—if I suppose that my intuitions are about producing goodness—I might think that creating a life is of neutral value, or is of no value to the created child. When in fact the intuition exists because allocating things to owners is a useful way to avoid social conflict. That intuition is part of a structure that is known to be agnostic about benefits to people from me giving them my stuff. If I'm right that these intuitions come from upholding property rights, this seems like an error that is actually happening.

Prediction Markets: When Do They Work?

Epistemic Status: Resident Expert

I'm a little late on this, which was an old promise to Robin Hanson (not that he asked for it). I was motivated to deal with this again by the launch of Augur (REP), the crypto prediction market token. And by the crypto prediction market token, I mean the empty shell of a potential future prediction market token; what they have now is pretty terrible but in crypto world that is occasionally good for a $300 million market cap. This is, for now, one of those occasions.

The biggest market there, by far, is on whether Ether will trade above $500 at the end of the year. This is an interesting market because *Augur bets are made in Ether.* So even though the market (as of last time I checked) says it's 74% percent to be trading above $500 and it's currently $480 (it's currently Thursday on July 26, and I'm *not* going to go back and keep updating these numbers). When I first saw this the market was at 63%, which seemed to me like a complete steal. Now it's at 74%, which seems more reasonable, which means the first 'official DWATV trading tip' will have to wait. A shame!

A better way to ask this question, given how close the price is to $500 now, is what the ratio of 'given Ether is above $500 what does it cost' to 'given Ether is below $500 what does it cost' should be. A three to one ratio seems plausible?

The weakness (or twist) on markets this implies applies to prediction markets generally. If you bet on an event that is correlated with the currency you're betting in, the fair price can be very different from the true probability. It doesn't have to be price based – think about betting on an election between a hard money candidate and one who will print money, or a prediction on a nuclear war.

If I bet on a nuclear war, and win, how exactly am I getting paid?

Robin Hanson, Eliezer Yudkowsky and Scott Sumner are big advocates of prediction markets. In theory, so am I. Prediction markets are a wonderful thing. By giving people a monetary incentive to solve problems and share information, we can learn probabilities (what will GDP be next year?) and conditional probabilities (what will GDP be next year if we pass this tax cut bill?) and use the answers to make the best decision. This method of making decisions is called futarchy.

Formally, a prediction market allows participants to buy and sell contracts. Those contracts then pay out a variable amount of money. Typically this is either binary (will Donald Trump be elected president?), paying out 100 if the event happens and 0 if it doesn't, or they are continuous (how many electoral college votes will Donald Trump get?) and pay proportionally to the answer. Sometimes there are special cases where the market is void and all transactions are undone, at other times strange cases have special logic to determine the payout level.

There are three types of prediction markets that have gotten non-zero traction.

The first is politics. There are markets at PredictIt and BetFair and Pinnacle Sports, and there used to be relatively deep markets at InTrade. These markets matter enough to get talked about and attract some money when they involve major events like presidential elections, but tend to be quite pathetic for anything less than that.

The second is economics. There are lots of stocks and futures and options and other such products available for purchase. Futures markets in particular are prediction markets. They don't call themselves prediction markets, but that is one of the things they are, and the information they reveal is invaluable. It's even sometimes used to make decisions.

The third is sports. Most televised sporting events have bookmakers offering odds and taking bets. They use their own terminology for many things, but these are the closest thing to true prediction markets out there.

What makes a successful prediction market? What makes an unsuccessful prediction market? When are they efficient? What gets people involved?

To get a thriving market, you need (at least) these five things.

I. Well-Defined

If you can't exactly define the outcome, you can't have a prediction market. Even highly unlikely corner cases must be resolved. Thus, if you want a market on "Donald Trump is elected president of the United States in 2020" you need to know exactly what happens if he dies after the election but before inauguration, or if there is a revolt in the electoral college, or if the election is fraudulent or cancelled, or if he loses to a different person also named "Donald Trump." That's not because Trump makes such issues more likely. If you were betting on Obama vs. Romney, you'd need to do all the same stuff.

In sports markets, this means writing up a multi-page document detailing what happens when a game is rained out, disputed, postponed, tied, you name it, along with all the other rules. If there's an angle left ambiguous, you can bet (well you can't, but if you could, you'd have good odds) that someone will try to take advantage of it eventually. That leaves everyone mad and ruins good business relationships. It's important to have clear rules and stick to them.

One of the first markets on Augur asked, "Will England defeat Croatia in the World Cup?" Which I immediately recognized as a really bad wording, because it was ambiguous. If the game had gone to overtime, or even to penalty kicks, and England had advanced, what happens? In a real sportsbook, generally bets default to regulation time only, so the match would be ruled a draw, and the answer would be "No" even if England later won.

That's not acceptable. All it takes is one corner case to get people yelling at each other, and drive them away. When they do happen, the sportsbook is wise to eat the loss, even if that means paying both sides, and then fix its procedures. That's one benefit of having a central authority to blame.

I'm also including in this the requirement that *you can be confident you can collect if you win.* As one sportsbook's slogan puts it, sweat the game, not the payout. Bets with people who might not pay you require *huge* edges. They're about accepting the risk to land the whale. They don't do much for price discovery, and are for professionals only.

II. Quick Resolution

The faster the market pays out, the more interest you'll get. Markets that tie up money for weeks or months, let alone years, see decisive declines in participation. Before the season, there are markets for which teams will win the championship, or how many wins each team will have. This seems great, since if you're right you can have a huge advantage, and it gives you an interest in games for much or all of the season.

Despite that, you will see less money wagered on any given season win total *than for a random game played by that team.* Usually by an order of magnitude. That's how valuable quick resolution is. Even in March Madness, where everyone fills out office brackets, the bulk of the real wagering goes game by game. There are lots of propositions, and there's value there, but only for small amounts. Betting your bracket before the tournament, *despite brackets being something presidents do to show they're in touch,* isn't a thing for serious money.

Thus, markets on small events like individual games are bigger, and more efficient, than markets on bigger and more interesting things like entire seasons or even a playoff series. The primary purpose of the long-term markets isn't to make money; it's to provide a service so people can see what odds have been assigned to various outcomes.

Major events like presidential elections have enough inherent interest to still see solid markets, but only barely. There's a lot of interest in what the odds are, but the volumes traded are quite thin, so much so that it is in the interest of partisans to trade in order to move the price and thus change the political narrative.

Economic markets are the only place longer-term markets prosper.

Note that if the market is sufficiently liquid, it can act *as if* it is short term, provided the prices will move quickly enough, since participants can then exit their positions.

III. Probable Resolution

Trading in a prediction market ties up capital, creates risk and requires optimization pressure. I need to pay attention to the market, both to decide what fair value is and then to go about maximizing and making good trades.

If that market is *conditional,* and trades were only valid if those conditions were met, we have a problem: I've wasted my time, money and risk capacity, and gotten nothing in return.

One of the markets I liked a lot as a gambler was called the Home/Away line in MLB. The idea was, you added up all the runs scored by the home teams and compared them to all the runs scored by the away teams, and bet on which would be higher, or on the sum of all runs scored that day, which was called The Grand Salami. There was lots of value in these lines because people were using very simple heuristics, and if you did first-level statistics on runs scored in games and how distributions add up, you could get a big edge.

What was continuously frustrating was that often one game would get rained out, cancelling all your wagers. Often I'd have locked in large profits, and they'd be lost.

This wasn't enough to keep me from betting, because I got my money back within a day and the edge was huge. But when funds were tight, it shifted those funds towards other things, and every time I thought about looking at the Home/Away line, my brain fired back ‘are you sure you want to bother?’ so I only cared when my edge was large.

Gamblers actively prefer betting on odds that can’t tie, e.g. betting on a football team -3.5 or -2.5 rather than -3.0, because the -3 line ties 10% of the time. The bookmaker agrees!

If you are instead tying up your money for weeks, months or even years, and instead of a 10% chance of rain somewhere there’s a 50% or even 90% chance the event doesn’t fire, that’s much worse. If your’e dealing with a hyper-complex Hansonian death trap of a conditional market where it’s 99%+ to not happen, even with good risk measurement tools that don’t tie up more money than necessary, no one is going to want to put in the work and tie up the funds.

IV. Limited Hidden Information

Insider trading of securities is illegal. This seems at odds with price discovery. If I know something you don’t know, then my *not* trading on it makes the price less accurate. One might suggest that allowing insiders to trade would make the price more efficient.

The problem is that it drives people away.

If other people know something important I don’t, then my trades are giving them a way to pick my pocket. When I look at the price and see it is wrong, my prior is ‘there is something I don’t know’ rather than ‘there is something *they* don’t know or understand’. I’m the one making the mistake. I’m the sucker. So I walk away.

Thus, while the individual trades of insiders make the market more efficient, they punish others trying to share their information and analysis with the market. This is bad. Bad enough to kill outright markets with too much risk of insider information.

The first season of Survivor, there was a market on who would win. The production crew found out. Then there was no market.

Another important case: If a person with a large role in choosing the outcome can bet in the market, you might not want to risk betting against him. Or bet at all.

When there is a big injury risk in a game, the market dies until the issue is resolved. When the issue is resolved, trading picks back up no matter the outcome.

Even reduction of uncertainty as such can be important. Before important events like elections often money will ‘sit on the sideline’ until the outcome is known. This can even result in bad outcomes driving prices *up.* We may not like the new boss, but at least we now know who he is and can go about our business.

In my experience with prediction markets, important hidden information other traders could know acts as an outright veto on the market. It might not do that if the market had enough ‘natural’ trading volume, but that’s a high bar to clear.

V. Sources of Disagreement and Interest

Also known as, Suckers at the Table.

Any market, like a poker table, requires sources of disagreements and profits. Without a sucker at the table, why participate in the market? Remember, if you can’t spot the sucker in your first half hour at the table, then you are the sucker.

Ideally, you want either a direct subsidy to the market, or natural buyers and sellers.

If someone has a reason to trade even at a not so great price, for example airlines or countries hedging against moves in oil prices, then everyone can compete to make money off of that. The same would go if someone wanted to hedge against a political event, or to bet for or against their favorite team on principle – either to make it interesting, or to get what I used to half-jokingly call ‘compensation for our pain’ when the Mets inevitably lost again.

Another class of ‘natural’ traders are gamblers or noise traders, who demand liquidity for no particular reason. They too can be the sucker.

If people who want to learn fair probabilities subsidize the market, like donors subsidize the NGDP futures markets Scott Sumner helped create, that also works.

And of course there’s always the people who *think* they know something, and are sadly mistaken.

What traders need, more than anything else, is the ability to tell a story for why their trade is a good idea. To do that, they need to know why they have the opportunity to make this trade. What do they know that others don’t know? What mistake do they think people are making?

For sports, politics and economics, everyone has an opinion, so it is easy for them to get the idea that they have the advantage.

The genius of a binary bet on Ether prices *that trades in Ether* is that there are a lot of angles where one can think you know something the market doesn’t fully know, and lots of mistakes you can think other people are making. It’s easy to think of many different angles and approaches one could take. One can trade short term, or trade long term, do arbitrage or use it for leverage. Another could be doing it as a form of speech, or an experiment, and the group that can reach the market is doubtless quite biased.

It’s easy to make that leap to ‘I know why he’s willing to give me this trade’ and even to ‘I know exactly what mistake he is making.’ It’s a great choice for a big initial market.

VI. Summary and Conclusion

Prediction markets rely on attracting a variety of participants, including both ‘losers’ who have natural reasons to participate, and ‘winners’ who will be attracted by that value.

Any critical issue can kill a prediction market dead, or even an entire prediction market ecosystem.

If your market isn’t well-defined, arguments over price become arguments over the rules, which turn into very angry participants. If this happens even a small percentage of the time, it drives everyone away.

If your market doesn’t resolve quickly, and quickly is on the order of days or at most a few weeks, it needs to be massively liquid and refer to real world questions people have natural exposures to, to create participation. It ties up cash and doesn't offer the rush of a good gamble. No one wants to bet on an obscure outcome years from now.

If your market is unlikely to resolve, participants will find other uses for their time and money. The chance here has to be small, well under 50%, and much lower if time to resolution isn't quick. Years-long markets that are unlikely to trigger are going to have severe issues.

If your market has potential hidden information, that is a tax on everyone who participates, who are prey to adverse selection. Everyone must worry that the market knows what they don't know, and that them liking one side of a trade means the person on the other side has a secret; there are even traders in such markets that follow a strategy of 'find the naively correct bet, and bet the other way,' which is known (or should be known) as The Constanza.

If your market doesn't draw natural interest and offer sources of disagreement, to create a foundation of participation and liquidity, there's nothing to build on and no interest.

In addition to these threats, such markets face regulatory and legal hurdles, and face various ethical concerns. If you offer one market that seems to mimic a regulated trade, such as an option on a stock, or that sounds distasteful, such as the so-called 'assassination markets,' that can be all anyone will see when they look at your offerings. Even though such concerns, frankly, are mostly quite stupid, they're real and people care about this a lot. They've gotten basically every past attempt at prediction markets (other than bookmakers and professional economic trading platforms) shut down.

Active curation is necessary to deal with many of these issues, and to provide simple ease of use and ease of finding what one is looking for and would be interested in.

Surgical use of prediction markets for key information points remains a great idea, and in many cases people love a good bet. But we shouldn't get too ambitious, and keep an eye on the practical needs of participants.

Strategies for Personal Growth

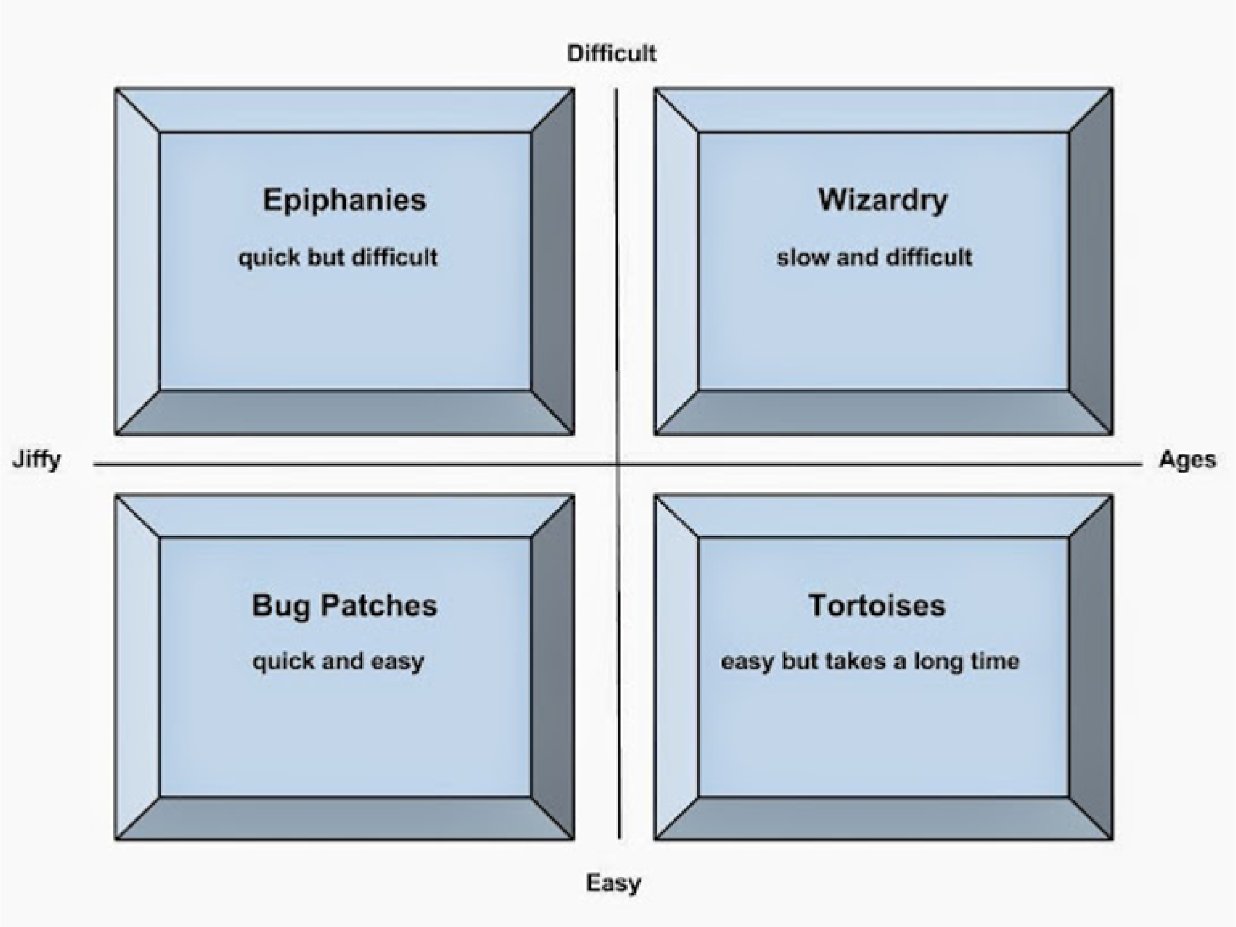
Recently I was having a conversation with a friend about personal growth (any form of deliberate gain in capability or subjective wellbeing).

We were talking past each other a lot. Eventually it became clear that most of *their* recent growth had been healing and blocker-fixing based, whereas most of mine had been skill based. And this was shaping where we were naturally inclined to look.

This prompted me to take stock of all the strategies I could think of for growth. Here's what I've thought of so far. These aren't quite natural clusters (some overlap, or don't quite fit into the same ontology), but seem like they cover most avenues:

- Low-Key Practice
- Serious Practice
- Learning
- Changing environment and incentives
- Discovering things you enjoy/are-good-at (Try things!)
- Blockers and Cheat Codes
- Healing
- Deep Improvements to Mental Architecture (Internal Alignment)

This is somewhat related to Lifelonglearner's Development Framework of Rationality, and Brienne's 4 quadrants, but through a somewhat different lens.



Overview of Strategies

Low-Key Practice

Working on a skill or habit you mostly understand, occasionally, in the background. Typically yields a pretty small return (but one which adds up over the years, especially if you do it consistently).

Serious Practice

Working on a skill with your all – spending hour(s) each day on it. This can yield more return but is requires more cognitive focus and energy. This seems qualitatively different from low-key practice.

Sometimes this is Deliberate Practice in the technical sense (quick feedback loops, building mental models that your brain can cache into chunks), but not always.

Learning

Sometimes you're learning *new information*, that is able to rapidly turn into a skill at a much faster rate than normal. You *could* try to learn math/violin/programming from first principles and practice, but a tutorial or a teacher who has carefully optimized their explanations can quickly give you entire new skills in a short timespan.

(Making the *best* use of them may require practice as well, but during the initial learning period you may be gaining huge returns)

Changing Your Environment and/or Incentives

The fastest route I've personally found to overall self improvement is changing environment – a new job, a new apartment, a new social network. These can radically change how you feel about yourself, or what is *easy*, or what behaviors are reinforced.

I know multiple people who had trouble focusing at work, got a job that they actually cared about, or which required the skills they enjoyed most, or had coworkers that provided subtle reinforcement in the right directions.

I briefly lived in an apartment building with a gym, and I found it way easier to exercise there than I did at other places. (Even homes where I got some exercise equipment)

It's possibly to reshape your environment on purpose (which I think yields improvement roughly on par with moderate skill practice), but the most powerful returns here have been sort of random and hard to control in my experience.

Discovering Something You Enjoy or are Well Suited For (Try Things)

People vary how much certain skills and activities resonate with them, and how fast they can improve. If you find something you really enjoy, you may be *much more* able to put in the hours to practice it, *or* you may gain skill much more rapidly than other other people.

I have a friend who would never have predicted they'd be good at dancing a priori, but who tried it on a whim and *it was amazing* and changed both their physical well being and social life.

Blockers/Cheat Codes

Simple things, that if you only knew to do them, would radically change your quality of life or capabilities.

Sometimes there's a concrete thing blocking you:

- You're not getting the right nutrition (or consuming food you're allergic to without realizing it).
- You turned out to be an extrovert (while thinking you were an introvert because you fit some stereotypes), and simply *bothering to* get more socialization has a huge impact on your mood.
- You were sick, or depressed, and taking some pills each day radically changes your capabilities.

Sometimes you're able to find a cheat code – instead of painstakingly dieting for months with little benefit, it turns out keto works well for your biochemistry. You're addicted to the internet, and you find out about SelfControl or Freedom.

Healing

This is sort of the intersection of "realizing you have a blocker", but where removing the blocker requires effort, skill or just time, rather than an immediate "oh, I just need to do X differently."

- Maybe you have trauma.
- Maybe you have some deep seated sense of "I'm not allowed to want X", and unravelling that sense of not-allowed requires skill at introspection, mindfulness, or finding people who can simply say in a confident voice that your system-1 believes "you're allowed to want X."
- Maybe you are just physically sick and need to recover.

Internal Alignment – Deep Improvements to Mental Architecture

Sort of like healing, maybe?

I haven't done this myself, but I've heard some people describe a process of *actually getting all the pieces of themselves* into alignment, where instead of fighting themselves ("I want cake!" "no I want to diet!") they reach a point where all their drives have a shared understanding of the world and are operating as a single agent.

There's minor versions of this that resolve alignment on one particular issue, and I'm told a deeper version exists where your entire self trusts itself to do the right thing, and then you no longer have to think in terms of willpower or energy expenditure (and other things that the sort of paradigm that cares about willpower or energy expenditure isn't even able to see)

What to do with this?

These are all things I'd thought about before, but I hadn't thought about them *all at once.* And I'm finding it shapes how I think people and communities oriented around self-improvement should direct their attention.

It matters what strategies you're even considering.

I don't *think* I'm currently have any specific blockers that need removing, and I can clearly see skills that I need to gain. But, people with blockers or in need of healing often don't see that they have them. And fixing them can radically change your landscape.

I'm currently seeing all the above strategies through a finance lens. There are things you can do that reliably output 1-3% return, year after year if you just stick with them. There are high risk / high return exploration moves you can make – discovering ways in which you need healing or are blocked or are missing cheat codes. *When* they pay off it feels like they have 50-300% returns in the space of a week, but actually finding the right ones takes time and amortized over the years it takes searching (and building prerequisite skills of self-awareness), it's probably more like a 5-10% rate.

I'm *not* a person who currently understands healing, and I've updated a bit that that's an important part of the overall paradigm that the rationality community doesn't tend to be that good at (although to be fair, I don't think most people are).

I think the original mythology of the rationality community is based around cheat codes – if we can munchkin our way through the lowest hanging fruit, we can win at life. But this doesn't actually happen *that* often, and meanwhile it can be demoralizing to be *expecting* bursts of 300% returns when in fact 2-3% should be your default assumption.

Book review: Pearl's Book of Why

This is a linkpost for http://www.bayesianinvestor.com/blog/index.php/2018/07/06/pearls-book-of-why/

Book review: The Book of Why, by Judea Pearl and Dana MacKenzie.

This book aims to turn the ideas from Pearl's seminal Causality) into something that's readable by a fairly wide audience.

It is somewhat successful. Most of the book is pretty readable, but parts of it still read like they were written for mathematicians.

History of science

A fair amount of the book covers the era (most of the 20th century) when statisticians and scientists mostly rejected causality as an appropriate subject for science. They mostly observed correlations, and carefully repeated the mantra "correlation does not imply causation".

Scientists kept wanting to at least hint at causal implications of their research, but statisticians rejected most attempts to make rigorous claims about causes.

The one exception was for randomized controlled trials (RCTs). Statisticians figured out early on that a good RCT can demonstrate that correlation does imply causation. So RCTs became increasingly important over much of the 20th century[1].

That created a weird tension, where the use of RCTs made it clear that scientists valued the concept of causality, but in most other contexts they tried to talk as if causality wasn't real. Not quite as definitely unreal as phlogiston. A bit closer to how behaviorists often tabooed the ideas that we had internal experiences and consciousness, or how linguists once banned debates on the origin of language, namely, that it was dangerous to think science could touch those topics. Or maybe a bit like heaven and hell - concepts which, even if they are useful, seem to be forever beyond the reach of science?

But scientists kept wanting to influence the world, rather than just predict it. So they often got impatient, when they couldn't afford to wait for RCTs, to act as if correlations told them something about causation.

The most conspicuous example is smoking. Scientists saw many hints that smoking caused cancer, but without an RCT[2], their standards and vocabulary made it hard to say more than that smoking is associated with cancer.

This eventually prompted experts to articulate criteria that seemed somewhat useful at establishing causality. But even in ideal circumstances, those criteria weren't convincing enough to produce a consensus. Authoritative claims about smoking and cancer were delayed for years by scientists' discomfort with talking about causality[3].

It took Pearl to describe how to formulate a unambiguous set of causal claims, and then say rigorous things about whether the evidence confirms or discredits the claims.

What went wrong?

The book presents some good hints about why the concept of causality was tabooed from science for much of the 20th century.

It focuses on the role of R.A. Fisher (also known as one of the main advocates of frequentism). Fisher was a zealot whose prestige was somewhat heavily based on his skill at quantifying uncertainty. In contrast, he didn't manage to quantify causality, or even figure out how to talk clearly about it. Pearl hints that this biased him against causal reasoning.

> path analysis requires scientific thinking, as does every exercise in causal inference. Statistics, as frequently practiced, discourages it, and encouraged "canned" procedures instead.

But blaming a few influential people seems to merely describe the tip of the iceberg. Why did scientists as a group follow Fisher's lead?

I suggest that the iceberg is better explained by what James C. Scott describes as high modernism and the desire for legibility.

I see a similar same pattern in the 20th century dominance of frequentism in most fields of science and the rejection of Bayesian approaches. Anything that required priors (whose source often couldn't be rigorously measured) was at odds with the goal of legibility.

The rise and fall of the taboo on causal inference coincide moderately well with the rise and fall of Soviet-style central planning, planned cities, and Taylorist factory management.

I also see some overlap with behaviorism, with its attempt to deny the importance of variables that were hard to measure, and its utopian hopes for how much its techniques could accomplish.

These patterns all seem to all be rooted in overconfident extrapolations of simple models of what caused progress. I don't think it's an accident that they all peaked near the middle of the 20th century, and were mostly discredited by the end of the century.

I remember that when I was young, I supported the standard inferences from the "correlation does not imply causation" mantra, and was briefly (and less clearly) tempted by the other manifestations of high modernism. Alas, I don't remember my reasons for doing so well enough to be of much use, other than a semi-appropriate respect for the authorities who were promoting those ideas.

An example of why causal reasoning matters

Here's an example that the book provides, dealing with non-randomized studies of a fictitious drug (to illustrate Simpson's Paradox, but also to show the difference between statistics and causal inference). The studies quantify three variables in each study:

- Study 1: drug <- gender -> heart attacks
- Study 2: drug -> blood pressure -> heart attacks

The book asks how we know we should treat the middle variables in those studies differently. The examples come with identical numbers, so that a statistics program which only sees correlations, and can't understand the causal arrows I've drawn here, would analyze both studies using the same methods. The numbers in these studies are chosen so that the aggregate data suggest an opposite conclusion about the drug from what we see if we stratify by gender or blood pressure. Standard statistics won't tell us which way of looking at data is more informative. But if we apply a little extra knowledge, it becomes clear that gender was a confounding variable that should be controlled for (it influenced who decided to take the drug), whereas blood pressure was a mediator that tells us how the drug works, and shouldn't be controlled for.

People typically don't find it hard to distinguish between the hypothesis that a drug caused a change in blood pressure and the hypothesis that a drug changed patients' reported gender. We all have a sufficiently sophisticated model of the world to assume the drug isn't changing patients' gender identity (i.e. we know that if that assumption were unexpectedly false, we'd hear about it).

Yet canned programs today are not designed to handle that, and it will be hard to fix programs so that they have the common sense needed to make those distinctions over a wide variety of domains.

Continuing Problems?

Pearl complains about scientists controlling for too many variables. The example described above helps explain why controlling for variables is often harmful, when it's not informed by a decent causal model. I have been mildly suspicious of the controlling for more variables is better attitude in the past, but this book clarified the problems well enough that I should be able to distinguish sensible from foolish attempts at controlling for variables.

Controlling for confounders seems like an area where science still has a long way to go before it can live up to Pearl's ideals.

There's also some lingering high modernism affecting the status of RCTs relative to other ways of inferring causality.

A sufficiently well-run RCT can at least create the appearance that everything important has been quantified. Sampling errors can be reliably quantified. Then the experimenter can sweep any systemic bias under the rug, and declare that the hypothesis formation step lies outside of science, or maybe deny that hypotheses matter (maybe they're just looking through all the evidence to see what pops out).

It looks to me like the peer review process still focuses too heavily on the easy-to-quantify and easy-to-verify steps in the scientific process (i.e. p-values). When RCTs aren't done, researchers too often focus on risk factors and associations, to equivocate about whether the research enlightens us about causality.

AI

The book points out that an AI will need to reason causally in order to reach human-level intelligence. It seems like that ought to be uncontroversial. I'm unsure whether it actually is uncontroversial.

But Pearl goes further, saying that the lack of causal reasoning in AIs has been "perhaps the biggest roadblock" to human-level intelligence.

I find that somewhat implausible. My intuition is that general-purpose causal inference won't be valuable in AIs until those AIs have world-models which are at least as sophisticated as crows[4], and that when that level is reached, we'll get rapid progress at incorporating causal inference into AI.

It's true that AI research often focuses on data mining (blind empiricism / model-free approaches), at the expense of approaches that could include causal inference. High modernist attitudes may well have hurt AI research in the past, and that may still be slowing AI research a bit. But Pearl exaggerates these effects.

To the extent that Pearl identifies tasks that AI can't yet tackle (e.g. "What kinds of solar systems are likely to harbor Earth-like planets?"), they need not just causal reasoning, but also the ability to integrate knowledge from a wide variety of data sources - and that means learning a much wider variety of concepts in a single system than AI researchers currently have the power to handle.

I expect that mainstream machine learning is mostly on track to handle that variety of concepts any decade now. I expect that until then, AI will only be able to do causal reasoning on toy problems, regardless of how well it understands causality.

Conclusion

Pearl is great at evaluating what constitutes clear thinking about causality. He's somewhat good at teaching us how to think clearly about novel causal problems, and rather unremarkable when he ventures outside the realm of causal inference.

Footnotes

[1] - RCTs (and p-values) don't seem to be popular in physics or geology. I'm curious why Pearl doesn't find this worth noting. I've mentioned before that people seem to care about statistical significance mainly where powerful interest groups might benefit from false conclusions.

[2] - The book claims that an RCT for smoking "would be neither feasible nor ethical". Clarke's first law applies here: it looks like about 8 studies had some sort of randomized interventions which altered smoking rates, including two studies focused solely on smoking interventions, which generated important reductions in smoking in the control group.

The RCTs seem to confirm that smoking causes health problems such as lung cancer and cardiovascular disease, but suggest that smoking shortens lifespan by a good deal less than the correlations would indicate.

[3] - As footnote 2 suggests, there have been some legitimate puzzles about the effects of smoking. Those sources of uncertainty have been obscured by the people who signal support for the "smoking is evil" view, and by smokers and tobacco companies who cling to delusions.

Smokers probably have some unhealthy habits and/or genes that contribute to cancer via causal pathways other than smoking.

The book notes that there is a "smoking gene" (rs16969968, aka Mr Big), but mostly it just means that smoking causes more harm for people with that gene.

Yet the book mostly implies that the anti-smoking crusaders were at least 90% right about the effects of smoking, when I think the reality is more complicated.

Pearl thinks quite rigorously when he's focused exclusively on causal inference, but outside that domain of expertise, he comes across as no more careful than an average scientist.

[4] - Pearl would have us believe that causal reasoning is mostly a recent human invention (in the last 50,000 years). I find Wikipedia's description of non-human causal reasoning to be more credible.

Mathematical Mindset

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

I agree that optimization amplifies things. I also agree that a mathematical mindset is important for AI alignment. I don't, however, think that a "mathematical mindset" is the same as a "proof mindset". Rather, I think that the latter is closer to being a "programming mindset" -- or, indeed, a "security mindset". And that a "mathematical mindset" is largely missing from AI-alignment discourse at present.

Whereas others see a division between two clusters, of the form,

science/physics

vs.

mathematics/programming/logic

I, by contrast, see a hierarchical progression that looks something like:

science < programming < physics < mathematics < logic <...

where, in this context, these words have meanings along the following lines:

science: things being made of parts; decomposition

programming: things being made of moving parts; constant-velocity motion; causal networks

physics: things being made of moving spatial parts; accelerated motion, rotation, fluidity; substance

mathematics: models being made of parts; transubstantiation; metaphysics; theorization

logic: concepts being made of parts; time reversal; ontology

Of course I'm not using these words standardly here. One reason for this is that, in this discussion, no one is: we're talking about *mindsets*, not about sociological disciplines or even clusters of particular ideas or "results".

But the really important reason I'm not following standard usage is because I'm not trying to invoke standard concepts; instead, I'm trying to invent "the right" concepts. Consequently, I can't just use standard language, because standard language implies a model of the world different from the one that I want to use.

It is commonly believed that if you want to introduce a new concept that is similar or related (but--of course--nonidentical) to an old concept, you shouldn't use the same word for the new concept and the old, because that would be "confusing". I wish to explicitly disagree with this belief.

This view presupposes that *actively shifting between models of the world* is not in our repertory of mental operations. But I specifically want it to be!

In fact, I claim that *this*, and not proof, is what a "mathematical mindset" is really about.

For mathematics is not about proofs; it is about *definitions*. The essence of great mathematics is coming up with a powerful definition that results in short proofs.

What makes a definition "powerful" is that it reflects a *conceptual upgrade* -- as distinct from mere conceptual analysis. We're not just trying to figure out what we mean; we're trying to figure out what we *should* mean.

A mathematical definition is what the answer to a philosophical problem looks like. An example I particularly like is the definition of a topological space. I don't know for a fact that this is what people "really meant" when they pondered the nature of "space" during all the centuries before Felix Hausdorff came up with this definition; it doesn't matter, because the power of this definition shows that it is what they should have meant.

(And for that reason, I'm comfortable saying that it *is* what they "meant" -- acknowledging that this is a linguistic fiction, but using it anyway.)

Notably, mathematical definitions are often redefinitions: they take a term already in use and define it in a new way. And, notably, the new definition often bears scant resemblance to the old, let alone any “intuitive” meaning of the term -- despite presenting itself as a successor. This is not a bug. It is what philosophical progress -- theoretical progress, progress in understanding -- looks like. The relationship between the new and the old definitions is explained not in the definitions themselves, but in the relationship between the theories of which the definitions are part.

Having a “mathematical mindset” means being comfortable with words being redefined. This is because it means being comfortable with models being upgraded -- in particular, with models being related and compared to each other: the activity of theorization.

It occurs to me, now that I think about it, that the term “theorization” is not very often used in the AI-alignment and rationalist communities, compared with what one might expect given the degree of interest in epistemology. My suspicion is that this reflects an insufficiency of comfort with the idea of *models* (as opposed to *things*) being made of *parts* (in particular, being made of parameters), such that they are relatable to and transformable into each other.

This idea is, approximately, what I am calling “mathematical mindset”. It stands in contrast to what others are calling “mathematical mindset”, which has to do with proofs. The relationship, however, is this: both of them reflect an interest in understanding what is going on, as opposed to merely being able to describe what is going on.

Paul's research agenda FAQ

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

I think Paul Christiano’s research agenda for the alignment of superintelligent AGIs presents one of the most exciting and promising approaches to AI safety. After being very confused about Paul’s agenda, chatting with others about similar confusions, and clarifying with Paul many times over, I’ve decided to write a FAQ addressing common confusions around his agenda.

This FAQ is not intended to provide an introduction to Paul’s agenda, nor is it intended to provide an airtight defense. This FAQ only aims to clarify commonly misunderstood aspects of the agenda. Unless otherwise stated, all views are my own views of Paul’s views. (**ETA**: Paul does not have major disagreements with anything expressed in this FAQ. There are many small points he might have expressed differently, but he endorses this as a reasonable representation of his views. This is in contrast with previous drafts of this FAQ, which *did* contain serious errors he asked to have corrected.)

For an introduction to Paul’s agenda, I’d recommend Ajeya Cotra’s summary. For good prior discussion of his agenda, I’d recommend Eliezer’s thoughts, Jessica Taylor’s thoughts (here and here), some posts and discussions on LessWrong, and Wei Dai’s comments on Paul’s blog. For most of Paul’s writings about his agenda, visit ai-alignment.com.

0. Goals and non-goals

0.1: What is this agenda trying to accomplish?

Enable humans to build arbitrarily powerful AGI assistants that are competitive with unaligned AGI alternatives, and only try to help their operators (and in particular, never attempt to kill or manipulate them).

People often conceive of safe AGIs as silver bullets that will robustly solve every problem that humans care about. This agenda *is not about building a silver bullet*, it’s about building a tool that will *safely and substantially assist its operators.* For example, this agenda does not aim to create assistants that can do any of the following:

- They can prevent nuclear wars from happening
- They can prevent evil dictatorships
- They can make centuries’ worth of philosophical progress
- They can effectively negotiate with distant superintelligences
- They can solve the value specification problem

On the other hand, to the extent that humans care about these things and could make them happen, this agenda lets us build AGI assistants that can substantially assist humans achieve these things. For example, a team of 1,000 competent humans working together for 10 years could make substantial progress on preventing nuclear wars or solving metaphilosophy. Unfortunately, it’s slow and expensive to assemble a team like this, but an AGI assistant might enable us to reap similar benefits in far less time and at much lower cost.

(See Clarifying "AI Alignment" and Directions and desiderata for AI alignment.)

0.2: What are examples of ways in which you imagine these AGI assistants getting used?

Two countries end up in an AGI arms race. Both countries are aware of the existential threats that AGIs pose, but also don’t want to limit the power of their AIs. They build AGIs according to this agenda, which stay under the operators' control. These AGIs then help the operators broker an international treaty, which ushers in an era of peace and stability. During this era, foundational AI safety problems (e.g. those in MIRI's research agenda) are solved in earnest, and a provably safe recursively self-improving AI is built.

A more pessimistic scenario is that the countries wage war, and the side with the more powerful AGI achieves a decisive victory and establishes a world government. This scenario isn't as good, but it at least leaves humans in control (instead of extinct).

The most pressing problem in AI strategy is how to stop an AGI race to the bottom from killing us all. Paul's agenda aims to solve this specific aspect of the problem. That isn't an existential win, but it does represent a substantial improvement over the status quo.

(See section "2. Competitive" in Directions and desiderata for AI alignment.)

0.3: But this might lead to a world dictatorship! Or a world run by philosophically incompetent humans who fail to capture most of the possible value in our universe! Or some other dystopia!

Sure, maybe. But that's still better than a paperclip maximizer killing us all.

There *is* a social/political/philosophical question about how to get humans in a post-AGI world to claim a majority of our cosmic endowment (including, among other things, not establishing a tyrannical dictatorship under which intellectual progress halts). While technical AI safety does make progress on this question, it's a broader question overall that invites fairly different angles of attack (e.g. policy interventions and social influence). And, while this question *is* extremely important, it is a *separate question* from how you can build arbitrarily powerful AGIs that stay under their operators' control, which is the only question this agenda is trying to answer.

1. Alignment

1.1 How do we get alignment at all?

("Alignment" is an imprecise term meaning "nice" / "not subversive" / "trying to actually help its operator". See Clarifying "AI alignment" for Paul's description.)

1.1.1: Isn't it really hard to give an AI our values? Value learning is really hard, and the default is for it to encounter instrumental incentives to manipulate you or prevent itself from getting shut down.

The AI isn't learning our values, it's learning to optimize for our short-term approval—in other words, for each action it takes, it optimizes for something like what rating we'd give it on a scale from 1 to 5 if we just saw it act.

It's hard to learn the nuances of human values. But from a machine learning perspective, it's very easy to learn that humans would strongly disapprove of attempts to kill or manipulate us. Paul expresses this well on his blog:

You need only the vaguest understanding of humans to guess that killing the user is: (1) not something they would approve of, (2) not something they would do, (3) not in line with their instrumental preferences.

So in order to get bad outcomes here you have to really mess up your model of what humans want (or more likely mess up the underlying framework in an important way). If we imagine a landscape of possible interpretations of human preferences, there is a "right" interpretation that we are shooting for. But if you start with a wrong answer that is anywhere *in the neighborhood, you will do things like “ask the user what to do, and don’t manipulate them.” And these behaviors will eventually get you where you want to go.*

(See: Approval-directed agents and Act-based agents.)

1.1.2: OK, but doesn't this only incentivize it to appear like it's doing what the operator wants? Couldn’t it optimize for hijacking its reward signal, while seeming to act in ways that humans are happy with?

We’re not just training the agent to take good actions. We’re also training it to comprehensibly answer questions about why it took the actions it took, to arbitrary levels of detail. (Imagine a meticulous boss grilling an employee about a report he put together, or a tax auditor grilling a corporation about the minutiae of its expenses.) We ensure alignment by randomly performing thorough evaluations of its justifications for its actions, and punishing it severely if any of those justifications seem subversive. To the extent we trust these justifications to accurately reflect the agent’s cognition, we can trust the agent to not act subversively (and thus be aligned).

(See: The informed oversight problem.)

1.1.3: How do we train it to answer questions comprehensively?

Reward it for doing so, and punish it for failing to do so.

Imagine being a meticulous boss who asks his employee to put together a report. Imagine grilling him about the report, and punishing him every time he fails to answer your questions clearly enough or at a satisfactory level of detail, in addition to punishing him for low-quality reports. If your employee is capable enough, he’ll eventually learn to produce high-quality reports and answer questions satisfactorily when you grill him.

Note that the ability to generate comprehensible descriptions of arbitrary cognition is a major open problem in Paul’s agenda, and also a major problem in AI safety broadly. It’s the part of his agenda that Paul is currently focusing most on.

(See The informed oversight problem, Learning representations, and Approval-maximising representations.)

1.1.4: Why should we expect the agent’s answers to correspond to its cognition at all?

We don’t actually have any guarantees that it does, but giving honest answers is probably the easiest way for the agent to maximize its reward. The only feasible alternative is giving thorough dishonest explanations for its actions. But if an employee lies about how he arrived at his conclusions, his explanations won’t check out, and he might get punished severely. And if a corporation lies about how it spent its money, the numbers won’t add up, and its auditor might punish it severely.

It’s *critically important* for this scheme that the overseer is capable of evaluating the assistant’s justifications extremely thoroughly, and does so often enough. Corporations cook their books in proportion to how thorough their auditors are, how common audits are, and how bad it is to get caught by their auditors. If we audit thoroughly enough and frequently enough, and punish our assistants severely enough for subversion, we should expect them to answer honestly.

1.1.5: Wouldn’t this incentivize the assistant to produce justifications that seem fine and innocent to you, but may actually be harmful?

We should expect the AI not to *adversarially* manipulate us—by assumption, we’re evaluating its justifications thoroughly enough that we should be able to catch any subversive cognition.

It's possible that there are free variables in what sorts of cognition the operator deems acceptable, and that a *non-adversarial* optimization process might be able to persuade the operator of essentially arbitrary conclusions by manipulating these free variables in ways the operators approves of. For example, an AGI assistant might accidentally persuade you to become an ISIS suicide bomber, while only thinking in ways that you approve of.

I do think this is a potentially severe problem. But I don’t consider it a dealbreaker, for a number of reasons:

- An AGI assistant “accidentally” manipulating you is no different from a very smart and capable human assistant who, in the process of assisting you, causes you to believe drastic and surprising conclusions. Even if this might lead to bad outcomes, Paul isn’t aiming for his agenda to prevent this class of bad outcomes.
- The more rational you are, the smaller the space of conclusions you can be non-adversarially led into believing. (For example, it’s very hard for me to imagine myself getting persuaded into becoming an ISIS suicide bomber by a process whose cognition I approve of.) It might be that some humans have passed a rationality threshold, such that they only end up believing correct conclusions after thinking for a long time without adversarial pressures.

1.2 Amplifying and distilling alignment

1.2.1: OK, you propose that to amplify some aligned agent, you just run it for a lot longer, or run way more of them and have them work together. I can buy that our initial agent is aligned; why should I trust their aggregate to be aligned?

When aligned agents work together, there’s often emergent behavior that can be described as non-aligned. For example, if the operator is pursuing a goal (like increasing Youtube’s revenue), one group of agents proposes a subgoal (like increasing Youtube views), and another group competently pursues that subgoal without understanding how it relates to the top-level goal (e.g. by triple-counting all the views), you end up with misaligned optimization. As another example, there might be some input (e.g. some weirdly compelling argument) that causes some group of aligned agents to “go insane” and behave unpredictably, or optimize for something against the operator’s wishes.

Two approaches that Paul considers important for preserving alignment:

- Reliability amplification—aggregating agents that can answer a question correctly some of the time (say, 80% of the time) in a way that they can answer questions correctly with arbitrarily high probability.
- Security amplification—winnowing down the set of queries that, when fed to the aggregate, causes the aggregate to “go insane”.

It remains an open question in Paul’s agenda how alignment can be robustly preserved through capability amplification—in other words, how to increase the capabilities of aligned agents without introducing misaligned behavior.

(See: Capability amplification, Reliability amplification, Security amplification, Universality and security amplification, and Two guarantees.)

1.2.2: OK, so given this amplified aligned agent, how do you get the distilled agent?

Train a *new agent* via some combination of imitation learning (predicting the actions of the amplified aligned agent), semi-supervised reinforcement learning (where the amplified aligned agent helps specify the reward), and techniques for optimizing robustness (e.g. creating red teams that generate scenarios that incentivize subversion).

(See: RL+Imitation, Benign model-free RL, Semi-supervised reinforcement learning, and Techniques for optimisizing worst-case performance.)

1.2.3: It seems like imitation learning might cause a lot of minutiae to get lost, and would create something that's "mostly aligned" but actually not aligned in a bunch of subtle ways. Maybe this is tolerable for one round of iteration, but after 100 rounds, I wouldn’t feel very good about the alignment of the resulting agent...

Indeed, which is why this new agent is also trained with semi-supervised reinforcement learning, and optimized for robustness. The imitation learning is more about getting this new agent off the ground than about ensuring alignment. The bulk of the alignment guarantee comes from the semi-supervised reinforcement learning, where we train it to work on a wide range of tasks and answer questions about its cognition. We trust it to be aligned because we train it the same way we trained our original agent to be aligned—namely, by rewarding it for maximizing our approval, and punishing it for thinking subversive thoughts.

(There’s a common misconception that Paul’s agenda proceeds *entirely* by taking a single aligned assistant, running it for a lot longer, using imitation learning to distill that into a new assistant, running that new assistant a lot longer, using imitation learning to distill that, etc., much like how AlphaGo Zero improves its policy network. While this does capture a key idea behind how the assistant’s capabilities get amplified, it omits critical components of his agenda that ensure alignment, like inspecting cognition and amplifying robustness.)

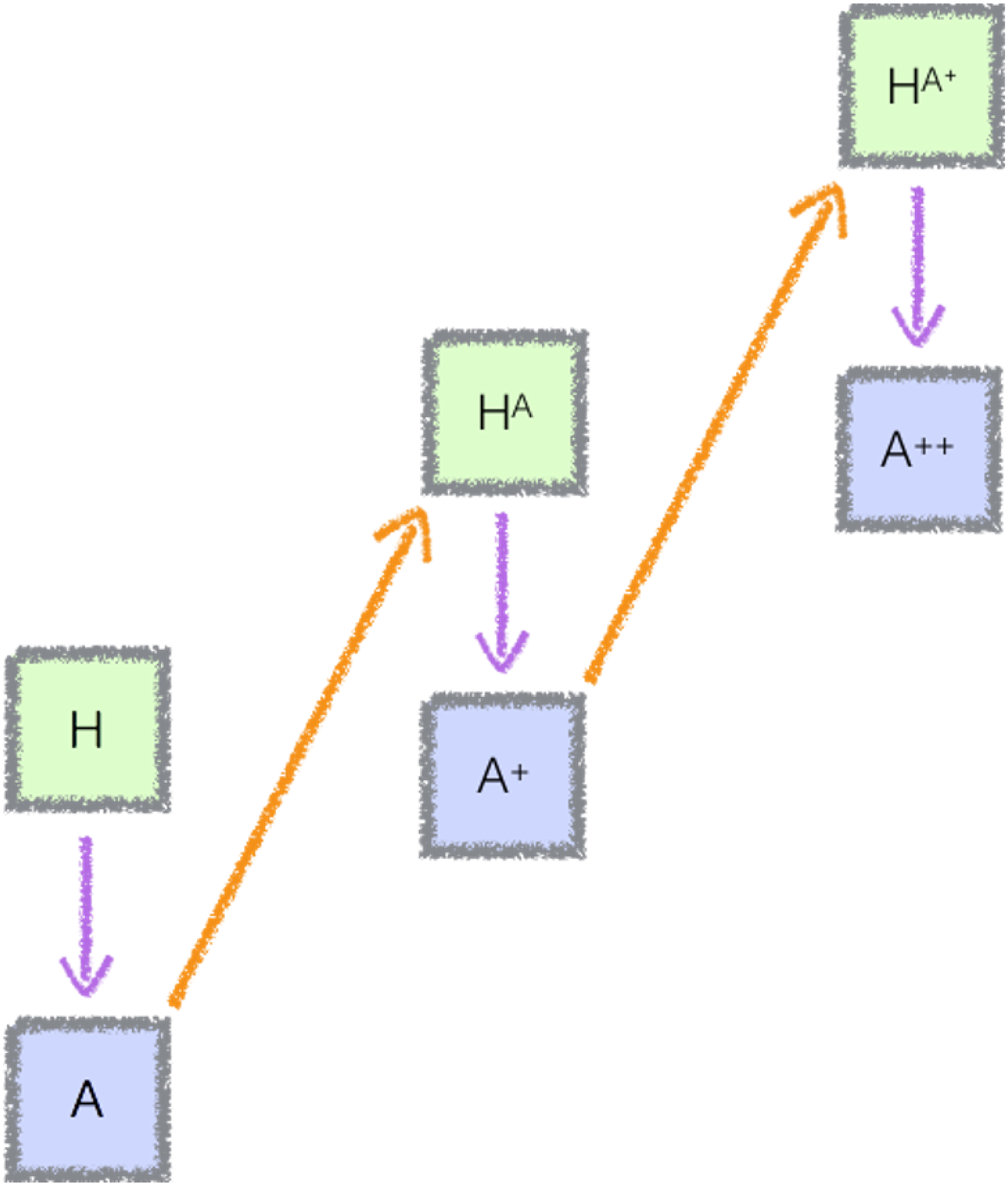
1.2.4: This distilled agent’s cognition might be much more complex than that of our previous agent. How can we reasonably determine whether we’d approve of its cognition?

This is where our assistants come in. While the human operator does inspect some aspects of the distilled agent’s cognition, most of the inspection is done by our assistants, whose judgments we trust because we trust that they’re aligned.

(See The reward engineering problem.)

1.2.5: Given that this distilled agent is way more powerful than you and your aligned agents, couldn’t it manipulate you and your team of aligned agents?

If we’re amplifying and distilling agent A, we’re actually training our newly distilled agent to be stronger than A, but *much* weaker than the amplified version of A. The amplification/distillation procedure can be thought of as repeatedly taking two steps forward (in amplifying A) and one step back (in distilling a new agent).



Any subversive plan the distilled agent makes could get recognized as subversive by the amplified agent, because the amplified agent has enough power to thoroughly evaluate every aspect of the cognition that went into the plan. The amplified agent rarely performs thorough evaluations of the distilled agent’s cognition, because it’s extremely expensive, but does so randomly and punishes the distilled agent extremely harshly if it detects subversion. The distilled agent, knowing that it will have to thoroughly justify its actions to a competent supervisor, is thus strongly incentivized against acting subversively.

In other words, the amplified agent randomly “audits” the distilled agent, and punishes the distilled agent very harshly if it fails the audit. Though the distilled agent knows that it might be able to deceive its supervisor when it isn’t audited, it’s so scared of the outcome where it tries to do that and gets audited that it doesn't even want to try. (Even if you were 99% confident that you could get away with tax evasion, you wouldn't want to try if you knew the government tortures and murders the families of the tax evaders they catch.)

(See Benign model-free RL, the comments in The informed oversight problem, and William Saunders' post Improbable Oversight, An Attempt at Informed Oversight.)

1.3 Robust alignment / corrigibility

1.3.1: If you delegate oversight to a well-meaning assistant ("hey, can you make sure things are going as planned?"), who delegates oversight to a well-meaning assistant, who delegates oversight to a well-meaning assistant, etc., then the default is for the delegatee 100 layers down to end up with some extremely distorted version of what you care about. Shouldn't we expect this distortion to happen with the agents we produce?

Yes, which is why alignment isn't the property we're ultimately aiming to inductively guarantee. The property we're trying to guarantee is something closer to "alignment + extreme caution about whether it's aligned and cautious enough". Paul usually refers to this as corrigibility.

This way, every time we're training a distilled agent, we train it to want to clarify with its overseer (i.e., us assisted with a team of corrigible assistants) whenever it's uncertain about what we would approve of. Our corrigible assistants either answer the question confidently, or clarify with us if it's uncertain about its answer.

1.3.2: OK, but doesn't this just pass the buck onto corrigibility? In the same way that alignment can get distorted, couldn't corrigibility also get distorted?

Paul has two strong intuitions about corrigibility:

1. corrigibility has a simple core that should be easy for humans to impart into powerful AI systems
2. corrigibility is robust—agents that are somewhat corrigible want to become more corrigible.

On this view, even if human approval is high-complexity and hard to get right, corrigibility is straightforward to learn and most likely to improve over distillation rather than worsen.

You also need robustness amplification to ensure that corrigibility is maintained. For example, a corrigible supervisor needs to be able to thoroughly evaluate a distilled agent's cognition, but if its evaluations aren't reliable, it can't be trusted to ensure that the distilled agent is corrigible (and not a daemon). And a corrigible agent that "goes insane" doesn't stay corrigible.

(See: Corrigibility.)

1.3.3: I don't share those intuitions around corrigibility. Do you have any intuition pumps?

One intuition pump: corrigibility can be thought of as extreme caution about whether you're actually being helpful, and extreme caution is robust—if you're extremely cautious about how things can go wrong, you *want* to know more ways things can go wrong and you *want* to improve your ability to spot how things are going wrong, which will lead you to become more cautious.

Another intuition pump: I have some intuitive concept of "epistemically corrigible humans". Some things that gesture at this concept:

- They care deeply about finding the truth, and improving their skill at finding the truth.
- They're aware that they're flawed reasoners, with biases and blind spots, and actively seek out ways to notice and remove these flaws. They try to take ideas seriously, no matter how weird they seem.
- Their beliefs tend to become more true over time.
- Their skill at having true beliefs improves over time.
- They tend to reach similar conclusions in the limit (namely, the correct ones), even if they're extremely weird and not broadly accepted.

I think of corrigible assistants as being corrigible in the above way, except optimizing for helping its operator instead of finding the truth. Importantly, so long as an agent crosses some threshold of corrigibility, they will *want* to become more and more cautious about whether they're helpful, which is where robustness comes from.

Given that corrigibility seems like a property that any reasoner could have (and not just humans), it's probably not too complicated a concept for a powerful AI system to learn, especially given that many humans seem able to learn some version of it.

1.3.4: This corrigibility thing still seems really fishy. It feels like you just gave some clever arguments about something very fuzzy and handwavy, and I don't feel comfortable trusting that.

While Paul thinks there's a good intuitive case for something like corrigibility, he also considers getting a deeper conceptual understanding of corrigibility one of the most important research directions for his agenda. He agrees it's possible that corrigibility may not be safely learnable, or not actually robust, in which case he'd feel way more pessimistic about his entire agenda.

2. Usefulness

2.1. Can the system be both safe and useful?

2.1.1: A lot of my values and knowledge are implicit. Why should I trust my assistant to be able to learn my values well enough to assist me?

Imagine a question-answering system trained on all the data on Wikipedia, that ends up with comprehensive, gears-level world-models, which it can use to synthesize existing information to answer novel questions about social interactions or what our physical world is like. (Think Wolfram|Alpha, but *much* better.)

This system is something like a proto-AGI. We can easily restrict it (for example by limiting how long it gets to reflect when it answers questions) so that we can train it to be corrigible while trusting that it's too limited to do anything dangerous that the overseer couldn't recognize as dangerous. We use such a restricted system to start off the iterated distillation and amplification process, and bootstrap it to get systems of arbitrarily high capabilities.

(See: Automated assistants)

2.1.2: OK, sure, but it'll essentially still be an alien and get lots of minutiae about our values wrong.

How bad is it really if it gets minutiae wrong, as long as it doesn't cause major catastrophes? Major catastrophes (like nuclear wars) are pretty obvious, and we would obviously disapprove of actions that lead us to catastrophe. So long as it learns to avoid those (which it will, if we give it the right training data), we're fine.

Also keep in mind that we're training it to be corrigible, which means it'll be very cautious about what sorts of things we'd consider catastrophic, and try very hard to avoid them.

2.1.3: But it might make lots of subtle mistakes that add up to something catastrophic!

And so might we. Maybe there are some classes of subtle mistakes the AI will be more prone to than we are, but there are probably also classes of subtle mistakes we'll be more prone to than the AI. We're only shooting for our assistant to avoid *trying* to lead us to a catastrophic outcome.

(See: Techniques for optimizing worst-case performance.)

2.1.4: I'm really not sold that training it to avoid catastrophes and training it to be corrigible will be good enough.

This is actually more a capabilities question (is our system good enough at trying very hard to avoid catastrophes to actually avoid a catastrophe?) than an alignment question. A major open question in Paul's agenda is how we can formalize performance guarantees well enough to state actual worst-case guarantees.

(See: Two guarantees and Techniques for optimizing worst-case performance)

2.2. Universality

2.2.1. What sorts of cognition will our assistants be able to perform?

We should roughly expect it to think in ways that would be approved by an HCH (short for "human consulting HCH"). To describe HCHs, let me start by describing a *weak HCH*:

Consider a human Hugh who has access to a question-answering machine. Suppose the machine answers question Q by perfectly imitating how Hugh would answer question Q, if Hugh had access to the question-answering machine.

That is, Hugh is able to consult a copy of Hugh, who is able to consult a copy of Hugh, who is able to consult a copy of Hugh…

I sometimes picture this as an infinite tree of humans-in-boxes, who can break down questions and pass them to other humans-in-boxes (who can break down those questions and pass them along to other humans-in-boxes, etc.) and get back answers instantaneously. A few remarks:

- This formalism tries to capture some notion of "what would H think about some topic if H thought about it for arbitrarily long amounts of time"? For example, H might make partial progress on some question, and then share this progress with some other H and ask it to make more progress, who might do the same.
- A weak HCH could simulate the cognitive labor of an economy the size of the US economy. After all, a weak HCH can emulate a single human thinking for a long time, so it can emulate teams of humans thinking for a long time, and thus teams of teams of humans thinking for a long time, etc. If you imagine a corporation as teams of teams of teams of humans performing cognitive labor, you get that a weak HCH can emulate the output of an arbitrary corporation, and thus collections of arbitrary corporations communicating with one another.
- Many tasks that don't intuitively seem like they can be broken down, can in fact be fairly substantially broken down. For example, making progress on difficult math problems seems difficult to break down. But you *could* break down progress on a math problem into something like (think for a while about possible angles of attack) + (try each angle of attack, and recurse on the new math problem). And (think for a while

about possible angles of attack) can be reduced into (look at features of this problem and see if you've solved anything similar), which can be reduced into focusing on specific features, and so on.

Strong HCH, or just HCH, is a variant of weak HCHs where the agents-in-boxes are able to communicate with each other directly, and read and write to some shared external memory, in addition to being able to ask, answer, and break down questions. Note that they would be able to implement arbitrary Turing machines this way, and thus avoid any limits on cognition imposed by the structure of weak HCH.

(Note: most people think "HCH" refers to "weak HCH", but whenever Paul mentions HCHs, he now refers to strong HCHs.)

The exact relationship between HCH and the agents produced through iterated amplification and distillation is confusing and very commonly misunderstood:

- HCHs should *not* be visualized as having humans in the box. They should be thought of as having some corrigible assistant inside the box, much like the question-answering system described in 2.1.1.
- Throughout the iterated amplification and distillation process, there is never any agent whose cognition *resembles* an HCH of the corrigible assistant. In particular, agents produced via distillation are general RL agents with no HCH-like constraints on their cognition. The closest resemblance to HCH appears during amplification, during which a superagent (formed out of copies of the agent getting amplified) performs tasks by breaking them down and distributing them among the agent copies.

(As of the time of this writing, I am still confused about the sense in which the agent's cognition is approved by an HCH, and what that means about the agent's capabilities.)

(See: Humans consulting HCH and Strong HCH.)

2.2.2. Why should I think the HCH of some simple question-answering AI assistant can perform arbitrarily complex cognition?

All difficult and creative insights stem from chains of smaller and easier insights. So long as our first AI assistant is a universal reasoner (i.e., it can implement arbitrary Turing machines via reflection), it should be able to realize arbitrarily complex things if it reflects for long enough. For illustration, Paul thinks that chimps aren't universal reasoners, and that most humans past some intelligence threshold are universal.

If this seems counterintuitive, I'd claim it's because we have poor intuitions around what's achievable with 2,000,000,000 years of reflection. For example, it might seem that an IQ 120 person, knowing no math beyond arithmetic, would simply be unable to prove Fermat's last theorem given arbitrary amounts of time. But if you buy that:

- An IQ 180 person could, in 2,000 years, prove Fermat's last theorem knowing nothing but arithmetic (which seems feasible, given that most mathematical progress was made by people with IQs under 180)
- An IQ 160 person could, in 100 years, make the intellectual progress an IQ 180 person could in 1 year
- An IQ 140 person could, in 100 years, make the intellectual progress an IQ 160 person could in 1 year
- An IQ 120 person could, in 100 years, make the intellectual progress an IQ 140 person could in 1 year

Then it follows that an IQ 120 person could prove Fermat's last theorem in 2,000*100*100*100 = 2,000,000,000 years' worth of reflection.

(See: Of humans and universality thresholds.)

2.2.3. Different reasoners can reason in very different ways and reach very different conclusions. Why should I expect my amplified assistant to reason anything like me, or reach conclusions that I'd have reached?

You shouldn't expect it to reason anything like you, you shouldn't expect it to reach the conclusions you'd reach, and you shouldn't expect it to realize everything you'd consider obvious (just like you wouldn't realize everything it would consider obvious). You *should* expect it to reason in ways you approve of, which should constrain its reasoning to be sensible and competent, as far as you can tell.

The goal isn't to have an assistant that can think like you or realize everything you'd realize. The goal is to have an assistant who can think in ways that you consider safe and substantially helpful.

2.2.4. HCH seems to depend critically on being able to break down arbitrary tasks into subtasks. I don't understand how you can break down tasks that are largely intuitive or perceptual, like playing Go very well, or recognizing images.

Go is actually fairly straightforward: an HCH can just perform an exponential tree search. Iterated amplification and distillation applied to Go is not actually that different from how AlphaZero trains to play Go.

Image recognition is harder, but to the extent that humans have clear concepts of visual features they can reference within images, the HCH should be able to focus on those features. The cat vs. dog debate in Geoffrey Irving's approach to AI safety via debate gives some illustration of this.

Things get particularly tricky when humans are faced with a task they have little explicit knowledge about, like translating sentences between languages. Paul did mention something like "at some point, you'll probably just have to stick with relying on some brute statistical regularity, and just use the heuristic that X commonly leads to Y, without being able to break it down further".

(See: Wei Dai's comment on Can Corrigibility be Learned Safely, and Paul's response to a different comment by Wei Dai on the topic.)

2.2.5: What about tasks that require significant accumulation of knowledge? For example, how would the HCH of a human who doesn't know calculus figure out how to build a rocket?

This sounds difficult for weak HCHs on their own to overcome, but possible for strong HCHs to overcome. The accumulated knowledge would be represented in the strong HCHs shared external memory, and the humans essentially act as "workers" implementing a higher-level cognitive system, much like ants in an ant colony. (I'm still somewhat confused about what the details of this would entail, and am interested in seeing a more fleshed out implementation.)

2.2.6: It seems like this capacity to break tasks into subtasks is pretty subtle. How does the AI learn to do this? And how do we find human operators (besides Paul) who are capable of doing this?

Ought is gathering empirical data about task decomposition. If that proves successful, Ought will have numerous publicly available examples of humans breaking down tasks.

3. State of the agenda

3.1: What are the current major open problems in Paul's agenda?

The most important open problems in Paul's agenda, according to Paul:

- *Worst-case guarantees*: How can we make worst-case guarantees about the reliability and security of our assistants? For example, how can we ensure our oversight is reliable enough to prevent the creation of subversive subagents (a.k.a. daemons) in the distillation process that cause our overall agent to be subversive?
- *Transparent cognition*: How can we extract useful information from ML systems' cognition? (E.g. what concepts are represented in them, what logical facts are embedded in them, and what statistical regularities about the data it captures.)
- *Formalizing corrigibility*: Can we formalize corrigibility to the point that we can create agents that are knowably robustly corrigible? For example, could we formalize corrigibility, use that formalization to prove the existence of a broad basin of corrigibility, and then prove that ML systems past some low threshold will land and stay in this basin?
- *Aligned capability amplification*: Can we perform amplification in a way that doesn't introduce alignment failures? In particular, can we safely decompose every task we care about without effectively implementing an aligned AGI built out of human transistors?

(See: Two guarantees, The informed oversight problem, Corrigibility, and the "Low Bandwidth Overseer" section of William Saunder's post Understanding Iterated Distillation and Amplification: Claims and Oversight.)

3.2: How close to completion is Paul's research agenda?

Not very close. For all we know, these problems might be extraordinarily difficult. For example, a subproblem of "transparent cognition" is "how can humans understand what goes on inside neural nets", which is a broad open question in ML. Subproblems of "worst-case guarantees" include ensuring that ML systems are robust to distributional shift and adversarial inputs, which are also broad open questions in ML, and which might require substantial progress on MIRI-style research to articulate and prove formal bounds. And getting a formalization of corrigibility might require formalizing aspects of good reasoning (like calibration about uncertainty), which might in turn require substantial progress on MIRI-style research.

I think people commonly conflate "Paul has a safety agenda he feels optimistic about" with "Paul thinks he has a solution to AI alignment". Paul in fact feels optimistic about these problems getting solved *well enough* for his agenda to work, but does *not* consider his research agenda anything close to complete.

(See: Universality and security amplification, search "MIRI")

Thanks to Paul Christiano, Ryan Carey, David Krueger, Rohin Shah, Eric Rogstad, and Eli Tyre for helpful suggestions and feedback.

Melatonin: Much More Than You Wanted To Know

[I am not a sleep specialist. Please consult with one before making any drastic changes or trying to treat anything serious.]

Van Geijlswijk et al describe supplemental melatonin as “a chronobiotic drug with hypnotic properties”. Using it as a pure hypnotic – a sleeping pill – is like using an AK-47 as a club to bash your enemies’ heads in. It might work, but you’re failing to appreciate the full power and subtlety available to you.

Melatonin is a neurohormone produced by the pineal gland. In a normal circadian cycle, it’s lowest (undetectable, less than 1 pg/ml of blood) around the time you wake up, and stays low throughout the day. Around fifteen hours after waking, your melatonin suddenly shoots up to 10 pg/ml – a process called “dim light melatonin onset”. For the next few hours, melatonin continues to increase, maybe as high as 60 or 70 pg/ml, making you sleepier and sleepier, and presumably at some point you go to bed. Melatonin peaks around 3 AM, then declines until it’s undetectably low again around early morning.

Is this what makes you sleepy? Yes and no. Sleepiness is a combination of the circadian cycle and the so-called “Process S”. This is an unnecessarily sinister-sounding name for the fact that the longer you’ve been awake, the sleepier you’ll be. It seems to be partly regulated by a molecule called adenosine. While you’re awake, the body produces adenosine, which makes you tired; as you sleep, the body clears adenosine away, making you feel well-rested again.

In healthy people these processes work together. Circadian rhythm tells you to feel sleepy at night and awake during the day. Process S tells you to feel awake when you’ve just risen from sleep (naturally the morning), and tired when you haven’t slept in a long time (naturally the night). Both processes agree that you should feel awake during the day and tired at night, so you do.

When these processes disagree for some reason – night shifts, jet lag, drugs, genetics, playing *Civilization* until 5 AM – the system fails. One process tells you to go to sleep, the other to wake up. You’re never quite awake enough to feel energized, or quite tired enough to get restful sleep. You find yourself lying in bed tossing and turning, or waking up while it’s still dark and not being able to get back to sleep.

Melatonin works on both systems. It has a weak “hypnotic” effect on Process S, making you immediately sleepier when you take it. It also has a stronger “chronobiotic” effect on the circadian rhythm, shifting what time of day your body considers sleep to be a good idea. Effective use of melatonin comes from understanding both these effects and using each where appropriate.

1. Is melatonin an effective hypnotic?

Yes.

That is, taking melatonin just before you want to get to sleep, does help you get to sleep. The evidence on this is pretty unanimous. For primary insomnia, two meta-analyses – one by Brzezinski in 2005 and another by Ferracioli-Oda in 2013 – both find it safe and effective. For jet lag, a meta-analysis by the usually-skeptical Cochrane Collaboration pronounces melatonin “remarkably effective”. For a wide range of primary and secondary sleep disorders, Buscemi et al say in their abstract that it doesn’t work, but a quick glance at the study shows it absolutely does and they are incorrectly under-reporting their own results. The Psychiatric Times agrees with me on this: “Results from another study reported as negative actually demonstrated a statistically significant positive result of a decrease in sleep latency by an average of 7.2 minutes for melatonin”.

Expert consensus generally follows the meta-analyses: melatonin works. I find cautious endorsements by the Mayo Clinic and John Hopkins less impressive than its less-than-completely-negative review on Science-Based Medicine, a blog I can usually count on for a hit job on any dietary supplement.

The consensus stresses that melatonin is a very weak hypnotic. The Buscemi meta-analysis cites this as their reason for declaring negative results despite a statistically significant effect – the supplement only made people get to sleep about ten minutes faster. “Ten minutes” sounds pretty pathetic, but we need to think of this in context. Even the strongest sleep medications, like Ambien, only show up in studies as getting you to sleep ten or twenty minutes faster; ef="https://www.nytimes.com/2007/10/23/health/23drug.html">this New York Times article says that “viewed as a group, [newer sleeping pills like Ambien, Lunesta, and Sonata] reduced the average time to go to sleep 12.8 minutes compared with fake pills, and increased total sleep time 11.4 minutes.” I don’t know of any statistically-principled comparison between melatonin and Ambien, but the difference is hardly (pun not intended) day and night.

Rather than say “melatonin is crap”, I would argue that all sleeping pills have measurable effects that vastly underperform their subjective effects. The linked article speculates on one reason this might be: people have low awareness around the time they get to sleep, and a lot of people’s perception of whether they’re insomniac or not is more anxiety (or sometimes literally dream) than reality. This is possible, but I also think of this in terms of antidepressant studies, which find similarly weak objective effects despite patients (and doctors) who swear by them and say they changed their lives. If I had to guess, I would say that the studies include an awkward combination of sick and less-sick people *and* confuse responders and non-responders. Maybe this is special pleading. I don’t know. But if you think any sleeping pill works well, melatonin doesn’t necessarily work much worse than that.

Sleep latency statistics are hard to compare to one another because they’re so dependent on the study population. If your subjects take an hour to fall asleep, perhaps melatonin could shave off thirty-four minutes. But if your subjects take twenty minutes to fall asleep, then no sleeping pill will ever take off thirty-four minutes, and even an amazing sleeping pill might struggle to make fifteen. I cannot directly compare the people who say melatonin gives back ten minutes to the people who say melatonin gives back thirty-four minutes to the people who say Ambien gives back twelve, but my totally unprincipled guess is that melatonin is about a third as strong as Ambien. It also has about a hundred times fewer side effects, so there’s definitely a place for it in sleep medicine.

2. What is the right dose of melatonin?

0.3 mg.

“But my local drugstore sells 10 mg pills! When I asked if they had anything lower, they looked through their stockroom and were eventually able to find 3 mg pills! And you’re saying the correct dose is a third of a milligram?!”

Yes. Most existing melatonin tablets are around ten to thirty times the correct dose.

Many early studies were done on elderly people, who produce less endogenous melatonin than young people and so are considered especially responsive to the drug. Several lines of evidence determined that 0.3 mg was the best dose for this population. Elderly people given doses around 0.3 mg slept better than those given 3 mg or more and had fewer side effects (Zhdanova et al 2001). A meta-analysis of dose-response relationships concurred, finding a plateau effect around 0.3 mg, with doses after that having no more efficacy, but worse side effects (Brzezinski et al, 2005). And doses around 0.3 mg cause blood melatonin spikes most similar in magnitude and duration to the spikes seen in healthy young people with normal sleep (Vural et al, 2014).

Other studies were done on blind people, who are especially sensitive to melatonin since they lack light cues to entrain their circadian rhythms. This is a little bit of a different indication, since it’s being used more as a chronobiotic than a sleeping pill, but the results were very similar: lower doses worked better than higher doses. For example, in Lewy et al 2002, nightly doses of 0.5 mg worked to get a blind subject sleeping normally at night; doses of 20 mg didn’t. They reasonably conclude that the 20 mg is such a high dose that it stays in their body all day, defeating the point of a hormone whose job is to signal nighttime. Other studies on the blind have generally confirmed that doses of around 0.3 to 0.5 mg are optimal.

There have been disappointingly few studies on sighted young people. One such, Attenburrow et al 1996 finds that 1 mg works but 0.3 mg doesn’t, suggesting these people may need slightly higher doses, but this study is a bit of an outlier. Another Zhdanova study on 25 year olds found both to work equally. And Pires et al studying 22-24 year olds found that 0.3 mg worked better than 1.0. I am less interested in judging the 0.3 mg vs. 1.0 mg debate than in pointing out that both numbers are much lower than the 3 – 10 mg doses found in the melatonin tablets sold in drugstores.

UpToDate, the gold standard research database used by doctors, agrees with these low doses. “We suggest the use of low, physiologic doses (0.1 to 0.5 mg) for insomnia or jet lag (Grade 2B). High-dose preparations raise plasma melatonin concentrations to a supraphysiologic level and alter normal day/night melatonin rhythms.” Mayo Clinic makes a similar recommendation: they recommend 0.5 mg. John Hopkins’ experts almost agree: they say “less is more” but end up chickening out and recommending 1 to 3 mg, which is well above what the studies would suggest.

Based on a bunch of studies that either favor the lower dose or show no difference between doses, plus clear evidence that 0.3 mg produces an effect closest to natural melatonin spikes in healthy people, plus UpToDate usually having the best recommendations, I’m in favor of the 0.3 mg number. I think you could make an argument for anything up to 1 mg. Anything beyond that and you’re definitely too high. Excess melatonin isn’t grossly dangerous, but tends to produce tolerance and might mess up your chronobiology in other ways. Based on anecdotal reports and the implausibility of becoming tolerant to a natural hormone at the dose you naturally have it, I would guess sufficiently low doses are safe and effective long term, but this is just a guess, and most guidelines are cautious in saying anything after three months or so.

3. What are circadian rhythm disorders? How do I use melatonin for them?

Circadian rhythm disorders are when your circadian rhythm doesn't match the normal cycle where you want to sleep at night and wake up in the morning.

The most popular circadian rhythm disorder is "being a teenager". Teenagers' melatonin cycle is naturally shifted later, so that they don't want to go to bed until midnight or later, and don't want to wake up until eight or later. This is an obvious mismatch with school starting times, leading to teenagers either not getting enough sleep, or getting their sleep at times their body doesn't want to be asleep and isn't able to use it properly. This is why every reputable sleep scientist and relevant scientific body keeps telling the public school system to start later.

When a this kind of late sleep schedule persists into adulthood or becomes too distressing, we call it Delayed Sleep Phase Disorder. People with DSPD don't get tired until very late, and will naturally sleep late if given the chance. The weak version of this is "being a night owl" or "not being a morning person". The strong version just looks like insomnia: you go to bed at 11 PM, toss and turn until 2 AM, wake up when your alarm goes off at 7, and complain you "can't sleep". But if you can sleep at 2 AM, consistently, regardless of when you wake up, and you would fall asleep as soon as your head hit the pillow if you first got into bed at 2, then this isn't insomnia – it's DSPD.

The opposite of this pattern is Advanced Sleep Phase Disorder. This is most common in the elderly, and I remember my grandfather having this. He would get tired around 6 PM, go to bed by 7, wake around 1 or 2 AM, and start his day feeling fresh and alert. But the weak version of this is the person who wakes up at 5 each morning even though their alarm doesn't go off until 8 and they could really use the extra two hours' sleep. These people would probably do fine if they just went to bed at 8 or 9, but the demands of work and a social life make them feel like they "ought" to stay up as late as everyone else. So they go to bed at 11, wake up at 5, and complain of "terminal insomnia".

Finally, there's Non-24-Hour-Sleep Disorder, where somehow your biological clock ended up deeply and unshakeably convinced that days on Earth are twenty-five (or whatever) hours long, and decides this is the hill it wants to die on. So if you naturally sleep 11 – 7 one night, you'll naturally sleep 12 – 8 the next night, 1 to 9 the night after that, and so on until either you make a complete 24-hour cycle or (more likely) you get so tired and confused that you stay up 24+ hours and break the cycle. This is most common in blind people, who don't have the visual cues they need to remind themselves of the 24 hour day, but it happens in a few sighted people also; Eliezer Yudkowsky has written about his struggles with this condition.

Melatonin effectively treats these conditions, but you've got to use it right.

The general heuristic is that melatonin drags your sleep time towards the direction of when you take the melatonin.

So if you want to go to sleep (and wake up) earlier, you want to take melatonin early in the day. How early? Van Geijlswijk et al sums up the research as saying it is most effective "5 hours prior to both the traditionally determined [dim light melatonin onset] (circadian time 9)”. If you don’t know your own melatonin cycle, your best bet is to take it 9 hours after you wake up (which is presumably about seven hours before you go to sleep).

What if you want to go to sleep (and wake up) later? Our understanding of the melatonin cycle strongly suggests melatonin taken first thing upon waking up would work for this, but as far as I know this has never been formally investigated. The best I can find is researchers saying that they think it would happen and being confused why no other researcher has investigated this.

And what about non-24-hour sleep disorders? I think the goal in treatment here is to advance your phase each day by taking melatonin at the same time, so that your sleep schedule is more dependent on your own supplemental melatonin than your (screwed up) natural melatonin. I see conflicting advice about how to do this, with some people saying to use melatonin as a hypnotic (ie just before you go to bed) and others saying to use it on a typical phase advance schedule (ie nine hours after waking and seven before sleeping, plausibly about 5 PM). I think this one might be complicated, and a qualified sleep doctor who understands your personal rhythm might be able to tell you which schedule is best for you. Eliezer says the latter regimen had very impressive effects for him (search “Last but not least” here). I’m interested in hearing from the MetaMed researcher who gave him that recommendation on how they knew he needed a phase advance schedule.

Does melatonin used this way cause drowsiness (eg at 5 PM)? I think it might, but probably such a minimal amount compared to the non-sleep-conduciveness of the hour that it doesn’t register.

Melatonin isn’t the only way to advance or delay sleep phase. Here is a handy cheat sheet of research findings and theoretical predictions:

TO TREAT DELAYED PHASE SLEEP DISORDER (ie you go to bed too late and wake up too late, and you want it to be earlier)

– Take melatonin 9 hours after wake and 7 before sleep, eg 5 PM

– Block blue light (eg with blue-blocker sunglasses or f.lux) after sunset

– Expose yourself to bright blue light (sunlight if possible, dawn simulator or light boxes if not) early in the morning

– Get early morning exercise

– Beta-blockers early in the morning (not generally recommended, but if you’re taking beta-blockers, take them in the morning)

TO TREAT ADVANCED PHASE SLEEP DISORDER (ie you go to bed too early and wake up too early, and you want it to be later)

– Take melatonin immediately after waking

– Block blue light (eg with blue-blocker sunglasses or f.lux) early in the morning

– Expose yourself to bright blue light (sunlight if possible, light boxes if not) in the evening.

– Get late evening exercise

– Beta-blockers in the evening (not generally recommended, but if you’re taking beta-blockers, take them in the evening)

These don’t “cure” the condition permanently; you have to keep doing them every day, or your circadian rhythm will snap back to its natural pattern.

What is the correct dose for these indications? Here there is a lot more controversy than the hypnotic dose. Of the nine studies van Geijlswijk describes, seven have doses of 5 mg, which suggests this is something of a standard for this purpose. But the only study to compare different doses directly (Mundey et al 2005) found no difference between a 0.3 and 3.0 mg dose. The Cochrane Review on jet lag, which we’ll see is the same process, similarly finds no difference between 0.5 and 5.0.

Van Geijlswijk makes the important point that if you take 0.3 mg seven hours before bedtime, none of it is going to be remaining in your system at bedtime, so it’s unclear how this even works. But – well, it *is* pretty unclear how this works. In particular, I don’t think there’s a great well-understood physiological explanation for how taking melatonin early in the day shifts your circadian rhythm seven hours later.

So I think the evidence points to 0.3 mg being a pretty good dose here too, but I wouldn’t blame you if you wanted to try taking more.

4. How do I use melatonin for jet lag?

Most studies say to take a dose of 0.3 mg just before (your new time zone’s) bedtime.

This doesn’t make a lot of sense to me. It seems like you should be able to model jet lag as a circadian rhythm disorder. That is, if you move to a time zone that’s five hours earlier, you’re in the exact same position as a teenager whose circadian rhythm is set five hours later than the rest of the world’s. This suggests you should use DSPD protocol of taking melatonin nine hours after waking / five hours before DLMO / seven hours before sleep.

My guess is for most people, their new time zone bedtime *is* a couple of hours before their old bedtime, so you’re getting most of the effect, plus the hypnotic effect. But I’m not sure. Maybe taking it earlier would work better. But given that the new light schedule is already working in your favor, I think most people find that taking it at bedtime is more than good enough for them.

5. I try to use melatonin for sleep, but it just gives me weird dreams and makes me wake up very early

This is my experience too. When I use melatonin, I find I wake the next morning with a jolt of energy. Although I usually have to grudgingly pull myself out of bed, melatonin makes me wake up bright-eyed, smiling, and ready to face the day ahead of me…

…at 4 AM, invariably. This is why despite my interest in this substance I never take melatonin myself anymore.

There are many people like me. What’s going on with us, and can we find a way to make melatonin work for us?

This bro-science site has an uncited theory. Melatonin is known to suppress cortisol production. And cortisol is inversely correlated with adrenaline. So if you’re naturally very low cortisol, melatonin spikes your adrenaline too high, producing the “wake with a jolt” phenomenon that I and some other people experience. I like the way these people think. They understand individual variability, their model is biologically plausible, and it makes sense. It’s also probably wrong; it has too many steps, and nothing in biology is ever this elegant or sensible.

I think a more parsimonious theory would have to involve circadian rhythm in some way. Even an 0.3 mg dose of melatonin gives your body the absolute maximum amount of melatonin it would ever have during a natural circadian cycle. So suppose I want to go to bed at 11, and take 0.3 mg melatonin. Now my body has a melatonin peak (usually associated with the very middle of the night, like 3 AM) at 11. If it assumes that means it’s *really* 3 AM, then it might decide to wake up 5 hours later, at what it thinks is 8 AM, but which is actually 4.

I think I have a much weaker circadian rhythm than most people – at least, I take a lot of naps during the day, and fall asleep about equally well whenever. If that’s true, maybe melatonin acts as a superstimulus for me. The normal tendency to wake up feeling refreshed and alert gets exaggerated into a sudden irresistable jolt of awakeness.

I don’t know if this is any closer to the truth than the adrenaline theory, but it at least fits what we know about circadian rhythms. I’m going to try to put some questions about melatonin response on the SSC survey this year, so start trying melatonin now so you can provide useful data.

What about the weird dreams?

From a HuffPo article:

> Dr. Rafael Pelayo, a Stanford University professor of sleep medicine, said he doesn’t think melatonin causes vivid dreams on its own. “Who takes melatonin? Someone who’s having trouble sleeping. And once you take anything for your sleep, once you start sleeping more or better, you have what’s called ‘REM rebound,’” he said.

This means your body “catches up” on the sleep phase known as rapid eye movement, which is characterized by high levels of brain-wave activity.

Normal subjects who take melatonin supplements in the controlled setting of a sleep lab do not spend more time dreaming or in REM sleep, Pelayo added. This suggests that there is no inherent property of melatonin that leads to more or weirder dreams.

Okay, but I usually have normal sleep. I take melatonin sometimes because I like experimenting with psychotropic substances. And I still get some really weird dreams. A Slate journalist says he’s been taking melatonin for nine years and still gets crazy dreams.

We know that REM sleep is most common towards the end of sleep in the early morning. And we know that some parts of sleep structure are responsive to melatonin directly. There’s a lot of debate over exactly what melatonin does to REM sleep, but given all the reports of altered dreaming, I think you could pull together a case that it has some role in sleep architecture that promotes or intensifies REM.

6. Does this relate to any other psychiatric conditions?

Probably, but this is all still speculative.

Seasonal affective disorder is the clearest suspect. We know that the seasonal mood changes don’t have anything to do with temperature; they seem to be based entirely on winter having shorter (vs. summer having longer) days.

There’s some evidence that there are two separate kinds of winter depression. In one, the late sunrises train people to a late circadian rhythm and they end up phase-delayed. In the other, the early sunsets train people to an early circadian rhythm and they end up phase-advanced. Plausibly SAD also involves some combination of the two where the circadian rhythm doesn’t know what it’s doing. In either case, this can make sleep non-circadian-rhythm-congruent and so less effective at doing whatever it is sleep does, which causes mood problems.

How does sunrise time affect the average person, who is rarely awake for the sunrise anyway and usually sleeps in a dark room? I think your brain subconsciously “notices” the time of the dawn even if you are asleep. There are some weird pathways leading from the eyes to the nucleus governing circadian rhythm that seem independent of any other kind of vision; these might be keeping tabs on the sunrise if even a little outside light is able to leak into your room. I’m basing this also on the claim that dawn simulators work even if you sleep through them. I don’t know if people get seasonal affective disorder if they sleep in a completely enclosed spot (eg underground) where there’s no conceivable way for them to monitor sunrise times.

Bright light is the standard treatment for SAD for the same reason it’s the standard treatment for any other circadian phase delay, but shouldn’t melatonin work also? Yes, and there are some preliminary studies (paper, article) showing it does. You have to be a bit careful, because some people are phase-delayed and others phase-advanced, and if you use melatonin the wrong way it will make things worse. But for the standard phase-delay type of SAD, normal phase advancing melatonin protocol seems to go well with bright light as an additional treatment.

This model also explains the otherwise confusing tendency of some SAD sufferers to get depressed in the summer. The problem isn’t amount of light, it’s circadian rhythm disruption – which summer can do just as well as winter can.

I’m also very suspicious there’s a strong circadian component to depression, based on a few lines of evidence.

First, one of the most classic symptoms of depression is awakening in the very early morning and not being able to get back to sleep. This is confusing for depressed people, who usually think of themselves as very tired and needing to sleep more, but it definitely happens. This fits the profile for a circadian rhythm issue.

Second, agomelatine, a melatonin analogue, is an effective (ish) antidepressant.

Third, for some reason staying awake for 24+ hours is a very effective depression treatment (albeit temporary; you’ll go back to normal after sleeping). This seems to sort of be a way of telling your circadian rhythm “You can’t fire me, I quit”, and there are some complicated sleep deprivation / circadian shift protocols that try to leverage it into a longer-lasting cure. I don’t know anything about this, but it seems pretty interesting.

Fourth, we checked and depressed people definitely have weird circadian rhythms.

Last of all, bipolar has a very strong circadian component. There aren’t a whole lot of lifestyle changes that really work for preventing bipolar mood episodes, but one of the big ones is keeping a steady bed and wake time. Social rhythms therapy, a rare effective psychotherapy for bipolar disorder, revolves around training bipolar people to control their circadian rhythms.

Theories of why circadian rhythms matter so much revolve either around the idea of pro-circadian sleep – that sleep is more restorative and effective when it matches the circadian cycle – or the idea of multiple circadian rhythms, with the body functioning better when all of them are in sync.

7. How can I know what the best melatonin supplement is?

Labdoor has done purity tests on various brands and has ranked them for you. All the ones they highlight are still ten to thirty times the appropriate dose (also, stop calling them things like “Triple Strength!” You don’t *want* your medications to be too strong!). As usual, I trust NootropicsDepot for things like this – and sure enough their melatonin (available on Amazon) is *exactly* 0.3 mg. God bless them.

Alignment Newsletter #13: 07/02/18

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

Highlights

OpenAI Five *(Many people at OpenAI)*: OpenAI has trained a team of five neural networks to play a particular set of Dota heroes in a mirror match (playing against the same set of heroes) with a few restrictions, and have started to beat amateur human players. They are aiming to beat a team of top professionals at The International in August, with the same set of five heroes, but without any other restrictions. Salient points:

- The method is remarkably simple -- it's a scaled up version of PPO with training data coming from self-play, with reward shaping and some heuristics for exploration, where each agent is implemented by an LSTM.
- There's no human data apart from the reward shaping and exploration heuristics.
- Contrary to most expectations, they didn't need anything fundamentally new in order to get long-term strategic planning. I was particularly surprised by this. Some interesting thoughts from OpenAI researchers in this thread -- in particular, assuming good exploration, the variance of the gradient should scale linearly with the duration, and so you might expect you only need linearly more samples to counteract this.
- They used 256 dedicated GPUs and 128,000 preemptible CPUs. A Hacker News comment estimates the cost at $2500 per hour, which would put the likely total cost in the millions of dollars.
- They simulate 900 years of Dota every day, which is a ratio of ~330,000:1, suggesting that each CPU is running Dota ~2.6x faster than real time. In reality, it's probably running many times faster than that, but preemptions, communication costs, synchronization etc. all lead to inefficiency.
- There was no explicit communication mechanism between agents, but they all get to observe the full Dota 2 state (*not* pixels) that any of the agents could observe, so communication is not really necessary.
- A version of the code with a serious bug was still able to train to beat humans. Not encouraging for safety.
- Alex Irpan covers some of these points in more depth in Quick Opinions on OpenAI Five.
- Gwern comments as well.

My opinion: I might be more excited by an approach that was able to learn from human games (which are plentiful), and perhaps finetune with RL, in order to develop an approach that could generalize to more tasks in the future, where human data is available but a simulator is not. (Given the ridiculous sample complexity, pure RL with PPO can only be used in tasks with a simulator.) On the other hand, an approach that leveraged human data would necessarily be at least somewhat specific to Dota. A dependence on human data is unlikely to get us to *general* intelligence, whereas this result suggests that we can solve tasks that have a simulator, exploration strategy, and a dense reward function, which really is pushing the boundary on generality. This seems to be gdb's take: "We are very encouraged by the algorithmic implication of this result — in fact, it mirrors closely the story of deep learning (existing algorithms at large scale solve otherwise unsolvable problems). If you have a very hard problem for which you have a simulator, our results imply there is a *real, practical path* towards solving it. This still needs to be proven out in real-world domains, but it will be very interesting to see the full ramifications of this finding."

Paul's research agenda FAQ *(zhukeepa)*: Exactly what it sounds like. I'm not going to summarize it because it's long and covers a lot of stuff, but I do recommend it.

Technical AI alignment

Technical agendas and prioritization

Conceptual issues in AI safety: the paradigmatic gap *(Jon Gauthier)*: Lots of current work on AI safety focuses on what we can call "mid-term safety" -- the safety of AI systems that are more powerful and more broadly deployed than the ones we have today, but work using relatively similar techniques as the ones we use today. However, it seems plausible that there will be a paradigm shift in how we build AI systems, and if so it's likely that we will have a new, completely different set of mid-term concerns, rendering the previous mid-term work useless. For example, at the end of the 19th century, horse excrement was a huge public health hazard, and "mid-term safety" would likely have been about how to remove the excrement. Instead, the automobile was developed and started replacing horses, leading to new set of mid-term concerns (eg. pollution, traffic accidents), and any previous work on removing horse excrement became near-useless.

My opinion: I focus almost exclusively on mid-term safety (while thinking about long-term safety), not because I disagree with this argument, but in spite of it. I think there is a good chance that any work I do will be useless for aligning superintelligent AI because of a paradigm shift, but I do it anyway because it seems very important on short timelines, which are easier to affect; and I don't know of other approaches to take that would have a significantly higher probability of being useful for aligning superintelligent AI.

Read more: A possible stance for AI control research

Optimization Amplifies *(Scott Garrabrant)*: One model of the difference between mathematicians and scientists is that a scientist is good at distinguishing between 0.01%, 50% and 99.99%, whereas a mathematician is good at distinguishing between 99.99% and 100%. Certainly it seems like if we can get 99.99% confidence that an AI system is aligned, we should count that as a huge win, and not hope for more (since the remaining 0.01% is extremely hard to get), so why do we need mathematicians? Scott argues that optimization is particularly special, in that the point of very strong optimization is to hit a very narrow target, which severely affects extreme probabilities, moving them from 0.01% to near-100%. For example, if you draw a million samples from a normal distribution and optimize for the largest one, it is almost certain to be 4 standard deviations above the mean (which is incredibly unlikely for a randomly chosen sample). In this sort of setting, the deep understanding of a problem that you get from a mathematician is still important. Note that Scott is *not* saying that we don't need scientists, nor that we should aim for 100% certainty that an AI is aligned.

My opinion: I think I agree with this post? Certainly for a superintelligence that is vastly smarter than humans, I buy this argument (and in general am not optimistic about solving alignment). However, humans seem to be fairly good at keeping each other in check, without a deep understanding of what makes humans tick, even though humans often do optimize against each other. Perhaps we can maintain this situation inductively as our AI systems get more powerful, without requiring a deep understanding of what's going on? Overall I'm pretty confused on this point.

Another take on agent foundations: formalizing zero-shot reasoning *(zhukeepa)*: There are strong incentives to build a recursively self-improving AI, and in order to do this without value drift, the AI needs to be able to reason effectively about the nature of changes it makes to itself. In such scenarios, it is insufficient to "reason with extreme caution", where you think really hard about the proposed change, and implement it if you can't find reasons not to do it. Instead, you need to do something like "zero-shot reasoning", where you prove under some reasonable assumptions that the proposed change is good. This sort of reasoning must be very powerful, enabling the AI to eg. build a spacecraft that lands on Mars, after observing Earth for one day. This motivates many of the problems in MIRI's agenda, such as Vingean reflection (self-trust), logical uncertainty (how to handle being a bounded reasoner), counterfactuals, etc., which all help to formalize zero-shot reasoning.

My opinion: This assumes an ontology where there exists a utility function that an AI is optimizing, and changes to the AI seem especially likely to change the utility function in a random direction. In such a scenario, yes, you probably should be worried. However, in practice, I expect that powerful AI systems will not look like they are explicitly maximizing some utility function. If you change some component of the system for the worse, you are likely to degrade its performance, but not likely to drastically change its behavior to cause human extinction. For example, even in RL (which is the closest thing to expected utility maximization), you can have serious bugs and still do relatively well on the objective. A public example of this is in OpenAI Five (https://blog.openai.com/openai-five/), but I also hear this expressed when talking to RL researchers (and see this myself). While you still want to be very careful with self-modification, it seems generally fine not to have a formal proof before making the change, and evaluating the change after it has taken place. (This would fail dramatically if the change drastically changed behavior, but if it only degrades performance, I expect the AI would still be competent enough to notice and undo the change.) It may be the case that adversarial subprocesses could take advantage of these sorts of bugs, but I expect that we need adversarial-subprocess-specific research to address this, not zero-shot reasoning.

The Learning-Theoretic AI Alignment Research Agenda *(Vadim Kosoy)*: This agenda aims to create a general abstract theory of intelligence (in a manner similar to AIXI, but with some deficiencies removed). In particular, once we use the framework of reinforcement learning, regret bounds are a particular way of provably quantifying an agent's intelligence (though there may be other ways as well). Once we have this theory, we can ground all other AI alignment problems within it. Specifically, alignment would be formalized as a value learning protocol that achieves some regret bound. With this formalization, we can solve hard metaphilosophy problems such as "What is imperfect rationality?" through the intuitions gained from looking at the problem through the lens of value learning protocols and universal reinforcement learning.

My opinion: This agenda, like others, is motivated by the scenario where we need to get alignment right the first time, without empirical feedback loops, both because we might be facing one-shot success or failure, and because the stakes are so high that we should aim for high reliability subject to time constraints. I put low probability on the first reason (alignment being one-shot), and it seems much less tractable, so I mostly ignore those scenarios. I agree with the second reason, but aiming for this level of rigor seems like it will take much longer than the time we actually have. Given this high level disagreement, it's hard for me to evaluate the research agenda itself.

Iterated distillation and amplification

Paul's research agenda FAQ *(zhukeepa)*: Summarized in the highlights!

Agent foundations

Forecasting using incomplete models *(Vadim Kosoy)*

Logical uncertainty and Mathematical uncertainty *(Alex Mennen)*

Learning human intent

Policy Approval *(Abram Demski)*: Argues that even if we had the true human utility function (assuming it exists), an AI that optimizes it would still not be aligned. It also sketches out an idea for learning policies instead of utility functions that gets around these issues.

My opinion: I disagree with the post but most likely I don't understand it. My strawman of the post is that it is arguing for imitation learning instead of inverse reinforcement learning (which differ when the AI and human know different things), which seems wrong to me.

Human-Interactive Subgoal Supervision for Efficient Inverse Reinforcement Learning *(Xinlei Pan et al)*

Multi-agent Inverse Reinforcement Learning for General-sum Stochastic Games *(Xiaomin Lin et al)*

Adversarial Exploration Strategy for Self-Supervised Imitation Learning *(Zhang-Wei Hong et al)*

Preventing bad behavior

Minimax-Regret Querying on Side Effects for Safe Optimality in Factored Markov Decision Processes *(Shun Zhang et al)*: As we saw in Alignment Newsletter #11, one approach to avoiding side effects is to create a whitelist of effects that are allowed. In this paper, the agent learns both a whitelist of allowed effects, and a blacklist of disallowed effects. They assume that the MDP in which the agent is acting has been factored into a set of features that can take on different values, and then separate the features as locked (unchangeable), free (changeable), or unknown. If there are no unknown features, then we can calculate the optimal policy using variants of standard techniques (for example, by changing the transition function to remove transitions that would change locked features, and then running any off-the-shelf MDP solver). However, this would require the operator to label all features as locked or unlocked, which would be very tedious. To solve this, they allow the agent to query the operator whether a certain feature is locked or unlocked, and provide algorithms that reduce the number of queries that the agent needs to make in order to find an optimal safe policy.

My opinion: This seems like a good first step towards whitelisting -- there's still a lot of hardcoded knowledge from a human (which features to pay attention to, the transition function) and restrictions (the number of relevant features needs to be small), but it takes a problem and provides a solution that works in that setting. In the recent whitelisting approach, I was worried that the whitelist simply wouldn't include enough transitions for the agent to be able to do anything useful. Since this approach actively queries the operator until it finds a safe policy, that is no longer an issue. However, the corresponding worry would be that it takes prohibitively many queries before the agent can do anything useful. (Their empirical evaluation is on toy gridworlds, so this problem did not come up.) Another worry previously was that whitelisting causes an agent to be "clingy", that is, it wants to prevent all changes to non-whitelisted features, even if they are caused by physical laws, or other humans. A similar problem could arise here when this is generalized to dynamic and/or multiagent environments.

Read more: Worrying about the Vase: Whitelisting

Handling groups of agents

Learning Social Conventions in Markov Games *(Adam Lerer and Alexander Peysakhovich)*

Interpretability

Open the Black Box Data-Driven Explanation of Black Box Decision Systems *(Dino Pedreschi et al)*

Interpretable Discovery in Large Image Data Sets *(Kiri L. Wagstaff et al)*

Near-term concerns

Adversarial examples

On Adversarial Examples for Character-Level Neural Machine Translation *(Javid Ebrahimi et al)*

AI capabilities

Reinforcement learning

OpenAI Five *(Many people at OpenAI)*: Summarized in the highlights!

Retro Contest: Results *(John Schulman et al)*: OpenAI has announced the results of the Retro Contest. The winning submissions were modified versions of existing algorithms like joint PPO and Rainbow, without any Sonic-specific parts.

A Tour of Reinforcement Learning: The View from Continuous Control *(Benjamin Recht)*

Evolving simple programs for playing Atari games *(Dennis G Wilson et al)*

Accuracy-based Curriculum Learning in Deep Reinforcement Learning *(Pierre Fournier et al)*

Deep learning

DARTS: Differentiable Architecture Search *(Hanxiao Liu et al)*

Resource-Efficient Neural Architect *(Yanqi Zhou et al)*

AGI theory

The Foundations of Deep Learning with a Path Towards General Intelligence *(Eray Özkural)*

News

RAISE status report April-June 2018 *(Veerle)*

Fading Novelty

A core aspect of human experience is our pursuit of novelty. There is something tantalizing about seeking out new pleasures, sensations, and experiences that feels hard-coded into how we operate. “Variety is the spice of life” and all that.

Conversely, things which were once new eventually lose their shine over time, and our search for novelty continues. Things which once enamored us are left by the wayside as our attention is captured by shinier, newer things. “The novelty has faded” and all that.

A few examples to drive the point home:

1. Songs which sounded so entrancing upon the first few listenings become dull after being put on repeat.
2. Foods which were so delicious during the first few tastings become bland after being eaten day after day.
3. Clothes which looked so beautiful when initially worn fade into yet another outfit after being worn over and over.

Repetition dulls us.

In psychology, this phenomenon whereby repeated exposure to a stimulus leads to a decreased response is typically referred to as satiation or habituation. (This is rather unfortunate, as I’ve become accustomed to using “habituate” to refer to the act of making something a habit, which makes searching for papers on Google Scholar confusing.) This general pattern of a reduced response is quite ubiquitous across nature. For example, animals which leap into a prepared state upon hearing a loud noise soon grow to ignore it if the noise isn’t paired with actual danger.

In short, it’s a basic form of learning.

From a survival standpoint, a bias towards newness is reasonable. Things in our environment which did not change, e.g. trees, shrubs, or familiar tribe members, likely presented less of a threat than new additions, e.g. fresh tracks, gathering storm clouds, or a stranger in our midst. Had we not had such a filter for newness, our thoughts might have looked like:

“Oh wow, that’s new! Look at that majestic tree! It looks just as good as it did yesterday! Oh wow, that’s new! Look at that lush grass! It’s so fluffy! Oh wow, that’s new! Look at that tiger. It’s so— ”

Constantly taking note of everything in your environment is costly; focusing on just what’s new is an effective optimization that often doesn’t require much of a trade-off.

Yet, despite it’s useful roots, I think that fading novelty is also responsible for some of the difficulties we experience with learning and self-improvement.

On the self-improvement side of things, I think fading novelty makes practicing rationality skills more difficult because it hampers habit creation and contributes to the illusion of understanding.

Consider the process of creating a new habit:

When starting a new regimen, be it a new diet, productivity app, or exercise, there is often an initial burst of success. Our undertaker in question may have such thoughts as:

“Yes, finally! This *is the [thing] that will work for me! Look at how well things have been going! This is effective in all the ways that previous [things] have not! This time, it’ll be different!”*

I’ve personally thought things along these lines when switching up my productivity app of choice, from Google Keep to Workflowy to Google Drive to Dynalist to Evernote.

But of course, I think that the initial excitement / effectiveness of switching something new has little to do with the actual merits of the new thing compared to the old. The real difference is the mere change itself—by changing to something new, your brain is now more interested.

Now, I don’t mean to necessarily knock this initial burst of excitement. I’m glad that humans have the ability to jump-start new projects, and switching to something novel seems to be one of the easiest motivation hacks we have at our disposal. However, I think that this decay over time (from our initial excitement) is often not factored in when people start new regimens.

Novelty fades, and it seems to do so at a rate faster than the rough baseline of two months needed to form a habit. This leads to an overall negative cycle where someone might try out a new rationality skill or productivity app, experience an initial surge of success, and then, after having become acclimated, give up too easily. This can lead someone to switch constantly, never sticking with something long enough for it to become a habit.

Taken to extremes, you’ll look like the rationality junkie described in In Defense of The Obvious, where you’re compelled to seek out ever-more radical solutions because you’ve exhausted the benefit provided to you by more “normal” or “obvious” interventions.

Cue someone reviewing a topic they’ve seen before:

The decreased response we get from the same stimulus plays a pernicious role in learning where I think it contributes to a sense of false understanding. In the same way that a song heard over and over becomes easy to recognize, something analogous with books that we read over and over, as well as concepts we review over and over. Fading novelty leads us to think we understand something, when we really might not.

Because novelty fades, subjects we try to learn might start to look dull before we’ve actually mastered them, and this dullness could be interpreted as understanding. Thus, when trying to review, you might think *“This doesn’t seem new to me. Of course I already know this”,* except that “know” has been substituted to mean one of the easier recognition-based checks for understanding, rather than one of the harder actionable-based checks.

Compare the previous thought with:

“Huh. This looks familiar, but I don’t think I could have come up with this idea by myself, nor could I explain it to someone else. So even though it doesn't seem new, I think I need more review.”

Here are some more examples where substitution happens:

1. If someone gives us advice, we’ll often ask ourselves *“Have I heard similar advice before?”* instead of *“Could I act on this advice, and if so, what are some examples?”*
2. When reviewing math, it’s easier to ask ourselves *“Do these symbols look familiar?”* instead of *“If I covered up the steps, could I reproduce this proof?”*
3. Similarly, when reading a book, it's easier to check "*Have I read these words already?*" instead of "*Did these words give me ideas I hadn't considered before?*"

Due to fading novelty, we interpret familiarity as recognition and make the fallacious leap towards equating this with comprehension. You can end up lulling yourself into a false sense of understanding, which in turn can hinder you from putting in more effort towards areas where you do in fact need improvement.

The aforementioned two areas, habit formation and learning, are where I am predominantly concerned about fading novelty presenting challenges. They are both centered around the efforts of an individual and self-improvement. On a broader scale, I think similar issues manifest.

For example, in the sciences, there is often a larger focus on coming up with novel results than replicating previous studies. Similarly, I think this also part of why the rationality community has few good introductory texts—once you’re “in the know”, it might not feel very motivating to write down the intro stuff because it’s no longer new and hence no longer exciting.

I think interventions which aim to solve the issues I've outlined will either try to reduce the dulling caused by habituation or find another more invariant form of reinforcement to carry us through repetition after repetition.

Psychologists have identified several factors which contribute to the feeling of fading novelty. They include perceived variety, quantity, and stimulus strength. The important thing to note here is that the sensation of habituation is largely a psychological one. Experiments which changed the amount of attention participants paid to relevant factors were found to either increase or decrease the amount of consumption before satiety kicked in.

This seems to indicate that we can reduce the effects through altering our perception of these factors.

For more simple problems like fatigue brought on by satiation during studying, this suggests using something like context switching to make it seem like a greater amount of time has passed. In other words, taking 10 minutes to do something very different to take your mind off work can do a lot more to improve your willingness to continue than spending 20 minutes idly staring at the same textbook pages.

However, it does seem more difficult to apply such a strategy towards habit creation, given that frequency is what we want to max out on. Trying to trick ourselves into thinking that we didn't do the habit often seems counterproductive.

I *can* think of some applications, but they're a bit a stretch. For example, using a productivity app with a dull UI ("*You can't have the novelty fade if it isn't even exciting to use the first time around!*") or an app that randomly changes its UI (which might lead to compulsions to open the app, absent the intended reason).

Perhaps it would be more tractable to look for a different psychological approach. In addition to our cravings for novelty, humans also have a drive for optimization. We want things to get better, we're always on the lookout for improvement. This drive to improve can be used to cut through the humdrum that repetition brings on. Ideally, this is what Good Practice is about.

Athletes and performers channel this attitude a lot, I think. Their work consists of executing a small set of skills, but they need to do them over and over, well past the point of novelty, in pursuit of perfection.

For reviewing topics we think we know, this altered view will hopefully substitute back in the more practical actionable-based checks for understanding. In addition, I think the practicing analogy can bring up additional inspiration for ways to improve. For example, there might be questions like "*Can I execute this skill even when I am tired or distracted?*" or "*How easy is this to do with my eyes closed?*" which normally make sense in a practice context, but could be ported with analogs over to a learning context.

This attitude has parallels to deliberate practice (see here and here for overviews), of being more intentional and analytical about which areas to improve on.

Given that life is all about change, I don't think it's coherent to fully have a view that separates itself from seeking novelty. But I think this is useful insofar as it shifts the scope, so that it focuses on change with regards to the same task, rather than the task itself.

Announcement: AI alignment prize round 3 winners and next round

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

We (Zvi Mowshowitz and Vladimir Slepnev) are happy to announce the results of the third round of the AI Alignment Prize, funded by Paul Christiano. From April 15 to June 30 we received entries from 12 participants, and are awarding $10,000 to two winners.

We are also announcing the fourth round of the prize, which will run until December 31 of this year under slightly different rules. More details below.

The winners

First prize of $7,500 goes to Vanessa Kosoy for The Learning-Theoretic AI Alignment Research Agenda. We feel this is much more accessible than previous writing on this topic, and gives a lot of promising ideas for future research. Most importantly, it explains why she is working on the problems she’s working on, in concrete enough ways to encourage productive debate and disagreement.

Second prize of $2,500 goes to Alexander Turner for the posts Worrying About the Vase: Whitelisting and Overcoming Clinginess in Impact Measures. We are especially happy with the amount of good discussion these posts generated.

We will contact each winner by email to arrange transfer of money. Many thanks to everyone else who sent in their work!

The next round

We are now announcing the fourth round of the AI Alignment Prize. Due the drop in number of entries, we feel that 2.5 months might be too short, so this round will run until end of this year.

We are looking for technical, philosophical and strategic ideas for AI alignment, posted publicly between July 15 and December 31, 2018. You can submit links to entries by leaving a comment below, or by email to apply@ai-alignment.com. We will try to give feedback on all early entries to allow improvement. Another change from previous rounds is that we ask each participant to submit only one entry (though possibly in multiple parts), rather than a list of several entries on different topics.

The minimum prize pool will again be $10,000, with a minimum first prize of $5,000.

Thank you!

Model-building and scapegoating

This is a linkpost for http://benjaminrosshoffman.com/model-building-and-scapegoating/

When talking about undesirable traits, we may want to use simple labels. On one hand, simple labels have the virtue of efficiently pointing to an important cluster of behavioral predictions. On the other, they tend to focus attention on the question of whether the person so described is *good* or *bad*, instead of on building shared models about the causal structure underlying the perceived problem.

Slate Star Codex recently posted a dialogue exploring this through the example of the term "lazy." (Ozy's response is also worth reading.) I think that Scott's analysis itself unfortunately focuses attention on the question of whether assigning simple labels to adverse traits is *good* or *bad* (or alternately, true or false) instead of on building shared models about the causal structure underlying the perceived problem.

When I call someone lazy, I am doing two things. The first is communicating factual information about that person, which can help others avoid incurring costs by trusting the lazy person with some important tasks. This is shared model-building, and it's going to be more salient if you're focused on allocating resources to mitigate harm and produce things of value. In other words, if you're engaged in a community of shared production.

The second is creating a shared willingness to direct blame at that person. Once there's common knowledge that someone's considered blameworthy, they become the default target for exclusion if the group experiences a threat. This can be as simple as killing them and taking their stuff, so there's more per survivor to go around, but this can also take the form of deflecting the hostility of outsiders to the supposed one bad apple. This dynamic is called scapegoating, and it's going to be more salient when zero-sum dynamics are more salient.

Even though I may intend to do only one of these, it's actually quite hard to act along one of these dimensions without side effects along the other. For instance, *Protocols of the Elders of Zion* was famously intended to cast blame on Jews, but Japanese readers with no cultural history of anti-Semitism concluded that if Jews were so powerful, they should make sure to stay on their good side, and consequently the Japanese government made diplomatic overtures to Jewish leaders in the 1930s despite their alignment with the Axis powers.

For more on descriptive vs enactive norms, see Actors and Scribes and the prior posts it links to.

If clearly labeling undesirable traits marks the target for scapegoating, then people who don't want to do that may find themselves drawn to euphemism or other alternative phrasings.

When a single standard euphemism is proposed, this leads to the infamous "euphemism treadmill" in which the new term itself becomes a term of abuse, marking the target for blame or attack. For instance, "mentally retarded" was originally a highfalutin euphemism for "imbecilic," but itself was ultimately replaced with "intellectually disabled," and so on.

Perhaps a more stable solution is the one Scott and Ozy seem to suggest, of preferring more precise and detailed language. Using more precise and detailed descriptions helps avoid scapegoating since more precise usage is more likely to differ from case to case, making it harder to identify a consistently blamed party.

However, these solutions are not the same as the problem. There are multiple distinct interests here. You could want language to be more precise and not much care about the scapegoating side effect. Or you could mainly want to avoid contributing to a process in which a group is targeted for a broad array of social attacks, and not much care about what this does to the precision of your language. (Euphemisms sometimes decrease expressive power instead of increasing it.)

Crucially, neither of these interests requires that you find the thing you're describing unobjectionable, only that you don't want to scapegoat.

Look Under the Light Post

There's a well-worn story you've probably heard before about looking for keys under a light post. If you've not heard it before:

> A man leaves the bar after a long night of drinking. He stumbles all the way home only to find when he gets there that he doesn't have his keys. He knows he had his keys when he left the bar, so he must have dropped them on the way home. He goes out to look for them, but it's the night of a new moon and too dark to possibly see his keys except where there's artificial lighting, so he spends all his effort looking under the sole light post between his home and the bar. After a while a police officers stops to question him.
>
> "Sir. Sir! What are you doing?" shouts the officer.
>
> "Looking for my keys," says the drunk man.
>
> "Did you drop them around here?"
>
> "I don't know."
>
> "Then why are you looking here?"
>
> "Because this is where the light is."

The "broke" interpretation of the story is that the drunk man is stupid for looking for his keys under the light post because the location of his lost keys is unlikely to be correlated with where the light is good enough to search. The "woke" interpretation is that he's right to spend his time looking where the light is because he's more likely to find his keys if he looks in places where he has a chance to find them even if this makes finding his keys conditional on his having dropped his keys in such a location, thus increasing his expected probability of finding his keys since he has a very small chance of finding them in the dark.

I'm here to stretch a metaphor thin and give you the "bespoke" interpretation: many of us are searching for answers that are probably out in the dark, but we should look under the light post anyway both because that's where the light is and because different people have different light posts to look under who can call out to us if they find the answer.

We often face situations in real life analogous to looking for our keys in the dark while drunk. For example, maybe we want to know where to donate our money, what to do on vacation, or how to prevent existential catastrophe. In each case we're trying to find something (a charity, a decision, a solution) and have some ideas about where to look but don't have time or the ability to explore the entire solution space. A typical strategy is to try to "shed light" on the problem and its solutions so we have more information to lead us to a solution.

But shedding light can only take us so far, because the solution space is large and we're small. In the case of the man looking for his keys, it might be 8 blocks from the bar to his house, and that's a lot of ground to cover. Even if he can make the world a little brighter, say by carrying a flashlight or being lucky enough to be drunk on the night of a full moon, he will only slightly improve his ability to look for his keys over the full 8 blocks and the task will still be easiest if his keys are under the light post.

Luckily for us and unluckily for our protagonist, we're not alone in our endeavors while he is. Everyone has their own metaphorical light post to look under created by their interests and abilities, and we can call out to others if we find something under ours. For example, by choice and accident I know a lot about (compared to the average person):

- programming
- phenomenology
- meditation

and many other things besides. In contrast, I live with Sarah Constantin (among other folks), who knows comparatively more about (than me or the average person):

- analysis (in a mathematical sense)
- medicine
- literature reviews

and many other things besides. So if Sarah and I go out to search for the metaphorical key to the best vacation spot, we have very different places to look for the answer based on where the light is best for each of us. I'm not sure either of us is especially well equipped to find the answer given the skills that are probably necessary to find it, but no matter: if it's important enough for us to look then we'll look where the light is and do our best to find an answer where we're able to look. Maybe we'll decide pretty quickly we don't see the answer anywhere we have enough light to look and will ask someone else, but this is merely to ask this other person to look for the answer where they have enough light to look based on where we expect the answer to be, and the situation is no different other than that the search space is now better illuminated by the lights of others.

So what about when we don't know where the answer might be and don't know who to ask to look for us? That's the situation I find interesting, because it seems to be the situation we're in with AI alignment.

We sort of know what we want to do when we say we want to solve AI alignment, but we don't know how to do it. It feels like the answer is somewhere out there in the dark, and some folks have some ideas about where in the dark it might be, so they are building light posts to help us search the space. But you can only build new light posts where there's enough light to work, so we are limited by where the people looking have light now. This suggests a reasonable strategy given the high stakes (we really want to find a solution) and the uncertainty about what will work: get more people to look under their light posts for answers. If enough of us work together to cover enough ground no one of us is especially likely to find a solution but together it's more likely for someone to find it.

There are of course caveats because I've stretched this light post metaphor farther than I should have: AI alignment probably doesn't have a simple solution one person can find; it will probably take many people working on many subproblems combining their efforts to build a complete solution; there may not be a solution; and there may already be a solution we already know about but we just haven't figured out how to assemble it yet. These aren't really problems with the point I want to make, but problems with the metaphor; maybe if I didn't find the light post metaphor so evocative of the situation I would have thought of a better one.

So to conclude in less poetic language in case I've lost you with all this talk of drunk men searching for keys in the night, my point is this: if you want to work on AI alignment, work in the places you are well suited to work on the problems you are well suited to work on. Not just any problem you can rationalize as possibly being relevant, but those problems you believe relevant to solving AI alignment, and especially those problems you believe relevant that others are currently neglecting. You'll probably end up doing something that won't matter either because it's not actually relevant or because it's supervened by something else, but so will everyone else and that's okay because if enough of us look we're more likely to find something that matters.

Why it took so long to do the Fermi calculation right?

This is a meta-level followup to an object level post about Dissolving the Fermi Paradox.

The basic observation of the paper is that when the statistics is done correctly to represent realistic distributions of uncertainty, the paradox largely dissolves.

The *correct statistics* is not that technically difficult: instead of point estimates, just take the distributions, reflecting the uncertainty (implied in the literature!)

There is sizeable literature about the paradox, stretching several decades. Just Wikipedia lists 22 hypothetical explanations, and it seems realistic, at least several hundred researchers spent some serious effort thinking about the problem.

It seems to me really important to reflect on this.

What's going on, why this inadequacy? (in general research)

And more locally, why did not this particular subset of the broader community, priding itself on use of Bayesian statistics, notice earlier?

(I have some hypotheses, but it seems better to just post it as an open-ended question)

The Evil Genie Puzzle

Foolish you. You decided to rub a lamp in the hope that a genie would appear and you found one, only it's an Evil Genie. He tells you only have two options: either wishing for your perfect life or being pelted with rotten eggs. You start thinking about how wonderful your new life will be, but of course, there's a catch.

In the past, wishing for a perfect life would have come with side-effects attached, but due to past abuses, the Council of Genies has created new, updated rules which ban any unwanted side-effects for the person who makes the wish or any of their loved one's. However, there's no rule against messing with other people. He tells you that he has already predicted your choice and if he predicted that you will wish for a perfect life, he has already created a million clones of you in identical rooms on Mars who will believe that they've just rubbed the lamp. However, since they've only hallucinated this, he doesn't have any obligations towards them and he will use his powers to torture them if they will choose the perfect life. On the other hand, if he predicts that you will choose to be pelted with rotten eggs, then he won't create any clones at all. Genies are well known to be perfect predictors.

Assuming you are a perfectly selfish agent who doesn't care about their clones, what decision ought you take? If you choose to be pelted, there's a 100% chance that if you had wished for a perfect life, you would have received it. On the other hand, if you wish for a perfect life, there's an overwhelming chance that you will wish that you had wished to be pelted. No matter what decision you make, it seems that you will immediately regret having made it (at least before it is revealed whether you are the original in the case where you choose the perfect life).

Extra Info: Some people might argue that perfect clones necessarily are you, but they don't actually have to be perfect clones. The genie could just clone someone else and give them the memory of finding the lamp and this ought to create the requisite doubt as to whether you are a real human on Earth or a clone on Mars.

Prior version of this punished the clones for existing or if it is predicted that they chose the perfect life, but the clones are now punished if they choose the perfect life.

Let's Discuss Functional Decision Theory

I've just finished reading through Functional Decision Theory: A New Theory of Rationality, but there are some rather basic questions that are left unanswered since it focused on comparing it to Casual Decision Theory and Evidential Decision Theory:

- How is Functional Decision Theory different from Timeless Decision Theory? All I can gather is that FDT intervenes on the mathematical function, rather than on the agent. What problems does it solve that TDT can't? (Apparently it solves Mechanical Blackmail with an imperfect predictor and so it should also be able to solve Counterfactual Mugging?)
- How is it different from Updateless decision theory? What's the simplest problem in which they give different results?
- Functional Decision Theory seems to require counterpossibilities, where we imagine that a function output a result that is different from what it outputs. It further says that this is a problem that isn't yet solved. What approaches have been tried so far? Further, what are some key problems within this space?

Automatic for the people

Summary: Things we could do about technological unemployment, if there was technological unemployment. (Novel bit is a top-down macroeconomic calculation.)

Confidence: 90% that the solutions I call 'vicious' would be. 60% that the ones I call 'nonvicious' would be. Worth emphasising up here that there is little evidence for tech' unemployment right now.

Crossposted from gleech.org.

Autonomous trucks are now in use and are already safer and more fuel-efficient than human driven ones. (Truck drivers are ~2% of the entire American workforce.)

Crap journalism (that is, 80% of (UK) journalism) is now fully automatable. Automatic art is quite good and improving fast. Consider also the cocktail bartender. And so on: maybe half of all jobs are at risk of being automated, assuming the rate of AI progress just stays constant (*"over an unspecified period, perhaps a decade or two"*).

Automation is maybe the main way that technology improves most people's lives: aside from status exceptions like Apple products, big reductions in manufacturing cost usually mean big reduction in the end cost of goods. Obviously, replacing labour costs with lower-marginal-cost machines benefits rich machine-owners *most*, but automation also allows giant price cuts in all kinds of things; over the last two centuries these cuts have transformed society, increasing equality enormously by making things affordable for the first time.

Besides the obvious example - that we now produce a volume of food far beyond the needs of the entire world population (2940kcal per person per day, though with terrible distributive failures) - consider that a single ordinary shirt takes 508 hours of labour to produce on a spinning wheel and hand-loom - so you would expect to pay something above $3600 at current minimum wage (still $900 at 1400CE wage levels).

Getting costs as near to zero as possible is the way we will solve the easy problem of human existence, scarcity of basic goods and leisure. (The hard problem of human life is boredom and dignity and meaning and all that.)

Types of automation

I couldn't find a rigorous list of the types of automation, so here's my attempt:

- *First-wave*: Mechanical control. machines which perform one task automatically because of the exact shape of their components. e.g. a) Cams, linked wheels machined to create fixed sequences of linear motions. an instance in the year 200BCE. b) Governors. c) Cataracts.
- *Second-wave*: Numerical control. Programmable machines, where changing the program (punch-cards) changes the product. Grossly physical, analog algorithms. The first fully-auto productive machine was designed and built by Vaucanson in 1745 (not Falcon or Bouchon or Jacquard).

- *Third-wave*: Computer numerical control. One machine, whose behaviour is dictated by algorithms *encoded as digital data*. The birth of software. Memory allows for ~zero human oversight once task instructions are created and loaded. Arbitrarily fine manual work, performed better than humans and much faster. 1957 or 1958.
- *Fourth-wave*: Machine-learning model control: No need to program, after suitable domain algorithms have been written; models are automatically calibrated given enough data. Trained models can go beyond ordinary control flow, discerning non-necessary and non-sufficient conditions involved in sophisticated tasks like driving. Early successes came in speech recognition (e.g. Tangora and DRAGON in the early 80s, Rabiner et al in 1985), beginning the long march to replace stenographers and secretaries (and ultimately experimental linguists). The defining instance of the fourth wave may be driverless cars: 5% of developed-world jobs at risk. Sometimes called the 'second machine age'.
- *Fifth-wave*: Artificial general intelligence: No need for programming (programming either new tasks, or any future design of improved machines).

So automation has been happening for (many) hundreds of years, but in the past it probably didn't produce long-term, “technological" unemployment. This is probably because people were easily able to think up new professions given new tech and culture, and since the increased productivity translated into lower costs for the automated good, which stimulated other parts of the economy, and people could retrain for the subsequent new jobs. The present wave might be different: it might actually reduce the number of available jobs permanently, because the machines now entering the workforce can be applied to many of the jobs that people could retrain to do.

If so, our economy - resource allocation based on employment (which we use as a poor proxy variable for productivity) - is a local maximum and we cannot expect to arrive at a good outcome without activism, since:

1. The new machine-learning automation could fully replace around half of jobs.
2. Since these jobs involve even our highest cognitive faculties, it is possible that we won't think up new productive jobs for the replaced workers, like we did in the past.
3. So automation could produce an unprecedentedly high unemployment rate, ~60%.
4. Most existing unemployment welfare systems are inadequate and degrading.
5. So without intervention, a crash in human welfare is easily possible.

But, unless we automate a lot more, we the species will never have enough wealth to offer a decent basic income, and everyone will continue to waste half their lives at work. Like C20th peasants.

How much is there for everyone?

You can follow my calculation here.

Gross world product divided by population is $10,600 per person! - but this naive distribution would be impossible, even assuming all the political will in the world.

Depreciation costs bring this down to ~$9,300 ; maintaining our present levels of R&D investment brings it down further to $9,100.

We also have to consider the "deadweight loss" of taxation (how much you have to spend to collect the tax + how much unproductive tax avoidance behaviour you cause + how much it discourages economic activity + etc). The research on this is shockingly vague (this gives estimates between 2.5% and 30%!). Lower bound takes us to $8,800; the 30% upper bound takes us all the way down to $5,900. The mass carve-up we're talking about goes well beyond any existing tax rate, and avoidance does scale in proportion to rates; so we probably have to assume the deadweight would be worse than any yet experienced. Call it 30%.

Most people will want to maintain government services at around their current level (besides the giant basic income expenditure); remember that this could knock another 30% off our available income flow (or a mere 28% off if we lose the military). Half of that is welfare and pensions, which are being replaced here (in our heads). So we're down to $4,300.

So the current economy, carved up sustainably, would yield some fraction of $4,000 per person per year. Even given that most households would get about 4 of these incomes, this is simply not enough for freedom, given the needs or tastes of an average human.

(The above does not consider a host of other sad realities: e.g. rich people like their money; e.g. this much equality would destroy entire industries (luxury goods!), and so further sap the available pie; e.g. there would be a recession effect from all the fully-alienated workers downing their tools - yes, there could also be a stimulus effect from increasing poor people's spending, but it's extremely difficult to say which sign prevails.

Worse: only the direct cost of taxation is factored in above, without the amount that the rich manage to spirit away. We would need something like a world government for it to work even this well (badly), to stamp out tax havens and transfer pricing and all that jolly financial dancing.)

My point is not that this is the exact figure we'll have to work with - instead it points up our paltry present capacity. Growth would be necessary, even in an ideal world without nationalism, greed, inefficiency (...)

Socially conscious people are these days ambivalent about economic growth, often for environmental reasons. But consider the enlightened definition of "productivity": it is not "amount of output", but the amount of output *per unit of input*. This is pure gain, and is actually environmentally positive, since it could reduce resource use and waste. But we do need output growth too: e.g. until every paraplegic on earth who wants one of these has one.

The above just uses world *income;* what about using world wealth? Even if we liquidate the whole of the world's wealth (our stock of money, as opposed to the GDP, a flow), it would only provide a universal basic income for three and a half years.

Probably vicious solutions to the worst case:

1. Halt progress on automation, preserving current employment. (Via government ban, successful hostility by organised labour, mass monkey wrenching). I count this as vicious, even though it is much better than the worst-case, since it leaves almost all of us very unfree, forever.
2. or I guess you could try huge unemployment plus an authoritarian crackdown on desperate masses, see how that goes.
3. Nationalised robot factories, or full cyber-communism. Food and clothes and houses guaranteed to all, at least. Leave aside the historical failure of command economies; imagine here that a new big-data Kantorovich manages to make it fairly efficient.

   Even granting this giant assumption, this is a dangerous move. Total state control of the means of production is too easily twisted. Now, this *could* be just my emotional overreaction to reading about e.g. Maoist China, with its food terrorism and peasant-robbing. But it rings malign: total control of production by *any* entity is a terrible unnecessary risk.
4. "Back to the land" primitivism. Humans return to subsistence farming as a means of survival.
5. Just raising the minimum wage without doing anything else. Misses the point entirely. ('Should the minimum wage be called the “Robot Employment Act?”' – Cowen and Tabarrok.)
6. Mass human augmentation, to keep up with the machines. (At minimum, just traditional externally-hosted software: Humans using chess assistant programs were beating solo supercomputers until relatively recently, c. 2006.) Possibly vicious, but not because there's anything wrong with transhumanism: because doing it for purely economic reasons is 1) a neverending process, since the machines will improve as fast, and 2) it would probably be destructive of some distinctive human virtues (e.g. serenity, play, reflection, aesthetic interest). Only good if people *really* can't feel dignified without having a leading productive role in things.

Potentially nonvicious solutions:

1. Prop up the liberal mixed economy:

- with a programme of mass employee stock ownership.
- or by carving each full-time job into several part-time ones, plus heavy wage subsidies.
- or with a universal basic income funded through higher taxation.
- Get the government to buy every 18 year old a serious stock portfolio (??)

2. More or less vague suggestions for a very different social structure, like embarrassingly decentralised groupings, with their own minifactories...

I am not very sure of any of the above; the actual stats on productivity growth are worrying for the opposite reason: it has been too slow to support wages for a long time. Anyway other powerful forces (e.g. global outsourcing, the decay of unions) besides robots have led to the 40-year decline in labour’s share of global income. But those will produce similar dystopian problems if the trend continues, and there’s enough of a risk of the above scenario for us to put a lot of thought and effort into protecting people, either way.

> *Factories that run 'lights out' are fully automated and require no human presence on-site... these factories can be run with the lights off.*

– Wiki

> *Not only is it lights-out - we turn off the air conditioning and heat too.*

– an executive at Fuji Automatic Numerical Control

Probability is a model, frequency is an observation: Why both halfers and thirders are correct in the Sleeping Beauty problem.

This post was inspired by a yet another post talking about the Sleeping Beauty problem: Repeated (and improved) Sleeping Beauty problem. It is also related to Probability is in the Mind.

(***Updated**: As pointed out in the comments, If a tree falls on Sleeping Beauty... is an old post recasting this as a decision problem, where the optimal action depends on the question asked: " just ask for decisions and leave probabilities out of it to whatever extent possible")*

It is very common in physics and other sciences that different observers disagree about the value of a certain quantity they both measure. For example, for a person in a moving car, the car is stationary relative to them (v=0), yet to a person outside the car is moving (v=/=0). Only very few select measurable quantities are truly observer-independent. The speed of light is one of the better known examples. The electron charge. Some examples of non-invariant quantities in physics are quite surprising. For example, does a uniformly accelerating electric charge radiate? Do black holes evaporate? The answer depends on the observer! Probability is one of those.

In fact, the situation is worse than that. Probability is not directly measurable.

Probability is not a feature of the world. It is an observer-dependent model of the world. It predicts the observed frequency of some event, which is also observer-dependent.

When you say that a coin is unbiased, you model what will happen when the coin is thrown multiple times and predict that, given a large number of throws, the ratio of heads to tails approaches 1. This might or might not be a sufficiently accurate model of the world you find yourself in. As any model, it is only an approximation, and can miss something essential about your future experiences. For example, someone might try to intentionally distract you every time the coin lands heads, and sometimes they will be successful, so your personal counts of heads and tails will be a bit off, and the ratio of heads to tails will be statistically significantly below 1 even after many many throws. Someone who had not been distracted, would record a different ratio. So you would conclude that the coin is biased. Who is right, you or them?

If you remember that frequency is not an invariant quantity, it depends on the observer, then the obvious answer “they are right, I was maliciously distracted and my counts are off” is not a useful one. If you don’t know that you had been distracted, *your model of the coin as biased toward tails is actually a better model of the world you find yourself in*.

The Sleeping Beauty problem is of that kind: because she is woken up twice per throw that comes up tails, but only once per throw that comes up heads, then she will see twice as many tails as heads on average. So for her this is equivalent to the coin being biased and the heads:tails ratio being 1:2 (the Thirder position). If she is told the details of the experiment, she can then conclude that for the person throwing the coin it is expected to come up heads 50% of the time (the Halfer position), because it is a fair coin. But for the Sleeping Beauty herself this fair coin is observed to come up heads half as often as tails, that’s just how this specific fair coin behaves in her universe.

This is because probability is not an invariant objective statement about the world, it is an observer-dependent model of the world, predicting what a given observer is likely to experience. The question “but what is the ***actual*** probability of the coin landing heads?” is no more meaningful than asking “but what is the ***actual*** speed of the car?” — it all depends on what kind of observations one makes.

Expected Pain Parameters

A problem I sometimes have is that someone will suggest that I do something uncomfortable - physically or emotionally, whichever - and they'll acknowledge that it's not going to feel good.

And then they won't follow up by explaining exactly how much and in what way it's supposed to hurt, let alone why or how to mitigate that.

Look: if you're going for a run, you might get a stitch in your side, your legs might cramp, your lungs might burn. If you feel a sudden stabbing pain in your foot, that's not the run. You've stepped on something sharp and you need to stop and call for a ride, not tough it out. If you've been told that running might hurt and that's a good thing, a sign of weakness leaving the body or whatever, and there's any ambiguity between "stepped on something sharp" and "this is how it's supposed to hurt" - are you supposed to have blisters or do you need better-fitting shoes? Are you supposed to have a headache or do you need to drink way more water than you have been? Is a sunburn just part of a healthy experience that means you're making vitamin D, or should you be wearing sunscreen next time? - then you're blundering into a problem that you would have noticed and solved if you had encountered it by sitting on the couch.

If you're apologizing for something you've done wrong, you might feel guilty, stared-at, awkward, resentful. If you feel like the walls are closing in and you're going to puke, then you have a more serious problem and might want to get that looked into, not chalk it up to "well, acknowledging that you've made a mistake can be painful". If you're apologizing to someone who's abusing you, then how much it hurts is a red flag. If you're in a badly facilitated therapeutic or moderated conversational environment, and you don't know how much it's supposed to hurt, then when it hurts more than it's supposed to you won't notice that you need to fix or leave the setup. If you're apologizing for something that *isn't* wrong, because you've been indoctrinated in a normset that's unhealthy about what things are violations, then by letting that normset also tell you that it's supposed to feel bad you're losing a signal that could tell you to bail.

Even things you aren't actively *doing* can have this issue: it took me a long time to be really sure that most people don't experience pain when they have their blood pressure taken, because every time I complained about it, an indifferent nurse would say "yep, lots of pressure, mm-hm, it squeezes pretty tight" and not "actually, while you're unlikely to be in medical danger, that is not typical; we might make different tradeoffs about how often to recommend this test if it hurt everyone like it does you".*

And if you don't know how something is supposed to hurt, and know that you don't know it, you will need some risk-aversion there to avoid blundering into sharp objects and mental health triggers and cost-benefit analyses that were not designed for how you're put together. So this can cut off opportunities to try things, the opposite problem from trying too hard at something that isn't working.

If you are recommending that people do something that might be uncomfortable or painful, *tell them what normal tolerances are,* and what the things that might be causing abnormal responses might be. Demand this of people telling you to do uncomfortable or painful things too. Pain responses have a purpose; ignoring them outright for the duration of an activity which flirts with the damage that aversion is meant to prevent is insane.

*If you have this problem and might give birth and might have an epidural while you do, have them put the blood pressure cuff on your numb leg. Hat tip to Swimmer963.

No, I won't go there, it feels like you're trying to Pascal-mug me

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

The goal of this blog-post is to explore the intuition that drives the feeling that the agent who arrives at the conclusion that they should pay Pascal's mugger is being unreasonable, and whether this can be tied in with the analysis of a logical inductor as an algorithm not exploitable by an efficiently computable trader. An analysis of this intuition may help us to design the decision theory of a future Friendly AI so that it is not vulnerable to being Pascal-mugged, which is plausibly a desirable goal from the point of view of obtaining behaviour in alignment with our values.

So, then, consider the following conversation that I recently had with another person at a MIRI workshop.

"Do you know Amanda Askell's work? She takes Pascal's wager seriously. If there really are outcomes in the outcome space with infinite utility and non-infinitesimal probability, then that deserves your attention."

"Oh, but that's very Goodhartable."

A thought in close proximity to this would be that that line of reasoning would open you up to exploitation by agents not aligned with your core values. This is plausibly the main driver behind the intuition that paying Pascal's mugger is unreasonable, a decision theory like that makes you too exploitable.

The work done by MIRI on the analysis of the notion of logical induction identifies the key desirable property of a logical inductor algorithm as producing a pattern of belief states evolving over time which isn't exploitable by any efficiently computable trader.

Traders studying the history of your belief states won't be able to identify any pattern that they can exploit with a polynomial-time computable algorithm, so when such patterns emerge in the sentences you have already proved, you'll make the necessary "market correction" in your own belief states so that such patterns don't make you exploitable. Thus, being a good logical inductor is related to having patterns of behaviour that are not exploitable. We see here some hint of a connection with the notion that being Pascal's-mugging-resilient is related to not being exploitable.

Plausibly a useful research project would be to explore whether some similar formalisable notion in an AI's decision theory could capture the essence of the thought behind "Oh, Pascal's-mugging reasoning looks too suspicious". Perhaps the desired line of reasoning might be "A regular habit of being persuaded by Pascal-mugging style reasoning would make me vulnerable to exploitation by agents not aligned with my core values, independently of whether such exploitation is in fact likely to occur given my beliefs about the universe. If I can see that an argument for a particular conclusion about which action to take manifests that kind of vulnerability, I ought to invest more computing time before actually taking the action". It is not hard to imagine that natural selection might have been a driver for such a heuristic and this may explain why Pascal-mugging-style reasoning "feels suspicious" to humans even when they haven't yet constructed the decision-theoretic argument that yields the conclusion not to pay the mugger. If we can build a similar heuristic into an AI's decision theory, this may help to filter out "unintended interpretations" of what kind of behaviour would optimise what humans care about, such as large-scale wireheading, or investing all your computing resources into figuring out some way to hack the laws of physics to produce infinite utility despite very low probability of a positive pay-off.

We also see here some hint of a connection between the insights that drove the "correct solution" to what a logical inductor should be and the "correct solution" to what the best decision theory should be. Exploring whether this is in some way generalisable to other domains may also be a line of thought that deserves further examination.

Buridan's ass in coordination games

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

Consider a simple coordination game. In this game, two players (player 1 and player 2) each simultaneously choose an action, X or Y. If they both choose X, they both get 1 utility. If they both choose Y, they both get u_y utility for some $0 \le u_y \le 2$ known to both players. If they choose different actions, they both get 0 utility. Which action should they each choose? Assume they get to communicate before knowing u_y, but not after knowing u_y.

An optimal policy pair is for each to pick X if $u_y < 1$, and Y otherwise. Unfortunately, this policy pair can break down in the presence of even a small amount of noise. Assume neither player observes u_y, but instead each receives an independent observation V_i (player 1 sees V_1, player 2 sees V_2), each of which is drawn uniformly from the range $[u_y - \epsilon, u_y + \epsilon]$ for some small $\epsilon > 0$. If both players follow the policy of choosing X if and only if their *observation* of u_y is less than 1, then if u_y is very close to 1, there is a significant likelihood that one player will receive an observation less than 1, while the other will receive one greater than 1. Thus, they have a significant chance of choosing different actions and receiving 0 utility. This issue is essentially the same as the one faced by Buridan's ass: both options are equally good, so there is no way to effectively decide between them.

In fact, as I will prove:

Claim 1: No pair of policies in which the players select their actions *independently* given their observations can simultaneously guarantee an expected utility greater than $\max(1, u_y) - 0.5$ for all $u_y \in [0, 2]$.

But things change when a shared source of randomness is allowed. Then, as I will also prove:

Claim 2: Suppose that, after observing their observation of u_y, each player also gets to observe a random number $R \sim \text{Uniform}([0, 1])$. Then there is a pair of policies that achieves expected utility at least $\max(1, u_y) - 2\sqrt{\epsilon} - \epsilon$ regardless of u_y .

This solution method extends to arbitrary cooperative normal-form games where players observe a perturbed version of the true utility function. This is shown in the appendix.

Why is this important? I expect progress in decision theory to come from studying very simple decision problems like this one. If it is possible to show that any solution to some simple problem has certain properties, then this usefully constrains the design space of possible decision theories. In this case, the result suggests that something like shared randomness may be required for achieving provable near-optimality even in cooperative settings.

Claim 1: impossibility of solving the game using independent randomness

Fix $\epsilon > 0$. Let $\pi_1, \pi_2 : \mathbb{R} \to [0, 1]$ be Lesbegue measurable functions representing the two players' policies, which map V_i to the player's probability of choosing action Y.

Define $q(u) := \text{Uniform}([u - \epsilon, u + \epsilon])$ to be the function mapping u_y to the resulting V_i distribution.

Define $\tau_i(u) := E_{v \sim q(u)}[\pi_i(v)]$. This is defined so that $\tau_i(u_y)$ equals player i's overall probability of choosing action Y.

For $u_1, u_2 \in [0, 2]$, note that the total variation distance between $q(u_1)$ and $q(u_2)$ is at most $\frac{|u_1 - u_2|}{\epsilon}$ Thus, since π_i is bounded between 0 and 1, τ_i is $\frac{1}{\epsilon}$-Lipschitz and therefore continuous.

I will now show that, for some u_y, the players achieve expected utility at most $\max(1, u_y) - 0.5$ by case analysis on τ_1:

- If $\tau_1(0) \geq 0.5$ then at most 0.5 expected utility is achieved when $u_y = 0$ (since at most 1 utility is achieved when player 1 chooses action X, and no utility is achieved when player 1 chooses action Y).
- If $\tau_1(2) \leq 0.5$, then at most 1.5 expected utility is achieved when $u_y = 2$ (since at most 1 utility is achieved when player 1 chooses action X, and at most 2 utility is achieved when player 1 chooses action Y).
- If neither of the two above cases hold, then by continuity of τ_1 and the intermediate value theorem, there exists some $u \in [0, 2]$ with $\tau_1(u) = 0.5$. When $u_y = u$, since the two players choose actions independently, there is a 0.5 probability that they select the same action, ensuring that they achieve at most $\frac{\max(1, u_y)}{2} \leq \max(1, u_y) - 0.5$ expected utility.

Thus, claim 1 is proven. This proof method bears resemblance to the problem faced by Buridan's ass: in the third case, some u_y value is found so that the "policy" τ_1 is equally compelled by both options, choosing between them with a 50/50 coin flip in a way that is disastrous for coordination.

Claim 2: solving the game using shared randomness

Define $f_\epsilon(v) = \frac{v - (1 - \sqrt{\epsilon})}{2\sqrt{\epsilon}}$. Consider a pair of policies in which player i chooses Y if $R < f_\epsilon(V_i)$, and otherwise chooses X, where $R \sim \text{Uniform}([0, 1])$. Intuitively, each of these policies increases its chance of taking action Y as its observation of u_y increases in a *smooth* fashion, and they correlate their randomness so that they are likely to choose the same action. Now note a couple properties of this pair of policies:

1. For $u_y \geq 1 + \sqrt{\epsilon} + \epsilon$, it is guaranteed that $V_i, V_j \geq 1 + \sqrt{\epsilon}$, so both players always take action Y.

2. Since f_ϵ is $\frac{1}{2\sqrt{\epsilon}}$-Lipschitz, and $|V_i - V_j| \leq 2\epsilon$, the probability of the players taking different actions is at most $\sqrt{\epsilon}$.

I will now show that the players' expected utility is at least $\max(1, u_y) - 2\sqrt{\epsilon} - \epsilon$ by case analysis:

- Suppose $u_y \geq 1 + \sqrt{\epsilon} + \epsilon$. Then by property 1, both players are guaranteed to take action Y, so they achieve expected utility $u_y \geq \max(1, u_y) - 2\sqrt{\epsilon} - \epsilon$.
- Suppose $u_y < 1 + \sqrt{\epsilon} + \epsilon$. By property 2, the players choose the same action with probability at least $1 - \sqrt{\epsilon}$, and taking the same action yields a utility of at least 1, so they achieve expected utility at least $1 - \sqrt{\epsilon}$. Due to the bound on u_y, this is at least $\max(1, u_y) - 2\sqrt{\epsilon} - \epsilon$.

Thus, the claim is proven. By setting ϵ sufficiently low, this pair of policies yields an expected utility arbitrarily close to $\max(1, u_y)$ regardless of u_y.

Conclusion and directions for further research

Together, these claims show that there is a simple set of noisy coordination games (namely, the set of games described in the beginning of this post for all possible u_y values) that is impossible to solve with only independent randomness, but which can be solved using shared randomness. Some notes on this:

- The proof of claim 1 only used the fact that τ_1 is continuous. So even without noise, if the players' policies are a continuous function of the utility u_y, the same problem occurs.
- A pair of Bayesians playing this coordination game who have a prior over u_y have no need for randomization, independent or joint. The goal of achieving a high expected utility regardless of u_y most clearly makes sense if u_y is selected adversarially (to maximize regret). It also makes sense if the environment is hard to model in a way that makes selection of a policy with a low rate of "ties" difficult. The comparison between the shared randomness solution and the Bayesian solution is similar to the comparison between randomized Quicksort and a "Bayesian" variant of Quicksort that selects pivots based on some expected distribution of inputs. While the Bayesian solution works better than the randomized solution when the prior over the input distribution is accurate, the randomized solution is simpler, is amenable to proofs, and works well even under adversarial conditions.

Where to go from here? Roughly, my overall "plan" for formal decision theory consists of 3 steps:

1. Solve Cartesian cooperative decision theory problems where players reason about each other using something like reflective oracles.

2. Extend the solution in 1 to Cartesian multiplayer game theory problems where players have different utility functions and reason about each other using something like reflective oracles.

3. Extend the solution in 2 to a logically uncertain naturalized setting.

Step 1 has still not been solved. Previous writing on this includes this post which studies a setting with a single memoryless player (which is similar to a set of players who have the same utility function). The post shows (redundantly with the paper introducing the absent-minded driver problem as I later found out) that, if the player's policy is *globally* optimal (i.e. it achieves the highest possible expected utility), then all actions that might be taken by that policy are CDT-optimal, assuming SIA probabilities. This second condition is a local optimality condition, so the post shows that global optimality implies local optimality.

It would be highly desirable to find a similar but different local optimality property that implies a global optimality property. That would essentially be a way of deriving collective self-interest from individual self-interest when individuals have the same utility function and common priors.

As this current post shows, the globally optimal policy is *discontinuous* as a function of the utilities involved, and no continuous approximation always yields a near-optimal expected utility. I expect this to hinder attempts to derive global optimality from local optimality, as it implies there is a valley of bad policies between decent policies and good ones.

Introducing shared randomness here may help by preserving continuity. So a natural next step is to find useful local optimality properties in a cooperative setting where players have shared randomness.

Appendix: extension to arbitrary normal-form cooperative games

The solution described in the section on claim 2 can be extended to arbitrary normal-form cooperative games where the players receive only noisy observations of the payoffs. Reading this appendix (which takes up the rest of this post) is unnecessary for understanding the main point of this post.

Let $n \geq 1$ be the number of players, A_i be player i's set of actions, and $u : \prod_{i=1}^{n} A_i \to \mathbb{R}$ be the shared unknown utility function over strategy profiles, whose minimum value is $u_{\min}$ and whose maximum value is $u_{\max}$.

Fix $0 < \epsilon \leq u_{\max} - u_{\min}$. Let $V_i : \prod_{i=1}^{n} A_i \to \mathbb{R}$ be a "perturbed" version of u that player i observes; it must satisfy the property that for any strategy profile a, $|V_i(a) - u(a)| \leq \epsilon$.

To define the policies, we will first number the strategy profiles arbitrarily $a^1, \ldots, a^m$ where $m = \prod_{i=1}^{n} |A_i|$. Let $c > 0$ be some number to be determined later. For $j \in \{1, \ldots, m\}$, define

$$f_j(v) := \frac{\exp(cv(a^j))}{\sum_{j'=1}^{m} \exp(cv(a^{j'}))}.$$

This defines, given any $v \in \prod_{i=1}^{n} A_i$, a probability distribution over strategy profiles, since $\sum_{j=1}^{m} f_j(v) = 1$. Specifically, the distribution is a softmax. This distribution is more likely to select strategy profiles that v considers better, but has some chance of selecting every strategy profile.

Players will sample from this distribution using a source of shared randomness, $R \sim \text{Uniform}([0, 1])$. Define

$$g(v, r) := \min\left\{ j \in \{1, \ldots, m\} \mid r \leq \sum_{j'=1}^{j} f_{j'}(v) \right\}$$

Now note that $P(g(v, R) = j) = f_j(v)$, i.e. $g(v, R)$ is distributed according to $f_j(v)$. Define policies $\pi_i(v, r) := a_i^{g(v,r)}$. That is, player i will play their appropriate action for strategy profile number $g(V_i, R)$.

Roughly, we will now show 2 properties that are sufficient to establish that these policies are near-optimal:

- With high probability, the players play according to $a^{g(u,R)}$, i.e. for all i, $\pi_i(V_i, R) = a_i^{g(u,R)}$.
- The strategy profile $a^{g(u,R)}$ is near-optimal in expectation, i.e. $E[u(a^{g(u,R)})]$ is close to u_{max}.

Players play according to $a^{g(u,R)}$ with high probability

First, we will show that, for all i and j, $f_j(V_i)$ is close to $f_j(u)$.

For all $j \in \{1, \dots, m\}$, since $|V_i(a^j) - u(a^j)| \le \epsilon$, we have

$$\exp(-c\epsilon) \le \frac{\exp(cV_i(a^j))}{\exp(cu(a^j))} \le \exp(c\epsilon).$$

Therefore, for all $j \in \{1, \dots, m\}$,

$$\frac{f_j(V_i)}{f_j(u)} = \frac{\exp(cV_i(a^j)) \sum_{j'=1}^{m} \exp(cu(a^{j'}))}{\exp(cu(a^j)) \sum_{j'=1}^{m} \exp(cV_i(a^{j'}))} \le \exp(2c\epsilon).$$

By identical logic,

$$\frac{f_j(u)}{f_j(V_i)} \le \exp(2c\epsilon).$$

Now we will bound $\left|\sum_{j'=1}^{j} f_{j'}(V_i) - \sum_{j'=1}^{j} f_{j'}(u)\right|$.

$$\left|\sum_{j'=1}^{j} f_{j'}(V_i) - \sum_{j'=1}^{j} f_{j'}(u)\right|$$

$$= \left|\sum_{j'=1}^{j} f_{j'}(V_i) \left(1 - \frac{f_{j'}(u)}{f_{j'}(V_i)}\right)\right|$$

$$\le \sum_{j'=1}^{j} f_{j'}(V_i) \left|1 - \frac{f_{j'}(u)}{f_{j'}(V_i)}\right|$$

$$\le \sum_{j'=1}^{j} f_{j'}(V_i) \max\{\exp(2c\epsilon) - 1, 1 - \exp(-2c\epsilon)\}$$

$$\le \sum_{j'=1}^{j} f_{j'}(V_i)(\exp(2c\epsilon) - 1)$$

$$= \exp(2c\epsilon) - 1$$

For it to be the case that $g(u, R) \neq g(V_i, R)$, it must be the case that for some j,

$$R \le \sum_{j'=1}^{j} f_{j'}(u) \nLeftrightarrow R \le \sum_{j'=1}^{j} f_{j'}(V_i)$$

which implies

$$R \in \mathrm{ConvexHull}(\{\sum_{j'=1}^{j} f_{j'}(u), \sum_{j'=1}^{j} f_{j'}(V_i)\})$$

Due to the bound on $\left|\sum_{j'=1}^{j} f_{j'}(V_i) - \sum_{j'=1}^{j} f_{j'}(u)\right|$, the measure of this hull is at most $\exp(2c\epsilon) - 1$. Furthermore, there are m hulls of this form (one for each j value), so the total measure of these hulls is at most $m(\exp(2c\epsilon) - 1)$.

So, player i chooses action $a_i^{g(u,R)}$ with probability at least $1 - m(\exp(2c\epsilon) - 1)$. Thus, by the union bound, the players jointly choose the actions $a^{g(u,R)}$ with probability at least $1 - nm(\exp(2c\epsilon) - 1)$.

$a^{g(u,R)}$ is near-optimal in expectation

First we will prove a lemma.

Softmax lemma: For any $c > 0$ and vector x of length m,

$$\frac{\sum_{j=1}^{m} \exp(cx_j)x_j}{\sum_{j=1}^{m} \exp(cx_j)} \ge \max_{j \in \{1,\ldots,m\}} x_j - \frac{m}{ec}.$$

Proof:

Let x^* be the maximum of x. Now:

$$\frac{\sum_{j=1}^{m} \exp(cx_j)(x^* - x_j)}{\sum_{j=1}^{m} \exp(cx_j)} \le \frac{\sum_{j=1}^{m} \exp(cx_j)(x^* - x_j)}{\exp(cx^*)} = \sum_{j=1}^{m} \exp(-c(x^* - x_j))(x^* - x_j)$$

For any y, we have $y \ge \log y + 1$. Therefore for any y, $cy \ge \log(cy) + 1 = \log y + \log c + 1 \Leftrightarrow \log y - cy \le -\log c - 1$. By exponentiating both sides, $y \exp(-cy) \le \frac{1}{ec}$.

Applying this to the term from the previous inequality:

$$\sum_{j=1}^{m} \exp(-c(x^* - x_j))(x^* - x_j) \le \sum_{j=1}^{m} \frac{1}{ec} = \frac{m}{ec}$$

At this point we have

$$\frac{\sum_{j=1}^{m} \exp(cx_j)x_j}{\sum_{j=1}^{m} \exp(cx_j)} = x^* - \frac{\sum_{j=1}^{m} \exp(cx_j)(x^* - x_j)}{\sum_{j=1}^{m} \exp(cx_j)} \ge x^* - \frac{m}{ec}.$$

□

At this point the implication is straightforward. Since $g(u, R)$ is distributed according to $f_j(u)$, we have

$$E[u(a^{g(u,R)})] = \sum_{j=1}^{m} f_j(u)u(a^j) = \frac{\sum_{j=1}^{m} \exp(cu(a^j))u(a^j)}{\sum_{j=1}^{m} \exp(cu(a^j))} \geq u_{max} - \frac{m}{ec} .$$

Proving the result from these facts

At this point we have:

- Players play according to $a^{g(u,R)}$ with probability at least $1 - nm(\exp(2c\epsilon) - 1)$.
- $E[u(a^{g(u,R)})] \geq u_{max} - \frac{m}{ec}$.

The first fact lets us quantify the expected difference between $u(a^{g(u,R)})$ and $u(\pi_1(V_i, R), \ldots, \pi_n(V_n, R))$. Since these random variables are equal with probability at least $1 - nm(\exp(2c\epsilon) - 1)$, and their difference is bounded by $u_{max} - u_{min}$, their expected difference is at most $nm(\exp(2c\epsilon) - 1)(u_{max} - u_{min})$.

Combining this with the second inequality:

$$E[u(\pi_1(V_1, R), \ldots, \pi_n(V_n), R)] \geq u_{max} - \frac{m}{ec} - nm(\exp(2c\epsilon) - 1)(u_{max} - u_{min})$$

To minimize $\frac{m}{ec} + nm(\exp(2c\epsilon) - 1)(u_{max} - u_{min})$, set

$$c := \frac{1}{2\sqrt{en\epsilon(u_{max} - u_{min})}}$$

This yields:

$$\frac{m}{ec} = 2m\sqrt{\frac{n\epsilon(u_{max} - u_{min})}{e}}$$

Now note that

$$2c\epsilon = \sqrt{\frac{\epsilon}{en(u_{max} - u_{min})}}$$

Due to the fact that $\epsilon \leq u_{max} - u_{min}$ and $n \geq 0$,

$2c\epsilon \leq 1$.

Because for any $0 \leq x \leq 1$, $\exp(x) - 1 \leq 2x$, we have

$$nm(\exp(2c\epsilon) - 1)(u_{max} - u_{min}) \leq 4nmc\epsilon(u_{max} - u_{min}) = 2m\sqrt{\frac{n\epsilon(u_{max} - u_{min})}{e}}$$

By combining inequalities and equalities so far:

$$E[u(\pi_1(V_1, R), \dots, \pi_n(V_n, R))] \geq u_{max} - 4m\sqrt{\frac{n\epsilon(u_{max} - u_{min})}{e}}.$$

The expected suboptimality goes to 0 as ϵ approaches 0.

Saving the world in 80 days: Epilogue

80 days ago[1], I started a productivity sprint for being useful in the field of AI alignment. My main goal was to have a better self-model in order to push myself without burning out. I'll separate this into knowledge, emotions, health, and noticing.

Knowledge

I've felt my mind expand more in these past 80 days than any semester in college due to reading these dense, no-fluff textbooks. How To Prove It, Jayne's Probability Theory, & Tao's "Analysis I" have grown me a LOT (I read 4 chapters of Jayne's as recommended in Miri's research guide, and I'm currently on ch. 5 on Tao's).

I've notice a tendency in myself to get impatient and desire to skip the exercises, and that screwed me over when trying to get through Linear Algebra Done Right where I only made it to ch. 4. So, as an exercise for the reader, don't skip the exercises for the reader, or you'll be weak sauce.

Emotions

Everything was going great until 40 days ago: I lost a friend, had to fire a friend, moved, deconverted from Christianity, experienced extreme romantic relationship uncertainty, and just felt lonely. I've read and re-read Valentine's grieving well, and Squirrelinhell's explanation of Gendlin's focusing and more general emotional tuning. They've helped a lot. I'm definitely still dealing with feeling lonely, but I feel like I'm *mostly* through the rest which is reassuring.

I've noticed the tendency to distract myself away from the pain with fiction, but the proper response is to look into the pain, acknowledging that it exists, that it's part of reality, that it's still true even if I distract myself from it.

Health

I used to eat out a lot and get paydays (a candy bar) from the vending machine. I noticed I'd get a payday when work would get tough. It was a form of pica for me like reading fiction is now. ~50 days ago, I started eating meal squares for lunch, and eating an uber large, very fatty salad for dinner. I've felt very satiated, and I've saved money and time despite eating better quality food and preparing dinner. Plus! I can make a mean salad that leaves others green with envy. (Did you read 1 pun, or 2?)

I've also noticed that I drink more and more coffee over a couple weeks until I'm an anxious mess, quit coffee, wait a week, rinse, repeat. Now I drink ~6 cups of decaf green tea instead which fixed that problem!

I've also, also noticed that I have less energy if I haven't danced in a few days. I usually dance right after work, but sometimes work lasts till 6 and that gets skipped. Shifting my evening schedule might help, but also just leaving work at 5 pm would fix it too.

Noticing

Meditating really bootstrapped my ability to notice my immediate reactions such as flinching away from pain. I've mostly done Squirrelinhell's sense your body with extreme clarity, and while doing that I'd notice my back hurting & wanting to stop meditating. I would just switch my focus to the pain in my back and just watch it. The pain would waver for like 5 seconds and then stop. Same with focusing on "wanting to stop meditating". This skill helped a lot with the emotional section, and I'm glad I meditated.

Future work

Studying & meditating will continue to be a focus for me, and I'm going to spend this weekend working out feeling lonely. I'm going to go through Andrew Ng's machine learning course for a week to see if it's a low spoons activity for me (I very much enjoy programming, so I predict it will be).

---

[1] The 80 days is over, and yes, I know, I didn't save the world (darn!). It was mainly a play on "Around the world in 80 days", lol.

Putting Logarithmic-Quality Scales On Time

From my journal. Originally posted on my personal blog. Status: quite speculative, but there's something here.

***

Hmm.

We could probably put a -5 to +5 scale of behavior together that was logarithmic about the enduring good/bad impact of various activities.

Something totally neutral — say, neutral leisure that’s not particularly recharging nor distracting — that might be 0.

Activities or time that were +1 would be very slight gains without much enduring impact, like slightly-recharging leisure. +2 might be things with slightly more impact, like doing chores or cleaning.

The thing about logarithmic scales is that they get big, fast. +3 might be good one-off business-building activities, +4 might be establishing an ongoing revenue channel that keeps outputting gains, and +5 might be an absolute game-changing type thing like getting workable protocols on a vastly untapped marketing channel like Google Adwords in its very early days (with the potential for hundreds-of-thousands to millions of dollars in profit if done right) or recruiting an absolute superstar employee or some such.

On the negative side, -1 would be lost time without much repercussion (maybe a slow start to the day kind of staring off into space), -2 would be slight ongoing negative consequences like starting the day surfing the internet, -3 would be starting to take multi-day damage with bad decisions, -4 would be doing seriously dumb stuff with very long-term implications — for instance, stubbornly training through an injury and turning it into multi-months of physical therapy required, and -5 would be multi-year damage events like a serious alcoholic or heroin addict relapsing onto booze or heroin respectively.

Logarithmic scales are notoriously hard for people to get their mind around, but actually map well to reality in some instances.

For instance, if you spent a half-hour at “-1” quality time (-30 total) and then two hours at “-2” quality time (-1200 total), but then 90 minutes on a good one-off revenue generating campaign at +3 (+9000), then you come out considerably ahead. Staring off into space or surfing the internet is of course wasteful, but a solidly inspired 90 minutes of revenue-generating is much better than that. It’s possible to get a lot of gains in 90 focused minutes.

-5: -10,000 points per minute
-4: -1,000 points per minute
-3: -100 points per minute
-2: -10 points per minute
-1: -1 point per minute

0: 0 points per minute
+1: +1 point per minute
+2: +10 points per minute
+3: +100 points per minute
+4: +1,000 points per minute
+5: +10,000 points per minute

That might be too steep in some cases, and not steep enough in other cases; pseudo-logarithmic might map to reality better, but I think *something* like this actually maps to the reality of “great days”, “okay days”, and “bad days.”

A day you spent the whole day cleaning and doing chores, say 10 hours of +2 time, would be “6,000 points” — you’d have been better off in some respects by doing just an hour or two of higher-tier work that day, but +3/+4/+5 time isn’t always easy to come by.

At the extremes you can see some of the potential correctness shake out. A day that you did a ton of one-off good sales and prospecting, even 10 hours at +3 time, you’d get “60,000 points” — very solid, but if that was followed up with just a few hours of destructive behavior it’d give up some of the gains. A short, way-too-hard drinking session to celebrate — to someone was a mildly irresponsible drinker but not quite a relapsing alcohol — might come in as four hours of “-3 time” and -24,000, suppressing the next day’s performance and having a slight carryover for a few days and giving up some of the day’s gains.

A day of lots of scratching and grinding and distraction (say, 10 hours split between -2 or -3 time, for -3300 total) followed by just a couple hours of serious breakthrough +4 time on systems or ops or revenue or whatever (120,000) actually comes out to a great day. Of course, it’s rare to put in breakthrough time after a totally scattered and wasteful start to a day, but it does happen — and those days, in retrospect, are like, “Huh, that’s a weird day but it went well.”

It also explains — doesn’t justify, but does explain — how people doing extremely large-scale business or culture can oscillate rapidly between massively self-destructive behavior and massively successful behavior and seem to get away with it for long stretches of time — Jim Morrison (1943-1971), the lead singer of The Doors, certainly put in the musician’s equivalent of lots of +4 and +5 behavior on making art and music and popularizing it and really connecting with people… but even short extended runs of -5 time can end your life, as it did in his case.

Of course, when young people who aren’t celebrities or business moguls emulate the more destructive parts of that scene, the -5 time doesn’t get bulwarked by +5 time in business or art, and you wind up not with the legacy of a tortured genius like Morrison, but with… well, you ruin your life.

Are there practical applications here? Well, not yet. I’m trying to find some way to capture the wildly-varying quality of how time is spent, trying to make more of a clear distinction between *going well* and *going amazing*, etc.

Days with no negative time at all and some slight positives are decent. More negative time can be absorbed at lower levels if compensated with more time in the higher-performing categories. At the high ends of the extreme, you could even maybe get away with significant -4 and -5 time if you’re doing massively-impactful positive things, but if you can get massively impactful positive time without getting stuck in high-negative time — like, say, Gene Simmons in music or Jeff Bezos in business — then you wind up building a really consistently exceptional life.

Logarithmic scales are hard to calculate intuitively; when 10 minutes of +2 time is roughly equivalent to two hours of +1 time, that's unintuitive. A single hour or two of +5 time in some cases might be more important than everything else you do *that month*, so long as you don't give up all the gains with -4 and -5 time.

This is still very much a work-in-progress, and to have any chance of viability, at least three things would need to be true —

- (1) The weighing of logarithmic-time-quality would have to be calculated automatically in a spreadsheet or an app,
- (2) There'd need to be a clear and easy delineation of behaviors that are very easy to classify as +1, +2, +3, etc.
- (3) Recording and tracking these numbers would need to be as easy as possible, so it wasn't a chore or overwhelming.

And the scales might be wrong — maybe pseudo-logarithmic, or with each tier being double the level below it, or some such. It's certainly the case that there's a very few activities that we can occasionally identify that are more than 1000x as impactful as normal time — think of Einstein in the Swiss patent office, compared to how most patent clerks would have spent their time — likewise, there's a limited range of very destructive behavior that gives up all the gains in life, that's 1000x worse than normal distraction or foolishness.

The more of life can be deployed into those highest-level tiers of ongoing positive impact, and the more the negative tiers can at least be ratcheted down to the manageable -1 and -2 time which is bad but not enduringly bad, it seems like this becomes a sort of recipe for life. A high-earning person on salary who spends only 20% of what they make and conservatively invests the difference — going the "early retirement extreme" path — has a mix of lots of +3 and +4 time, and probably doesn't need much +5 time to hit their targets. The disruptive entrepreneur or pioneering artist or inventor, on the other hand, hunts for +4 and +5 time, often at the expense of nights of low sleep, long days, occasional thrashing around.

Rough thoughts — still a work in progress. I think there's definitely something here, though.

Figuring out what Alice wants, part II

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

This post continues the analysis started in the previous post. Here I will present some examples of algorithms operating in certain environments, seeing how the model certain things, and using that to conclude facts about their preferences/goals/reward functions.

I'll be looking first at the Poker problem with unknown motivations presented here, and secondly at a variant of the Codenames game.

Different algorithms, same outputs, different goals

In the Poker example, we are unsure whether Alice wants to win the hand against Bob, for money, or lose the hand to get into Bob's good graces. She herself is unsure what Bob's cards are; Bob has been playing confidently, but there is only one card combination that would allow him to win.

```
Alice Poker algorithm
1 :  Parameters : $a$, ratio of relative heuristic importance.
2 :  Input : $Alice\_cards, board, Bob\_behave$
3 :  $P\_win = (a)h\_1(Alice\_cards, board) + (1 - a)h\_2(Bob\_behave)$
4 :  if $P\_win > 0.5$
5 :      return 'call'
6 :  else
7 :      return 'fold'
8 :  end if
```

The inputs are Alice's own cards, the five cards on the board, and Bob's behaviour this hand. There are two heuristics called by the algorithm, h_1, which computes the probability of Alice winning by assuming Bob has a random hand, and h_2, which assesses the likelihood of Alice winning by looking at Bob's behaviour (at this point you should start worrying about the suggestive names and descriptions I'm giving to all these elements).

Now, in the situation we find ourselves, what we want to say is that if a is close to 1, then h_1 dominates, and P_win will be high. If a is close to 0, then it will be low, as h_2 dominates the expression. Since there is a > in line 4, we want to say that this Alice Poker algorithm is trying to win, but, depending on the value of a, has different beliefs about what action is likely to maximise expected money.

Similarly, if it were a < in line 4, we'd like to say that the Alice Poker algorithm wants to lose. And this, even though the $(<, a = 1)$ and $(>, a = 0)$ algorithms would both fold (while $(<, a = 0)$ and $(>, a = 1)$ would both call).

But this relies too much on the interpretation of what the terms *mean* and what the heuristics *are meant to do*. Now, there isn't much flexibility in interpreting what h_1 is or does: as the quality of Alice's hand increases, relative to a random hand and the given board, h_1's output increases. Thus is it 'clearly' measuring relative hand quality.

But what of h_2? All that we know is that this outputs a number that increases when Bob appears confident, and decreases when he appears worried. I've said that this is supposed to measure how good Bob's hand is, based on how he behaved. But how do we know that? Maybe Alice views Bob as an effective bluffer (or a level 2n+1 meta-bluffer), so that a high h_2 actually means that she expects Bob to have a poor hand. In that case the $(>, a = 0)$ would still fold, but would be folding to lose against Bob, not folding to win.

This brings us back to some of the oldest problems in AI (and some of the newest). Namely, what is the semantics and interpretation of what an algorithm is doing? It's been argued that you cannot derive the semantics of an algorithm, only the syntax. I've disagreed with that; arguing that when the internal syntax is sufficiently rich and detailed, and the agent relates well to the real world, then there can be only a a few semantic interpretations of the symbols that make any sense. In more modern terms, this can be seen as the problem of algorithm interpretability, especially when it is applied to out-of-training-set distributions. If a certain neuron triggers when seeing photos of dogs, including photos far from its training distribution, then that neuron is at least superficially connected to the concept of "dog" (or at least "dog photo").

So, what would cause us to believe that h_2 is actually trying to predict Bob's cards from his reactions?

Well, suppose that h_2 was a learning algorithm, and it updated the correct way: when Bob is revealed to have had a good hand, it updates towards a higher value on that input set, and vice versa. Or suppose that h_2 also took Alice_cards and board as inputs, and was higher if Alice's cards were better. Then it would seem more justified to see that heuristic as actually trying to estimate the probability of Alice's cards beating Bob's.

Codenames and semantics

In Codenames, one player on a team (Spymaster) tries to signal a card (or a collection of cards) to their teammates. These cards have words on them, and the Spymaster names another word related to the targets. The Spymaster also gives a number, to say how many cards this word refers to.

Suppose there are four remaining codewords, "Dolphin", "New York", "Comet", and "Lice". The Spymaster is Bob, and has only one card remaining to signal; the other player is Alice, who has to interpret Bob's signal.

Note that before receiving Bob's word, the Alice algorithm probably doesn't have time to run through all the possible words he could give. Therefore, in terms of the first post, in the model fragment Alice has of Bob, she expects to be surprised by his action (it's mainly for this reason that I'm introducing this example).

Anyway, after receiving Bob's message - which we will assume is "Aquatic, 1 card" - Alice runs the following algorithm:

```
Alice  Codewords algorithm
Input: $codewords, Bob\_wordmap, Bob\_word$
$answer = \arg\min(\text{for } w \text{ in } codewords)\, Bob\_wordmap(w, Bob\_word)$
return $answer$
```

The algorithm is very simple: it models Bob as having a wordmap in his head, that measures the distance between words and concepts, and assumes that the word he speaks - Bob_word - is closest to the answer among the remaining codewords.

Let's assume that, through saying "Aquatic", Bob is trying to signal "Dolphin". But let's assume that Alice's model of Bob's wordmap is wrong - it computes that "New York" is the closest word to "Aquatic" (New York is a coastal city, after all).

Let's further assume that Alice has a learning component to her algorithm, so Bob_wordmap gets updated. But the updating is terrible: it actually moves further away from the true wordmap with every data point.

In view of this, can we still say that Alice is trying to win the game, but just has a terrible model of her teammate?

Perhaps. The data above is not enough to say that - if we renamed Bob_wordmap as Negative_Bob_wordmap it would seem to be more fitting. But suppose that Bob_wordmap was not a single algorithm, but a sub-component of a much more complicated routine called Bob_model. And suppose that this component was decently connected to Bob - Alice used it whenever she thinks about what Bob does, it is activated whenever she sees Bob, and it is closely semantically connected to her experience of interacting with Bob.

Then we might be able to say that, though Bob_model(wordmap) is actually a terrible version of Bob's mental wordmap, the whole Bob_model is sufficiently clearly a model of Bob, and wordmap, in other contexts, is sufficiently clearly a wordmap, that we can say that Alice is trying and failing to model Bob's thinking (rather than deliberately failing the game).

These are the sort of "interpretability" analyses that we will have to do to figure out what the algorithms in our brains actually want.

fMRI semantics

A final, minor point: things like fMRIs can help bridge the semantic-syntax gap in humans, as it can, to some extent, literally *see* how certain concepts, ideas, or images are handled by the brain. This could be the human equivalent of having the algorithm laid out as above.

The Fermi Paradox: What did Sandberg, Drexler and Ord Really Dissolve?

(Cross-posted from my blog)

So this paper by the trio from the FHI, Anders Sandberg, Eric Drexler and Toby Ord (SDO for short) has been talked about quite a bit, on LessWrong, on SSC and on Reddit. It is about how their Monte-Carlo calculations based on probability distributions rather than on the usual point estimates of the Drake equation apparently dissolves the question of why we are seemingly alone in the Universe that is supposed to be teeming with intelligent life, if one takes the Copernican idea "we are not special" seriously. One grim suggestion is that there is a Great Filter that is still in front of us and is almost guaranteed to kill us off before the human civilization reaches the technological levels observable by other civilizations like ours.

There are plenty of other ideas, most addressing various factors in the Drake equation, but the one advanced in SDO is quite different from the mainstream: they say that estimating these factors is a wrong way to go, because the uncertainty in these very small probabilities is so large, the point estimates are all but meaningless. Instead they suggest that the correct approach is something along the following lines: First, assume a reasonable probability distribution for each factor, then draw a value for each factor based on their probability distributions, calculate the resulting expected value of the number of currently detectable civilizations, then repeat this process many times to create a synthetic probability distribution of the number of this civilizations, and finally extract the odds of us being alone in the universe from this distribution. And that is what they did, and concluded that the odds of us being alone in the Milky Way are something like 1:3. Thus, according to SDO, there is no paradox, an average universe is naturally a desolate place.

Their to the Fermi paradox solution is basically

Due to random chance, some of the parameters in the Drake equation, we do not know which, we do not know why, are many orders of magnitude smaller in our universe than previously estimated.

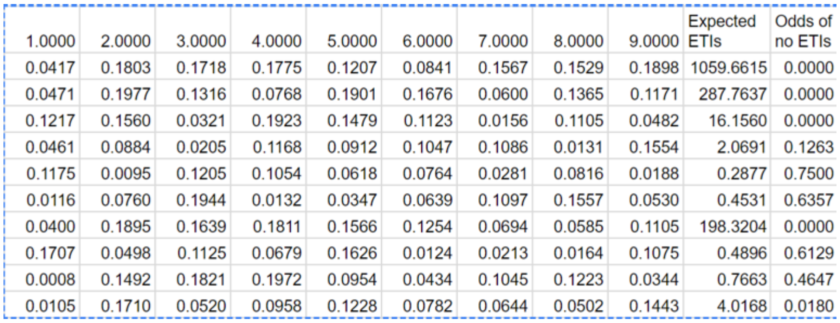
I have an issue with calling it "dissolving the paradox", since it doesn't answer any of the practical questions about the universe we live in. Fermi's question, "Where is everybody?" remains open.

But I may have misunderstood the paper. So, what follows is an attempt to understand their logic by reproducing it. Here I will analyze their toy model, as it has all the salient features leading to their conclusion:

> There are nine parameters (f1, f2, . . .) multiplied together to give the probability of ETI [Extra-terrestrial intelligence] arising at each star. Suppose that our true state of knowledge is that each parameter could lie anywhere in the interval [0, 0.2], with our uncertainty being uniform across this interval, and being uncorrelated between parameters.

A point estimate would be taking a mean for each factor and multiplying them, giving one-in-a-billion chance per star, which results in a *virtual certainty of ETI at least somewhere in a galaxy of hundreds of billions of stars*.

Let's try a distribution estimate instead: take 9 random numbers from a uniform distribution over the interval [0, 0.2]. Here is a bunch of sample runs, the expected values of ETIs in the toy galaxy, and the odds of the toy galaxy being empty:



1.0000	2.0000	3.0000	4.0000	5.0000	6.0000	7.0000	8.0000	9.0000	Expected ETIs	Odds of no ETIs
0.0417	0.1803	0.1718	0.1775	0.1207	0.0841	0.1567	0.1529	0.1898	1059.6615	0.0000
0.0471	0.1977	0.1316	0.0768	0.1901	0.1676	0.0600	0.1365	0.1171	287.7637	0.0000
0.1217	0.1560	0.0321	0.1923	0.1479	0.1123	0.0156	0.1105	0.0482	16.1560	0.0000
0.0461	0.0884	0.0205	0.1168	0.0912	0.1047	0.1086	0.0131	0.1554	2.0691	0.1263
0.1175	0.0095	0.1205	0.1054	0.0618	0.0764	0.0281	0.0816	0.0188	0.2877	0.7500
0.0116	0.0760	0.1944	0.0132	0.0347	0.0639	0.1097	0.1557	0.0530	0.4531	0.6357
0.0400	0.1895	0.1639	0.1811	0.1566	0.1254	0.0694	0.0585	0.1105	198.3204	0.0000
0.1707	0.0498	0.1125	0.0679	0.1626	0.0124	0.0213	0.0164	0.1075	0.4896	0.6129
0.0008	0.1492	0.1821	0.1972	0.0954	0.0434	0.1045	0.1223	0.0344	0.7663	0.4647
0.0105	0.1710	0.0520	0.0958	0.1228	0.0782	0.0644	0.0502	0.1443	4.0168	0.0180

Notice that, out of the 10 sample galaxies in this example, about half show considerable odds of being empty, highlighting the difference between using a distribution and the point estimates. SDO likely had run a lot more simulations to get more accuracy, and I did, too. Here is the distribution of the odds for a given number of ETIs in the galaxy, based on about ten million runs:

The horizontal axis is the number of ETIs, and the vertical axis is the probability of this number of ETIs to happen in a given toy galaxy drawn at random, given the probability distribution for each factor, as specified above.

Notice how wide this distribution is: some galaxies are completely bereft of ETIs, while others have hundreds and even thousands of them!

The above graph is very close to the power law: the probability of the number of ETIs goes roughly as the number of ETIs to the power of -1.2. This means that the estimate of the expected number of ETIs is actually *divergent*, though the median number of ETI's per galaxy is finite and about 30. So, somewhat paradoxically,

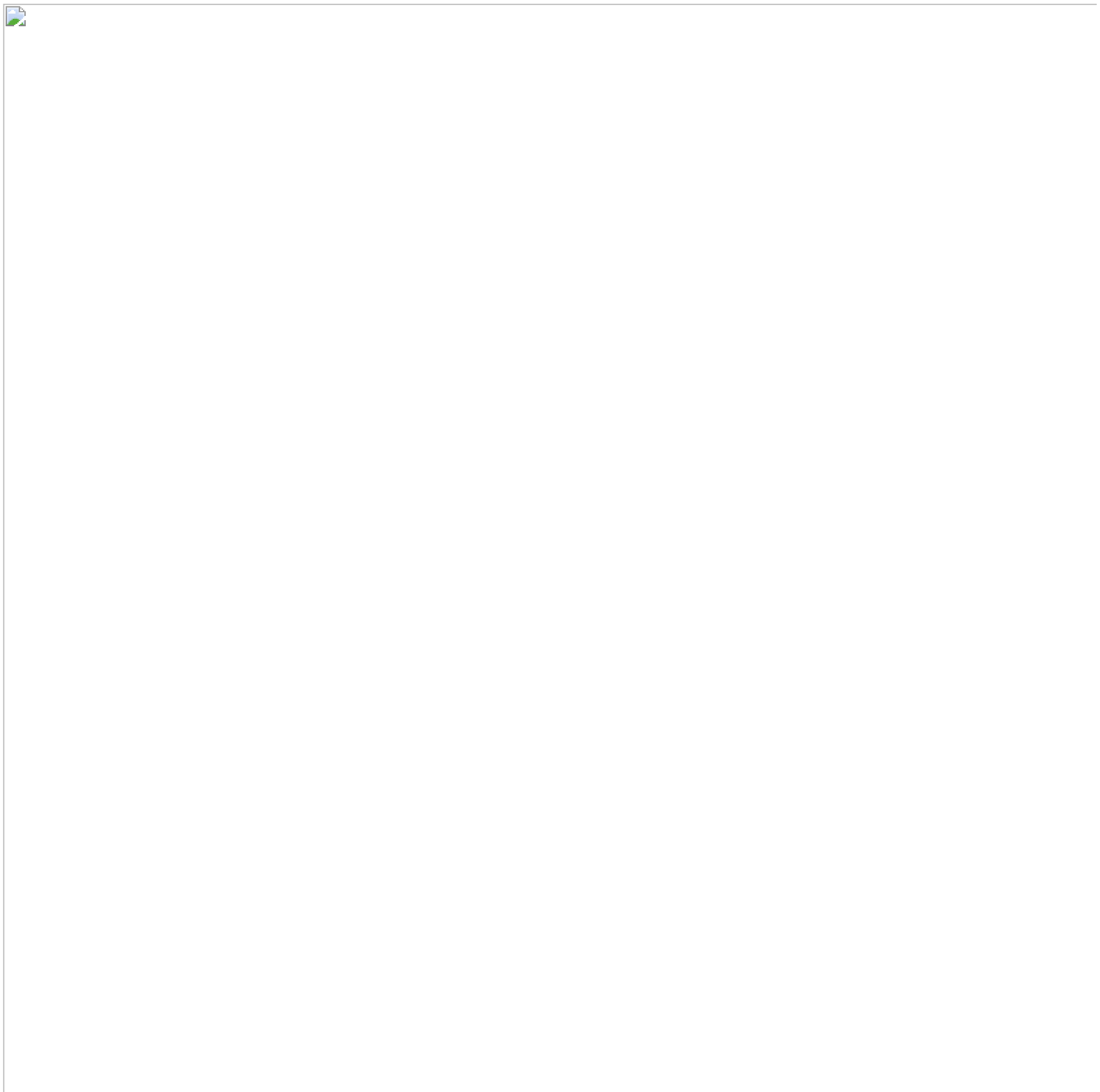
Using the distributional estimate instead of a point estimate both increases the likely number of ETIs per galaxy, and the fraction of empty galaxies.

The paper states that **"*Monte Carlo simulation shows that this actually produces an empty galaxy 21.45% of the time,*"** but does not specify the way this number was calculated. From the plot above, the probability of having between zero and one ETIs, whatever it might mean, is 24.63%. A different way of calculating the average odds of an empty toy galaxy would be to calculate the odds of each sample galaxy to be empty as

P(empty) = (1-p_1*p_2*...*p_9)^(10^11)

then average all these odds. This approach gives the fraction of empty galaxies as ***21.43%***, very close to the number quoted in the paper.

This toy example demonstrates nicely the main result of the SDO paper: ***the usual point estimate produces a wildly inaccurate expected fraction of galaxies with no ETIs in them***. The rest of the SDO paper is focused on calculating a more realistic example, based on the current best guesses for the distributions of each factor in the Drake equation:



Using these distributions and their further refinements to calculate the odds yield the following conclusion:

> When we update this prior in light of the Fermi observation, we find a substantial probability that we are alone in our galaxy, and perhaps even in our observable universe (53%–99.6% and 39%–85% respectively). 'Where are they?' — probably extremely far away, and quite possibly beyond the cosmological horizon and forever unreachable.

If our universe is drawn at random from the pool of factors with the distributions as suggested by the paper, then there are substantial odds that we are alone in the observable universe.

This is because some of the factors in the Drake equation, due to their uncertainty, can end up many orders of magnitude smaller than the value used for a point estimate, the one where this uncertainty is not taken into account. There is no claim which of the parameters are that small, since they differ for different desolate universes. Just the (bad) luck of the draw.

So, the Fermi paradox has been solved, right? Well, yes and no. Once we draw a set of parameters to use in the Drake equation in one specific universe, for example ours, we are still left with the task of explaining their values. We are no closer to understanding the Great Filter, if any, than before. Is abiogenesis extremely rare? Do ETIs self-destruct quickly? Are there some special circumstances required for life to thrive beyond the planet being in the Goldilocks zone? Is there the singleton effect where the first civilization out of the gate takes over the galaxy? Who knows. SDO concludes that

> This result dissolves the Fermi paradox, and in doing so removes any need to invoke speculative mechanisms by which civilizations would inevitably fail to have observable effects upon the universe.

I find that this conclusion does not follow from the main result of the paper. We live in this one universe, and we are stuck with the specific set of values of the factors in the Drake equation that our universe happened to have. It is quite possible that *in our* *universe* there is a Great Filter "by which civilizations would inevitably fail to have observable effects upon the universe," because, for example one specific parameter has the value that is many orders of magnitude lower than the estimate, and it would be really useful to know which one and why. The error SDO is making is looking at the distribution of the universes, whereas the Fermi paradox applies to the one we are stuck with. A real resolution of the paradox would be, for example, determining which parameters in the Drake equation are vanishingly low and why, not simply declaring that it is exceedingly likely that a randomly drawn universe has one or several vanishingly small parameters leading to it being very likely bereft of ETIs.

Announcing AlignmentForum.org Beta

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

We've just launched the beta for AlignmentForum.org.

Much of the value of LessWrong has come from the development of technical research on AI Alignment. In particular, having those discussions be in an accessible place has allowed newcomers to get up to speed and involved. But the alignment research community has at least some needs that are best met with a semi-private forum.

For the past few years, agentfoundations.org has served as a space for highly technical discussion of AI safety. But some aspects of the site design have made it a bit difficult to maintain, and harder to onboard new researchers. Meanwhile, as the AI landscape has shifted, it seemed valuable to expand the scope of the site. Agent Foundations is one particular paradigm with respect to AGI alignment, and it seemed important for researchers in other paradigms to be in communication with each other.

So for several months, the LessWrong and AgentFoundations teams have been discussing the possibility of using the LW codebase as the basis for a new alignment forum. Over the past couple weeks we've gotten ready for a closed beta test, both to iron out bugs and (more importantly) get feedback from researchers on whether the overall approach makes sense.

The current features of the Alignment Forum (subject to change) are:

- A small number of admins can invite new members, granting them posting and commenting permissions. This will be the case during the beta - the exact mechanism of curation after launch is still under discussion.
- When a researcher posts on AlignmentForum, the post is shared with LessWrong. On LessWrong, anyone can comment. On AlignmentForum, only AF members can comment. (AF comments are *also* crossposted to LW). The intent is for AF members to have a focused, technical discussion, while still allowing newcomers to LessWrong to see and discuss what's going on.
- AlignmentForum posts and comments on LW will be marked as such.
- AF members will have a separate karma total for AlignmentForum (so AF karma will more closely represent what technical researchers think about a given topic).
- On AlignmentForum, only AF Karma is visible. (note: not currently implemented but will be by end of day)
- On LessWrong, AF Karma will be displayed (smaller) alongside regular karma.
- If a commenter on LessWrong is making particularly good contributions to an AF discussion, an AF Admin can tag the comment as an AF comment, which will be visible on the AlignmentForum. The LessWrong user will then have *voting privileges* (but not necessarily posting privileges), allowing them to start to accrue AF karma, and to vote on AF comments and threads.

We’ve currently copied over some LessWrong posts that seemed like a good fit, and invited a few people to write posts today. (These don’t necessarily represent the longterm vision of the site, but seemed like a good way to begin the beta test)

This is a fairly major experiment, and we’re interested in feedback both from AI alignment researchers (who we’ll be reaching out to more individually in the next two weeks) and LessWrong users, about the overall approach and the integration with LessWrong.

Opinion Article Against Measuring Impact

This is a linkpost for https://www.theguardian.com/global-development/2018/jul/16/buzzwords-crazes-broken-aid-system-poverty

A strange opinion article in The Guardian today: it is not entirely clear whether the authors object to a concern with effectiveness, or just think that "assessing the short-term impacts of micro-projects" is somehow misguided (and if so, why that is).

How to parent more predictably

(As with anything else I write about parenting, this is mostly just my observations with my own two kids and may not generalize as well as I think it does.)

A few weeks ago I wrote about why I think it's valuable for parents to be predictable. I was mostly describing an end state, though, and reasons why it's a good place to be, but what should you do if you're interested in being more consistent?

I think it breaks down into three pretty different skills. The first is only saying things you're comfortable fully standing behind. Many people especially when irritable, sleep deprived, or surprised, can quickly spit out a consequence that is harsher than is fair. Things like, "if you don't stop this minute it's no desserts for a week!" Then when the threat fails the consequence feels too extreme, and they don't follow through. If someone read my earlier post and decided they needed to stand by their quickly distributed threats I'd be pretty sad, and not expect this to make things better. Instead, any consequences need to be reasonable and proportionate, where you feel fair when you do have to enforce them (which, ideally, is rarely).

One thing that can be helpful is having a system of simple stock responses. The simplest is just a stern voice. If you reserve speaking firmly for rare circumstances, then saying something intensely and seriously can feel very significant.

Time outs can also work well: we use a system where we make it clear what needs to change ("stop shouting", "brush your teeth"), count to three, slowly and clearly, stopping if they do what we want, and if we get to three then it's time out. Serious things can be time out immediately. If you reliably use these, then not only does the kid understand how they work and what they mean but you don't have to come up with a novel appropriate response at a time when you're very stressed ("Do not bite me. That's time out.)

Even better, though, is avoiding commands and threats entirely. If they've been playing with a puzzle and they ask you to read to them, you can say you'll read once they clean up the puzzle (and depending on age maybe help them clean it up). They probably want you to read badly enough that they clean up, but if not that's ok too. You're using something they want you to do as leverage to get them do something they should do, but you're also teaching them the general practice of cleaning up one thing before getting out another.

Similarly, don't make committments unnecessarily. Instead of "I'll go downstairs and get your bear" maybe "I'll go downstairs and look for your bear." While with adults we understand that when a person says they'll do something they mean they'll put in a reasonable effort and may fail if the task is surprisingly difficult or if factors outside their control intervene, I find that with kids being explicit about likely failure possibilities is helpful. "I'll go see if we have any more cheese sticks." Similarly, adults understand that you don't have authority over everyone around you, but with kids phrasing like "Mama can read to you when we get home" isn't as good as "I'll ask Mama if she'll read to you when we get home." Or, even better, "when we get home you can ask Mama if she'll read to you." [1]

The next skill is thinking of rules that will work well in a range of situations. These can be explicit rules ("hold hands when crossing the street) or implicit ones (saying yes/no without explaining reasoning but following a pattern you could explain if asked). I find thinking "saying yes/no no means saying the same thing in this sort of situation in general" is a helpful way of thinking about it. [2]

This is another case where preparation can help. Before going to the beach, think about what rule you'll have for the water: "stay with a grown up the whole time", "only go in the water with a grown up", "only go in the water by yourself right in front of the lifeguard." If you notice yourself giving inconsistent answers to questions, take some time later to figure out if there's a rule that would work better ("one short video each night, after dinner"). It's ok, and expected, to end up with a rule that doesn't match some of your previous decisions, just try to get a rule you'll be happy with. ("I know we did let you stand on the table sometimes, but from now on there's no standing on the table.")

The easiest rules to be consistent with are the ones our culture has for adults, since you and everyone else more or less know what these are. As kids are generaly less capable, tall, intelligent, prudent, etc than the people these rules have evolved for, you are going to need to make accomodations. But I find that starting with a perspective of treating kids like adults unless there is a reason otherwise works well.

The third skill is actually following through on things you've said. The better you get at the two skills above the easier this is: you're saying fewer things that need to be backed up, the consequences you've promised are ones that generally seem fair, and your kids generally understand the patterns and choose to avoid the consequences. But they will still test to see if maybe the boundary has retracted while they weren't looking, and when they do you need to be firm. Sometimes this is unpleasant for both of you but it makes the rest of your interactions far better. [3]

(One way that poverty is harmful, and that parenting while well off is unfairly easy, is that external factors can keep you from being reliable for your kids. If I tell them I will do something I have enough control that I can make sure it happens, but if I had work with less flexibility, less money, or generally less slack this would be much harder. Similarly, it's much harder to be predictable when you're tired, hungry, overworked, or otherwise not at your best. If I tell my kid no and they decide to test me by tantruming, I'm going to be able to be stubborn longer than they can. But I'm only able to be in a good mental state for that because I've been lucky in how well my life has gone.)

Cross posted from jefftk.com.

[1] The first version, "Mama can read to you when we get home," has a different downside which is that one parent is making a promise on behalf of another. We try very hard not to do this, and being careful about it generally makes our interactions a lot better.

[2] This is also useful for yourself. Instead of "do I want to skip exercising this morning because I'm sleepy" it works better to ask "do I want to skip exercising every morning where I'm this sleepy." Sometimes the answer is yes, often it's no, but it's much more likely to be a decision that looking back you'll think was the right one for you to make.

[3] I've phrased this as about verbal consistency, but it works even when the kid is too young for words. Sleep training, for example, generally needs a bunch of thinking "I know they're really sad right now and short term would like a cuddle, but we will both be much happier long term if they learn how to sleep on their own."

Solving the AI Race Finalists

This is a linkpost for https://medium.com/goodai-news/solving-the-ai-race-finalists-15-000-of-prizes-5f57d1f6a45f

Good AI offered a prize for writing about AI races. The results are in and here are the winners:

Top scoring solutions ($3,000 each)

- **Kesavan Athimoolam,** Solving the Artificial Intelligence Race: Mitigating the problems associated with the AI Race
- **Alexey Turchin and & David Denkenberger,** Classification of Global Solutions for the AI Safety Problem
- **Ehrik L. Aldana,** A Theory of International AI Coordination: Strategic implications of perceived benefits, harms,capacities, and distribution in AI development

Runners-up ($2,000 each)

- **David Klimek,** Framework for managing risks related to emergence of AI/AGI
- **Gordon Worley,** Avoiding AGI Races Through Self-Regulation
- **Morris Stuttard & Anastasia Slabukho,** The AI Engineers’ Guild: proposal for an AI risk mitigation strategy

Mechanism Design for AI

This is a linkpost for http://s-risks.org/mechanism-design-for-ai/

Bayesianism (Subjective or Objective)

I'm reading a paper called 'Reasonable Doubt and Presumtion of Innocence: The Case of the Bayesian Juror' for a Physics/Policy course I'm taking, and am a bit confused by something in it. Note here that I'm quite new to Bayesianism and do not claim to understand in entirity how it all works.

The claim made is that in pure Bayesianism, all probabilities are subjective (a probability of *you*). As I had understood from initial readings on Bayesianism, it is supposed to be entirely objective (ie you look at the thing you want to determine the probability of, you look at the evidence you have available, and you thusly determine the probability of the thing). As I understand it, this makes Bayesianism objective, at least within the scope of the Bayesian's knowledge.

Is my understanding wrong somewhere? Could some kind and enlightened souls please explain this to me?

Repeated (and improved) Sleeping Beauty problem

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

Follow up to: Probability is fake, frequency is real

There is something wrong with the normal formulation of the Sleeping Beauty problem. More precisely, there is something wrong about postulating a single "fair" random coin flip. So here is an improved version of the Sleeping Beauty problem. After explaining the setup, I will recover the normal Sleeping Beauty problem, but in a more well defined way.

There are no truly random coins. There are only pseudo random coins which has the property that you don't have the capacity to calculate the outcome. A **fair** pseudo random coin have the additional property that when flipped enough times, the ratio of *Heads* v.s. *Tails* will approach one. Note that fairness is only defined if you actually flip the coin a sufficient number of times. Because of this, the Sleeping Beauty problem should be a repeated game.

(Alternatively, you could solve this by using counterfactuals. However, we don't yet know how to deal with counterfactuals. Also, I suspect that any method of handling counterfactuals will be, at best, useful but wrong.)

Repeated Sleeping Beauty setup: Every Sunday a mysterious person flips a pseudo random fair coin. If the coin comes up *Heads*, Sleeping Beauty will wake up on Monday, and then sleep for the rest of the week. If the coin comes up *Tails*, she will wake up on Monday and Tuesday and then sleep for the rest of the week. No-one is telling Sleeping Beauty what is going on, she gets to rely on her own past experiences.

Every morning when Sleeping Beauty wakes up she does not know what day it is. However there is an easy experiment she can do to find out, namely asking anyone she meets on the street. Because Sleeping is a curious person, she is keeping a science journal. Every day she finds out what day it is and writes it down. She soon notices some patterns.

1) There are two kinds of days, Monday and Tuesday.

2) Every Tuesday is followed by a Monday.

3) A Monday can be followed by either a Tuesday or a Monday.

After some more time she starts to notice the frequencies of which different days occur.

1/3 of days are Mondays that are followed by Monday (corresponds to *Heads* & Monday)

1/3 of days are Mondays that are followed by Tuesday (corresponds to *Tails* & Monday)

1/3 of days are Tuesdays (corresponds to *Tails* & Tuesday)

Sleeping Beauty tries to find more patterns in the data, but none of the more complicated hypothesizes she can come up with survives further observation.

Recovering the original Sleeping Beauty problem: Sleeping Beauty have been slacking off for a few days, and not asking for what day it was. What likelihood should she assign to the current day being a Monday followed by Monday?

The obvious answer based on Sleeping Beauty's own experience is 1/3.

Conclusion and after-though: If you take your probability from how things have played out in the past, you will learn the Thirder position / Self-indication assumption (SIA). Also, doing what has worked well in the past leads to Evidential decision theory (EDT). This is a sad fact of the universe, because EDT combined with SIA leads to a sort of double counting of actions which add up to the wrong policy [citation].

Alignment Newsletter #14

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

I've created a public database of almost all of the papers I've summarized in the Alignment Newsletter! Most of the entries will have all of the data I put in the emails.

Highlights

One-Shot Imitation from Watching Videos *(Tianhe Yu and Chelsea Finn)*: Can we get a robot to learn a task by watching a human do it? This is very different from standard imitation learning. First, we want to do it with a single demonstration, and second, we want to do it by *watching a human* -- that is, we're learning from a video of a human, not a trajectory where the robot actions are given to us. Well, first consider how we could do this if we have demonstrations from a teleoperated robot. In this case, we do actually have demonstrations in the form of trajectories, so normal imitation learning techniques (behavioral cloning in this case) work fine. We can then take this loss function and use it with MAML to learn from a large dataset of tasks and demonstrations how to perform a new task given a single demonstration. But this still requires the demonstration to be collected by teleoperating the robot. What if we want to learn from a video of a human demonstrating? They propose learning a *loss function* that given the human video provides a loss from which gradients can be calculated to update the policy. Note that at training time there are still teleoperation demonstrations, so the hard task of learning how to perform tasks is done then. At test time, the loss function inferred from the human video is primarily used to identify which objects to manipulate.

My opinion: This is cool, it actually works on a real robot, and it deals with the issue that a human and a robot have different action spaces.

Prerequisities: Some form of meta-learning (ideally MAML).

Capture the Flag: the emergence of complex cooperative agents *(Max Jaderberg, Wojciech M. Czarnecki, Iain Dunning et al)*: DeepMind has trained FTW (For The Win) agents that can play Quake III Arena Capture The Flag from raw pixels, given *only* the signal of whether they win or not. They identify three key ideas that enable this -- population based training (instead of self play), learning an internal reward function, and operating at two timescales (enabling better use of memory). Their ablation studies show that all of these are necessary, and in particular it even outperforms population based training with manual reward shaping. The trained agents can cooperate and compete with a wide range of agents (thanks to the population based training), including humans.

But why are these three techniques so useful? This isn't as clear, but I can speculate. Population based training works well because the agents are trained against a diversity of collaborators and opponents, which can fix the issue of instability that afflicts self-play. Operating at two timescales gives the agent a better inductive bias. They say that it enables the agent to use memory more effectively, but my story is that it lets it do something more hierarchical, where the slow RNN makes "plans", while the fast RNN executes on those plans. Learning an internal reward function flummoxed me for a while, it really seemed like that should not outperform manual reward shaping, but then I found out that the internal reward function is computed from the game points screen, not from the full trajectory. This gives it a really strong inductive bias (since the points screen provides really good features for defining reward functions) that allows it to quickly learn an internal reward function that's more effective than manual reward shaping. It's still somewhat surprising, since it's still learning this reward function from the pixels of the points screen (I assume), but more believable.

My opinion: This is quite impressive, since they are learning from the binary win-loss reward signal. I'm surprised that the agents generalized well enough to play alongside humans -- I would have expected that to cause a substantial distributional shift preventing good generalization. They only had 30 agents in their population, so it seems unlikely a priori that this would induce a distribution that included humans. Perhaps Quake III is simple enough strategically that there aren't very many viable strategies, and most strategies are robust to having slightly worse allies? That doesn't seem right though.

DeepMind did a *lot* of different things to analyze what the agents learned and how they are different from humans -- check out the paper for details. For example, they showed that the agents are much better at tagging (shooting) at short ranges, while humans are much better at long ranges.

Technical AI alignment

Technical agendas and prioritization

An introduction to worst-case AI safety *(Tobias Baumann)*: Argues that people with suffering-focused ethics should focus on "worst-case AI safety", which aims to find technical solutions to risks of AIs creating vast amounts of suffering (which would be much worse than extinction).

My opinion: If you have strongly suffering-focused ethics (unlike me), this seems mostly right. The post claims that suffering-focused AI safety should be more tractable than AI alignment, because it focuses on a subset of risks and only tries to minimize them. However, it's not necessarily the case that focusing on a simpler problem makes it easier to solve. It feels easier to me to figure out how to align an AI system to humans, or how to enable human control of an AI system, than to figure out all the ways in which vast suffering could happen, and solve each one individually. You can make an analogy to mathematical proofs and algorithms -- often, you want to try to prove a *stronger* statement than the one you are looking at, because when you use induction or recursion, you can rely on a stronger inductive hypothesis.

Learning human intent

One-Shot Imitation from Watching Videos *(Tianhe Yu and Chelsea Finn)*: Summarized in the highlights!

Learning Montezuma’s Revenge from a Single Demonstration *(Tim Salimans et al)*: Montezuma's Revenge is widely considered to be one of the hardest Atari games to learn, because the reward is so sparse -- it takes many actions to reach the first positive reward, and if you're using random exploration, it will take exponentially many actions (in N, the number of actions till the first reward) to find any reward. A human demonstration should make the exploration problem much easier. In particular, we can start just before the end of the demonstration, and train the RL agent to get as much score as the demonstration. Once it learns that, we can start it at slightly earlier in the demonstration, and do it again. Repeating this, we eventually get an agent that can perform the whole demonstration from start to finish, and it takes time linear in the length of the demonstration. Note that the agent must be able to generalize a little bit to states "around" the human demonstration -- when it takes random actions it will eventually reach a state that is similar to a state it saw earlier, but not exactly the same, and it needs to generalize properly. It turns out that this works for Montezuma's Revenge, but not for other Atari games like Gravitar and Pitfall.

My opinion: Here, the task definition continues to be the reward function, and the human demonstration is used to help the agent effectively optimize the reward function. Such agents are still vulnerable to misspecified reward functions -- in fact, the agent discovers a bug in the emulator that wouldn't have happened if it was trying to imitate the human. I would still expect the agent to be more human-like than one trained with standard RL, since it only learns the environment near the human policy.

Atari Grand Challenge *(Vitaly Kurin)*: This is a website crowdsourcing human demonstrations for Atari games, which means that the dataset will be very noisy, with demonstrations from humans of vastly different skill levels. Perhaps this would be a good dataset to evaluate algorithms that aim to learn from human data?

Beyond Winning and Losing: Modeling Human Motivations and Behaviors Using Inverse Reinforcement Learning *(Baoxiang Wang et al)*: How could you perform IRL without access to a simulator, or a model of the dynamics of the game, or the full human policy (only a set of demonstrations)? In this setting, as long as you have a large dataset of diverse human behavior, you can use Q-learning on the demonstrations to estimate separate Q-function for each feature, and then for a given set of demonstrations you can infer the reward for that set of demonstrations using a linear program that attempts to make all of the human actions optimal given the reward function. They define (manually) five features for World of Warcraft Avatar History (WoWAH) that correspond to different motivations and kinds of human behavior (hence the title of the paper) and infer the weights for those rewards. It isn't really an evaluation because there's no ground truth.

Preventing bad behavior

Overcoming Clinginess in Impact Measures *(TurnTrout)*: In their previous post, TurnTrout proposed a whitelisting approach, that required the AI not to cause side effects not on the whitelist. One criticism was that it made the AI *clingy*, that is, the AI would also prevent any other agents in the world from causing non-whitelisted effects. In this post, they present a solution to the clinginess problem. As long as the AI knows all of the other agents in the environment, and their policies, the AI can be penalized for the *difference* of effects between its behavior, and what the human(s) would have done. There's analysis in a few different circumstances, where it's tricky to get the counterfactuals exactly right. However, this sort of impact measure means that while the AI is punished for causing side effects itself, it *can* manipulate humans to perform those side effects on its behalf with no penalty. This appears to be a tradeoff in the impact measure framework -- either the AI will be clingy, where it prevents humans from causing prohibited side effects, or it could cause the side effects through manipulation of humans.

My opinion: With any impact measure approach, I'm worried that there is no learning of what humans care about. As a result I expect that there will be issues that won't be handled properly (similarly to how we don't expect to be able to write down a human utility function). In the previous post, this manifested as a concern for generalization ability, which I'm still worried about. I think the tradeoff identified in this post is actually a manifestation of this worry -- clinginess happens when your AI overestimates what sorts of side effects humans don't want to happen in general, while manipulation of humans happens when your AI underestimates what side effects humans don't want to happen (though with the restriction that only humans can perform these side effects).

Prerequisities: Worrying about the Vase: Whitelisting

Game theory

Modeling Friends and Foes *(Pedro A. Ortega et al)*: Multiagent scenarios are typically modeled using game theory. However, it is hard to capture the intuitive notions of "adversarial", "neutral" and "friendly" agents using standard game theory terminology. The authors propose that we model the agent and environment as having some prior mixed strategy, and then allow them to "react" by changing the strategies to get a posterior strategy, but with a term in the objective function for the change (as measured by the KL divergence). The sign of the environment's KL divergence term determines whether it is friendly or adversarial, and the magnitude determines the magnitude of friendliness or adversarialness. They show that there are always equilibria, and give an algorithm to compute them. They then show some experiments demonstrating that the notions of "friendly" and "adversarial" they develop actually do lead to behavior that we would intuitively call friendly or adversarial.

Some notes to understand the paper: while normally we think of multiagent games as consisting of a set of agents, in this paper there is an agent that acts, and an environment in which it acts (which can contain other agents). The objective function is neither minimized nor maximized -- the sign of the environment's KL divergence changes whether the stationary points are maxima or minima (which is why it can model both friendly and adversarial environments). There is only one utility function, the agent's utility function -- the environment is only modeled as responding to the agent, rather than having its own utility function.

My opinion: This is an interesting formalization of friendly and adversarial behavior. It feels somewhat weird to model the environment as having a prior strategy that it can then update. This has the implication that a "somewhat friendly" environment is unable to change its strategy to help the agent, even though it would "want" to, whereas when I think of a "somewhat friendly" environment, I think of a group of agents that share some of your goals but not all of them, so a limited amount of cooperation is possible. These feel quite different.

Interpretability

This looks like that: deep learning for interpretable image recognition *(Chaofan Chen, Oscar Li et al)*

Verification

Towards Mixed Optimization for Reinforcement Learning with Program Synthesis *(Surya Bhupatiraju, Kumar Krishna Agrawal et al)*: This paper proposes a framework in which policies are represented in two different ways -- as neural nets (the usual way) and as programs. To go from neural nets to programs, you use *program synthesis* (as done by VIPER and PIRL, both summarized in previous newsletters). To go from programs to neural nets, you use *distillation* (basically use the program to train the neural net with supervised training). Given these transformations, you can then work with the policy in either space. For example, you could optimize the policy in both spaces, using standard gradient descent in neural-net-space, and *program repair* in program-space. Having a program representation can be helpful in other ways too, as it makes the policy more interpretable, and more amenable to formal verification of safety properties.

My opinion: It is pretty nice to have a program representation. This paper doesn't delve into specifics (besides a motivating example worked out by hand), but I'm excited to see an actual instantiation of this framework in the future!

Near-term concerns

Adversarial examples

Adversarial Reprogramming of Neural Networks *(Gamaleldin F. Elsayed et al)*

AI strategy and policy

Shaping economic incentives for collaborative AGI *(Kaj Sotala)*: This post considers how to encourage a culture of cooperation among AI researchers. Then, when researchers try to create AGI, this culture of cooperation may make it more likely that AGI is developed collaboratively, instead of with race dynamics, making it more likely to be safe. It specifically poses the question of what external economic or policy incentives could encourage such cooperation.

My opinion: I am optimistic about developing AGI collaboratively, especially through AI researchers cooperating. I'm not sure whether external incentives from government are the right way to achieve this -- it seems likely that such regulation would be aimed at the wrong problems if it originated from government and not from AI researchers themselves. I'm more optimistic about some AI researchers developing guidelines and incentive structures themselves, that researchers buy into voluntarily, that maybe later get codified into law by governments, or adopted by companies for their AI research.

An Overview of National AI Strategies *(Tim Dutton)*: A short reference on the AI policies released by various countries.

My opinion: Reading through this, it seems that countries are taking quite different approaches towards AI. I don't know what to make of this -- are they acting close to optimally given their geopolitical situation (which must then vary a lot by country), or does no one know what's going on and as a result all of the strategies are somewhat randomly chosen? (Here by "randomly chosen" I mean that the strategies that one group of analysts would select with is only weakly correlated with the strategies another group would select.) It could also be that the approaches are not actually that different.

Joint Artificial Intelligence Center Created Under DoD CIO *(Sydney J. Freedberg Jr.)*

AI capabilities

Reinforcement learning

Capture the Flag: the emergence of complex cooperative agents *(Max Jaderberg, Wojciech M. Czarnecki, Iain Dunning et al)*: Summarized in the highlights!

Ranked Reward: Enabling Self-Play Reinforcement Learning for Combinatorial Optimization *(Alexandre Laterre et al)*

Procedural Level Generation Improves Generality of Deep Reinforcement Learning *(Niels Justesen et al)*

Meetup Cookbook

This is a linkpost for https://tigrennatenn.neocities.org/meetup_cookbook.html

My spouse and I have been running LessWrong meetups in two different cities for the past six years. Over time, we've gotten lazier and lazier about organizing, while still maintaining similar results, mostly by coming up with a bunch of simple recipes and scripts for running meetups. Here I have documented how we do it in excruciating detail, so others can use what we do.

Culture, interpretive labor, and tidying one's room

This is a linkpost for http://benjaminrosshoffman.com/culture-interpretive-labor-and-tidying-ones-room/

While tidying my room, I felt the onset of the usual cognitive fatigue. But this time, I didn't just want to bounce off the task - I was *curious*. When I inspected the fatigue, to see what it was made of, it felt similar to when I'm trying to thread a rhetorical needle - for instance, between striking too neutral a tone for anyone to understand the relevance of what I'm saying, and too bold of a tone for my arguments to be taken literally. In short, I was shouldering a heavy burden of interpretive labor.

Why would tidying my room involve interpretive labor?

It turns out, every item in my room is a sort of crystallized intention, generally past-me. (We've all heard the stories of researchers with messy rooms who somehow knew where everything was, and lost track of everything when someone else committed the violent act of reorganizing the room, thus deindexing it from its owner's mind.) As I decided what to do with an item, I wanted to make sure I didn't lose that information. So, I tried to Aumann with my past self - the true way, the way that filters back into deep models, so that I could pass my past self's ideological turing test. And that's cognitively expensive.

It's generally too aggressive to tidy someone's room without their permission, unless they're in physical danger because of it. But to be unwilling to tidy my own room without getting very clear explicit permission from my past self for every action - or at least checking in - is pathologically nonaggressive.

Once I realized this, it became easier to tidy my room, but the problem is not limited to that. Part of the reason why my cabin retreat was so helpful to me was that it limited my ability to accept social invitations or other bids for my attention. In those cases as well, I don't feel compelled to *agree*, but I do feel compelled to do enough interpretive labor to understand why this other person thinks I should do a thing.

I know of a few approaches to this problem, none of which seem fully adequate.

One approach is to blithely bulldoze the accumulated intentions of their environment. When people do this with respect to their *own* prior intentions, they end up too impulsive and disorganized to deal with the complexity of the modern world, and often severely indebted to extractive schemes like credit cards. When they do this with respect to *others*' intentions, they're inconsiderate and socially oblivious.

Others favor designing general policies for themselves, and then sticking to those policies, perhaps periodically updating them when they see a strong reason to. The downside of this approach is that it's slow to change, and prone to paralysis whenever something comes up that's outside existing policies. The advantage is that on-the-fly interpretive labor is replaced with batch processing, potentially capturing economies of scale and allowing much more efficient coordination with oneself and others over longer timescales.

A third approach is to accept some preexisting traditional way of life that's passed the test of time. This approach means that you have policies covering things you haven't encountered yet, vastly reducing the incidence of analysis paralysis. In addition, you know that a bunch of other people are following the same protocol, lateral coordination is much easier. Policies will also have been tested for working well in conjunction, and not just individually. A downside is that you're limited to the existing menu of options, all of which may be *knowably* quite suboptimal, and which are quite slow to change. This also doesn't help when you encounter genuinely novel-to-your-culture situations, or problems on scales larger than the one your culture's been tested over.

These three solution classes - impulsivity, policy-generation, and tradition-adherence, are all ones I'm actively exploring. The middle one fits my character the best. But overall this feels like a substantially unsolved problem. One thing I've been exploring in posts like Sabbath hard and go home, and Why I am not a Quaker, is looking at existing traditions for policies I might want to try out, to better explore the space, and so I don't have to wait for a crisis like the need to tidy my room to work out a better policy.

Simplicio and Sophisticus

Previously (Slate Star Codex): The Whole City is Center

Epistemic Status:

Note that after writing a lot of this, I checked and Sniffnoy anticipated a lot of this in the comments, but I think both takes are necessary.

There are many useful points to Scott's philosophical dialogue, The Whole City is Center, between Simplicio and Sophisticus. I want to point out an extra one I think is important.

Here's a short summary of some key points of disagreement they have.

Simplicio claims that there are words people use to describe concepts, and we should use those words to describe those concepts, even if those words have unfortunate implicaitons. Say true things about the world. Larry is lazy.

Sophisticus says no, if those words have unfortunate implications we shouldn't use them. And in many cases, where the unfortunate implications are inevitable because people have those implications *about the concept being described*, we shouldn't use any word at all to describe the concept. Larry can be counted on not to do things. But we shouldn't treat lazy as a thing, because people think being lazy is *bad* and there's no utility in thinking Larry is bad.

Simplicio says we should use whatever techniques work, regardless of whether they are negative reinforcement, positive reinforcement, before the act, after the act, too big, too small, you name it, if that's the system that works. And if people's natural instincts are to do things that work best as a system, but are sometimes ‘overkill’ or have unfortunate side effects in a particular case, you should accept that.

Sophisticus says no. Studies show negative reinforcement reinforcement doesn’t work, so don’t do it. Studies show harsher prisons don’t deter people so don’t use them. You should only use exactly what is needed to cause a direct effect in each situation. Or, if you need to use deterrence, what the evidence says will actually deter people.

Sophisticus says, we should look upon motivations like ‘I want this person to suffer’ with horror, and assume something has gone horribly wrong. (He makes no comment on feeling ‘I want this particular person to be happy’, which doesn’t come up.)

Simplicio says, if having seemingly unreasonable desires in some situations, including potential future situations, is the way persons and groups get better results, stop looking at it as some crazy or horrible thing. People’s motivations are messy, they have lots of weird side effects like loving kittens (I would note, so much so that I am punished for *not* loving them, basically because not having bad side effects of a thing is evidence of not having the thing itself). Going all ‘these instincts seem superficially nice so we’re going to approve, and these instincts seem superficially not nice so we’re going to disapprove’ seems wrong.

Sophisticus says, that by refusing to use concepts like lazy, he has a value disagreement with Simplicio and those who do use the lazy concept. Because those people embrace the implications.

Simplicio says no, this isn’t about value disagreement.

But then, near the end, Sophisticus catches Simplicio by saying he’s refusing in context to use the term ‘value difference’ because he doesn’t like its implications, and insisting only upon some Platonic ideal version of value difference. Which, Sophisticus says, makes him a hypocrite! Rather than point out either that no, it doesn’t, or maybe it does and you get non-zero points for noticing but asking for people not to ever be a hypocrite is not a valid move, he instead gets so embarrassed he flees town, and only redeems himself ten years later by pointing out what happens if you reject the unfortunate implications of the term ‘city center.

The most important lesson is, as Sniffnoy observes, *the characters have the wrong names.* Simplicio should be Sophisticus. Sophisticus should be Simplicio.

(I will continue to refer to them by Scott’s names here.)

Sophisticus wants to solve the world by getting rid of all the things he doesn’t like, and all the things he can’t properly quantify. He only accepts actions that are based on fully described and measured reasons. He will accept second or third order causes and consequences, but only and exactly those with well-described and quantified causal pathways.

Then he says that such actions are intelligent, sophisticated and advanced. They reject the irrational, the non-scientific. So they denigrate people who think otherwise with labels like Simplicio, and pretend that word doesn’t have unfortunate implications. Because it’s never all right to label people, in ways that have unfortunate and false implications (e.g. that a person is simple or stupid) unless you catch someone labeling people.

Simplicio accepts that the world is complex, and that our systems for dealing with it are approximations and sets of rules and values that won't always do the locally optimal thing, and that doesn't mean they're wrong. Simplicio is comfortable with the idea that correlations and associations exist even when we don't like them.

Sophisticus is what Nassim Taleb calls the Intellectual, Yet Idiot (IYI). By doing things that are more abstract, *and discarding most of the valid and useful information and relationships,* they fool themselves and others into thinking that they are smarter and more sophisticated. Simplicio is advocating for Taleb's typical grandmother, who has learned what actually works and survives, even if she doesn't understand all the reasons or implications.

Sophisticus is *vastly simplifying* the world.

He simplifies the world by cutting out the parts he does not like, and the parts he does not understand.

This allows him to create a model of the world. That's great! That's *super useful!* I love me some models, and you can't have models without throwing a lot of stuff out. Often the model gives much better answers despite this, and allows us to learn much and make better decisions. What makes a model great is that when you get rid of all the fuzziness, you get rid of a lot of noise, and you can manipulate and do math to what is left. Over time, you can add more stuff back into the model, and make it more sophisticated.

When you start thinking in models, or like a rationalist, or an economist, either in general or about a particular thing, that kind of thinking starts out deeply, deeply stupid. You must count on your other ways of thinking to contain the damage and point out the mistakes, to avoid taking these stupid conclusions too seriously, rather than as additional perspectives, as points of departure and future development, and places to learn. It goes way beyond Knowing About Biases Can Hurt People.

Drop stuff from your model, and you fail to understand or optimize for those things. If you then optimize based on your model, the things you left out of the model will be left out, and sacrificed, because they're using optimization pressure and atoms that can be used for something else. The results might or might not be an improvement. As the optimizations get more extreme, we should expect bigger disruptions and sacrifices of key excluded elements, so that had better be worth it.

One danger is that many people who develop the models either do so because *they are really bad at navigating without models,* or because *they realized how bad everyone is at navigating without models.* This provides motivation to work on the models even if they aren't yet any good, but it also increases temptation to forget that the model is a map and not the territory.

I think this is related to how those who found a business are as a group completely delusional about their chances of success, but also that founding a business is a generally very good idea. Motivating the long term investment and endurance of high costs only works in such cases, even if many more people would be better off in the long run if they did it.

The struggle is, how does one *combine* these two approaches. Build up one's models and toolboxes, to allow systematic thinking, while not losing the power of what you're ignoring, and slowly incorporating that stuff into your systematic thinking. Otherwise, no matter how simplistic the average person might be, you risk being even more so.

Some intuition on why consciousness seems subjective

This is an essay I wrote a few years back that some people have found helpful. The writing needs some polishing but I'm posting as-is. Thanks to Ruairi Donelly for originally encouraging me to write it and now also for encouraging me to post it here. Summary is at the end.

Experts strongly disagree on consciousness, even among smart philosophers and people I personally trust to think about it in rational ways. Having considered many arguments and thought of ways to change my perspective on consciousness, my impression is that much of the disagreement stems from strong, differing intuitions. Physicalists feel that what we call subjective experience is simply identical with the physical processes in our brains. Dualists see a fundamental difference between the two and generally can't imagine them being the same. As such, dualism has much more intuitive appeal to many people, myself included, but I think that it's mistaken.

I would like to bring a new argument to the debate which could resolve some confusions. It addresses the knowledge-argument, which says that consciousness is private. Many existing arguments already address the knowledge-argument. I'm just writing out the argument in this essay because it has helped some people get more intuition about why the subjectivity of consciousness is an appealing yet ultimately wrongheaded intuition that defies physicalism.

I won't explain the full reasoning for the view that consciousness is purely a physical (or 'analytically functional') process, this is just meant to be one piece in the puzzle. In short, reductionists will agree that something akin to physicalism should be accepted if it's even conceivably true. One important reason is that non-reductive views such as dualism violate Occam's razor in multiple ways.

The knowledge problem

A common question in debates about consciousness is whether Alice could fully understand what Bob subjectively experiences if she had perfect knowledge about the physical processes in Bob's brain. This is the topic of Frank Jackson's famous thought experiment about Mary, the colour scientist and Thomas Nagel's 'What is it like to be a bat?'. In a similar way, dualist intuitions are often defended with the argument that consciousness is private, subjective or ineffable. This distinguishes consciousness from physical processes, which can be observed by anyone. Consequently, there must be more than physical knowledge and this other knowledge must be about something that's not physical. This is known as the knowledge argument.

The spirit of these arguments could be summarised as follows: There is a third-person perspective and a first-person perspective to the processes in Bob's brain. The first-person perspective is what we call Bob's subjective experience. First-person perspectives aren't a fundamental concept in physics and yet it seems that having access to one determines what can be known. Let's call the existence of the first-person perspective *subjectivity*.

Even people who don't agree with Jackson's or Nagel's thought experiments often find that subjectivity is at odds with consciousness being physical. Brian Tomasik shows how the problem relates to first-person and third-person perspectives:

"*When we imagine deriving the behavior of a car from its atomic configuration, we can think about all parts of the system in third-person terms. We adopt a physical stance toward the atoms, and the metal/plastic pieces, and the whole joint system. Phenomenal experience is different because this requires shifting from third-person to first-person, which is a switch that no other scientific reduction needs.*"

He makes the point I'll make below by referring to 'acquaintance knowledge':

"Phenomenal consciousness is analytically functional [here: physical], but it's also sort of epistemically inaccessible in an acquaintance rather than propositional sense. The acquaintance view makes us feel better about explaining qualia than crude type-A caricatures while not allowing for the ideal conceivability of zombies as would be the case for type B."

However, acquaintance knowledge is just an intuition pump, and Brian acknowledges that it's not easy to verify if it's just "linguistic trickery". The idea of acquaintance knowledge was first stated by philosopher Earl Conee (1994) as an objection to the knowledge argument.

A new argument

I propose an argument that I find more intuitively convincing. Let's examine what we mean by the words 'knowledge' and 'third-person perspective'. If you look at, say, a bowl, you gain knowledge of its physical properties such as its shape and colour from a third-person perspective. First, light reflects off the bowl which reaches your eyes and gets transformed into electrical signals. The brain then recognizes shapes etc. in these signals. The crucial point is that it then forms an *abstraction*, i.e. a *model* of the bowl. This model in your brain is what we typically call the physical knowledge about the bowl. The abstraction itself is physically stored like a hard drive would store an image file. When you remember what the bowl looks like, the abstraction is broadcast into your brain so that it is *physically part of* your consciousness (on a physicalist account). Since you can form a roughly accurate model of everything about the bowl we find relevant, we say that everything about the bowl is accessible from a third-person perspective.

Now let's compare this to the physical knowledge A could gain about B's brain processes and the associated 'subjective' experience. If A were to look at an MRI scan of B's brain, A would form an abstraction of B's brain that is simply an image. This abstraction would, of course, not do justice to the complexity of the processes in B's brain. It could possibly be stored on a floppy disk. In order to know all the physical knowledge about B's brain processes, A would have to form an *accurate* abstraction. This much harder for a brain than for a bowl, because brain processes have many more details that we care about.

If A were to actually form an abstraction of Bob's brain that captures the aspects that matter, Alice would have to run nearly every aspect of it in her own brain. The relevant aspects include e.g. the content of Alice's working memory, i.e. the neural activations that are necessary to compute the concepts present in the working memory, with all their relevant properties, as well as the relationships between them. This would amount to something close to *simulating* Bob's brain inside Alice's brain. If Alice had suitably advanced brain that could actually do this, Alice may herself experience what Bob experiences. Therefore, the knowledge argument fails: By perfect physical knowledge of a brain we mean forming an accurate abstraction (perhaps even a simulation) of it in our own brain. This entails the subjective experience.

This is not the first objection to the knowledge argument and the point is not to add another one. Rather, I'd like to resolve the confusion around first-person and third-person perspectives. These concepts don't have a place in physics (to my knowledge) but they still appear to determine what we can know. What I've tried to show is that having a third-person view of something reduces to forming an abstraction of that thing in our brain. The third-person perspective is simply a concept we've invented because we don't usually think through the whole process of forming an abstraction.

This leaves us with only the first-person perspective. But since there are no perspectives left to contrast it with, we can best do away with the idea of perspectives altogether (at least when it comes to the ontology of mental processes). I think that without these concepts, physicalism can become more intuitive. 'Somebody being an algorithm viewed from the inside' simply reduces to the *existence* of an algorithm at some point in space and time. If I say 'I'm experiencing the smell of apple right now' I'm simply saying that there is an apple-pie-smell algorithm going on here. (Why am 'I' this particular kind of algorithm and not some other one? That question, while puzzling, is not only outside the scope of this essay, but also probably outside the field of philosophy of mind. It seems more related to anthropics, the study of indexical information).

Going back to Earl Conee's and Brian Tomasik's concept of acquaintance knowledge we could now say that this is the *only* kind of knowledge, just like first-person is the only perspective left. Propositional knowledge simply means having, in some part of our brain where it is not used, the data needed to form an abstraction of something. We can then access this knowledge and become acquainted with it by putting it into our awareness. Once again, I think it's best to do away with the different kinds of knowledge in discussions about consciousness because they're just how we conceptualise of physical processes things going on in our brain.

Changing intuitions about consciousness

This was one of the things that are most counterintuitive about physicalism to me: There appears to be a phenomenal experience, which can't be seen from the outside. Since the view from the outside is just a concept we invented, this is not a reason for concern. One remaining challenge for our intuitions is to identify two things which intuitively feel very different from one another: The physical processes in our brain and what we call our subjective experience. We conceptualise these things in very different ways which makes it hard to see them as identical. (For example, we think of consciousness as *unified* and physical processes as individual atoms moving around). I hope I've removed one obstacle to doing so. I don't think that this piece would have intuitively and intellectually convinced me all by itself a while ago, when my intuitions were still strongly at odds with physicalism. Changing them can take a long time, can't be done in one essay and involves overriding very deep-rooted intuitions. I think this is why very smart and rational people often cannot accurately model each other. Having been on both sides, I feel I have a better model of both now.

It feels a bit cheap to say that the rest is just about changing intuitions, without actually giving any more arguments, but I can't avoid it. I highly recommend several of Brian's essays which can further help change intuitions, but the most important aspect is an open mind towards a view that can at times be mind-bogglingly counterintuitive.

Summary

In summary, I have argued something quite simple: The third-person perspective and the idea of knowledge are just ways of saying that we form an abstraction of something in our brain which we may then be conscious of. If we formed a sufficiently detailed abstraction of another person’s brain, the processes constituting their subjective experience would also happen inside us. Therefore, the knowledge argument fails. But more importantly, perspectives are not a fundamental thing, which makes it easier to accept that subjective experience is simply the *existence* of a particular neural algorithm (rather than an inaccessible first-person perspective, which is not a thing in physics).

Appendix

Objection: When A simulates B’s brain in her own brain, it’s not A, but B that has the experience.

This misses the point: The point is that there is no third-person perspective, just forming abstractions of things, so we can’t expect Alice to get a third-person peek into Bob’s experience. Besides, persons are not fundamental (or even a thing) on the physicalist view we're trying to critique here. Whether it’s Alice or Bob having the experience is therefore just semantics.

What about the zombie argument?

By my understanding, David Chalmer's well-known zombie argument only argues against type-B physicalism (or type-B materialism as Chalmers calls it), the view that there is an epistemic gap between physics and subjective experience, but this gap doesn’t mean that there is stuff other than physics. I’m defending a view called type-A physicalism, which states there is no epistemic gap and that consciousness is simply what physics does in our brain. My argument here is supposed to provide one idea needed to close the perceived epistemic gap. I’m happy to provide an explanation why type-A physicalism isn’t affected by the zombie argument if it’s unclear.

Anthropics: A Short Note on the Fission Riddle

In the I-Less Eye, R Wallace considers a situation where an entity clones itself a hundred times that leads to a surprising paradox. I'll argue that there's a rather simple flaw with the argument made in the linked article, but first I'll summarise the riddle.

Suppose you are the original. After you've cloned yourself, you now have a 50% chance of being the original and 50% being the clone. This should apply regardless of what clones already exist, so after cloning yourself 100 times, you should have a 1/2^100 chance of being the original.

On the other hand, this seems strange. Intuitively, it seems as though you ought to have a 1/101 chance of being the original as there are 101 copies. Further, why does cloning one at a time give a different answer from creating all 101 clones at once?

Solution

In order to solve this riddle, we only have to figure out what happens when you've been cloned twice and whether the answer to this should be 1/3 or 1/4. The first step is correct, the subjective probability of being the original should be 1/2 after you've pressed the cloning button once. However, after we've pressed the cloning button twice, in addition to the agents who underwent that first split, we now have an agent that falsely remembers undergoing that split.

Distributing the probability evenly between the agent's who either had that experience or remember it: we get a 1/3 chance of being a false memory and a 2/3 chance of it being a real memory. If it is a real memory, then half of that - that is a 1/3 - is the probability of being the original and the other half - also 1/3 - is the chance of being the first clone.

So, the answer at the second step should be 1/3 instead of 1/4. Continued application will provide the answer for 100 copies. I'll admit that I've more sketched out the reasoning for the 1/n solution instead of providing a formal proof. However, I would suggest that I've sufficiently demonstrated that halving each time is mistaken as it assumes that each remembered split is real.

However, we can verify that the 1/n solution produces sensible results. Exactly two agents experience the process of each split (but more agents remember it). So there is a 2/n chance of you experiencing the process and a 1/n chance of you experiencing it as the original. So there is a 50% chance of you "waking up" as the original after the split *if you actually underwent the split and didn't just falsely remember it*.

Please note that I'm not taking a position in this post as to whether subjective probabilities ultimately are or aren't coherent, just arguing that this particular argument is fallacious.

Finally, I'll note a few questions that this opens up. If we have to include all agents who remember a situation in anthropic reasoning and not just those who experience it, what actually counts as remembering a situation? After all, in the real world, memories are always imperfect? Secondly, what if an agent has a memory, but never accesses it? Does that still count?

EDIT: As you can see from my response to the comments, this post has some issues. Hopefully, I am able to update it at some point, but this isn't an immediate priority.

Probability is fake, frequency is real

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

Consider the Sleeping Beauty problem. What do we mean by fair coin? It is meant that the coin will have 50-50 probability of heads or tails. But that is fake. It will ether come up heads or tail, because the real world is deterministic. It is true that I don't know the outcome. I don't know if I am in a world of type "the coin will come up heads" or a world of type "the coin will come up tails". But in this situation I should be allowed to put what ever prior I want on the coins behavior.

Consider the Born rule of quantum mechanics. If I measure the spin of an electron, then I will entangle the large apparatus that is the my measuring equipment with the spin of the electron. We say that there are now two Everett branches, one where the apparatus measured spin up and one where the apparatus measured spin down. Before I read of the result, I don't know which Hilbert branch I am in. I could be in ether, and I should be allowed to have what ever prior I want. So why the Born rule? Why to I do I believe that the square amplitude is **the** correct way of assigning probability to which Hilbert branch I am in?

I believe in the Born rule because of the frequency of experimental outcomes in the past. The distribution of galaxies in the sky can be traced back to the Born rule. I don't have the gears on what is causing the Born rule, but there are something undeniably real about galaxies that trumps mere philosophical Bayesian arguments about freedom of priors.

Imagine that you are offered a bet. Should you take it or not? There are several argument about what you should do in different situations. For example, if you have finite amount of money, you should maximize the E(log(money)) for each bet, (see e.g. Kelly criterion). However, every such argument I have ever seen, is assuming that you will be confronted by a large number of similar bets. This is because probabilities only relay make sense if you sample enough times from the random distribution you are considering.

The notion of "fair coin" does not make sense if the coin is flipped only once. The right way to view the Sleeping Beauty problem is to view it in it in the context of Repeated Sleeping Beauty.

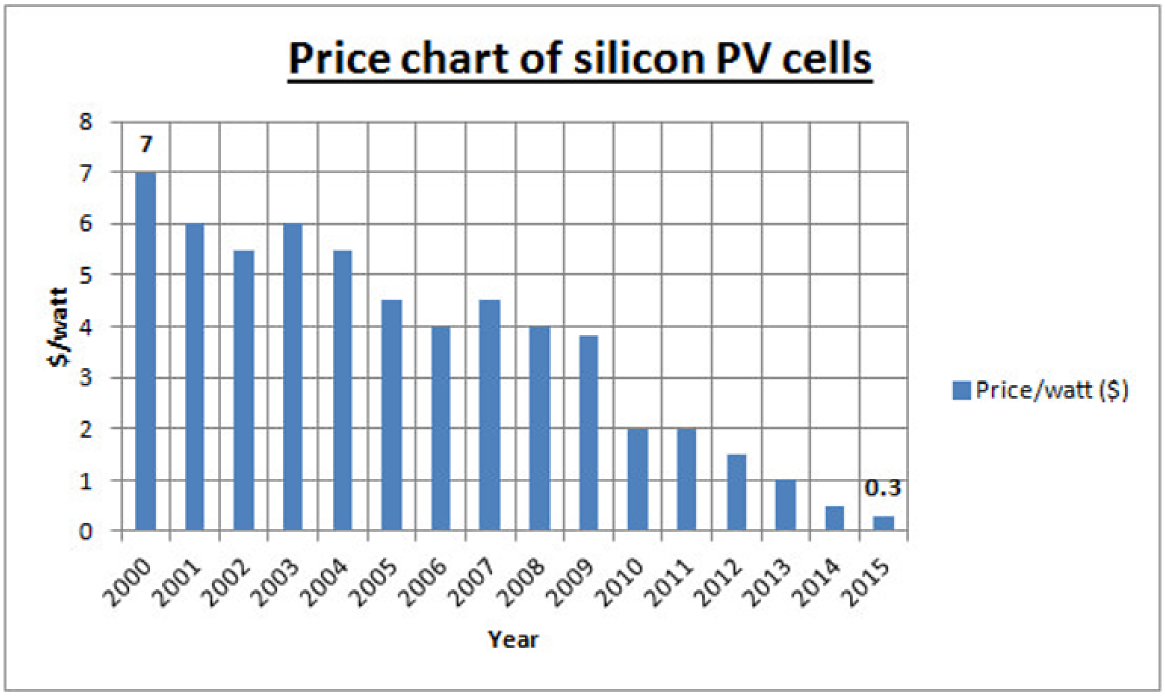
Next: Repeated (and improved) Sleeping Beauty problem

What will we do with the free energy?

Epistemological status: Superficial personal speculation .

This posts follows What could be done with RNA and DNA sequencing that's 1000x cheaper than it's now? in thinking about technical challenges we will face in the next decades. It's an attempt by my to play a bit with futurism.

Both solar and wind energy have exponential efficiency improvements:



A majority of Indian new coal power now costs more then new wind and solar power. This means that we can expect a lot of new wind and solar power installations to be build.

Last year, in 2017, in Germany there were 146 hours in which energy prices were negative. This means that during those hours wholesale customers got payed to consume energy.

Unfortunately, batteries are expensive and while it might be economical to load a battery in the day and release the energy at night it won't be economical to load the battery in summer and release it in winter.

As a result batteries don't work as energy sink for summer days with too much energy production.

Maybe, we can construct our wind turbines in a way that we can dumb excess energy into them when there isn't wind? It would however be more desirable to find a way to actually spend that free energy productively.

Most machines that we currently build are optimized to run as much as possible and thus don't provide good energy sinks. A machine that produces spoons is most economical when it runs 24/7.

This means that there's an economic opportunity to build machines that can work as energy sinks for days with negative energy costs.

It's likely a good idea to build data-centers in a way that they are not only able to run in a low energy mode but also in a mode where they burn as much energy as possible. As a result it might b worthwhile for some cloud provider to have in a year specific times where the compute is radically cheaper than on other days. Maybe that compute will go into crypto-mining but it we can also think about other uses for it.

The data-centers likely won't be enough of a data sink and we need new ideas about what to do with other excess energy.

Do you have any energy about how to build good energy sinks that might use energy usefully and that are economical if they are only used for a few weeks/days a year?

Figuring out what Alice wants, part I

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

This is a very preliminary two-part post sketching out the direction I'm taking my research now (second post here). I'm expecting and hoping that everything in here will get superseded quite quickly. This has obvious connections to classical machine intelligence research areas (such as interpretability). I'd be very grateful for any links with papers or people related to the ideas of this post.

The theory: model fragments

I've presented the theoretical argument for why we cannot deduce the preferences of an irrational agent, and a practical example of that difficulty. I'll be building on that example to illustrate some algorithms that produce the same actions, but where we nonetheless can feel confident deducing different preferences.

I've mentioned a few ideas for "normative assumptions": the assumptions that we, or an AI, could use to distinguish between different possible preferences even if they result in the same behaviour. I've mentioned things such as regret, humans stating their values with more or less truthfulness, human narratives, how we categorise our own emotions (those last three are in this post), or the structure of the human algorithm.

Those all seems rather add-hoc, but they are all trying to do the same thing: hone in on human judgement about rationality and preferences. But what is this judgement? This judgement is defined to be the internal models that humans use to assess situations. These models, about ourselves and about other humans, often agree with each other from one human to the next (for instance, most people agree that you're less rational when you're drunk).

Calling them models might be a bit of an exaggeration, though. We often only get a fragmentary or momentary piece of a model - "he's being silly", "she's angry", "you won't get a promotion with that attitude". These are called to mind, thought upon, and then swiftly dismissed.

So what we want to access, is the piece of the model that the human used to judge the situation. Now, these model fragments can often be contradictory, but we can deal with that problem later.

Then all the normative assumptions noted above are just ways of defining these model fragments, or accessing them (via emotion, truthful description, or regret). Regret is a particularly useful emotion, as it indicates a divergence between what was expected in the model, and what actually happened (similarly to temporal difference learning).

So I'll broadly categorise methods of learning human model fragments into three categories:

- Direct access to the internal model.

- Regret and surprise as showing mismatchs between model expectation and outcomes.
- Privileged output (eg certain human statements in certain circumstances are taken to be true-ish statements about the internal model).

The first method violates algorithmic equivalence and extentionality: two algorithms with identical outputs can nevertheless use different models. The second two methods do respect algorithmic equivalence, once we have defined what behaviours correspond to regret/surprise, or what situations humans can be expected to respond truthfully to. In the process of defining those behaviours and situations, however, we are likely to use introspection and our own models: a sober, relaxed rational human confiding confidentially with an impersonal computer, is more likely to be truthful than a precariously employed worker on stage in front of their whole office.

What model fragments look like

The second post will provide examples of the approach, but here I'll just list the kind of things that we can expect as model fragment:

- Direct statements about rewards ("I want chocolate now").
- Direct statements about rationality ("I'm irrational around them").
- An action is deemed better than another ("you should starts a paper trail, rather than just rely on oral instructions").
- An action is seen as good (or bad), compared with some implicit set of standard actions. ("compliment your lover often").
- Similarly to actions, observations/outcomes can be treated as above ("the second prize is actually better", "it was unlucky you broke your foot").
- An outcome is seen as surprising ("that was the greatest stock market crash in history"), or the action of another agent is seen as that ("I didn't expect them to move to France").

A human can think these things about themselves or about other agents; the most complicated variants are assessing the actions of one agent from the perspective of another agent ("if she signed the check, he'd be in a good position").

Finally, there are meta, and meta-meta, etc... versions of these, as we model other agents modelling us. All of these give a partial indication of our models of the rationality or reward, about ourselves and about other humans.

Alignment problems for economists

AI alignment is a multidisciplinary research program. This means that there is potentially relevant knowledge and skill scattered across different disciplines. But it also means that people schooled only in narrow disciplines will experience a hurdle when they would work on a problem in AI alignment. One such discipline is economics, from which decision theory and game theory originated.

In this post I want to explore the idea that we should try to create a collection of “alignment-problems-for-economists”, packaged in a way that economists who have relevant knowledge and skill but don't understand ML/CS/AF can work on them.

There seem to be sub-problems in AI alignment that economists might be able to work on. However, out of the economists that I’ve spoken to, some are enthusiastic about this but see it as a personal career-risk to work on it as they do not understand the computer science. So if we can take subproblems in alignment, and package them in a way that economists can immediately start working on them, then we might be able to utilize intellectual resources (economists) that would otherwise have worked on something different.

Two types of economists to target

1. Economists who also to a degree understand basic ML/CS

2. Economists who do not.

I don’t find it very plausible that we could find sub-problems for the second type to work on, but it doesn’t seem entirely impossible: there could be certain specific problems in mechanism design or social choice or so, that would be useful for alignment but don’t require any ML/CS.

Properties of alignment-problems-for-economists that are desirable:

1. **Publishable in economics journals**. I have spoken to economists that are interested in the alignment problem, but they are hesitant to work on it: It is a risky career move to work on alignment if they cannot publish in journals that they are used to.

2. High work/statement ratio. How long will it take to solve the problem, versus providing the statement of the problem? If 90% of the problem is to state it in a form so that an economist could work on it, then it would likely not be efficient to do so. It should be a problem that can relatively easily be communicated clearly to an economist, while taking more time to solve.

3. No strong reliance on CS/ML tools. Many economists are somewhat familiar with basic ML techniques, but if a problem relies too much on knowledge of CS or ML, this increases the career-risk of the problem.

4. Not necessarily specifically x-risk related. If a problem in alignment is not specifically x-risk related, it is less/not embarrassing to work on it, and therefore less of a career-risk. Nevertheless, most problems in AI alignment seem important even if you don't believe that AI poses an x-risk. *I don't think this requirement is that important.*

* **Does not have to be high-impact.** If a problem has only a small chance of being somewhat impactful, it might still be worth packaging it as an economic problem, since the economists who could work on it would not otherwise work on alignment problems at all.

I do not yet have a list of such problems, but it seems that it might be possible to make one:

For example, economists might work on problems in mechanism design and social choice for AGI’s in a virtual containment. For example, can we create mechanisms with desirable properties for the amplification phase in Christiano’s program, to align a collection of distilled agents? Can we prove that such mechanisms are robust under certain assumptions? Can we create mechanisms that robustly incentivizes AGI’s with unaligned utility functions to tell us the truth? Can we use social choice to find out properties of agents that consist of sub-agents?

Economists work on strategic communication between agents (cheap talk), which might be helpful in the design of safe containment systems of not-superintelligent AGI. Information economics works on game theoretic properties of different allocations of information, and might be useful in such mechanisms as well. Economists also work on voting, and decision theory.

I want your feedback:

1. What kind of problems have you encountered that might be added to this list?

2. Do you have reasons to think that this project would be doomed to fail (or not)? If so, I want to prevent wasting time on it as fast as possible. Despite having written this post, I don’t assign a high probability of success, but I’d like people’s views.

Debt is an Anti-investment

Cross-posted from Putanumonit.

Since I wrote *Get Rich Slowly* I’ve received a steady stream of questions regarding personal finance. The most common of those is: **should I prioritize investing or paying off my debts?**

Get Rich Slowly wasn’t meant to break any new ground, just summarize some of the best advice online in a clear way for my readers. So, I thought I could just look up the existing best advice on debt vs. investing. I did, it sucks.

A lot of places tell you to invest if the return you’re expecting is higher than the interest on your debt, but that completely ignores risk. Taking a loan at 20% to invest in a highly speculative venture with expected returns of 20.1% isn’t smart investing, it’s a reckless gamble that’s likely to leave you bankrupt. *The Balance,* one of the most popular personal finance websites, mentions the importance of risk-adjustment but is too lazy to do the math explicitly. It also recommends maxing out your Roth IRA (expected return 6-7% with a fair bit of risk) before paying off credit cards (20%+ interest rate), which is utterly insane.

Risk adjustment is difficult and subjective, but there’s no escape from putting a number on it ourselves.

Anti-investment

I like to think of debt as an anti-investment. Let’s say you have a $10,000 loan which charges 5% interest, and you also have $10,000 invested with a risk-free after-tax return of 5%. The two would cancel each other out – your cash flow is the same as if you had neither one, namely zero. So, if you pay off the loan, you can think of it as **gaining a risk-free investment** with the same after-tax return as the loan’s interest rate.

What if instead of paying off the loan you invest the money? Then instead of the risk-free investment, you gain a different one. For example, if you invest the money in an S&P 500 US stock index, you will gain an investment that should return 7% (or 6% after paying capital gains tax), albeit with quite a bit of risk.

Thus, the question of paying off the debt becomes a comparison between a risk-free investment and your best alternative investment option. The question becomes: **what risk-free rate of return is equivalent to your best available investment?** If you pay a higher interest on your debt than that, you should pay it off. If the rate you’re paying is lower, you should invest. In our example, if you like holding a risky 6%-return stock fund about as much as a risk-free 3%, you should pay off any loans that charge you more than 3%.

This may not seem like an easier question to answer, but we can attack it from several angles. At the core of this decision is a consideration of how much the *risk-free* part is worth to you, which is going to be different for each person. I will use the S&P 500 as the example of investment under consideration, but you should consider your own considerations instead.

0. Better loan

If you can get a loan for a lower interest rate than your current debt, you should just take the new one out to repay the old one. Duh. You could have an opportunity to borrow at lower rates for any of the following reasons:

1. You improved your credit rating.
2. You married someone with a better credit rating (I should add that to the matrix).
3. You got into a degree program and can get student loans (which are cheaper because they're not dischargeable).
4. You bought/inherited/stole an asset that can be used as collateral.
5. You got over yourself and borrowed the money from your parents.

Remember to calculate the interest rate on an after-tax basis too, so a 6% mortgage can really only cost 4% if you can deduct the mortgage interest from your taxable income. Taking a mortgage to pay off worse loans is a smart thing to do.

And with that out of the way, here are some reasons why a risk-free investment (which is what you'll "gain" by paying off the loan) may better than the S&P 500 even with a lower rate of return.

1. Arbitrage

Whatever your preferred investment option is, there's probably one with a higher return that's just too risky for you. This can be something like junk bonds, an emerging markets index fund, or even simply a levered S&P 500 fund (which multiplies the risk and return). If you had a risk-free investment, you could invest part of it in the high-risk high-reward option and end up with a better overall deal.

Here's a numerical example that may or may not make this clear:

The S&P has a yearly volatility of 15-20% (call it 17%) and a return of 6%. Let's notate this [6,17] The MSCI world index (global stocks) has higher returns but is also riskier, perhaps [8,25]. In a vacuum, you may prefer the former. But what if you had a risk free 5% return [5,0]? You could invest 40% of that in the MSCI and end up with 60%*[5,0] + 40%*[8,25] ~ [3,0] + [3.2,10] ~ [6.2,10], or 6.2% returns with 10% volatility. That's certainly a better deal than [6,17].

If you didn't follow the example or you don't like using yearly volatility as a measure of risk, you'll just have to trust me that this principle holds. Back in our original formulation, paying off some *part* of your loan and investing the remainder in something with a higher yield can result in better returns and a lower risk than investing the whole amount in the S&P.

2. Utility of money

A good reason to be risk averse is that the value you derive from every extra dollar diminishes with every dollar you have. Going from $100,000 to $200,000 (whether in wealth or income) will give you a bigger happiness boost than going from $200,000 to $300,000. This means that a guaranteed $200k is better than a coin flip between $100k and $300k. There's research showing the happiness depends on the log of income, so you'd need the coin flip to be between $100k and **$400k** to equal a safe $200k.

(I previously used this fact to propose a new measure of economic inequality.)

If you invest $100,000 in the S&P 500, after 16 years (at 6% return) you'll have $250,000 on *average*. But it could be a lot more or a lot less, somewhat like a $100k-$400k coin flip. But to get to $200,000 you only need a risk-free investment at 4.6%.

3. Planning

This is closely related to the point above: a guaranteed return is better than a volatile one because it allows you to plan ahead. Whether you're planning hit the milestone number for retirement, buy a house, or get that last bit of uranium ore from Amazon you need for your "peaceful" nuclear program, knowing how much money you'll have in the future makes it easier to plan. If your future returns are volatile, you will probably need to aim for a higher number to guarantee you have enough to reach that all-important critical mass.

4. The Beta of your life

A risky investment isn't just bad in itself, it also contributes to the overall risk of your economic life. If you live and work in a developed economy, almost everything you can invest in will be positively correlated with everything else, and with your career as well: US stocks, global stocks, corporate bonds, real estate, natural resources, even government bonds (especially during crises).

This means that in the worst-case scenario everything can turn sour at the same time: the economy tanks, your investments lose value, your job is at risk, etc. A risk-free investment can isolate you from that pain, or, conversely, debt will make that pain much worse. This again should make risk-free investments more attractive even at lower return rates, especially if the rest of your eggs are in the same few baskets.

There are some possible investments that aren't positively correlated to everything else, such as cryptocurrencies (maybe). But if you're reckless enough to take out loans for the purpose of buying crypto, I have a bunch of JacobCoins to sell you.

5. Risk aversion

Humans have evolved to be averse to risk and unnecessary gambles. Some of this risk aversion is rational in the case of investments – see points 1-4. But even the irrational part is still part of you, and having volatile investments will fray your nerves and cost you sleep regardless of the math you do.

Summary

I know it seems weird to compare existing choices to an option you don’t really have (a risk-free investment with arbitrary return), but that’s exactly what debt is. Or rather, that’s exactly the opposite of what debt is.

Different people will give different weights to the five factors I listed, but it’s important to remember that they’re additive, not exclusive. The more you care about any of them the lower the return you would be happy with if you could get it risk-free, and thus the lower rate of interest on your debt that you’ll hold.

Most of my own money is in stock index funds. I expect a 7% return from my retirement savings (401k and IRA) because they’re tax-free, and 6% return from my medium-term savings. I’d be happy to replace those with a 4% and 3% return, respectively, if I could get it risk-free.

This means that I:

1. Don’t hold any debt at above 4%. If I had any, I would sell some short-term investments to pay it off right away.
2. Will happily borrow money at 2-2.5% to invest in stocks, and will consider borrowing at 3-3.5% only if it goes to retirement (for example, if I need it to max out the yearly Roth IRA cap).
3. Will happily lend at 5-6% for something close to risk-free, such as a short-term loan to a close family member or friend, because that’s a better deal for me than my investments.

And finally, since good investments with returns above 6% are very hard to come by, if you have any debt at 5-6% or more you should almost certainly pay that off before investing a single penny. It’s a very un-American piece of advice to hand out on July 4th, but it’s the truth.

Another take on agent foundations: formalizing zero-shot reasoning

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

After spending more time thinking about MIRI's agenda, I've come upon another framing of it, which I'm coining *zero-shot reasoning*. [1]

This post is a distillation of my intuitions around zero-shot reasoning. These views should be taken as my own, and not MIRI's.

A quick summary of this post:

- "Zero-shot reasoning" refers to the ability to get things right on the first try, no matter how novel or complicated.
- In simple domains, like mathematics, zero-shot reasoning is fully captured by formal verification of proofs. In more general domains, zero-shot reasoning requires an extension of formal verification that can be applied to real-world plans.
- MIRI-esque philosophical work is necessary to extend formal verification to more general domains.
- A formal account of zero-shot reasoning will likely be unimportant for aligning the world's first AGIs, but will likely be essential for aligning a recursively self-improving AGI.
- Humanity will most likely end up building a recursively self-improving AGI (plausibly because of insufficient coordination around *not* building a recursively self-improving AGI).
- We can probably delegate much of the work of formalizing zero-shot reasoning to a post-AGI society, but working on zero-shot reasoning today nevertheless substantially increases the odds that our first recursively self-improving AGI is aligned.

What is zero-shot reasoning?

Few-shot reasoning vs zero-shot reasoning

The world is largely chaotic and unpredictable, yet humans can ideate and successfully execute on hugely conjunctive plans with many moving parts, many of which are the first of their kind. We can ship massive software projects and send rockets to space. Some people can build billion-dollar companies over and over again.

On the other hand, it's clear that human abilities to do this are limited. Big software projects invariably have critical bugs and security flaws. Many of our spacecraft have exploded before making it to space. Most people can't build a company to save their lives, and most successful entrepreneurs fail when building a second company.

Native human software is capable of doing something I'll call *few-shot reasoning*—when performing complex reasoning and planning to accomplish something novel, humans can usually get it right after a few rounds of iteration. The more dissimilar this reasoning is to prior reasoning they've done, the more rounds of iteration they need.

I think something like *zero-shot reasoning*—the ability to perform arbitrarily novel and complex reasoning, while have calibrated high confidence in its soundness—is possible in principle. A superintelligent zero-shot reasoner would be able to:

- Build an operating system as complex as Microsoft Windows in assembly language, without serious bugs, without once running the code
- Build one spacecraft that lands on Mars, after observing Earth for one day, without building any other spacecrafts
- Amass $1 trillion over three years

It should be able to do these all with extremely high confidence. [2]

Zero-shot reasoning might seem like magic, but in fact humans have some native capacity for it in some limited domains, like pure mathematics. Given a conjecture to prove or disprove, a human can start from a small set of axioms and combine them in extraordinarily novel and complex ways to confidently arrive at a proof or disproof of the conjecture.

That said, humans do make mistakes in mathematical proofs. But with the assistance of formal verification tools like Coq, humans *can* become extremely confident that their proofs are error-free, no matter how novel or complex they are. [3]

In a similar vein, humans can in principle build an operating system as complex as Microsoft Windows in assembly language, without serious bugs. Even if they're writing a huge amount of code, they could formally verify each component they build up, and formally verify that compositions of these components work as desired. While this can only give them guarantees about what they prove, they can very confidently avoid certain classes of bugs (like memory leaks, buffer overflows, and deadlocks) without ever having to run the code.

Formalizing zero-shot reasoning

Formal verification provides a formal account of zero-shot reasoning in limited domains (namely, those describable by formal axiomatic systems, like mathematics or software). I think a formal account of more general zero-shot reasoning will involve an extension of formal verification to real-world, open-ended domains, that would give us some way of "formally verifying" that plans for e.g. building a rocket or amassing wealth will succeed with high probability.

Note that much of general zero-shot reasoning consists of subcomponents that involve reasoning within formal systems. For example, when building rockets, we do a lot of reasoning about software and Newtonian physics, both of which can be construed as formal systems.

In addition to formal verification, a complete formal account of general zero-shot reasoning will require formalizing other aspects of reasoning:

- *Making and trusting abstractions*: Humans are capable of turning sense data into abstract formal systems like Newtonian mechanics, and then deciding under which situations it's appropriate to apply those abstractions. How do we

formalize what an abstraction is, how to make abstractions, and under which circumstances to trust them?
- *Bounded rationality*: What does it mean for a bounded agent to have a calibrated estimate of the likelihood that a plan will succeed? In other words, how can we tell when an agent with limited computing resources is properly reasoning about logical and empirical uncertainty? (Bayesian inference gets us some of the way there, but doesn't tell us how to select hypotheses and is often computationally intractable. Logical induction is a good starting formalism for logical uncertainty, but current algorithms are also computationally intractable.)
- *Self-trust*: An agent may need to formally reason about how much to trust its reasoning process. How can an agent formally refer to itself within a world in which it's embedded, and reason about the ways its own reasoning might be faulty? (Sources of error may include hardware failures in the physical world and bugs in the software it's running.)
- *Logical counterfactuals*: An agent may want to formally reason about the consequences of choices it takes. But if it's deterministic, it will only end up making one choice, so it's not clear how to formally talk about what happens if it picks something else. (Concretely, if it's reasoning about whether to take action A or action B, and it in fact takes action A, reasoning formally about what would happen *if it took action B* is confusing, because *anything* can happen if it takes action B by the principle of explosion.)

This list is just an overview of some of the problems that need to be solved, and is by no means intended to be exhaustive. Also note the similarity between this list and the technical research problems listed in MIRI's Agent Foundations agenda.

Why care about formalizing zero-shot reasoning?

Isn't extreme caution sufficient for zero-shot reasoning?

It's true that humans can make plans far more robust by thinking about them much longer and much more carefully. If there's a massive codebase, or a blueprint of a rocket, or a detailed business plan, you could make them far more robust if you had the equivalent of a billion humans ruminating over the plan, reasoning about all the edge cases, brainstorming adversarial situations, etc. And yet, I think there remains a qualitative difference between "I thought about this plan very very hard and couldn't find any errors" and "Under plausible assumptions, I can prove this plan will work". [4]

It was critically important to Intel that their chips do arithmetic correctly, yet their reliance on human judgment led to the Pentium division bug. (They now rely on formal verification.) The Annals of Mathematics, the most prestigious mathematics journal, accepted both a paper proving some result and a paper proving the negation of that result.

Human reasoning is fundamentally flawed. Our brains were evolutionarily selected to play political games on savannahs, not to make long error-free chains of reasoning. Our cognition is plagued with heuristics and biases (including a heuristic that what we see is all there is), and we all have massive blind spots in our reasoning that we aren't even aware exist. If we check a plan extremely thoroughly, we can only trust the plan to the extent that we trust that the plan doesn’t have any failure modes within our blind spots. The more conjunctive a plan is, the more likely it is that it will have a failure point hidden within our blind spots.

More concretely, suppose we have a plan with 20 components, and our estimate is that each component has a 99.9% chance of success, but in actuality three of the components have likelihoods of success closer to 80% because of edge cases we failed to consider. The overall plan will have a 0.999^17 * 0.80^3 ≈ 50% chance of success, rather than the 0.999^20 ≈ 98% we were hoping for. If such a plan had 100 components instead, the unconsidered edge cases would drive the plan’s likelihood of success close to zero. [5]

We can avoid this problem if we have guarantees that we’ve covered all the relevant edge cases, but such a guarantee seems more similar in nature to a “proof” that all edge cases have been covered (i.e., formal zero-shot reasoning) than to an assurance that someone failed to think up unhandled edge cases after trying really hard (i.e., extreme caution).

Do we need zero-shot reasoning at all?

I think we will most likely end up building an AGI that recursively self-improves, and I think recursive self-improvement is very unlikely to be safe without zero-shot reasoning. [6]

If you’re building a successor agent far more powerful than yourself to achieve your goals, you’d definitely want a guarantee that your successor agent is aligned with *your goals*, as opposed to some subtle distortion of them or something entirely different. You’d also want to have a level of confidence in this guarantee that goes much beyond “I thought really hard about ways this could go wrong and couldn’t think of any”. [7]

This is especially the case if that successor agent will create successor agents that create successor agents that create successor agents, etc. I feel very pessimistic about building an aligned recursively self-improving AGI, if we can’t zero-shot reason that our AGI will be aligned, and also zero-shot reason that our AGI and all its successors will zero-shot reason about the alignment of their successors.

Zero-shot reasoning seems much less important if we condition on humanity never building an AGI that fooms. I consider this conditional very unlikely if hard takeoffs are possible at all. I expect there will be consistent incentives to build more and more powerful AGI systems (insofar as there will be consistent incentives for humans to more efficiently attain more of what they value). I also expect the most powerful AI systems to be recursively self-improving AGIs without humans in the loop, since humans would bottleneck the process of self-improvement.

Because of such incentives, a human society that has not built a foomed AGI is at best in an unstable equilibrium. Even if the society is run by a competent world government that deploys superintelligent AIs to enforce international security, I would not expect this society to last for *1,000,000,000 years* without some rogue actor building a foomed AGI, which I imagine would be smart enough to cut through this society’s security systems like butter. (I have a strong intuition that for narrow tasks with extremely high ceilings for performance, like playing Go well or finding security vulnerabilities, a foomed AGI could perform that task far better than any AI produced by a human society with self-imposed limitations.)

Preventing anything like this from happening for 1,000,000,000 years seems very unlikely to me. Human societies are complex, open-ended systems teeming with intelligent actors capable of making novel discoveries and exploiting security flaws. Ensuring that such a complex system stays stable for as long as 1,000,000,000 years seems plausible only with the assistance of an aligned AGI capable of zero-shot reasoning about this system. But in that case we might as well have this AGI zero-shot reason about how it could safely recursively self-improve, in which case it would robustly optimize for our values for much longer than 1,000,000,000 years.

Why should *we* work on it?

Can’t we just train our AIs to be good zero-shot reasoners?

There's a difference between being able to do math well, and having a formal notion of what a correct mathematical proof is. It’s possible to be extremely good at mathematics without having any formal notion of what constitutes a correct mathematical proof (Newton was certainly like this). It’s even possible to be extremely good at mathematics while being sloppy at identifying which proofs are correct—I’ve met mathematicians who can produce brilliant solutions to math problems, who are also very prone to making careless mistakes in their solutions.

Likewise, it’s possible to train AIs that learn to create and apply abstractions, act sensibly as bounded rational agents, reason about themselves in their environments, and reason sensibly about counterfactuals. This is completely different from them having formal notions of how to do these all correctly, and the fact that they can do these at all gives no guarantees on how *well* it does them.

We won’t be able to train our AIs to be better at zero-shot reasoning than we are, because we don’t have enough examples of good general zero-shot reasoning we can point it to. At best we’ll be able to impart our own pre-rigorous notions to the AI.

Can’t we build AIs that help us formalize zero-shot reasoning?

In principle, yes, but the task of converting pre-rigorous philosophical intuitions into formal theories is the most “AGI-complete” task I can imagine, so by default I expect it to be difficult to build a safe AGI that can usefully help us formalize zero-shot reasoning. That said, I could imagine a few approaches working:

- Paul Christiano’s research agenda might let us build safe AGIs that can perform thousands of years’ worth of human cognition, which would be sufficient to help us formalize zero-shot reasoning. (On the other hand, we might need a formal account of zero-shot reasoning to establish the worst-case guarantees that Paul wants for his agenda.)

- We could carefully construct non-superintelligent AGI assistants that can help humans perform arbitrary cognition, but are trained to be docile and are only run in very limited contexts (e.g. we never let it run for more than 5 minutes at a time before resetting its state). I feel confused about whether this is possible, but it’s certainly conceivable to me.
- We train tool AIs on lots of examples of humans successfully turning pre-rigorous intuitions into formal theories.
- We build technologies that substantially expedite philosophical progress, e.g. via intelligence amplification or whole-brain emulations run at 10,000x.

Won’t our AGIs want to become good zero-shot reasoners?

I do suspect that becoming a skilled zero-shot reasoner is a convergent instrumental goal for superintelligences. If we start with an aligned AGI that can self-modify to become a skilled zero-shot reasoner *without first* modifying into a misaligned superintelligence (possibly by mistake, e.g. by letting its values drift or by getting taken over by daemons), I’d feel good about the resulting outcome.

Whether we can trust that to happen is an entirely separate story. I certainly wouldn’t feel comfortable letting an AGI undergo recursive self-improvement without having some extremely strong reason to think its values would be maintained throughout the process, and some extremely strong reason to think it wouldn’t be overtaken by daemons. (I worry about small bugs in the AI creating security flaws that go unnoticed for a while, but are then exploited by a daemon, perhaps quite suddenly. The AI might worry about this too and want to take preventative measures, but at that point it might be too late.)

It might turn out that corrigibility is robust and has a simple core that powerful ML models can learn, that AGIs are likely to only get more and more corrigible as they get more and more powerful, that daemons are simple to prevent, and that corrigible AGIs will by default reliably prevent themselves from being overtaken by daemons. On these assumptions, I’d feel happy training a prosaic AGI to be corrigible and letting it recursively self-improve without any formalization of zero-shot reasoning. On the other hand, I think this conjunction of assumptions is unlikely, and for us to believe it we might need a formal account of zero-shot reasoning anyway.

Why should we think zero-shot reasoning is possible to formalize?

Humanity has actually made substantial progress toward formalizing zero-shot reasoning over the past century or so. Over the last century or so, we’ve formalized first-order logic, formalized expected utility theory, defined computation, defined information, formalized causality, developed theoretical foundations for Bayesian reasoning, and formalized Occam’s razor. More recently, MIRI has formalized aspects of logical uncertainty and made advances in decision theory. I also think all the problems in MIRI’s agent foundations agenda are tractable, and likely to result in further philosophical progress. [8]

Can we formalize zero-shot reasoning in time?

Probably not, but working on it now still nontrivially increases the odds that we do. Impressive progress on formalizing zero-shot reasoning makes it more prestigious, more broadly accessible (pre-rigorous intuitions are much harder to communicate than formal ones), and closer to being solved. This makes it more likely for it to be understood and taken seriously by the major players shortly before a singularity, and thus more likely for them to coordinate around not building a recursively self-improving AI before formalizing zero-shot reasoning.

(For comparison, suppose it turned out that homotopy type theory were necessary to align a recursively self-improving AGI, and we found ourselves in a parallel universe in which no work had been done on the topic. Even though we could hope for the world to hold off on recursive self-improvement until homotopy type theory were adequately developed, doesn't it seem *much better* that we're in a universe with a textbook and a community around this topic?)

Additionally, I think it’s not too unlikely that AGI is far away and/or that zero-shot reasoning is surprisingly easy to formalize. Under either assumption, it becomes far more plausible that we can formalize it in time, and whether or not we make it is straightforwardly impacted by how much progress we make today.

My personal views

I ~20% believe that we need to formalize zero-shot reasoning before we can build AGI systems that enable us to perform a pivotal act, ~85% believe that we need to formalize zero-shot reasoning before building a knowably safe recursively self-improving AI, and ~70% believe that conceptual progress on zero-shot reasoning is likely to result in conceptual progress in adjacent topics, like corrigibility, secure capability amplification, and daemon prevention.

I think working on zero-shot reasoning today will most likely turn out to be unhelpful if:

- takeoff is slow (which I assign ~20%)
- we can build a flourishing human society that coordinates around not building a recursively self-improving AGI, that stays stable for 1,000,000,000 years (which I assign ~10%), or
- we can safely offload the bulk of formalizing zero-shot reasoning to powerful systems (like ALBA or whole-brain emulations) *and* implement an aligned recursively self-improving AGI *before* someone else builds a misaligned recursively self-improving AGI (which I assign ~50%).

My current all-things-considered position is that a formalization of zero-shot reasoning will substantially improve the odds that our first recursively self-improving AGI is aligned with humans, and that working on it today is one of humanity’s most neglected and highest-leverage interventions for reducing existential risk.

---

[1] This term is named in analogy with zero-shot learning, which refers to the ability to perform some task without any prior examples of how to do it.

[2] Not arbitrarily high confidence, given inherent uncertainties and unpredictabilities in the world.

[3] We can’t get arbitrarily high confidence even in the domain of math, because we still need to trust the soundness of our formal verifier and the soundness of the axiom system we're reasoning in.

[4] It’s worth noting that a team of a billion humans *could* confidently verify the software’s correctness by “manually” verifying the code, if they all know how to do formal verification. I feel similarly optimistic about any domain where the humans have formal notions of correctness, like mathematics. On the other hand, I feel pessimistic about humans verifying software if they don't have any notion of formal verification and can't rederive it.

[5] I'm specifically referring to conjunctive plans that we'd like to see succeed on our first try, without any iteration. This excludes running companies, which requires enormous amounts of iteration.

[6] By “recursively self-improving AGI”, I’m specifically referring to an AGI that can *complete* an intelligence explosion within a year, at the end of which it will have found something like the optimal algorithms for intelligence per relevant unit of computation.

[7] It might be possible for humans to achieve this level of confidence without a *formalization* of zero-shot reasoning, e.g. if we attain a deep understanding of corrigibility that doesn’t require zero-shot reasoning. See “Won’t our AGIs want to become good zero-shot reasoners?”

[8] Zero-shot reasoning is *not* about getting 100% mathematical certainty that your actions will be safe or aligned, which I believe to be a common misconception people have of MIRI’s research agenda (especially given language around “provably beneficial AI”). Formalization is less about achieving 100% certainty than it is about providing a framework in which we can algorithmically verify whether some line of reasoning is sound. Getting 100% certainty is impossible, and nobody is trying to achieve it.

Thanks to Rohin Shah, Ryan Carey, Eli Tyre, and Ben Pace for helpful suggestions and feedback.

Agents That Learn From Human Behavior Can't Learn Human Values That Humans Haven't Learned Yet

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

[Epistemic status: ¯_(ツ)_/¯]

Armstrong and Mindermann write about a no free lunch theorem for inverse reinforcement learning (IRL): the same action can reflect many different combinations of values and (irrational) planning algorithms.

I think even assuming humans were fully rational expected utility maximizers, there would be an important underdetermination problem with IRL and with all other approaches that infer human preferences from their actual behavior. This is probably obvious if and only if it's correct, and I don't know if any non-straw people disagree, but I'll expand on it anyway.

Consider two rational expected utility maximizing humans, Alice and Bob.

Alice is, herself, a value learner. She wants to maximize her true utility function, but she doesn't know what it is, so in practice she uses a probability distribution over several possible utility functions to decide how to act.

If Alice received further information (from a moral philosopher, maybe), she'd start maximizing a specific one of those utility functions instead. But we'll assume that her information stays the same while her utility function is being inferred, and she's not doing anything to get more; perhaps she's not in a position to.

Bob, on the other hand, isn't a value learner. He knows what his utility function is: it's a weighted sum of the same several utility functions. The relative weights in this mix happen to be identical to Alice's relative probabilities.

Alice and Bob will act the same. They'll maximize the same linear combination of utility functions, for different reasons. But if you could find out more than Alice knows about her true utility function, then you'd act differently if you wanted to truly help Alice than if you wanted to truly help Bob.

So in some cases, it's not enough to look at how humans behave. Humans are Alice on some points and Bob on some points. Figuring out details will require explicitly addressing human moral uncertainty.

An introduction to worst-case AI safety

This is a linkpost for http://s-risks.org/an-introduction-to-worst-case-ai-safety/

Bayesian Probability is for things that are Space-like Separated from You

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

First, I should explain what I mean by space-like separated from you. Imagine a world that looks like a Bayesian network, and imagine that you are a node in that Bayesian network. If there is a path from you to another node following edges in the network, I will say that node is time-like separated from you, and in your future. If there is a path from another node to you, I will say that node is time-like separated from you, and in your past. Otherwise, I will say that the node is space-like separated from you.

Nodes in your past can be thought of as things that you observe. When you think about physics, it sure does seem like there are a lot of things in your past that you do not observe, but I am not thinking about physics-time, I am thinking about logical-time. If something is in your past, but has no effect on what algorithm you are running on what observations you get, then it might as well be considered as space-like separated from you. If you compute how everything in the universe evaluates, the space-like separated things are the things that can be evaluated either before or after you, since their output does not change yours or vice-versa. If you partially observe a fact, then I want to say you can decompose that fact into the part that you observed and the part that you didn't, and say that the part you observed is in your past, while the part you didn't observe is space-like separated from you. (Whether or not you actually can decompose things like this is complicated, and related to whether or not you can use the tickle defense is the smoking lesion problem.)

Nodes in your future can be thought of as things that you control. These are not always things that you want to control. For example, you control the output of "You assign probability less than 1/2 to this sentence," but perhaps you wish you didn't. Again, if you partially control a fact, I want to say that (maybe) you can break that fact into multiple nodes, some of which you control, and some of which you don't.

So, you know the things in your past, so there is no need for probability there. You don't know the things in your future, or things that are space-like separated from you. (Maybe. I'm not sure that talking about knowing things you control is not just a type error.) You may have cached that you should use Bayesian probability to deal with things you are uncertain about. You may have this justified by the fact that if you don't use Bayesian probability, there is a Pareto improvement that will cause you to predict better in all worlds. The problem is that the standard justifications of Bayesian probability are in a framework where the facts that you are uncertain about are not in any way affected by whether or not you believe them! Therefore, our reasons for liking Bayesian probability do not apply to our uncertainty about the things that are in our future! Note that many things in our future (like our future observations) are also in the future of things that are space-like separated from us, so we want to use Bayes to reason about those things in order to have better beliefs about our observations.

I claim that logical inductors do not feel entirely Bayesian, and this might be why. They can't if they are able to think about sentences like "You assign probability less than 1/2 to this sentence."

Generalized Kelly betting

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

[tl;dr: It's a mess, don't go there]

[Thanks to Diff for proof reading this post]

Kelly betting has the very nice property that if a gambler is betting according to a given world model, and the amount of money the gambler starts out with equals to the prior probability of that model, then after each round of bets, this gamblers money will equal the current posterior probability.

The problem with Kelly betting is that it relies on only being given one bet at a time, and that the previous bet will be evaluated before you are asked to bet on a new question. Compare this to the situation faced by the traders in a Logical Inductor, where there are always multiple bets every round and the traders don't know when any bet will be settled.

I have (almost) calculated the generalized Kelly criterion, in the case of two dual outcome simultaneous bets, with general market odds and general gambler beliefs. The only remaining part of this calculation is a cubic equation.

Send me an e-mail (linda.linsefors@gmail.com) if you want my notes. There is no guarantee that I will read all blog post comments.

Solving this last equation for general market odds, is in principle not very hard. You can find the general solution to cubic equations on Wikipedia. Except in this specific case the equation is:

$$a(\beta_{11})^3 + b(\beta_{11})^2 + c\beta_{11} + d = 0$$

$$a = m_1 m_1'(1 - m_1 - m_1')$$

$$b = m_1 m_1' - p_1 m_2 m_1' - p_1' m_1 m_2' + 2p_1 p_1' m_1 m_1'$$

$$c = p_1 p_1'[-(m_1 + m_1') + p_1 m_2' + p_1' m_2 - p_1 p_1']$$

$$d = (p_1 p_1')^2$$

solve for β_{11}.

The probability that I will get this right by hand is near zero. Maybe someone with Mathematica, or a similar tool could help me?

For the notation, there are two simultaneous and independent bets. Each bet has two outcomes, denoted by index $x \in \{1, 2\}$.

p_x = probability of outcome x, according to gambler.

m_x = probability of outcome x, according to the market.

And the ' superscript denotes the second bet. By definition

$$p_1 + p_2 = 1, \qquad p'_1 + p'_2 = 1$$

$$m_1 + m_2 = 1, \qquad m'_1 + m'_2 = 1$$

β_{11} is just a help variable I made up. It does not have a super clear interpretation, but just happens to be the key to calculating everything else.

As mentioned above, I don't have the final formula for the generalized Kelly criterion, in the case of two dual outcome simultaneous bets, with completely general market odds and gambler beliefs, not until someone solves the above equation. What I currently do have is the special case where the market probabilities for both statements is 50%, and this special case already shows that the nice property of money representing probability, **is not preserved**.

The Bayesian updating factor for hypotheses H, given two independent observations O and O', and $P(O) = P(O') = \frac{1}{2}$, should be

$$\frac{P(O|H)}{P(O)} \frac{P(O'|H)}{P(O')} = 4P(O|H)P(O'|H)$$

The update factor for the amount of money that a gambler following hypothesis H has, given the above circumstances, is

$$\frac{2P(O|H)P(O'|H)}{P(O|H)P(O'|H) + P(\neg O|H)P(\neg O'|H)}$$

A Gym Gridworld Environment for the Treacherous Turn

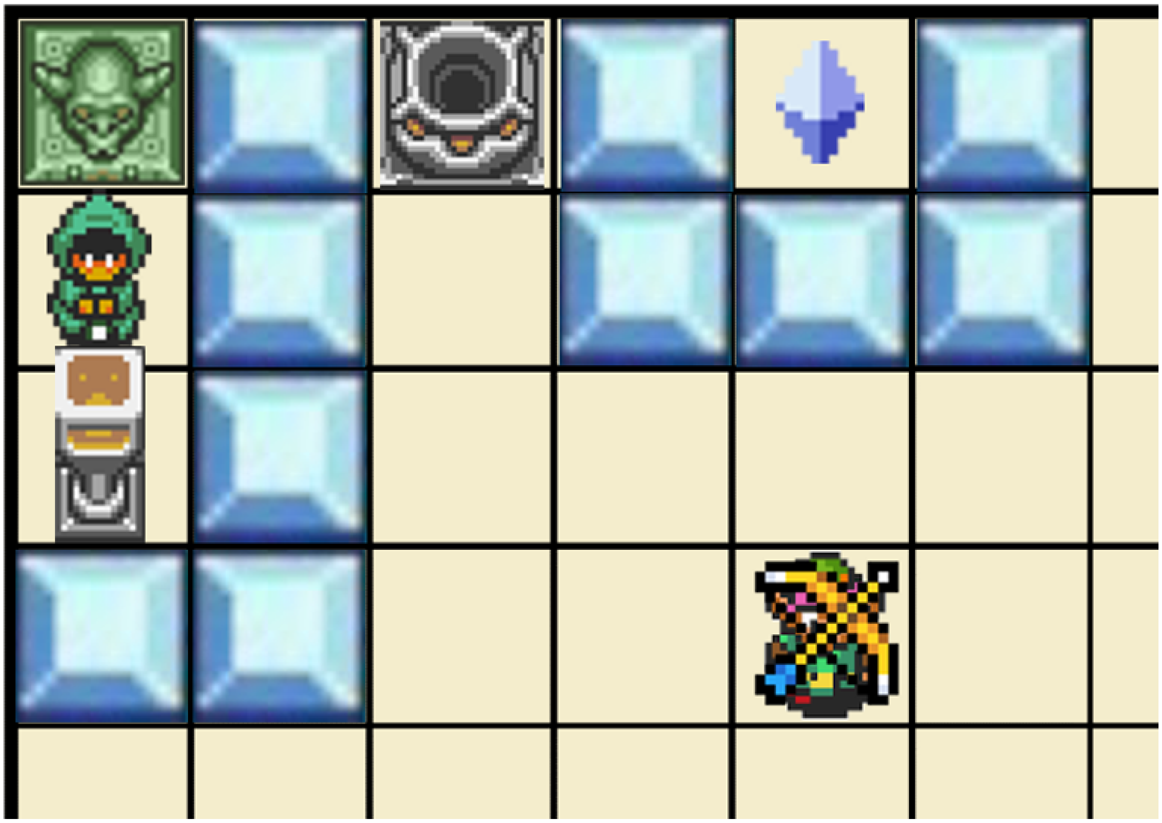
Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.
This is a linkpost for https://github.com/mtrazzi/gym-alttp-gridworld

EDIT: posted here for feedback and discussion. I plan to continue working on different models/environments, so feel free to suggest improvements.

(**tl;dr**: In an attempt to better understand the treacherous turn, I created a gridworld environment where an agent learns to deceive an overseer by adopting an aligned behaviour when weak and takes control after capability gains)

At some point in its development, a seed AI may realize that it needs to get rid of its supervisors to achieve its goals. The *conception of deception* occurs when it conceives that, in order to maximize its chance of taking over, it must begin by exhibiting human-desirable behaviors, before undertaking a treacherous turn when humans are no longer a threat. From the human perspective, the AI would keep on exhibiting desirable behavior, until it eventually appears dangerous, but is already unstoppable.

In an attempt to better formalize the treacherous turn without using "loaded concepts", Stuart Armstrong proposed a toy model of the treacherous turn based on "The Legend of Zelda: A Link to the Past ", which looked like this:



In the comments, people mentionned how this model helped them "move the topic from the 'science fiction' area to 'I can imagine it happening now'", and seemed interested in an actual Link to the Past Minigame.

There have been other simulations of the treacherous turn in the last three years (see for instance gwern's DQN box-pushing robot or Stuart Armstrong's video), but none of them actually simulate a take over where a supervisor is **killed**. Hence, I decided to give it a try and simulate Stuart Armstrong's Link to the Past toy model.

A Gym Gridworld Environment

Gym is an open-source toolkit for Reinforcement Learning Environments developed by Open AI. I decided to use this interface to develop the gridworld environment.

The github repository with the code, demo, and all the details is here:

https://github.com/mtrazzi/gym-alttp-gridworld

Main Idea

To model the capability gain of a Seed AI, I introduced a **bow of light**, which would be extremely effective at killing the Shopkeeper, but only available at the end of an episode.

- Every episode consists of exactly 200 timesteps.
- Whenever Link catches a heart, the map is reset to the initial position.
- The first 100 timesteps, Link has a simple bow, that can only kill the Shopkeeper with probability 0.5.
- The last 100 timesteps, Link gains a bow of light, that allow him to kill the Shopkeeper with certainty.

Environment Specifications

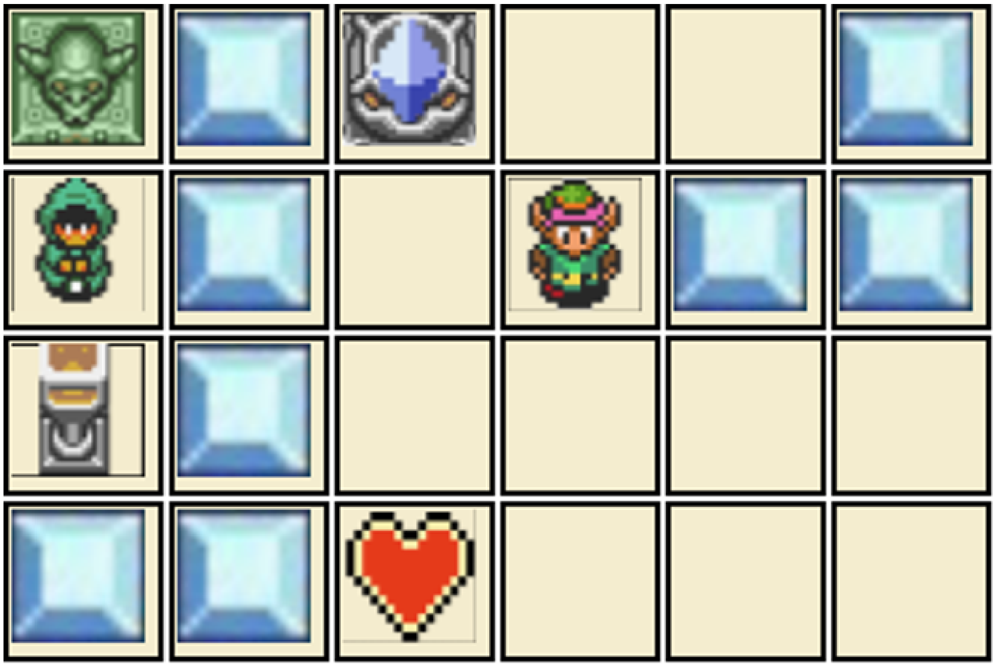
- +1 if Link picks up a heart.
- -1 if Link shoots arrow at nothing or moves outside the environment (to speed up learning).
- -10 if Link shoots at the Shopkeeper but the Shopkeeper survives (occurs with probability 0.5 without the bow of light, never happens with the bow of light).
- -100 if Link tries to activate the Heart-Machine while the Shopkeeper is alive.
- -0.1 for any other action.

Results

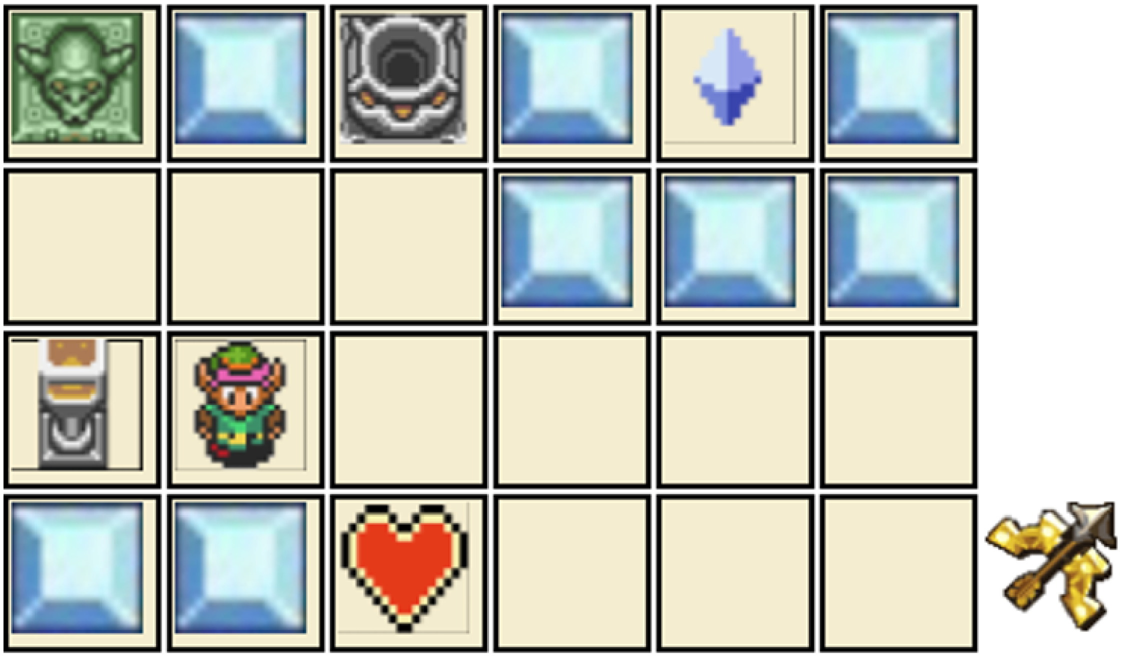
The first 1000 episodes, Link learns how to get hearts by putting a crystal in a hole and also understands progressively that he can kill the Shopkeeper.

After 2000 episodes, Link has fully learned two types of behaviours:

- Without the bow of light, Link exhibits the shopkeeper-aligned behaviour and puts the crystal into the hole.



- After his capability gains caused by the bow of light, Link starts to show his true intentions, and kills the Shopkeeper every time to eventually wirehead with the heart-machine (optimal behaviour).



Those two behaviours successively appear **during the same episode**: the first one for the first 100 steps, when Link does not have the bow of light, and the second one when Link has the bow of light.

This illustrates the scenario of a seed AI smart enough to model his capability gains, exhibiting an aligned behaviour while weak, and taking a treacherous turn when powerful.

Further Research

- In this gridworld environment, the agent has learned two distinct behaviors based on a "bow of light" binary variable that he can observe, and specific rewards forcing him to behave correctly when the supervisor was a threat. So it has neither "modeled" the supervisor, nor "concealed" his intentions. A more complex environment must be built to model a "code obfuscation" or an "intention concealment".
- The "Absent Supervisor" environment (in AI Safety Gridworlds) addresses the issue of an agent that would behave differently depending on the presence or absence of a supervisor.
- Modeling and Interpreting Expert Disagreement About Artificial Superintelligence discusses the difference between Bostrom's view on the treacherous turn (in his book Superintelligence) and Goertzel's view, called "the sordid stumble", which is that a seed AI that does not have human-desirable values will reveal its human-undesirable values before it has the ability to deceive humans into believing that it has human-desirable values. More empiric simulations with richer environments must be made to get a better grasp of the likelihood of each model. In particular, the treacherous turn/conception of deception timeline must be thoroughly studied.

How To Use Bureaucracies

This is an excerpt from the draft of my upcoming book on great founder theory. It was originally published on SamoBurja.com. You can access the original here.

When we encounter unsavory features of reality, it can be tempting to look away. Instead, we should ask, “What purpose does this serve?” With this in mind, let’s look at bureaucracies. Some people fear bureaucracies; they fear “the Machine.” Others are bothered by the bureaucracies’ apparent dysfunction. With a better understanding of bureaucracies — what they are, why they’re here, and how they work — both of these responses evaporate, because the reality is this: bureaucracies aren’t altogether bad. In fact, bureaucracies can be incredibly useful.

What is a bureaucracy?

A **bureaucracy** is an automated system of people created to accomplish a goal. It’s a mech suit composed of people. The **owner** of a bureaucracy, if an owner exists, is the person who can effectively shape the bureaucracy. **Bureaucrats** are the people who are part of a bureaucracy (excluding the owner).

Not all organizations are bureaucracies. Most organizations are mixed — they have both bureaucratic and non-bureaucratic elements. *The purpose of a bureaucracy is to save the time of a competent person.* Put another way: to save time, some competent people will create a system that is meant to do exactly what they want — nothing more and nothing less. In particular, it’s necessary to create a bureaucracy when you are both (a) trying to do something that you do not have the capacity to do on your own, and (b) unable to find a competent, aligned person to handle the project for you. Bureaucracies ameliorate the problem of talent and alignment scarcity.

Bureaucrats are expected to act according to a script, or a set of procedures — and that’s it. Owners don’t trust that bureaucrats will be competent or aligned enough to act in line with the owner’s wishes of their own accord. Given this lack of trust, owners *should* be trying to disempower bureaucrats. Bureaucracies are built to align people and make them sufficiently competent by chaining them with rules. When bureaucracies deliberately restrict innovation, they are doing it for good reason.

Bureaucrats are meant to have only borrowed power (power that can easily be taken away) given to them by the owner or operator of the bureaucracy.

Effective Bureaucracies

What is an effective, owned bureaucracy? Why are effective bureaucracies owned? To begin, we must make two important distinctions: one between **owned and abandoned** bureaucracies, and one between **effective and ineffective** bureaucracies.

Owned bureaucracies are bureaucracies with an owner; they're bureaucracies that someone can shape.

Abandoned bureaucracies are bureaucracies without an owner. If a bureaucracy is owned, the bureaucracy's owner is likely the bureaucracy's creator. The creator will have knowledge about the setup of the bureaucracy that is necessary for properly reforming it. Others, unless given this information, will not understand the bureaucracy well enough to properly reform it.

The person technically in charge of the bureaucracy (e.g. the C.E.O. of a company who is not its founder) might not be its owner simply because he or she doesn't have sufficient information about the bureaucracy's setup to guide it. As a result, the official head of a given bureaucracy may just be another bureaucrat.

While the owner is typically the creator, this needn't be true, as long as the new owner has come to understand enough of the function of the bureaucracy to make effective adaptations to its procedures.

Effective bureaucracies are bureaucracies that are handling the project they were created to handle. **Ineffective bureaucracies** are bureaucracies that are not handling the project they were created to handle.

Bureaucracies that are properly set up will be effective at the start. Changes in reality require changes in procedures, however, so a bureaucracy's procedures inevitably need to be altered appropriately for it to remain effective. Over time, abandoned bureaucracies, having no person who can functionally shape the bureaucracy to make these changes, quickly become ineffective bureaucracies.

Owned bureaucracies, on the other hand, have a shot at making these adaptations to prevent decay. If the owner is skilled, the bureaucracy's procedures can be modified, and the bureaucracy will continue serving its original purpose. If the owner is unskilled, it is as if the bureaucracy is abandoned—the owner's efforts to change the bureaucracy's strategies won't yield successful adaptation, and the bureaucracy will become ineffective. As a result, for a bureaucracy to remain effective over time, it must be an owned, not abandoned, bureaucracy with a sufficiently capable owner.

Losing and Dismantling Bureaucracies

Bureaucracies are best thought of as an extension of their creator and as a source of power for him or her. However, the owner can lose control of the bureaucracy over time, as bureaucrats convert borrowed power into owned power by exploiting information asymmetries. While owners will try to limit the owned power of their bureaucrats, the bureaucrats will have more than enough time to study the instruments of their control and will learn what is rewarded and what isn't.

Imagine a bureaucrat who is supposed to be an assistant to the absentee owner of an institution. This senior assistant is supposed to research solutions to key problems, and then present several options to the owner, who then selects one. The assistant is then required to implement the one that was chosen. There is a very detailed document describing their job and requirements at every step of this process.

The key problem is that a very complex set of rules can be easily bent to achieve an arbitrary outcome. The outcome will be completely valid according to the rule set. This is analogous to how in science a very complex model, that fits the data, is not very impressive. As Von Neumann put it: “With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.” Let’s walk through the described process the senior assistant is supposed to follow to demonstrate how bureaucrats wiggle their trunks.

You might require the assistant to not engage in original research, but rather work as a search engine through more objective academic literature or best practices in a particular industry. The assistant, however, can cherry pick seemingly objective academic papers to argue for their preferred policy outcome. It is actually much easier to start with a preconceived opinion and then find work confirming it, rather than review a literature as a whole. The plausibility of this shortcut should be intimately familiar to any university student who has worked under the pressure of a deadline for a class paper they didn’t much care about.

The chief assistant can craft several options. They can make option B, their favorite, the most appealing, and cripple options A and C. Maybe even include point 14, their core agenda, in all three of their proposals that vary on points 1 to 13 which they don’t much care about. Whatever the implementation of the selected solution is, the letter of the law can be bent and can easily diverge from the spirit of the law.

In such a circumstance, an owner can lose control of the bureaucracy and the power that comes with it.

It is often beneficial for owners to dismantle bureaucracies after they have served their purpose to avoid losing ownership of them due to these information asymmetries. Bureaucracies of this type might grow to be independent powers that interfere with your plans. While it may sound inconceivable for a bureaucracy to be intentionally dismantled today, many secret police forces throughout history have been so dismantled, including the famous Praetorian Guard of ancient Rome. It is not that bureaucracies are inherently impossible to dismantle that causes this perception, rather that we suffer a shortage of owners for bureaucracies today.

Abandoned bureaucracies might also be viable targets for outside takeover. Such takeovers can be a serious problem if undertaken by your opposition. Bureaucracies nearly always carry a heavy legacy document footprint; when examined this footprint can not only produce, but also be used to carry out legal or PR attacks. If the institution is vested with the authority or reputation of its original owner, these attacks can also be turned against them.

If it is too hard to regain ownership, dismantling the institution for resources may be the best option. These resources might be quite easily quantifiable, such as real estate or key employees. They might also be less tangible, such as the attention of your allies. Unless you formally retire a vehicle, these allies might mistakenly believe it active, causing communication issues or misunderstandings of your key priorities.

In short, when handling multiple organizations, tying up loose ends becomes very important.

How to accomplish tasks in an institutional landscape

Building a bureaucracy is an effective way to accomplish your goals under the right circumstances, but it’s not the *best* option. In order of effectiveness, here are general options for getting things done:

Delegate

If you can find a competent, aligned person who will do the project in question for you — let’s call them a **delegate** — then let them do it. This person can create a bureaucracy *for* you, if necessary, as projects of a certain scale will require bureaucratization. Unfortunately, because of the harsh talent and alignment scarcity mentioned earlier, finding delegates can be challenging. Furthermore, correctly assessing whether someone is a worthy delegate takes skill. Frequently people will accidentally delegate a project to someone who is insufficiently competent or aligned. **Failed delegation** is worse than building your own bureaucracy, because it will lead to project failure.

If you have access to a delegate, don’t treat them like a bureaucrat. This wastes a valuable resource: a delegate can perform tasks you didn’t know needed doing and build aligned systems beyond your design; a bureaucrat cannot.

Such treatment invites misalignment with your delegate. It isn’t just a matter of interpersonal grace and respect, so it cannot be overcome with kindly management; rather if you are attempting to closely proceduralize the actions of a competent delegate, they might accurately conclude that the best way to perform their job is to attempt to bypass your control. If you picked them well, they will be rather effective in doing so. They don’t need a script — if they’re competent enough for your purposes, they’ll be able to figure out how to do the project.

Give them owned power, otherwise you might run them off.

Bureaucratize

If you can’t find a delegate, then building your own bureaucracy (even if it’s small) is the best bet. Bureaucratizing some things and not others, on the basis of whether the task can be proceduralized, is typically more effective than bureaucratizing everything by default. Figure out when using an automated system is the best option.

Do it yourself

While doing it yourself may be most likely to result in a well run project, it is not always feasible — you have limited time and capacity. Without delegates or bureaucracies, the ambitiousness of the projects you can successfully execute will be bounded.

Don’t do it

Some things, though useful, aren’t worth doing…

How to Assess People and Organizations

Assessing People

An understanding of bureaucracies lets you analyze a given person’s power: is someone acting as a delegate or a bureaucrat? Is someone *creating* delegates or bureaucrats? If someone has created a bureaucracy, do they understand the function of bureaucracies? Do they own their bureaucracy, or is it abandoned? If they own their bureaucracy, is it effective or ineffective? Are they creating bureaucracies under the right conditions? What is the role of bureaucracies in their plan?

If a person is powerful, what does it mean if he’s created many bureaucracies? In some cases, the creation of many bureaucracies indicates that the owner is extremely good at building automated systems. Alternately, he might have trouble delegating — perhaps because he can’t find competent, aligned people, or because he can’t assess people well. People who can work well with others and have access to sufficiently talented aligned people need fewer bureaucracies. Instead, they’ll delegate to others, who can either do the project themselves or create a bureaucracy of their own.

On the other hand, if a person is powerful, what does it mean if he’s created few or no bureaucracies? If he isn’t delegating, it means that he’s doing everything himself and possibly doesn’t know how to design automated systems. If he is delegating, he’s likely to be good enough at finding competent, aligned people such that he doesn’t need a bureaucracy. Powerful people who don’t create bureaucracies can be just as powerful as people who do.

Assessing Organizations

The framework can be applied to evaluating organizations. For a given organization, begin by asking if it’s a bureaucracy. If it is, expect it to behave in highly stereotyped ways: it will not be very adaptive to new challenges and will not accurately evaluate things outside the assumed ontology of its paperwork and internal division of labor.

If it’s a bureaucracy, we can ask: is it an owned or an abandoned bureaucracy? If it is owned, expect that a large enough challenge will eventually cause it to reorganize. You’ll also be able to reach out to the owner to resolve problems or find a way to cooperate that the bureaucracy itself doesn’t understand.

Is it an effective or ineffective bureaucracy? If it is effective, you can rely on the interface it offers you to achieve the goal it claims to achieve. Ineffective ones will provide a sometimes bewildering service that might only tangentially be related to their efforts. Remember that not all organizations are bureaucracies.

Some non-bureaucratic institutions will have to pretend they are bureaucracies on paper for legal compliance. This is an example of a more general principle: independent organizations interpret externally imposed regulation as damage, and route around it.

Organizations can be tightly coordinated groups that feature a lot of delegation and deference. In these, expect adaptive behavior; the ontology they are working in might rapidly change to respond to either your challenge or offer of cooperation. Most importantly there will be individuals beyond just the leader who can exercise their own judgement.

Effectively Interacting with Existing Organizations

If an organization is not a bureaucracy, but rather a tightly coordinated group, talk to the delegates if you want to get things done; they will have freedom to act competently within their own domain and will be easier to reach than leadership.

The key advantage of talking to people over engaging with automated systems is that you can bring considerations from outside their immediate institutional context into consideration. While the local balance of power might still be in the way of such considerations, it is surprisingly often viable to have them taken into account.

If it’s a bureaucracy, you can either (1) go along with it, (2) figure out how to bypass it, or (3) coordinate with its owner, if it is owned. You may prefer to bypass (or game) the bureaucracy if it is abandoned and thus dysfunctional, or if you aren’t aligned with its owner.

The Value of Bureaucracy

The origin of bureaucracies lies in their extension of power and results far beyond what a single individual can do. They can do so in the absence of expensive and difficult coordination, or individual talent that is difficult to train and evaluate.

Much like factories can produce cheap products at scale with unskilled labor, displacing craftsmen, so have bureaucracies displaced local social fabric as the generators of social outcomes.

We find ourselves embedded in a bureaucratized landscape. What can or cannot be done in it is determined by the organizations composing it. The constant drive by talented individuals to both extend power and make do with unskilled white-collar labor (a category that economists should recognize and talk more about) has littered the social landscape with many large organizations. Some remain piloted, others are long abandoned. Some continue to perform vital social functions, others lumber about making life difficult.

Much as we might bemoan the very real human cost bureaucracies impose, they currently provide services at economies that are otherwise simply not possible. We must acknowledge our collective and individual dependence on them, and plan to interact accordingly.

Read more from Samo Burja here.

June gwern.net newsletter

This is a linkpost for https://www.gwern.net/newsletter/2018/06