Best of LessWrong: July 2013

Best of LessWrong: July 2013

"Stupid" questions thread

r/Fitness does a weekly "Moronic Monday", a judgment-free thread where people can ask questions that they would ordinarily feel embarrassed for not knowing the answer to. I thought this seemed like a useful thing to have here - after all, the concepts discussed on LessWrong are probably at least a little harder to grasp than those of weightlifting. Plus, I have a few stupid questions of my own, so it doesn't seem unreasonable that other people might as well.

A New Interpretation of the Marshmallow Test

I've begun to notice a pattern with experiments in behavioral economics. An experiment produces a result that's counter-intuitive and surprising, and demonstrates that people don't behave as rationally as expected. Then, as time passes, other researchers contrive different versions of the experiment that show the experiment may not have been about what we thought it was about in the first place. For example, in the dictator game, Jeffrey Winking and Nicholas Mizer changed the experiment so that the participants didn't know each other and the subjects didn't know they were in an experiment. With this simple adjustment that made the conditions of the game more realistic, the "dictators" switched from giving away a large portion of their unearned gains to giving away nothing. Now it's happened to the marshmallow test.

In the original Stanford marshmallow experiment, children were given one marshmallow. They could eat the marshmallow right away; or, if they waited fifteen minutes for the experimenter to return without eating the marshmallow, they'd get a second marshmallow. Even more interestingly, in follow-up studies two decades later, the children who waited longer for the second marshmallow, i.e. showed delayed gratification, had higher SAT scores, school performance, and even improved Body Mass Index. This is normally interpreted as indicating the importance of self-control and delayed gratification for life success.

Not so fast.

In a new variant of the experiment entitled (I kid you not) "Rational snacking", Celeste Kidd, Holly Palmeri, and Richard N. Aslin from the University of Rochester gave the children a similar test with an interesting twist.

They assigned 28 children to two groups asked to perform art projects. Children in the first group each received half a container of used crayons, and were told that if they could wait, the researcher would bring them more and better art supplies. However, after two and a half minutes, the adult returned and told the child they had made a mistake, and there were no more art supplies so they'd have to use the original crayons.

In part 2, the adult gave the child a single sticker and told the child that if they waited, the adult would bring them more stickers to use. Again the adult reneged.

Children in the second group went through the same routine except this time the adult fulfilled their promises, bringing the children more and better art supplies and several large stickers.

After these two events, the experimenters repeated the classic marshmallow test with both groups. The results demonstrated children were a lot more rational than we might have thought. Of the 14 children in group 1, who had been shown that the experimenters were unreliable adults, 13 of them ate the first marshmallow. 8 of the 14 children in the reliable adult group, waited out the fifteen minutes. On average children in unreliable group 1 waited only 3 minutes, and those in reliable group 2 waited 12 minutes.

So maybe what the longitudinal studies show is that children who come from an environment where they have learned to be more trusting have better life outcomes. I make absolutely no claims as to which direction the arrow of causality may run, or whether it's pure correlation with other factors. For instance, maybe breastfeeding increases both trust and academic performance. But any way you interpret these results, the case for the importance and even the existence of innate self-control is looking a lot weaker.

How I Became More Ambitious

Follow-up to How I Ended Up Non-Ambitious

Living with yourself is a bit like having a preteen and watching them get taller; the changes happen so slowly that it's almost impossible to notice them, until you stumble across an old point of comparison and it becomes blindingly obvious. I hit that point a few days ago, while planning what I might want to talk about during an OkCupid date. My brain produced the following thought: "well, if this topic comes up, it might sound like I'm trying to take over the world, and that's intimidating- Wait. What?"

I'm not trying to take over the world. It sounds like a lot of work, and not my comparative advantage. If it seemed necessary, I would point out the problems that needed solving and delegate them to CFAR alumni with more domain-specific expertise than me.

However, I went back and reread the post linked at the beginning, and I no longer feel much kinship with that person. This is a change that happened maybe 25-50% deliberately, and the rest by drift, but I still changed my mind, so I will try to detail the particular changes, and what I think led to them. Introspection is unreliable, so I'll probably be at least 50% wrong, but what can you do?

1. Idealism versus practicality

I would still call myself practical, but I no longer think that this comes at the expense of idealism. Idealism is absolutely essential, if you want to have a world that changes because someone wanted it to, as opposed to just by drift. Lately in the rationalist/CFAR/LW community, there's been a lot of emphasis on agency and agentiness, which basically mean the ability to change the world and/or yourself deliberately, on purpose, through planned actions. This is hard. The first step is idealism-being able to imagine a state of affairs that is different and better. Then comes practicality, the part where you sit down and work hard and actually get something done.

It's still true that idealism without practicality doesn't get much done, and practicality without idealism can get a lot done, but it matters what problems you're working on, too. Are you being strategic? Are you even thinking, at all, about whether your actions are helping to accomplish your goals? One of the big things I've learned, a year and a half and two CFAR workshops later, is how automatic and easy this lack of strategy really is.

I had a limited sort of idealism in high school; I wanted to do work that was important and relevant; but I was lazy about it. I wanted someone to tell me what was important to be doing right now. Nursing seemed like an awesome solution. It still seems like a solution, but recently I've admitted to myself, with a painful twinge, that it might not be the best way for me, personally, to help the greatest number of people using my current and potential skill set. It's worth spending a few minutes or hours looking for interesting and important problems to work on.

I don't think I had the mental vocabulary to think that thought a year and a half ago. Some of the change comes from having dated an economics student. Come to think of it, I expect some of his general ambition rubbed off on me, too. The rest of the change comes from hanging out with the effective altruism and similar communities.

I'm still practical. I exercise, eat well, go to bed on time, work lots of hours, spend my money wisely, and maintain my social circle mostly on autopilot; it requires effort but not deliberate effort. I'm lucky to have this skill. But I no longer think it's a virtue over and above idealism. Practical idealists make the biggest difference, and they're pretty cool to hang out with. I want to be one when I grow up.

2. Fear of failure

Don't get me wrong. If there's one deep, gripping, soul-crushing terror in my life, one thing that gives me literal nightmares, it's failure. Making mistakes. Not being good enough. Et cetera.

In the past few years, the main change has been admitting to myself that this terror doesn't make a lot of sense. First of all, it's completely miscalibrated. As Eliezer pointed out during a conversation on this, I don't fail at things very often. Far from being a success, this is likely a sign that the things I'm trying aren't nearly challenging enough.

My threshold for what constitutes failure is also fairly low. I made a couple of embarrassing mistakes during my spring clinical. Some part of my brain is convinced that this equals *permanent* failure; I wasn't perfect during the placement, and I can't go back and change the past, thus I have failed. Forever.

I passed the clinical, wrote the provincial exam (results aren't in but I'm >99% confident I passed) (EDIT: Passed! YEAAHHH!!!), and I'm currently working in the intensive care unit, which has been my dream since I was about fifteen. The part of my brain that keeps telling me I failed permanently obviously isn't saying anything useful.

I think 'embarrassing' is a keyword here. The first thing I thought, on the several occasions that I made mistakes, was "oh my god did I just kill someone... Phew, no, no harm done." The second thought was "oh my god, my preceptor will think I'm stupid forever and she'll never respect me and no one wants me around, I'm not good enough..." This line of thought never goes anywhere good. It says something about me, though, that "I'm not good enough" is very directly connected to people wanting me around, to belonging somewhere. For several personality-formative years of my life, people *didn't* want me around. Probably for good reason; my ten-year-old self was prickly and socially inept and miserable. I think a lot of my determination not to seek status comes from the "uncool kids trying to be cool are pathetic" meme that was so rampant when I was in sixth grade.

Oh, and then there's the traumatic swim team experience. Somewhere, in a part of my brain where I don't go very often nowadays, there a bottomless whirlpool of powerless rage and despair around the phrase "no matter how hard I try, I'll never be good enough." So when I make an embarrassing mistake, my ten-year-old self is screaming at me "no wonder everyone hates you!" and my fourteen-year-old self is sadly muttering that "you know, maybe you just don't have enough natural talent," and none of it is at all useful.

The thing about those phrases is that they refer to complex and value-laden concepts, in a way that makes them seem like innate attributes, à la Fundamental Attribution Error. "Not good enough" isn't a yes-or-no attribute of a person; it's a magical category that only sounds simple because it's a three-word phrase. I've gotten somewhat better at propagating this to my emotional self. Slightly. It's a work in progress.

During a conversation about this with Anna Salamon, she noted that she likes to approach her own emotions and ask them what they want. It sounds weird, but it's helpful. "Dear crushing sense of despair and unworthiness, what do you want? ...Oh, you're worried that you're going to end up an outcast from your tribe and starve to death in the wilderness because you accidentally gave an extra dose of digoxin? You want to signal remorse and regret and make sure everyone knows you're taking your failure seriously so that maybe they'll forgive you? Thank you for trying to protect me. But really, you don't need to worry about the starving-outcast thing. No one was harmed and no one is mad at you personally. Your friends and family couldn't care less. This mistake is data, but it's just as much data about the environment as it is about your attributes. These hand-copied medication records are the perfect medium for human error. Instead of signalling remorse, let's put some mental energy into getting rid of the environmental conditions that led to this mistake."

Rejection therapy and having a general CoZE [Comfort Zone Expansion] mindset helped remove some of the sting of "but I'll look stupid if I try something too hard and fail at it!" I still worry about the pain of future embarrassment, but I'm more likely to point out to myself that it's not a valid objection and I should do X anyway. Making "I want to become stronger" an explicit motto is new to the last year and a half, too, and helps by giving me ammunition for why potential embarrassment isn't a reason not to do something.

In conclusion: failure still sucks. I'm a perfectionist. But I failed in a lot of small ways during my spring clinical, and passed/got a job anyway, which seems to have helped me propagate to my emotional self that *it's okay to try hard things*, where I'm almost certain to make mistakes, because mistakes don't equal instant damnation and hatred from all of my friends.

3. The morality of ambition

While I was in San Francisco a month ago, volunteering at the CFAR workshop and generally spending my time surrounded by smart, passionate, and ambitious people (thus convincing my emotional system that this is normal and okay), I had a conversation with Eliezer. He asked me to list ten areas in which I was above average.

This was a lot more painful than it had any reason to be. After bouncing off various poorly-formed objections in my mind, I said to myself "you know, having trouble admitting what you're good at doesn't make you virtuous." This was painful; losing a source of feeling-virtuous always is. But it was helpful. Yeah, talking all the time about how awesome you are at X, Y, Z makes you a bit of a bore. People might even avoid you (oh! the horror!). However, this doesn't mean that blocking even the *thought* of being above average makes you a good person. In fact, it's counterproductive. How are you supposed to know what problems you're capable of solving in the world if you can't be honest with yourself about your capabilities?

This conversation helped. (Even if some of the effect was "high status person says X -> I believe X," who cares? I endorsed myself changing my mind about this a year and a half ago. It's about time.)

HPMOR helped, too; specifically, the idea that there are four houses which have different positive qualities. Slytherins are demonized in canon, but in HPMOR their skills are recognized as essential. I can easily recognize the Ravenclaw and Hufflepuff and even the Gryffindor in myself, but not much of Slytherin. Having a word for the ambition-cunning-strategic concept cluster is helpful. I can ask myself "now what would a Slytherin do with this information:?" I can think thoughts that feel very un-virtuous. "I'm young and prettier than average. What's a Slytherin way to use this... Oh, I suppose I can leverage it to get high-status men to pay attention to me long enough for me to explain the merits of an idea I have." This thought feels yuck, but the universe doesn't explode.

Probably the biggest factor was going to the CFAR workshops in the first place. Not from any of the curriculum, particularly, although the mindset of goal factoring helped me to realize that the mental action of "feeling unvirtuous for thinking in ambitious or calculating ways" wasn't accomplishing anything I wanted. Mostly the change came from social normalization, from hanging out with people who talked openly about their strengths and weaknesses, and no one got shunned.

[Silly plan for taking over the world: Arrange to meet high status-people and offer to give their children swimming lessons. Gain their trust. Proceed from there.]

4. Laziness

Nope. Still lazy. If anything, akrasia and procrastination are more of a problem now that I'm trying to do harder things more deliberately.

I've been keeping written goals for about a year now. This means I actually notice when I don't accomplish them.

I use Remember the Milk as a GTD system, and some other productivity/organization software (rescuetime, Mint.com, etc). I finally switched to Gmail, where I can use Boomerang and other useful tools. My current openness to trying new organization methods is high.

My general interest in trying things is higher, mainly because I have lots of community-endorsed-warm-fuzzies positive affect around that phrase. I want to be someone who's open to new experiences; I've had enough new experiences to realize how exhilarating they can be.

Conclusion

I now have a wider range of potentially high-value personal projects ongoing. I now have an explicit goal of being well-known for non-fiction writing, probably in a blog form, in the next five years. (Do I have enough interesting things to say to make this a reality? We'll see. Is this goal vague? Yes. Working on it. I used to reject goals if they weren't utterly concrete, but even vague goals are something to build on).

I'm more explicit with myself about what I want from CFAR curriculum skills. (The general problem of critical thinking in nursing? Solvable! Why not?)

I think I've finally admitted to myself that "well, I'll just live in a cozy little house near my parents and work in the ICU and raise kids for the next forty years" might not be particularly virtuous *or* fun. There are things I would prefer to be different in the world, even if I can only completely specify a few of them. There are exciting scary opportunities happening all the time. I'm lucky enough to belong to a community of people that can help me find them.

I don't have plans for much beyond the next year. But here's to the next decade being interesting!

Model Combination and Adjustment

The debate on the proper use of inside and outside views has raged for some time now. I suggest a way forward, building on a family of methods commonly used in statistics and machine learning to address this issue — an approach I'll call "model combination and adjustment."

Inside and outside views: a quick review

1. There are two ways you might predict outcomes for a phenomenon. If you make your predictions using a detailed visualization of how something works, you're using an *inside view*. If instead you ignore the details of how something works, and instead make your predictions by assuming that a phenomenon will behave roughly like other similar phenomena, you're using an *outside view* (also called *reference class forecasting*).

Inside view examples:

- "When I break the project into steps and visualize how long each step will take, it looks like the project will take 6 weeks"
- "When I combine what I know of physics and computation, it looks like the serial speed formulation of Moore's Law will break down around 2005, because we haven't been able to scale down energy-use-per-computation as quickly as we've scaled up computations per second, which means the serial speed formulation of Moore's Law will run into roadblocks from energy consumption and heat dissipation somewhere around 2005."

Outside view examples:

- "I'm going to ignore the details of this project, and instead compare my project to similar projects. Other projects like this have taken 3 months, so that's probably about how long my project will take."
- "The serial speed formulation of Moore's Law has held up for several decades, through several different physical architectures, so it'll probably continue to hold through the next shift in physical architectures."

See also chapter 23 in Kahneman (2011); Planning Fallacy; Reference class forecasting. Note that, after several decades of past success, the serial speed formulation of Moore's Law did in fact break down in 2004 for the reasons described (Fuller & Millett 2011).

2. An outside view works best when using a reference class with a *similar causal structure* to the thing you're trying to predict. An inside view works best when a phenomenon's causal structure is well-understood, and when (to your knowledge) there are very few phenomena with a similar causal structure that you can use to predict things about the phenomenon you're investigating. See: The Outside View's Domain.

When writing a textbook that's much like other textbooks, you're probably best off predicting the cost and duration of the project by looking at similar textbook-writing projects. When you're predicting the trajectory of the serial speed formulation of Moore's Law, or predicting which spaceship designs will successfully land humans on the moon for the first time, you're probably best off using an (intensely *informed*) inside view.

3. Some things aren't very predictable with *either* an outside view or an inside view. Sometimes, the thing you're trying to predict seems to have a significantly different causal structure than other things, *and* you don't understand its causal structure very well. What should we do in such cases? This remains a matter of debate.

Eliezer Yudkowsky recommends a weak inside view for such cases:

> On problems that are drawn from a barrel of causally similar problems, where human optimism runs rampant and unforeseen troubles are common, the Outside View beats the Inside View... [But] on problems that are new things under the Sun, where there's a huge change of context and a structural change in underlying causal forces, the Outside View also fails - try to use it, and you'll just get into arguments about what is the proper domain of "similar historical cases" or what conclusions can be drawn therefrom. In this case, the best we can do is use the Weak Inside View — visualizing the causal process — to produce *loose qualitative conclusions about only those issues where there seems to be lopsided support*.

In contrast, Robin Hanson recommends an outside view for difficult cases:

> It is easy, way too easy, to generate new mechanisms, accounts, theories, and abstractions. To see if such things are *useful*, we need to vet them, and that is easiest "nearby", where we know a lot. When we want to deal with or understand things "far", where we know little, we have little choice other than to rely on mechanisms, theories, and concepts that have worked well near. Far is just the wrong place to try new things.
>
> There are a bazillion possible abstractions we could apply to the world. For each abstraction, the question is not whether one *can* divide up the world that way, but whether it "carves nature at its joints", giving *useful* insight not easily gained via other abstractions. We should be wary of inventing new abstractions just to make sense of things far; we should insist they first show their value nearby.

In Yudkowsky (2013), sec. 2.1, Yudkowsky offers a reply to these paragraphs, and continues to advocate for a weak inside view. He also adds:

> the other major problem I have with the “outside view” is that everyone who uses it seems to come up with a different reference class and a different answer.

This is the problem of "reference class tennis": each participant in the debate claims their own reference class is most appropriate for predicting the phenomenon under discussion, and if disagreement remains, they might each say "I’m taking my reference class and going home."

Responding to the same point made elsewhere, Robin Hanson wrote:

> [Earlier, I] warned against over-reliance on “unvetted” abstractions. I wasn’t at all trying to claim there is one true analogy and all others are false. Instead, I argue for preferring to rely on abstractions, including categories and similarity maps,

> that have been found useful by a substantial intellectual community working on related problems.

Multiple reference classes

Yudkowsky (2013) adds one more complaint about reference class forecasting in difficult forecasting circumstances:

> A final problem I have with many cases of 'reference class forecasting' is that... [the] final answers [generated from this process] often seem more specific than I think our state of knowledge should allow. [For example,] I don’t think you *should* be able to tell me that the next major growth mode will have a doubling time of between a month and a year. The alleged outside viewer claims to know too much, once they stake their all on a single preferred reference class.

Both this comment and Hanson's last comment above point to the vulnerability of relying on any *single* reference class, at least for difficult forecasting problems. Beware brittle arguments, says Paul Christiano.

One obvious solution is to use *multiple* reference classes, and weight them by how relevant you think they are to the phenomenon you're trying to predict. Holden Karnofsky writes of investigating things from "many different angles." Jonah Sinick refers to "many weak arguments." Statisticians call this "model combination." Machine learning researchers call it "ensemble learning" or "classifier combination."

In other words, we can use *many* outside views.

Nate Silver does this when he predicts elections (see Silver 2012, ch. 2). Venture capitalists do this when they evaluate startups. The best political forecasters studied in Tetlock (2005), the "foxes," tended to do this.

In fact, most of us do this regularly.

How do you predict which restaurant's food you'll most enjoy, when visiting San Francisco for the first time? One outside view comes from the restaurant's Yelp reviews. Another outside view comes from your friend Jade's opinion. Another outside view comes from the fact that you usually enjoy Asian cuisines more than other cuisines. And so on. Then you *combine* these different models of the situation, weighting them by how robustly they each tend to predict your eating enjoyment, and you grab a taxi to Osha Thai.

(Technical note: I say "model combination" rather than "model averaging" on purpose.)

Model combination and adjustment

You can probably do even better than this, though — if you know some things about the phenomenon and you're very careful. Once you've combined a handful of models to arrive at a qualitative or quantitative judgment, you should still be able to "adjust" the judgment in some cases using an inside view.

For example, suppose I used the above process, and I plan to visit Osha Thai for dinner. Then, somebody gives me my first taste of the *Synsepalum dulcificum* fruit. I happen to know that this fruit contains a molecule called miraculin which binds to one's tastebuds and makes sour foods taste sweet, and that this effect lasts for about an hour (Koizumi et al. 2011). Despite the results of my earlier model combination, I predict I won't particularly enjoy Osha Thai at the moment. Instead, I decide to try some tabasco sauce, to see whether it now tastes like doughnut glaze.

In some cases, you might also need to adjust for your prior over, say, "expected enjoyment of restaurant food," if for some reason your original model combination procedure didn't capture your prior properly.

Against "the outside view"

There is a *lot* more to say about model combination and adjustment (e.g. this), but for now let me make a suggestion about language usage.

Sometimes, small changes to our language can help us think more accurately. For example, gender-neutral language can reduce male bias in our associations (Stahlberg et al. 2007). In this spirit, I recommend we retire the phrase "the outside view..", and instead use phrases like "some outside view*s*..." and "*an* outside view..."

My reasons are:

1. Speaking of "the" outside view privileges a particular reference class, which could make us overconfident of that particular model's predictions, and leave model uncertainty unaccounted for.

2. Speaking of "the" outside view can act as a conversation-stopper, whereas speaking of multiple outside views encourages further discussion about how much weight each model should be given, and what each of them implies about the phenomenon under discussion.

Four Focus Areas of Effective Altruism

It was a pleasure to see all major strands of the effective altruism movement gathered in one place at last week's Effective Altruism Summit.

Representatives from GiveWell, The Life You Can Save, 80,000 Hours, Giving What We Can, Effective Animal Altruism, Leverage Research, the Center for Applied Rationality, and the Machine Intelligence Research Institute either attended or gave presentations. My thanks to Leverage Research for organizing and hosting the event!

What do all these groups have in common? As Peter Singer said in his TED talk, effective altruism "combines both the heart and the head." The heart motivates us to be empathic and altruistic toward others, while the head can "make sure that what [we] do is effective and well-directed," so that altruists can do not just *some* good but *as much good as possible*.

Effective altruists (EAs) tend to:

1. **Be globally altruistic**: EAs care about people equally, regardless of location. Typically, the most cost-effective altruistic cause won't happen to be in one's home country.
2. **Value consequences**: EAs tend to value causes according to their consequences, whether those consequences are happiness, health, justice, fairness and/or other values.
3. **Try to do as much good as possible**: EAs don't just want to do *some* good; they want to do (roughly) *as much good as possible*. As such, they hope to devote their altruistic resources (time, money, energy, attention) to unusually cost-effective causes. (This doesn't necessarily mean that EAs think "explicit" cost effectiveness calculations are the best *method for figuring out* which causes are likely to do the most good.)
4. **Think scientifically and quantitatively**: EAs tend to be analytic, scientific, and quantitative when trying to figure out which causes *actually* do the most good.
5. **Be willing to make significant life changes to be more effectively altruistic**: As a result of their efforts to be more effective in their altruism, EAs often (1) change which charities they support financially, (2) change careers, (3) spend significant chunks of time investigating which causes are most cost-effective according to their values, or (4) make other significant life changes.

Despite these similarities, EAs are a diverse bunch, and they focus their efforts on a variety of causes.

Below are four popular focus areas of effective altruism, ordered roughly by how large and visible they appear to be at the moment. Many EAs work on several of these focus areas at once, due to uncertainty about both facts and values.

Though labels and categories have their dangers, they can also enable chunking, which has benefits for memory, learning, and communication. There are many other ways we might categorize the efforts of today's EAs; this is only one categorization.

Focus Area 1: Poverty Reduction

Here, "poverty reduction" is meant in a broad sense that includes (e.g.) economic benefit, better health, and better education.

Major organizations in this focus area include:

- GiveWell is home to the most rigorous research on charitable causes, especially poverty reduction and global health. Their current charity recommendations are the Against Malaria Foundation, GiveDirectly, and the Schistosomiasis Control Initiative. (Note that GiveWell also does quite a bit of "meta effective altruism"; see below.)
- Good Ventures works closely with GiveWell.
- The Life You Can Save (TLYCS), named after Peter Singer's book on effective altruism, encourages people to pledge a fraction of their income to effective charities. TLYCS currently recommends GiveWell's recommended charities and several others.
- Giving What We Can (GWWC) does some charity evaluation and also encourages people to pledge 10% of their income effective charities. GWWC currently recommends two of GiveWell's recommended charities and two others.
- AidGrade evaluates the cost effectiveness of poverty reduction causes, with less of a focus on individual organizations.

In addition, some well-endowed foundations seem to have "one foot" in effective poverty reduction. For example, the Bill & Melinda Gates Foundation has funded many of the most cost-effective causes in the developing world (e.g. vaccinations), although it also funds less cost-effective-seeming interventions in the developed world.

In the future, poverty reduction EAs might also focus on economic, political, or research-infrastructure changes that might achieve poverty reduction, global health, and educational improvements more indirectly, as when Chinese economic reforms lifted hundreds of millions out of poverty. Though it is generally easier to evaluate the cost-effectiveness of direct efforts than that of indirect efforts, some groups (e.g. GiveWell Labs and The Vannevar Group) are beginning to evaluate the likely cost-effectiveness of these causes.

Focus Area 2: Meta Effective Altruism

Meta effective altruists focus less on specific causes and more on "meta" activities such as raising awareness of the importance of evidence-based altruism, helping EAs reach their potential, and doing research to help EAs decide which focus areas they should contribute to.

Organizations in this focus area include:

- 80,000 Hours highlights the importance of helping the world effectively through one's career. They also offer personal counseling to help EAs choose a career and a set of causes to support.
- Explicitly, the Center for Applied Rationality (CFAR) just trains people in rationality skills. But *de facto* they are especially focused on the application of rational thought to the practice of altruism, and are deeply embedded in the effective altruism community.

- Leverage Research focuses on growing and empowering the EA movement, e.g. by running Effective Altruism Summit, by organizing the THINK student group network, and by searching for "mind hacks" (like the memory palace) that can make EAs more effective.

Other people and organizations contribute to meta effective altruism, too. Paul Christiano examines effective altruism from a high level at Rational Altruist. GiveWell and others often write about the ethics and epistemology of effective altruism in addition to focusing on their chosen causes. And, of course, most EA organizations spend *some* resources growing the EA movement.

Focus Area 3: The Long-Term Future

Many EAs value future people roughly as much as currently-living people, and think that nearly all potential value is found in the well-being of the astronomical numbers of people who could populate the long-term future (Bostrom 2003; Beckstead 2013). Future-focused EAs aim to somewhat-directly capture these "astronomical benefits" of the long-term future, e.g. via explicit efforts to reduce existential risk.

Organizations in this focus area include:

- The Future of Humanity Institute at Oxford University is the primary hub of research on existential risk mitigation within the effective altruism movement. (CSER may join it soon, if it gets funding.)
- The Machine Intelligence Research Institute focuses on doing the research needed for humanity to one day build Friendly AI that could make astronomical numbers of future people enormously better off. It also runs the Less Wrong group blog and forum, where much of today's EA analysis and discussion occurs.

Other groups study particular existential risks (among other things), though perhaps not explicitly from the view of effective altruism. For example, NASA has spent time identifying nearby asteroids that could be an existential threat, and many organizations (e.g. GCRI) study worst-case scenarios for climate change or nuclear warfare that *might* result in human extinction but are more likely to result in "merely catastrophic" damage.

Some EAs (e.g. Holden Karnofsky, Paul Christiano) have argued that even if nearly all value lies in the long-term future, focusing on nearer-term goals (e.g. effective poverty reduction or meta effective altruism) may be more likely to realize that value than more direct efforts.

Focus Area 4: Animal Suffering

Effective animal altruists are focused on reducing animal suffering in cost-effective ways. After all, animals vastly outnumber humans, and growing numbers of scientists believe that many animals consciously experience pleasure and suffering.

The only organization of this type so far (that I know of) is Effective Animal Activism, which currently recommends supporting The Humane League and Vegan Outreach.

Edit: There is now also Animal Ethics, Inc.

Major inspirations for those in this focus area include Peter Singer, David Pearce, and Brian Tomasik.

Other focus areas

I could perhaps have listed "effective environmental altruism" as focus area 5. The environmental movement *in general* is large and well-known, but I'm not aware of many effective altruists who take environmentalism to be the most important cause for them to work on, after closely investigating the above focus areas. In contrast, the groups and people named above tend to have influenced each other, and have considered all these focus areas explicitly. For this reason, I've left "effective environmental altruism" off the list, though perhaps a popular focus on effective environmental altruism could arise in the future.

Other focus areas could later come to prominence, too.

Working together

I was pleased to see the EAs from different strands of the EA movement cooperating and learning from each other at the Effective Altruism Summit. Cooperation is crucial for growing the EA movement, so I hope that even if it’s not always easy, EAs will "go out of their way" to cooperate and work together, no matter which focus areas they’re sympathetic to.

The Robots, AI, and Unemployment Anti-FAQ

Q. Are the current high levels of unemployment being caused by advances in Artificial Intelligence automating away human jobs?

A. Conventional economic theory says this shouldn't happen. Suppose it costs 2 units of labor to produce a hot dog and 1 unit of labor to produce a bun, and that 30 units of labor are producing 10 hot dogs in 10 buns. If automation makes it possible to produce a hot dog using 1 unit of labor instead, conventional economics says that some people should shift from making hot dogs to buns, and the new equilibrium should be 15 hot dogs in 15 buns. On standard economic theory, improved productivity - including from automating away some jobs - should produce increased standards of living, not long-term unemployment.

Q. Sounds like a lovely theory. As the proverb goes, the tragedy of science is a beautiful theory slain by an ugly fact. Experiment trumps theory and in reality, unemployment is rising.

A. Sure. Except that the happy equilibrium with 15 hot dogs in buns, is *exactly* what happened over the last four centuries where we went from 95% of the population being farmers to 2% of the population being farmers (in agriculturally self-sufficient developed countries). We don't live in a world where 93% of the people are unemployed because 93% of the jobs went away. The first thought of automation removing a job, and thus the economy having one fewer job, has *not* been the way the world has worked since the Industrial Revolution. The parable of the hot dog in the bun is how economies really, actually worked in real life for centuries. Automation followed by re-employment went on for literally centuries in exactly the way that the standard lovely economic model said it should. The idea that there's a limited amount of work which is destroyed by automation is known in economics as the "lump of labour fallacy".

Q. But now people *aren't* being reemployed. The jobs that went away in the Great Recession aren't coming back, even as the stock market and corporate profits rise again.

A. Yes. And that's a *new* problem. We didn't get that when the Model T automobile mechanized the entire horse-and-buggy industry out of existence. The difficulty with supposing that automation is producing unemployment is that automation isn't new, so how can you use it to explain this new phenomenon of increasing long-term unemployment?

Baxter robot

Q. Maybe we've finally reached the point where there's no work left to be done, or where all the jobs that people can easily be retrained into can be even more easily automated.

A. You talked about jobs going away in the Great Recession and then not coming back. Well, the Great Recession wasn't produced by a sudden increase in productivity, it was produced by... I don't want to use fancy terms like "aggregate demand shock" so let's just call it problems in the financial system. The point is, in previous recessions the jobs came back strongly once NGDP rose again. (Nominal Gross Domestic Product - roughly the total amount of money being spent in face-value dollars.) *Now* there's been a recession and the jobs *aren't* coming back (in the US and EU), even though NGDP has risen back to its previous level (at least in the US). If the problem is automation, and we didn't experience any sudden leap in automation in 2008, then why can't people get back at least the jobs they used to have, as they did in previous recessions? Something has gone wrong with the engine of reemployment.

Q. And you don't think that what's gone wrong with the engine of reemployment is that it's easier to automate the lost jobs than to hire someone new?

A. No. That's something you could say just as easily about the 'lost' jobs from hand-weaving when mechanical looms came along. Some new obstacle is preventing jobs lost in the 2008 recession from coming back. Which may indeed mean that jobs eliminated by automation are *also* not coming back. And new high school and college graduates entering the labor market, likewise usually a good thing for an economy, will just end up being sad and unemployed. But this must mean something new and awful is happening to the processes of employment - it's not because the kind of automation that's happening today is different from automation in the 1990s, 1980s, 1920s, or 1870s; there were skilled jobs lost then, too. It should also be noted that automation has been a comparatively small force this decade next to shifts in global trade - which have *also* been going on for centuries and have *also* previously been a hugely positive economic force. But if something is generally wrong with reemployment, then it might be possible for increased trade with China to result in permanently lost jobs within the US, in *direct contrast* to the way it's worked over all previous economic history. But just like new college graduates ending up unemployed, something else must be going very wrong - that *wasn't* going wrong in 1960 - for anything so unusual to happen!

Q. What if what's changed is that we're out of new jobs to create? What if we've already got enough hot dog buns, for every kind of hot dog bun there is in the labor market, and now AI is automating away the *last* jobs and the *last* of the demand for labor?

A. This does not square with our being unable to recover the jobs that existed before the Great Recession. Or with lots of the world living in poverty. If we imagine the situation being much more extreme than it actually is, there was a time when professionals usually had personal cooks and maids - as Agatha Christie said, "When I was young I never expected to be so poor that I could not afford a servant, or so rich that I could afford a motor car." Many people would hire personal cooks or maids if we could afford them, which is the sort of new service that ought to come into existence if other jobs were eliminated - the reason maids became less common is that they were offered better jobs, not because demand for that form of human labor stopped existing. Or to be less extreme, there are lots of businesses who'd take nearly-free employees at various occupations, if those employees could be hired literally at minimum wage and legal liability wasn't an issue. *Right now* we haven't run out of *want* or *use* for human labor, so how could "The End of Demand" be producing unemployment *right now?* The fundamental fact that's driven employment over the course of previous human history is that it is a very strange state of affairs for somebody sitting around doing nothing, to have nothing better to do. We do not literally have nothing better for unemployed workers to do. Our civilization is not that advanced. So we must be doing something wrong (which we weren't doing wrong in 1950).

Q. So what *is* wrong with "reemployment", then?

A. I know less about macroeconomics than I know about AI, but even I can see *all sorts* of changed circumstances which are much more plausible sources of novel employment dysfunction than the relatively steady progress of automation. In terms of developed countries that seem to be doing okay on reemployment, Australia hasn't had any drops in employment and their monetary policy has kept *nominal* GDP growth on a much steadier keel - using their central bank to regularize the number of face-value Australian dollars being spent - which an increasing number of influential econbloggers think the US and even more so the EU have been getting catastrophically wrong. Though that's a long story.[1] Germany saw unemployment drop from 11% to 5% from 2006-2012 after implementing a series of labor market reforms, though there were other things going on during that time. (Germany has twice the number of robots per capita as the US, which probably isn't significant to their larger macroeconomic trends, but would be a strange fact if robots were the leading cause of unemployment.) Labor markets and monetary policy are both major, obvious, widely-discussed candidates for what could've changed between now and the 1950s that might make reemployment harder. And though I'm not a leading econblogger, some other obvious-seeming thoughts that occur to me are:

* Many industries that would otherwise be accessible to relatively less skilled labor, have much higher barriers to entry now than in 1950. Taxi medallions, governments saving us from the terror of unlicensed haircuts, fees and regulatory burdens associated with new businesses - all things that could've plausibly changed between now and the previous four centuries. This doesn't apply only to unskilled labor, either; in 1900 it was a lot easier, legally speaking, to set up shop as a doctor. (Yes, the average doctor was substantially worse back then. But ask yourself whether some simple, repetitive medical surgery should really, truly require 11 years of medical school and residency, rather than a 2-year vocational training program for someone with high dexterity and good focus.) These sorts of barriers to entry allow people who are currently employed in that field to extract value from people trying to get jobs in that field (and from the general population too, of course). In any one sector this wouldn't hurt the whole economy too much, but if it happens everywhere at once, that could be the problem.

* True effective marginal tax rates on low-income families have gone up today compared to the 1960s, after all phasing-out benefits are taken into account, counting federal and state taxes, city sales taxes, and so on. I've seen figures tossed around like 70% and worse, and this seems like the sort of thing that could easily trash reemployment.[2]

* Perhaps companies are, for some reason, less willing to hire previously unskilled people and train them on the job. Empirically this seems to be something that is more true today than in the 1950s. If I were to guess at why, I would say that employees moving more from job to job, and fewer life-long jobs, makes it less rewarding for employers to invest in training an employee; and also college is more universal now than then. Which means that employers might try to *rely on* colleges to train employees, and this is a function colleges can't actually handle because:

* The US educational system is either getting worse at training people to handle new jobs, or getting so much more expensive that people can't afford retraining, for various other reasons. (Plus, we are really stunningly stupid about matching educational supply to labor demand. How completely ridiculous is it to ask high school students to decide what they want to do with the rest of their lives and give

them nearly no support in doing so? Support like, say, spending a day apiece watching twenty different jobs and then another week at their top three choices, with salary charts and projections and probabilities of graduating that subject given their test scores? The more so considering this is a central allocation question for the entire economy? But I have no particular reason to believe this part has gotten *worse* since 1960.)

* The financial system is staring much more at the inside of its eyelids now than in the 1980s. This could be making it harder for expanding businesses to get loans at terms they would find acceptable, or making it harder for expanding businesses to access capital markets at acceptable terms, or interfering with central banks' attempts to regularize nominal demand, or acting as a brake on the system in some other fashion.

* Hiring a new employee now exposes an employer to more downside risk of being sued, or risk of being unable to fire the new employee if it turns out to be a bad decision. Human beings, including employers, are very averse to downside risk, so this could plausibly be a major obstacle to reemployment. Such risks are a plausible major factor in making the decision to hire someone *hedonically unpleasant* for the person who has to make that decision, which could've changed between now and 1950. (If your sympathies are with employees rather than employers, please consider that, nonetheless, if you pass any protective measure that makes the decision to hire somebody *less pleasant* for the hirer, fewer people will be hired and this is not good for people seeking employment. Many labor market regulations transfer wealth or job security to the already-employed at the expense of the unemployed, and these have been increasing over time.)

* Tyler Cowen's Zero Marginal Product Workers hypothesis: Anyone long-term-unemployed has now been swept into a group of people who have less than zero *average* marginal productivity, due to some of the people in this pool being negative-marginal-product workers who will destroy value, and employers not being able to tell the difference. We need some new factor to explain why this wasn't true in 1950, and obvious candidates would be (1) legal liability making past-employer references unreliable and (2) expanded use of college credentialing sweeping up more of the positive-product workers so that the average product of the uncredentialed workers drops.

* There's a thesis (whose most notable proponent I know is Peter Thiel, though this is not exactly how Thiel phrases it) that real, material technological change has been dying. If you can build a feature-app and flip it to Google for $20M in an acqui-hire, why bother trying to invent the next Model T? Maybe working on hard technology problems using math and science until you can build a liquid fluoride thorium reactor, has been made to seem less attractive to brilliant young kids than flipping a $20M company to Google or becoming a hedge-fund trader (and this is truer today relative to 1950).[3]

* Closely related to the above: Maybe change in atoms instead of bits has been regulated out of existence. The expected biotech revolution never happened because the FDA is just too much of a roadblock (it adds a great deal of expense, significant risk, and most of all, delays the returns beyond venture capital time horizons). It's plausible we'll never see a city with a high-speed all-robotic all-electric car fleet because the government, after lobbying from various industries, will require human attendants on every car - for safety reasons, of course! If cars were invented nowadays, the horse-and-saddle industry would surely *try* to arrange for them to be regulated out of existence, or sued out of existence, or limited to the same speed as

horses to ensure existing buggies remained safe. Patents are also an increasing drag on innovation in its most fragile stages, and may shortly bring an end to the remaining life in software startups as well. (But note that this thesis, like the one above, seems hard-pressed to account for jobs not coming back after the Great Recession. It is not conventional macroeconomics that re-employment after a recession requires macro sector shifts or new kinds of technology jobs.   The above is more of a Great Stagnation thesis of "What happened to productivity growth?" than a Great Recession thesis of "Why aren't the jobs coming back?"[4])

Q. Some of those ideas sounded more plausible than others, I have to say.

A. Well, it's not like they could all be true simultaneously. There's only a fixed effect size of unemployment to be explained, so the more likely it is that any one of these factors played a big role, the less we need to suppose that all the other factors were important; and perhaps what's Really Going On is something else entirely. Furthermore, the 'real cause' isn't always the factor you want to fix. If the European Union's unemployment problems were 'originally caused' by labor market regulation, there's no rule saying that those problems couldn't be mostly fixed by instituting an NGDP level targeting regime. This might or might not work, but the point is that there's no law saying that to fix a problem you have to fix its original historical cause.

Q. Regardless, if the engine of re-employment is broken *for whatever reason,* then AI really is killing jobs - a marginal job automated away by advances in AI algorithms won't come back.

A. Then it's odd to see so many news articles talking about AI killing jobs, when plain old non-AI computer programming and the Internet have affected many more jobs than that. The buyer ordering books over the Internet, the spreadsheet replacing the accountant - these processes are not strongly relying on the sort of algorithms that we would usually call 'AI' or 'machine learning' or 'robotics'. The main role I can think of for actual AI algorithms being involved, is in computer vision enabling more automation. And many manufacturing jobs were already automated by robotic arms even before robotic vision came along. Most computer programming is not AI programming, and most automation is not AI-driven. And then on near-term scales, like changes over the last five years, trade shifts and financial shocks and new labor market entrants are more powerful economic forces than the slow continuing march of computer programming. (Automation is a weak economic force in any given year, but cumulative and directional over decades. Trade shifts and financial shocks are stronger forces in any single year, but might go in the opposite direction the next decade. Thus, even generalized automation via computer programming is still an unlikely culprit for any sudden drop in employment as occurred in the Great Recession.)

Q. Okay, you've persuaded me that it's ridiculous to point to AI while talking about modern-day unemployment. What about *future* unemployment?

A. Like after the next ten years? We might or might not see robot-driven cars, which would be genuinely based in improved AI algorithms, and would automate away another bite of human labor. Even then, the total number of people driving cars for money would just be a small part of the total global economy; most humans are not paid to drive cars most of the time. Also again: for AI or productivity growth or increased trade or immigration or graduating students to increase unemployment, instead of resulting in more hot dogs and buns for everyone, you must be doing something terribly wrong that you weren't doing wrong in 1950.

Q. How about timescales longer than ten years? There was one class of laborers permanently unemployed by the automobile revolution, namely horses. There are a lot fewer horses nowadays because there is literally nothing left for horses to do that machines can't do better; horses' marginal labor productivity dropped below their cost of living. Could that happen to humans too, if AI advanced far enough that it could do *all* the labor?

A. If we imagine that in future decades machine intelligence is slowly going past the equivalent of IQ 70, 80, 90, eating up more and more jobs along the way... then I defer to Robin Hanson's analysis in Economic Growth Given Machine Intelligence, in which, as the abstract says, "Machines complement human labor when [humans] become more productive at the jobs they perform, but machines also substitute for human labor by taking over human jobs. At first, complementary effects dominate, and human wages rise with computer productivity. But eventually substitution can dominate, making wages fall as fast as computer prices now do."

Q. Could we already be in this substitution regime -

A. No, no, a dozen times no, for the dozen reasons already mentioned. That sentence in Hanson's paper has *nothing to do* with what is going on *right now*. The future cannot be a cause of the past. Future scenarios, even if they seem to associate the concept of AI with the concept of unemployment, cannot rationally increase the probability that current AI is responsible for current unemployment.

Q. But AI will inevitably become a problem later?

A. Not necessarily. We only get the Hansonian scenario if AI is *broadly, steadily* going past IQ 70, 80, 90, etc., making an increasingly large portion of the population fully obsolete in the sense that there is literally no job anywhere on Earth for them to do instead of nothing, because for *every* task they could do there is an AI algorithm or robot which does it more cheaply. That scenario isn't the only possibility.

Q. What other possibilities are there?

A. Lots, since what Hanson is talking about is a *new unprecedented phenomenon* extrapolated over *new future circumstances which have never been seen before* and there are all kinds of things which could potentially go differently within that. Hanson's paper may be the first obvious extrapolation from conventional macroeconomics and steady AI trendlines, but that's hardly a sure bet. Accurate prediction is hard, especially about the future, and I'm pretty sure Hanson would agree with that.

Q. I see. Yeah, when you put it that way, there are other possibilities. Like, Ray Kurzweil would predict that brain-computer interfaces would let humans keep up with computers, and then we wouldn't get mass unemployment.

A. The future would be more uncertain than that, even granting Kurzweil's hypotheses - it's not as simple as picking one futurist and assuming that their favorite assumptions correspond to their favorite outcome. You might get mass unemployment *anyway* if humans with brain-computer interfaces are more expensive or less effective than pure automated systems. With today's technology we could design robotic rigs to amplify a horse's muscle power - maybe, we're still working on that tech for humans - but it took around an extra century after the Model T to get to that point, and a plain old car is much cheaper.

Q. Bah, anyone can nod wisely and say "Uncertain, the future is." Stick your neck out, Yoda, and state your opinion clearly enough that you can later be proven wrong. Do *you* think we will eventually get to the point where AI produces mass unemployment?

A. My own guess is a moderately strong 'No', but for reasons that would sound like a complete subject change relative to all the macroeconomic phenomena we've been discussing so far. In particular I refer you to "Intelligence Explosion Microeconomics: Returns on cognitive reinvestment", a paper recently referenced on Scott Sumner's blog as relevant to this issue.

Q. Hold on, let me read the abstract and... what the heck is this?

A. It's an argument that you don't get the Hansonian scenario *or* the Kurzweilian scenario, because if you look at the historical course of hominid evolution and try to assess the inputs of marginally increased cumulative evolutionary selection pressure versus the cognitive outputs of hominid brains, and infer the corresponding curve of returns, then ask about a reinvestment scenario -

Q. English.

A. Arguably, what you get is I. J. Good's scenario where once an AI goes over some threshold of sufficient intelligence, it can self-improve and increase in intelligence far past the human level. This scenario is formally termed an 'intelligence explosion', informally 'hard takeoff' or 'AI-go-FOOM'. The resulting predictions are strongly distinct from traditional economic models of accelerating technological growth (we're *not* talking about Moore's Law here). Since it should take advanced general AI to automate away *most or all* humanly possible labor, my guess is that AI will intelligence-explode to superhuman intelligence *before* there's time for moderately-advanced AIs to crowd humans out of the global economy. (See also section 3.10 of the aforementioned paper.) Widespread economic adoption of a technology comes with a delay factor that wouldn't slow down an AI rewriting its own source code. This means we *don't* see the scenario of human programmers gradually improving broad AI technology past the 90, 100, 110-IQ threshold. An explosion of AI self-improvement utterly derails that scenario, and sends us onto a completely different track which confronts us with wholly dissimilar questions.

Q. Okay. What effect do you think a superhumanly intelligent self-improving AI would have on unemployment, especially the bottom 25% who are already struggling now? Should we really be trying to create this technological wonder of self-improving AI, if the end result is to make the world's poor even poorer? How is someone with a high-school education supposed to compete with a machine superintelligence for jobs?

A. I think you're asking an overly narrow question there.

Q. How so?

A. You might be thinking about 'intelligence' in terms of the contrast between a human college professor and a human janitor, rather than the contrast between a human and a chimpanzee. Human intelligence more or less created the entire modern world, including our invention of money; twenty thousand years ago we were just running around with bow and arrows. And yet on a biological level, human intelligence has stayed roughly the same since the invention of agriculture. Going past human-level intelligence is change on a scale much larger than the Industrial Revolution, or even the Agricultural Revolution, which both took place at a constant level of intelligence; human nature didn't change. As Vinge observed, building something *smarter than you* implies a future that is *fundamentally* different in a way that you wouldn't get from better medicine or interplanetary travel.

Q. But what *does* happen to people who were already economically disadvantaged, who don't have investments in the stock market and who aren't sharing in the profits of the corporations that own these superintelligences?

A. Um... we appear to be using substantially different background assumptions. The notion of a 'superintelligence' is not that it sits around in Goldman Sachs's basement trading stocks for its corporate masters. The concrete illustration I often use is that a superintelligence asks itself what the fastest possible route is to increasing its real-world power, and then, rather than bothering with the digital counters that humans call money, the superintelligence solves the protein structure prediction problem, emails some DNA sequences to online peptide synthesis labs, and gets back a batch of proteins which it can mix together to create an acoustically controlled equivalent of an artificial ribosome which it can use to make second-stage nanotechnology which manufactures third-stage nanotechnology which manufactures diamondoid molecular nanotechnology and then... well, it doesn't really matter from our perspective what comes after that, because from a human perspective any technology more advanced than molecular nanotech is just overkill. A superintelligence with molecular nanotech does not wait for you to buy things from it in order for it to acquire money. It just moves atoms around into whatever molecular structures or large-scale structures it wants.

Q. How would it get the energy to move those atoms, if not by buying electricity from existing power plants? Solar power?

A. Indeed, one popular speculation is that optimal use of a star system's resources is to disassemble local gas giants (Jupiter in our case) for the raw materials to build a Dyson Sphere, an enclosure that captures all of a star's energy output. This does not involve buying solar panels from human manufacturers, rather it involves self-replicating machinery which builds copies of itself on a rapid exponential curve -

Q. Yeah, I think I'm starting to get a picture of your background assumptions. So let me expand the question. If we grant that scenario rather than the Hansonian scenario or the Kurzweilian scenario, what sort of effect does *that* have on humans?

A. That depends on the *exact* initial design of the first AI which undergoes an intelligence explosion. Imagine a vast space containing all possible mind designs. Now imagine that humans, who all have a brain with a cerebellum, thalamus, a cerebral cortex organized into roughly the same areas, neurons firing at a top speed of 200 spikes per second, and so on, are one tiny little dot within this space of all possible minds. Different kinds of AIs can be vastly more different from each other than you are different from a chimpanzee. What happens after AI, depends on what kind of AI you build - the exact selected point in mind design space. If you can solve the technical problems and wisdom problems associated with building an AI that is nice to humans, or nice to sentient beings in general, then we all live happily ever afterward. If you build the AI incorrectly... well, the AI is unlikely to end up with a specific hate for humans. But such an AI won't attach a positive value to us either. "The AI does not hate you, nor does it love you, but you are made of atoms which it can use for something else." The human species would end up disassembled for spare atoms, after which human unemployment would be zero. In *neither* alternative do we end up with poverty-stricken unemployed humans hanging around being sad because they can't get jobs as janitors now that star-striding nanotech-wielding superintelligences are taking all the janitorial jobs. And so I conclude that advanced AI causing mass human unemployment is, all things considered, unlikely.

Q. Some of the background assumptions you used to arrive at that conclusion strike me as requiring additional support beyond the arguments you listed here.

A. I recommend Intelligence Explosion: Evidence and Import for an overview of the general issues and literature, Artificial Intelligence as a positive and negative factor in global risk for a summary of some of the issues around building AI correctly or incorrectly, and the aforementioned Intelligence Explosion Microeconomics for some ideas about analyzing the scenario of an AI investing cognitive labor in improving its own cognition. The last in particular is an important open problem in economics if you're a smart young economist reading this, although since the fate of the entire human species could well depend on the answer, you would be foolish to expect there'd be as many papers published about that as squirrel migration patterns. Nonetheless, bright young economists who want to say something important about AI should consider analyzing the microeconomics of returns on cognitive (re)investments, rather than post-AI macroeconomics which may not actually *exist* depending on the answer to the first question. Oh, and Nick Bostrom at the Oxford Future of Humanity Institute is supposed to have a forthcoming book on the intelligence explosion; that book isn't out yet so I can't link to it, but Bostrom personally and FHI generally have published some excellent academic papers already.

Q. But to sum up, you think that AI is definitely not the issue we should be talking about with respect to unemployment.

A. Right. From an economic perspective, AI is a completely odd place to focus your concern about modern-day unemployment. From an AI perspective, modern-day unemployment trends are a moderately odd reason to be worried about AI. Still, it is scarily true that increased automation, like increased global trade or new graduates or *anything else* that ought properly to produce a stream of employable labor to the benefit of all, might perversely operate to increase unemployment *if the broken reemployment engine is not fixed*.

Q. And with respect to future AI... what is it you think, exactly?

A. I think that with respect to moderately more advanced AI, we probably won't see *intrinsic unavoidable* mass unemployment in the economic world as we know it. *If* re-employment stays broken and new college graduates continue to have trouble finding jobs, then there are plausible stories where future AI advances far enough (but not *too* far) to be a significant part of what's freeing up new employable labor which bizarrely cannot be employed. I wouldn't consider this my main-line, average-case guess; I wouldn't expect to see it in the next 15 years or as the result of just robotic cars; and if it did happen, I wouldn't call AI the 'problem' while central banks still hadn't adopted NGDP level targeting. And then with respect to *very* advanced AI, the sort that might be produced by AI self-improving and going FOOM, asking about the effect of *machine superintelligence* on the conventional human labor market is like asking how US-Chinese trade patterns would be affected by the Moon crashing into the Earth. There would indeed be effects, but you'd be missing the point.

Q. Thanks for clearing that up.

A. No problem.

ADDED 8/30/13: Tyler Cowen's reply to this was one I hadn't listed:

> Think of the machines of the industrial revolution as getting underway sometime in the 1770s or 1780s. The big wage gains for British workers don’t really come until the 1840s. Depending on your exact starting point, that is over fifty years of labor market problems from automation.

See here for the rest of Tyler's reply.

Taken at face value this might suggest that if we wait 50 years everything will be all right. Kevin Drum replies that in 50 years there might be no human jobs left, which is possible but wouldn't be an effect we've seen *already*, rather a prediction of novel things yet to come.

Though Tyler also says, "A second point is that now we have a much more extensive network of government benefits and also regulations which increase the fixed cost of hiring labor" and this of course was already on my list of things that could be trashing modern reemployment unlike-in-the-1840s.

'Brett' in MR's comments section also counter-claims:

> The spread of steam-powered machinery and industrialization from textiles/mining/steel to all manner of British industries didn’t really get going until the 1830s and 1840s. Before that, it was mostly piece-meal, with some areas picking up the technology faster than others, while the overall economy didn’t change that drastically (hence the minimal changes in overall wages).

---

[1] The core idea in market monetarism is *very* roughly something like this: A central bank can control the total amount of money and thereby control any single economic variable measured in money, i.e., control one *nominal* variable. A central bank can't directly control how many people are employed, because that's a real variable. You could, however, try to control Nominal Gross Domestic Income (NGDI) or the total amount that people have available to spend (as measured in your currency). If the central bank commits to an NGDI *level target* then any shortfalls are made up the next year - if your NGDI growth target is 5% and you only get 4% in one year then you try for 6% the year after that. NGDI level targeting would mean that all the companies would know that, collectively, all the customers in the country would have 5% more money (measured in dollars) to spend in the next year than the previous year. This is usually called "NGDP level targeting" for historical reasons (NGDP is the other side of the equation, what the earned dollars are being spent on) but the most advanced modern form of the idea is probably "Level-targeting a market forecast of per-capita NGDI". Why this is the *best* nominal variable for central banks to control is a longer story and for that you'll have to read up on market monetarism. I will note that if you were worried about hyperinflation back when the Federal Reserve started dropping US interest rates to almost zero and buying government bonds by printing money... well, you really should note that (a) most economists said this wouldn't happen, (b) the market spreads on inflation-protected Treasuries said that the market was anticipating very low inflation, and that (c) we then *actually got* inflation *below* the Fed's 2% target. You can argue with economists. You can even argue with the market forecast, though in this case you ought to bet money on your beliefs. But when your fears of hyperinflation are disagreed with by economists, the market forecast *and observed reality,* it's time to give up on the theory that generated the false prediction. In this

case, market monetarists would have told you not to expect hyperinflation because NGDP/NGDI was collapsing and this constituted (overly) tight money regardless of what interest rates or the monetary base looked like.

[2] Call me a wacky utopian idealist, but I wonder if it might be genuinely politically feasible to reduce marginal taxes on the bottom 20%, if economists on both sides of the usual political divide got together behind the idea that income taxes (including payroll taxes) on the bottom 20% are (a) immoral and (b) do economic harm far out of proportion to government revenue generated. This would also require some amount of decreased taxes on the next quintile in order to avoid high *marginal* tax rates, i.e., if you suddenly start paying $2000/year in taxes as soon as your income goes from $19,000/year to $20,000/year then that was a 200% tax rate on that particular extra $1000 earned. The lost tax revenue must be made up somewhere else. In the current political environment this probably requires higher income taxes on higher wealth brackets rather than anything more creative. But if we allow ourselves to discuss economic dreamworlds, then income taxes, corporate income taxes, and capital-gains taxes are all very inefficient compared to consumption taxes, land taxes, and basically *anything but* income and corporate taxes. This is true even from the perspective of equality; a rich person who earns lots of money, but invests it all instead of spending it, is benefiting the economy rather than themselves and should not be taxed until they try to *spend* the money on a yacht, at which point you charge a consumption tax or luxury tax (even if that yacht is listed as a business expense, which should make no difference; consumption is not more moral when done by businesses instead of individuals). If I were given unlimited powers to try to fix the unemployment thing, I'd be reforming the entire tax code from scratch to present the minimum possible obstacles to exchanging one's labor for money, and as a second priority minimize obstacles to compound reinvestment of wealth. But trying to change anything on this scale is probably not politically feasible relative to a simpler, more understandable crusade to "Stop taxing the bottom 20%, it *harms our economy* because they're customers of all those other companies and it's *immoral* because they get a raw enough deal already."

[3] Two possible forces for significant technological change in the 21st century would be robotic cars and electric cars. Imagine a city with an all-robotic all-electric car fleet, dispatching light cars with only the battery sizes needed for the journey, traveling at much higher speeds with no crash risk and much lower fuel costs... *and* lowering rents by greatly extending the effective area of a city, i.e., extending the physical distance you can live from the center of the action while still getting to work on time because your average speed is 75mph. What comes to mind when you think of robotic cars? Google's prototype robotic cars. What comes to mind when you think of electric cars? Tesla. In both cases we're talking about ascended, post-exit Silicon Valley moguls trying to create industrial progress out of the goodness of their hearts, using money they earned from Internet startups. Can you sustain a whole economy based on what Elon Musk and Larry Page decide are cool?

[4]   Currently the conversation among economists is more like "Why has total factor productivity growth slowed down in developed countries?" than "Is productivity growing so fast due to automation that we'll run out of jobs?" Ask them the latter question and they will, with justice, give you very strange looks. Productivity *isn't* growing at high rates, and if it *were* that ought to cause employment rather than unemployment. This is why the Great Stagnation in productivity is one possible explanatory factor in unemployment, albeit (as mentioned) not a very good explanation for why we can't get back the jobs lost in the Great Recession. The idea would have to be that some natural rate of productivity growth and sectoral shift is

necessary for *re*-employment to happen after recessions, and we've lost that natural rate; but so far as I know this is not conventional macroeconomics.

Inferential credit history

Here's an interview with Seth Baum. Seth is an expert in risk analysis and a founder of the Global Catastrophic Risk Institute. As expected, *Bill O'Reilly* caricatured Seth as extreme, and cut up his interview with dramatic and extreme events from alien films. As a professional provocateur, it is his job to lay the gauntlet down to his guests. Also as expected, Seth put on a calm and confident performance. Was the interview net-positive or negative? It's hard to say, even in retrospect. Getting any publicity for catastrophic risk reduction is good, and difficult. Still, I'm not sure just *how bad* publicity has to be before it really *is* bad publicity…

Explaining catastrophic risks to the audience of Fox News is perhaps equally difficult to explaining the risk of artificial intelligence to anyone. This is a task that frustrated Eliezer Yudkowsky so deeply that he was driven to write the epic LessWrong sequences. In his view, the *inferential distance* was too large to be bridged by a single conversation. There were too many things that he knew that were prerequisites to understanding his current plan. So he wrote this *sequence* of online posts that set out everything he knew about cognitive science and probability theory, applied to help readers think more clearly and live out their scientific values. He had to write a thousand words per day for about two years before talking about AI explicitly. Perhaps surprisingly, and as an enormous credit to Eliezer's brain, these sequences formed the founding manifesto of the quickly growing rationality movement, many of whom now share his concerns about AI. Since he wrote these, his Machine Intelligence Research Institute (formely the singularity Institute) has grown precipitously and spun off the Center for Applied Rationality, a teaching facility and monument to the promotion of public rationality.

Why have Seth and Eliezer had such a hard time? Inferential distance explains a lot, but I have a second explanation, Seth and Eliezer had to build an *inferential credit history*. By the time you get to the end of the sequences, you have seen Eliezer bridge many an inferential distance, and you trust him to span another! If each time I loan Eliezer some attention, and suspend my disbelief, he has paid me back (in the currency of interesting and useful insight), then I will listen to him saying things that I don't yet believe for a long time.

When I watch Seth on *The Factor*, his interview is coloured by his *Triple A credit rating*. We have talked before, and I have read his papers. For the rest of the audience, he had no time to build *intellectual rapport.* It's not just that the inferential distance was large, it's more that he didn't have a credit rating of sufficient quality to take out a loan of that magnitude!

I contend that if you want to explain something abstract and unfamiliar, first you have to give a bunch of small and challenging chunks of insight, some of which must be practically applicable, and ideally you will lead your audience on a trek across a series of inferential distances, each slightly bigger than the last. It'll sure be helpful fills in some of the steps toward understanding the bigger picture, but not necessary.

This proposal could explain why historical explanations are often effective. Explanations that go like:

Initially I wanted to help people. And then I read The Life You Can Save. And then I realised I had been neglecting to think about large numbers of people. And then I read about scope insensitivity, which made me think *this*, and then I read Bostrom's *Fable of the Dragon Tyrant*, which made me think *that, and so on…*

This kind of explanation is often disorganised, with frequent detours, and false turns – steps in your ideological history that turned out to be wrong or unhelpful. The good thing about historical explanations is that they are stories, and that they have a main character – you – and this all makes the story more compelling. I will argue that a further advantage is that they give you the opportunity to borrow lots of small amounts of your audience's attention, and accrete a good credit rating, that you will need to make your boldest claims.

Lastly, let me present an alternative philosophy to overcoming inferential distances. It will seem to contradict what I have said so far, although I find it also useful.

If you say that X idea is crazy, then this can often become a self-fulfilling prophesy.

On this view, those who publicise AI risk should never complain about, and rarely talk about the large inferential distance before them, least of all publicy. They should normalise their proposal by treating it as normal. I still think it's important for them to acknowledge any intuitive reluctance on the part of their audience to entertain an idea. It's like how if you don't appear embarrassed after committing a faux-pas, you're seen as untrustworthy. But after acknowledging this challenge, they had best get back to their subject material, as any normal person would!

So if you believe in *inferential distance*, *inferential credit history* (building trust), and *acting normal*, then explain hard things by first beginning with lots of easy things, build larger and larger bridges, and acknowledge, but beware overemphasising any difficulties.

[also posted on my blog]

Low-hanging fruit: improving wikipedia entries

Many people are likely stumble across the Wikipedia entry for topics of interest relevant to those of us who frequent LessWrong: rationality, artificial intelligence, existential risks, decision theory, etc. These pages often shape one’s initial impressions of how interesting, important, or even credible a given topic is, and may have the potential to direct people towards productive resources (reading material, organizations like CFAR, notable figures such as Eliezer, etc.). As a result, ensuring that the Wikipedia entries on these topics are of better quality than some of them presently are presents an opportunity for investing relatively little effort in an activity with potentially substantial payoffs relative to the cost of time and effort put in.

I have already decided to improve some of the pages, beginning with the rather sloppy page that’s currently serving as the entry for existential risks, though of course others are welcome to contribute and may be more suited to the task than I am:

https://en.wikipedia.org/wiki/Risks_to_civilization,_humans,_and_planet_Earth

If you look at the section on risks posed by AI, for instance, it's notably inadequate, while the page includes a bizarre section referencing Mayan doomsday forecasts and Newton's predictions about the end of the world, neither of which seem adequately distinguished from rigorous attempts to actually assess legitimate existential risks.

I’m also constructing a list of other pages that are or are potentially in need of updating it and organizing it by my rough estimates of their relative importance (which I’m happy to share, modify, or discuss).

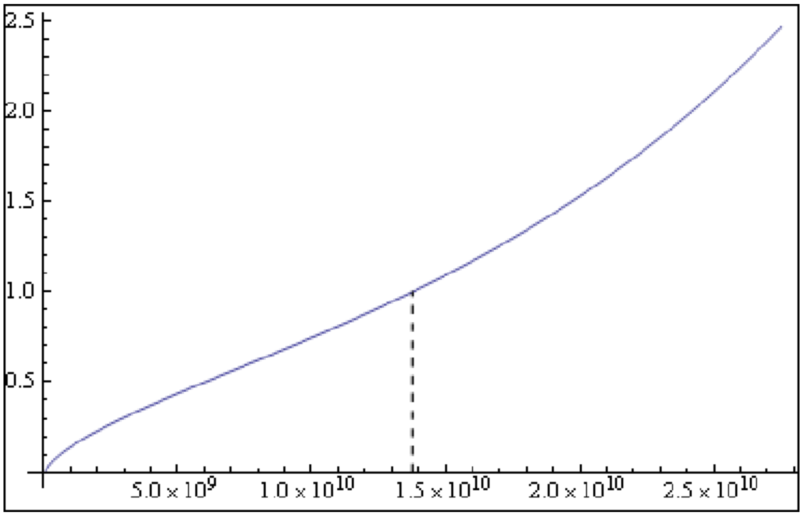
Turning this into a collaborative effort would be far more effective than doing it myself. If you think this is a worthwhile project and want to get involved I’d definitely like to hear from you and figure out a way to best coordinate our efforts.

To reduce astronomical waste: take your time, then go very fast

While we dither on the planet, are we losing resources in space? Nick Bostrom has an article on astronomical waste, talking about the vast amounts of potentially useful energy that we're simply not using for anything:

> As I write these words, suns are illuminating and heating empty rooms, unused energy is being flushed down black holes, and our great common endowment of negentropy is being irreversibly degraded into entropy on a cosmic scale. These are resources that an advanced civilization could have used to create value-structures, such as sentient beings living worthwhile lives.
>
> The rate of this loss boggles the mind. One recent paper speculates, using loose theoretical considerations based on the rate of increase of entropy, that the loss of potential human lives in our own galactic supercluster is at least ~10^{46} per century of delayed colonization.

On top of that, galaxies are slipping away from us because of the exponentially accelerating expansion of the universe (x axis in years since Big Bang, cosmic scale function arbitrarily set to 1 at the current day):



At the rate things are going, we seem to be losing slightly more than one galaxy a year. One entire galaxy, with its hundreds of billions of stars, is slipping away from us each year, never to be interacted with again. This is many solar systems a second; poof! Before you've even had time to grasp that concept, we've lost millions of times more resources than humanity has even used.

So it would seem that the answer to this desperate state of affairs is to rush thing: start expanding as soon as possible, greedily grab every hint of energy and negentropy before they vanish forever.

Not so fast! Nick Bostrom's point was not that we should rush things, but that we should be very very careful:

> However, the lesson for utilitarians is not that we ought to maximize the pace of technological development, but rather that we ought to maximize its safety, i.e. the probability that colonization will eventually occur.

If we rush things and lose the whole universe, then we certainly don't come out ahead in this game.

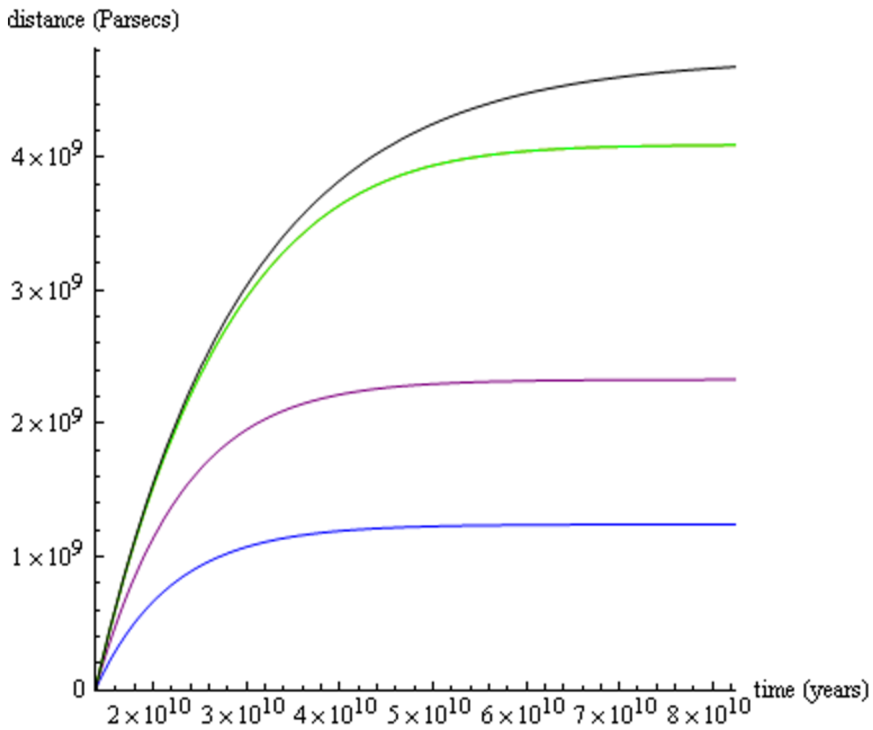
But let's ignore that; let's pretend that we've solved all risks, that we can expand safely, without fear of messing things up. Right. Full steam ahead to the starships, no?

No. Firstly, though the losses are large in absolute terms, they are small in relative terms. Most of the energy of a star is contained in its mass. The light streaming through windows in empty rooms? A few specks of negentropy, that barely diminish the value of the huge hoard that is the stars' physical beings (which can be harvested by eg dropping them slowly into small black holes and feeding off the Hawking radiation). And we lose a galaxy a year - but there are still billions out there. So waiting a while isn't a major problem, if we can gain something by doing so. Gain what? Well, maybe just going a tiny bit faster.

In a paper published with Anders Sandberg, we looked at the ease and difficulty of intergalactic or universal expansion. It seems to be surprisingly easy (which has a lot of implications for the Fermi Paradox), given sufficient automation or AI. About six hours of the sun's energy would be enough to launch self-replicating probes to every reachable galaxy in the entire universe. We could get this energy by constructing a Dyson swarm around the sun, by, for instance, disassembling Mercury. This is the kind of task that would be well within the capacities of an decently automated manufacturing process. A video overview of the process can be found in this talk (and a longer exposition, with slightly older figures, can be found here).

How fast will those probes travel? This depends not on the acceleration phase (which can be done fine with quench guns or rail guns, or lasers into solar sales), but on the deceleration. The relativistic rocket equation is vicious: it takes a lot of reaction mass to decelerate even a small payload. If fission power is used, decelerations from 50%c is about all that's reasonable. With fusion, we can push this to 80%c, while with matter-anti-matter reactions, we can get to 99%c. The top speed of 99%c is also obtainable if have more exotic ways of decelerating. This could be somehow using resources from the target galaxy (cunning gravitational braking or Bussard ramjets or something), or using the continuing expansion of the universe to bleed speed away (this is most practical for the most distant galaxies).

At these three speeds (and at 100% c), we can reach a certain distance into the universe, in current comoving coordinates, as shown by this graph (the x axis is in years since the Big Bang, with the origin set at the current day):



The maximum reached at 99%c is about 4 GigaParsecs - not a unit often used is casual conversation! If we can reach these distances, we can claim this many galaxies, approximately:

Speed	Distance (Parsecs)	# of Galaxies
50%c	1.24x10^9	1.16x10^8
80%c	2.33x10^9	7.62x10^8
99%c	4.09x10^9	4.13x10^9

These numbers don't change much if we delay. Even wasting a million years won't show up on these figure: it's a rounding error. Why is this?

Well, a typical probe will be flying through space, at quasi-constant velocity, for several billion years. Gains in speed make an immense difference, as they compound over the whole duration of the trip; gains from an early launch, not so much. So if we have to wait a million years to squeeze an extra 0.1% of speed, we're still coming out ahead. So waiting for extra research is always sensible (apart from the closest galaxies). If we can get more efficient engines, more exotic ways of shielding the probe, or new methods for deceleration, the benefits will be immense.

So, in conclusion: To efficiently colonise the universe, take your time. Do research. Think things over. Go to the pub. Saunter like an Egyptian. Write long letters to mum. Complain about the immorality of the youth of today. Watch dry paint stay dry.

But when you do go... go very, very fast.

Instrumental rationality/self help resources

I took part in a recent discussion in the current Open Thread about how instrumental rationality is under-emphasized on this website. I've heard other people say similar things, and I am inclined to agree. Someone suggested that there should be a "Instrumental Rationality Books" thread, similar to the "best textbooks on every subject" thread. I thought this sounded like a good idea.

The title is "resources" because in addition to books, you can post self-help websites, online videos, whatever.

The decorum for this thread will be as follows:

- One resource per comment
- Place your comment in the appropriate category
- Only post resources you've actually used. Write a short review of your resource and if possible, a short summary of the key points. Say whether or not you would recommend the resource.
- Mention approximately how long it's been since you first used the resource and whether or not you have made external improvements in the subject area. On the other hand, keep in mind that there are a myriad of confounding factors that can be present when applying self-help resources to your life, and therefore it is perfectly acceptable to say "I would recommend this resource, but I have not improved" or "I do not recommend this resource, but I have improved".

I think depending on how this thread goes, in a few days I might make a meta post on this subject in an attempt to inspire discussion on how the LessWrong community can work together to attempt to reach some sort of a consensus on what the best instrumental rationality methods and resources might be. lukeprog has already done great work in his The Science of Winning at Life sequence, but his reviews are uber-conservative and only mention resources with lots of scientific and academic backing. I think this leaves out a lot of really good stuff, and I think that we should be able to draw distinctions between stuff that isn't necessarily drawing on science but is reasonable, rational, and helps a lot of people, and *The Secret*.

But I thought we should get the ball rolling a little before we have that conversation. In the meantime, if you have a meta comment, you can just go ahead and post it as a reply to the top-level post.

Why Eat Less Meat?

Previously, I wrote on LessWrong about the preliminary evidence in favor of using leaflets to promote veganism as a way of cost-effectively reducing suffering. In response, there was a large discussion with 530+ comments.   In this discussion, I found that a lot of people wanted me to write about why I think nonhuman animals deserve our concern anyway.

Therefore, I wrote this essay with an attempt to defend the view that if one cares about suffering, one should also care about nonhuman animals, since (1) they are capable of suffering, (2) they do suffer quite a lot, and (3) we can prevent their suffering.   I hope that we can have a sober, non mind-killing discussion about this topic, since it’s possibly quite important.

Introduction

For the past two years, the only place I ate meat was at home with my family. As of October 2012, I've finally stopped eating meat altogether and can't see a reason why I would want to go back to eating meat. This kind of attitude toward eating is commonly classified as "vegetarianism" where one refrains from eating the flesh of all animals, including fish, but still will consume animal products like eggs and milk (though I try to avoid egg as best I can).

Why might I want to do this? And why might I see it as a serious issue? It's because I'm very concerned about the reality of suffering done to our "food animals" in the process of making them into meat, because I see vegetarianism as a way to reduce this suffering by stopping the harmful process, and because vegetarianism has not been hard at all for me to accomplish.

Animals Can Suffer

Back in the 1600s, Réné Descartes thought nonhuman animals were soulless automatons that could respond to their environment and react to stimuli, but could not feel anything — humans were the only species that were truly conscious. Descartes hit on an important point — since feelings are completely internal to the animal doing the feeling, it is impossible to demonstrate that anyone is truly conscious.

However, when it comes to humans, we don’t let that stop us from assuming other people feel pain. When we jab a person with a needle, no matter who they are, where they come from, or what they look like, they share a rather universal reaction of what we consider to be evidence of pain. We also extend this to our pets — we make great strides to avoid harming kittens, puppies, or other companion animals, and no one would want to kick a puppy or light a kitten on fire just because their consciousness cannot be directly observed. That’s why we even go as far as having laws against animal cruelty.

The animals we eat are no different. Pigs, chickens, cows, and fish all have incredibly analogous responses to stimuli that we would normally agree cause pain to humans and pets. Jab a pig with a needle, kick a chicken, or light a cow on fire, and they will react aversively like any cat, dog, horse, or human.

The Science

But we don't need to rely on just our intuition -- instead, we can look at the science. Animal scientists Temple Grandin and Mark Deesing conclude that "[o]ur review of the literature on frontal cortex development enables us to conclude that all mammals, including rats, have a sufficiently developed prefrontal cortex to suffer from pain". An interview of seven different scientists concludes that animals can suffer.

Dr. Jane Goodall, famous for having studied animals, writes in her introduction to **The Inner World of Farm Animals** that "farm animals feel pleasure and sadness, excitement and resentment, depression, fear, and pain. They are far more aware and intelligent than we ever imagined…they are individuals in their own right." Farm Sanctuary, an animal welfare organization, has a good overview documenting this research on animal emotion.

Lastly, among much other evidence, in the "Cambridge Declaration On Consciousness", prominent international group of cognitive neuroscientists, neuropharmacologists, neurophysiologists, neuroanatomists and computational neuroscientists states:

> Convergent evidence indicates that non-human animals have the neuroanatomical, neurochemical, and neurophysiological substrates of conscious states along with the capacity to exhibit intentional behaviors. Consequently, the weight of evidence indicates that humans are not unique in possessing the neurological substrates that generate consciousness. Nonhuman animals, including all mammals and birds, and many other creatures, including octopuses, also possess these neurological substrates.

Factory Farming Causes Considerable Suffering

However, the fact that animals can suffer is just one piece of the picture; we next have to establish that animals *do* suffer as a result of people eating meat. Honestly, this is easier shown than told -- there's an extremely harrowing and shocking 11-minute video about the cruelty available. Watching that video is perhaps the easiest way to see the suffering of nonhuman animals first hand in these "factory farms".

In making the case clear, Vegan Outreach writes "Many people believe that animals raised for food must be treated well because sick or dead animals would be of no use to agribusiness. This is not true."

They then go on to document, with sources, how virtually all birds raised for food are from factory farms where "resulting ammonia levels [from densely populated sheds and accumulated waste] commonly cause painful burns to the birds' skin, eyes, and respiratory tracts" and how hens "become immobilized and die of asphyxiation or dehydration", having been "[p]acked in cages (usually less than half a square foot of floor space per bird)". In fact, 137 million chickens suffer to death each year before they can even make it to slaughter -- more than the number of animals killed for fur, in shelters and in laboratories combined!

Farm Sanctuary also provides an excellent overview of the cruelty of factory farming, writing "Animals on factory farms are regarded as commodities to be exploited for profit. They undergo painful mutilations and are bred to grow unnaturally fast and large for the purpose of maximizing meat, egg, and milk production for the food industry."

It seems clear that factory farming practices are truly deplorable, and certainly are not worth the benefit of eating a slightly tastier meal. In "An Animal's Place", Michael Pollan writes:

> To visit a modern CAFO (Confined Animal Feeding Operation) is to enter a world that, for all its technological sophistication, is still designed according to Cartesian principles: animals are machines incapable of feeling pain. Since no thinking person can possibly believe this any more, industrial animal agriculture depends on a suspension of disbelief on the part of the people who operate it and a willingness to avert your eyes on the part of everyone else.

Vegetarianism Can Make a Difference

Many people see the staggering amount of suffering in factory farms, and if they don't aim to dismiss it outright will say that there's no way they can make a difference by changing their eating habits. However, this is certainly not the case!

How Many Would Be Saved?

Drawing from the 2010 Livestock Slaughter Animal Summary and the Poultry Slaughter Animal Summary, 9.1 billion land animals are either grown in the US or imported (94% of which are chickens!), 1.6 billion are exported, and 631 million die before anyone can eat them, leaving 8.1 billion land animals for US consumption **each year**.

A naïve average would divide this total among the population of the US, which is 311 million, assigning 26 land animals for each person's annual consumption. Thus, by being vegetarian, you are saving 26 land animals a year you would have otherwise eaten. And this doesn't even count fish, which could be quite high given how many fish need to be grown just to be fed to bigger fish!

Yet, this is not quite true. It's important to note that supply and demand aren't perfectly linear. If you reduce your demand for meat, the suppliers will react by lowering the price of meat a little bit, making it so more people can buy it. Since chickens dominate the meat market, we'll adjust by the supply elasticity of chickens, which is 0.22 and the demand elasticity of chickens, which is -0.52, and calculate the change in supply, which is 0.3. Taking this multiplier, it's more accurate to say **you're saving 7.8 land animals a year or more**. Though, there are a lot of complex considerations in calculating elasticity, so we should take this figure to have some uncertainty.

Collective Action

One might critique this response by responding that since meat is often bought in bulk, reducing meat consumption won't affect the amount of meat bought, and thus the suffering will still be the same, except with meat gone to waste. However, this ignores the effect of many different vegetarians acting together.

Imagine that you're supermarket buys cases of 200 chicken wings. It would thus take 200 people together to agree to buy 1 less wing in order for the supermarket to buy less wings. However, you have no idea if you're vegetarian #1 or vegetarian #56 or vegetarian #200, making the tipping point for 200 less wings to be bought. You thus can estimate that by buying one less wing you have a 1 in 200 chance of reducing 200 wings, which is equivalent to reducing the supply by one wing. So the effect basically cancels out. See here or here for more.

Every time you buy factory farmed meat, you are creating demand for that product, essentially saying "Thank you, I liked what you are doing and want to encourage you to do it more". By eating less meat, we can stop our support of this industry.

Vegetarianism Is Easier Than You Think

So nonhuman animals can suffer and do suffer in factory farms, and we can help stop this suffering by eating less meat. I know people who get this far, but then stop and say that, as much as they would like to, there's no way they could be a vegetarian because they like meat too much! However, such a joy for meat shouldn't count much compared to the massive suffering each animal undergoes just to be farmed -- imagine if someone wouldn't stop eating your pet just because they like eating your pet so much!

This is less than a problem than you might think, because being a vegetarian is really easy. Most people only think about what they would have to give up and how good it tastes, and don't think about what tasty things they could eat instead that have no meat in them. When I first decided to be a vegetarian, I simply switched from tasty hamburgers to tasty veggieburgers and there was no problem at all.

A Challenge

To those who say that vegetarianism is too hard, I’d like to simply challenge you to just try it for a few days. Feel free to give up afterward if you find it too hard. But I imagine that you should do just fine, find great replacements, and be able to save animals from suffering in the process.

If reducing suffering is one of your goals, there’s no reason why you must either be a die-hard meat eater or a die-hard vegetarian. Instead, feel free to explore some middle ground. You could be a vegetarian on weekdays but eat meat on weekends, or just try Meatless Mondays, or simply try to eat less meat. You could try to eat bigger animals like cows instead of fish or chicken, thus getting the same amount of meat with significantly less suffering.

-

(This was also cross-posted on my blog.)

Repository repository

A few weeks ago, Adele_L suggested that the repositories were underutilized and looked for suggestions on how to improve that. In that spirit, I added the following links to the Special Threads wiki page.

Solved Problems Repository - A collection of "solved problems in instrumental rationality."
Useful Concepts Repository - A collection of concepts that Less Wrong users have "found particularly useful for understanding the world."
Boring Advice Repository - A collection of advice that is optimized for helpfulness rather than depth of insight.
Useful Questions Repository - Questions that are useful to keep in mind in various situations.
Bad Concepts Repository - A collection of useless or harmful concepts
Grad Student Advice Repository - A collection of advice for graduate students.
Textbook Repository - The Best Textbooks on Every Subject
Reference repository - List of references and resources for LessWrong
Procedural Knowledge Gaps - How to do things that are "common sense" but that you may not know.
Mistakes Repository - A list of life-course altering mistakes that LW members have made.
Good things to have learned - A collection of skills and life lessons LWers have learned
Financial Effectiveness Repository - Tips for maximizing financial returns on (not necessarily market) investments.

In a similar vein, there is also a wiki page for the LessWrong Communities How-To's and Recommendations.

If there are other repositories that I've missed or a better way to collect these things, please link to it in a top level comment so that I get a direct message. A year and a half after this was originally posted, I still get suggestions and still add them or explain why I don't add them.

Edit: Added a few more to the list.

Making Rationality General-Interest

Introduction

Less Wrong currently represents a tiny, tiny, tiny segment of the population. In its current form, it might only *appeal* to a tiny, tiny segment of the population. Basically, the people who have a strong need for cognition, who are INTx on the Myers-Briggs (65% of us as per 2012 survey data), etc.

Raising the sanity waterline seems like a generally good idea. Smart people who believe stupid things, and go on to invest resources in stupid ways because of it, are *frustrating.* Trying to learn rationality skills in my 20s, when a bunch of thought patterns are already overlearned, is even more frustrating.

I have an intuition that a better future would be one where the concept of rationality (maybe called something different, but the same idea) is *normal*. Where it's as obvious as the idea that you shouldn't spend more money than you earn, or that you should live a healthy lifestyle, etc. The point isn't that *everyone* currently lives debt-free, eats decently well and exercises; that isn't the case; but they are normal things to do if you're a minimally proactive person who cares a bit about your future. No one has ever told me that doing taekwondo to stay fit is weird and culty, or that keeping a budget will make me unhappy because I'm overthinking thing.

I think the questions of "whether we should try to do this" and "if so, how do we do it in practice?" are both valuable to discuss, and interesting.

Is making rationality general-interest a good goal?

My intuitions are far from 100% reliable. I can think of a few reasons why this might be a *bad* idea:

1. A little bit of rationality can be damaging; it might push people in the direction of too much contrarianism, or something else I haven't thought of. Since introspection is imperfect, knowing a bit about cognitive biases and the mistakes that *other people* make might make people actually less likely to change their mind–they see other people making those well-known mistakes, but not themselves. Likewise, rationality taught only as a tool or skill, without any kind of underlying philosophy of why you should want to believe true things, might cause problems of a similar nature to martial art skills taught without the traditional, often non-violent philosophies–it could result in people abusing the skill to win fights/debates, making the larger community worse off overall. (Credit to Yan Zhang for martial arts metaphor).

2. Making the concepts general-interest, or just growing too fast, might involve watering them down or changing them in some way that the value of the LW microcommunity is lost. This could be worse for the people who currently enjoy LW even if it isn't worse overall. I don't know how easy it would be to avoid, or whether

3. It turns out that rationalists don't actually win, and x-rationality, as Yvain terms it, just isn't that amazing over-and-above already being proactive and doing stuff like keeping a budget. Yeah, you can say stuff like "the definition of rationality is that it helps you win", but if in real life, all the people who deliberately try to increase their

rationality do worse off overall, by their own standards (or even equally well, but with less time left over for other fun pursuits) than the people who aim for their life goals directly, I want to know that.

4. Making rationality general-interest is a good idea, but not the best thing to be spending time and energy on right now because of Mysterious Reasons X, Y, Z. Maybe I only think it is because of my personal bias towards liking community stuff (and wishing all of my friends were also friends with each other and liked the same activities, which would simplify my social life, but probably shouldn't happen for good reasons).

Obviously, if any of these are the case, I want to know about it. I also want to know about it if there are *other* reasons, off my radar, why this is a terrible idea.

What has to change for this to happen?

I don't really know, or I would be doing those things already (maybe, akrasia allowing). I have some ideas, though.

1. The jargon thing. I'm currently trying to compile a list of LW/CFAR jargon as a project for CFAR, and there are lots of terms I don't know. There are terms that I've realized in retrospect that I was using incorrectly all along. This presents both a large initial effort for someone interested in learning about rationality via the LW route, and also might contribute to the looking-like-a-cult thing.

2. The gender ratio thing. This has been discussed before, and it's a controversial thing to discuss, and I don't know how much arguing about it in comments will present any solutions. It seems pretty clear that if you want to appeal to the whole population, and a group that represents 50% of the general population only represents 10% of your participants (also as per 2012 survey data, see link above), there's going to be a problem somewhere down the road.

My data point: as a female on LW, I haven't experienced any discrimination, and I'm a bit baffled as to why the gender ratio is so skewed in the first place. Then again, I've already been through the filter of not caring if I'm the only girl at a meetup group. And I do hang out in female-dominated groups (i.e. the entire field of nursing), and fit in okay, but I'm probably not all that good as a typical example to generalize from.

3. LW currently appeals to intelligent people, or at least people who self-identify as intelligent; according to the 2012 survey data, the self-reported IQ median is 138. This wouldn't be surprising, and isn't a problem until you want to appeal to more than 1% of the population. But intelligence and rationality are, in theory, orthogonal, or at least not the same thing. If I suffered a brain injury that reduced my IQ significantly but didn't otherwise affects my likes and dislikes, I expect I would still be interested in improving my rationality and think it was important, perhaps even more so, but I also think I would find it frustrating. And I might feel horribly out of place.

4. Rationality in general has a bad rap; specifically, the Spock thing. And this isn't just affecting whether or not people thing Less Wrong the site is weird; it's affecting whether they want to think about their own decision-making.

This is only what I can think of in 5 minutes...

What's already happening?

Meetup groups are happening. CFAR is happening. And there are groups out there practicing skills similar or related to rationality, whether or not they call it the same thing.

Conclusion

Rationality, Less Wrong and CFAR have, gradually over the last 2-3 years, become a big part of my life. It's been fun, and I think it's made me stronger, and I would prefer a world where as many other people as possible have that. I'd like to know if people think that's a) a good idea, b) feasible, and c) how to do it practically.

Gains from trade: Slug versus Galaxy - how much would I give up to control you?

Edit: Moved to main at ThrustVectoring's suggestion.

A suggestion as to how to split the gains from trade in some situations.

The problem of Power

A year or so ago, people in the FHI embarked on a grand project: to try and find out if there was a single way of resolving negotiations, or a single way of merging competing moral theories. This project made a lot of progress in finding out how hard this was, but very little in terms of solving it. It seemed evident that the correct solution was to weigh the different utility functions, and then for everyone maximise the weighted sum, but all ways of weighting had their problems (the weighting with the most good properties was a very silly one: use the "min-max" weighting that sets your maximal attainable utility to 1 and your minimal to 0).

One thing that we didn't get close to addressing is the concept of power. If two partners in the negotiation have very different levels of power, then abstractly comparing their utilities seems the wrong solution (more to the point: it wouldn't be accepted by the powerful party).

The New Republic spans the Galaxy, with Jedi knights, battle fleets, armies, general coolness, and the manufacturing and human resources of countless systems at its command. The dull slug, ARthUrpHilIpDenu, moves very slowly around a plant, and possibly owns one leaf (or not - he can't produce the paperwork). Both these entities have preferences, but if they meet up, and their utilities are normalised abstractly, then ARthUrpHilIpDenu's preferences will weigh in far too much: a sizeable fraction of the galaxy's production will go towards satisfying the slug. Even if you think this is "fair", consider that the New Republic is the merging of countless individual preferences, so it doesn't make any sense that the two utilities get weighted equally.

The default point

After looking at various blackmail situations, it seems to me that it's the concept of default, or status quo, that most clearly differentiates between a threat and an offer. I wouldn't want you to make a credible threat, because this worsens the status quo, I would want you to make a credible offer, because this improves it. How this default is established is another matter - there may be some super-UDT approach that solves it from first principles. Maybe there is some deep way of distinguishing between threats and promises in some other way, and the default is simply the point between them.

In any case, without going any further into it's meaning or derivation, I'm going to assume that the problem we're working on has a definitive default/disagreement/threat point. I'll use the default point terminology, as that is closer to the concept I'm considering.

Simple trade problems often have a very clear default point. These are my goods, those are your goods, the default is we go home with what we started with. This is what I could build, that's what you could build, the default is that we both build purely for ourselves.

If we imagine ARthUrpHilIpDenu and the New Republic were at opposite ends of a regulated wormhole, and they could only trade in safe and simple goods, then we've got a pretty clear default point.

Having a default point opens up a whole host of new bargaining equilibriums, such as the Nash Bargaining Solution (NBS) and the Kalai-Smorodinsky Bargaining Solution (KSBS). But neither of these are really quite what we'd want: the KSBKS is all about fairness (which generally reduced expected outcomes), while the NBS uses a *product* of utility values, something that makes no intrinsic sense at all (NBS has some nice properties, like independence of irrelevant alternatives, but this only matters if the default point is reached through a process that has the same properties - and it can't be).

What am I *really* offering you in trade?

When two agents meet, especially if they are likely to meet more in the future (and most especially if they don't know the number of times and the circumstances in which they will meet), they should merge their utility functions: fix a common scale for their utility functions, add them together, and then both proceed to maximise the sum.

This explains what's really being offered in a trade. Not a few widgets or stars, but the possibility of copying your utility function into mine. But why would you want that? Because that will change my decisions, into a direction you find more pleasing. So what I'm actually offering you, is access to my decision points.

> *What is actually on offer in a trade, is access by one player's utility function to the other player's decision points.*

This gives a novel way of normalising utility functions. How much, precisely, is access to my decision points worth to you? If the default point gives a natural zero, then complete control over the other player's decision points is a natural one. "Power" is a nebulous concept, and different players may disagree as to how much power they each have. But power can only be articulated through making decisions (if you can't change any of your decisions, you have no power), and this normalisation allows each player to specify exactly how much they value the power/decision points of the other. Outcomes that involve one player controlling the other player's decision points can be designated the "utopia" point for that first player. These are what would happen if everything went exactly according to what they wanted.

What does this mean for ARthUrpHilIpDenu and the New Republic? Well, the New Republic stands to gain a leaf (maybe). From it's perspective, the difference between default (all the resources of the galaxy and no leaf) and utopia (all the resources of the galaxy plus one leaf) is tiny. And yet that tiny difference will get normalised to one: the New Republic's utility function will get multiplied by a huge amount. It will weigh heavily in any sum.

What about ARthUrpHilIpDenu? It stands to gain the resources of a galaxy. The difference between default (a leaf) and utopia (all the resources of a galaxy dedicated to making leaves) is unimaginably humongous. And yet that huge difference will get normalised to one: the ARthUrpHilIpDenu's utility function will get divided by a huge amount. It will weigh very little in any sum.

Thus if we add the two normalised utility functions, we get one that is nearly totally dominated by the New Republic. Which is what we'd expect, given the power differential between the two. So this bargaining system reflects the relative power of the players. Another way of thinking of this is that each player's utility is normalised taking into account how much they would give up to control the other. I'm calling it the "Mutual Worth Bargaining Solution" (MWBS), as it's the worth to players of the other player's decision points that are key. Also because I couldn't think of a better title.

Properties of the Mutual Worth Bargaining Solution

How does the MWBS compare with the NBS and the KSBS? The NBS is quite different, because it has no concept of relative power, normalising purely by the players' preferences. Indeed, one player could have no control at all, no decision points, and the NBS would still be unchanged.

The KSBS is more similar to the MWBS: the utopia points of the KSBS are the same as those of the MWBS. If we set the default point to (0,0) and the utopia points to (1,-) and (-,1), then the KSBS is given by the highest h such that (h,h) is a possible outcome. Whereas the MWBS is given by the outcome (x,y) such that x+y is highest possible.

Which is preferable? Obviously, if they knew exactly what the outcomes and utilities were on offer, then each player would have preferences as to which system to use (the one that gives them more). But if they didn't, if they had uncertainties as to what players and what preferences they would face in the future, then MWBS generally comes out on top (in expectation).

How so? Well, if a player doesn't know what other players they'll meet, they don't know in what way their decision points will be relevant to the other, and vice versa. They don't know what pieces of their utility will be relevant to the other, and vice versa. So they can expect to face a wide variety of normalised situations. To a first approximation, it isn't too bad an idea to assume that one is equally likely to face a certain situation as it's symmetric complement. Using the KSBS, you'd expect to get a utility of h (same in both case); under the MWBS, a utility of (x+y)/2 (x in one case, y in the other). Since x+y ≥ h+h = 2h by the definition of the MWBS, it comes out ahead in expectation.

Another important distinction between the MWBS is that while the KSBS and the NBS only allow Pareto improvements from the default point, MWBS does allow for some situation where one player will lose from the deal. It is possible, for instance, that (1/2,-1/4) is a possible outcome (summed utility 1/4), and there are no better options possible. Doesn't this go against the spirit of the default point? Why would someone go into a deal that leaves them poorer than before?

First off all, that situation will be rare. All MWBS must be in the triangle bounded by x<1, y<1 and x+y>0. The first bounds are definitional: one cannot get more expected utility that one's utopia point. The last bound comes from the fact that the default point is itself an option, with summed utility 0+0=0, so all summed utilities must be above zero. Sprinkle a few random outcome points into that triangle, and it very likely that the one with highest summed utility will be a Pareto improvement over (0,0).

But the other reason to accept the risk of losing, is because of the opportunity of gain. One could modify the MWBS to only allow Pareto improvements over the default: but in expectation, this would perform worse. The player would be immune from losing 1/4 utility from (1/2,-1/4), but unable to gain 1/2 from the (-1/4,1/2): the argument is the same as above. In ignorance as to the other player's preferences, accepting the possibility of loss improves the expected outcome.

It should be noted that the maximum that a player could theoretically lose by using the MWBS is equal to the maximum they could theoretically win. So the New Republic could lose at most a leaf, meaning that even powerful players would not be reluctant to trade. For less powerful players, the potential losses are higher, but so are the potential rewards.

Directions of research

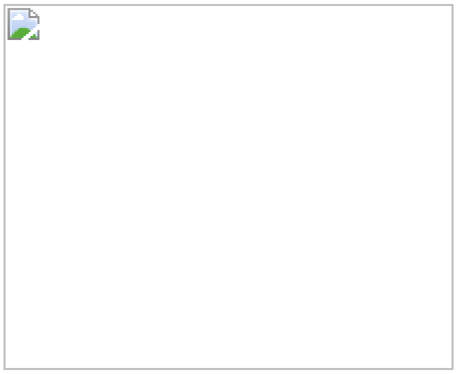
The MWBS is somewhat underdeveloped, and the explanation here isn't as clear as I'd have liked. However, me and Miriam are about to have a baby, so I'm not expecting to have any time at all soon, so I'm pushing out the idea, unpolished.

Some possible routes for further research: what are the other properties of MWBS? Are they properties that make MWBS feel more or less likely or acceptable? The NBS is equivalent with certain properties: what are the properties that are necessary and sufficient for the MWBS (and can they suggest better Bargaining Solutions)? Can we replace the default point? Maybe we can get a zero by imagining what would happen if the second player's decision nodes were under the control of an anti-agent (an agent that's the opposite of the first player), or a randomly selected agent?

The most important research route is to analyse what happens if several players come together at different times, and repeatedly normalise their utilities using the MWBS: does it matter the order in which they meet? I strongly feel that it's exploring this avenue that will reach "the ultimate" bargaining solution, if such a thing is to be found. Some solution that is stable under large numbers of agents, who don't know each other or how many they are, coming together in a order they can't predict.

Seed Study: Polyphasic Sleep in Ten Steps

(Update on this project now available here.)



A handful of Bay Area folks will be going polyphasic over the next month. By that, I mean we'll be adopting a sleep schedule that gets us 4 extra hours of productive work or play time per day, or two whole months per year. (Or a decade over 60 years.)

If you want to tell me about why it's a bad idea, feel free to post comments. I don't plan to use this space to sell you on polyphasic sleep. That might be another post, depending on how this goes.

I'm going to be collecting some very simple data **through this here form**. I invite you to join us!

This will be hard. It will hurt. You'll probably need a buddy to follow you around and keep you awake. If you don't have a lot of self-discipline, I don't recommend even trying.

Still with me? If you want in by the time you're done reading this, message me (or comment below) with your name so I know who you are. Here's the plan.

1. Stop using caffeine **right now**. If you try to maintain a caffeine addition during this process, you will fail. I promise.

2. Data collection began on July 10th. Start submitting daily reports at any point as soon as you want to participate, especially if you can begin in the next couple of days and then stick to our schedule. Fill out the above form once every 24hrs (whenever it's convenient) until August 10th.

3. Pick a time to take a 20min nap each day from Monday, July 15th through Sunday, July 21st. You probably won't actually sleep during this time, but you can use it for mindfulness meditation if you stay awake. The goal is to practice napping. This is important.

4. On Monday, July 22nd, begin fasting immediately after lunch.

5. On the night of Monday, July 22nd, skip sleep. No naps, then an all-nighter. **This is the official adaptation start date.** The idea is to make you sleep deprived so your naps the next day are more likely to take.

6. Eat breakfast on the morning of Tuesday, July 23rd. This should be the first time you've eaten anything since Monday lunch.

7. Starting on the morning of Tuesday, July 23rd, take a 20min nap every 2hrs (for a total of 12 naps per day). **Do not** oversleep. Use an obnoxious alarm or

whatever other means necessary. "Nap" counts as lying down trying to sleep; take your naps on a strict schedule regardless of how long you successfully sleep.

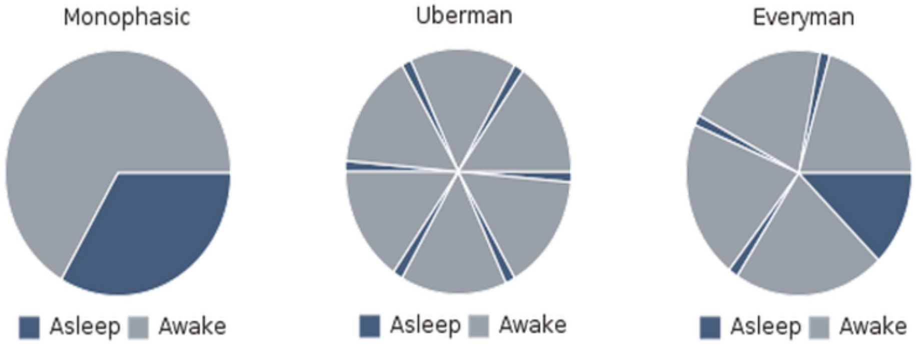
8. Start to cut your naps down toward 6 a day as quickly as you can without it hurting too much. Beginning to dream during your naps is a good indicator that you're ready for this part.*

9. Once you're down to one nap every 4 hours, you're on what's known as the Uberman schedule.

10. Matt Fallshaw informs me that the next part is a little tricky.

    10.1. If you managed to reach the Uberman feeling good, you'll probably start getting really tired again shortly thereafter. This flavor of tired will be different from what you've suffered for the past week, and by that flavor you will know that you have hit SWS deprivation. If this is what happens to you, the new kind of sleepy is your cue to transition straight to the Everyman 3 schedule, which means a 3 hour block of core sleep plus three 20 minute naps spaced evenly throughout the day. And that's it!

    10.2. If you're unlucky, you'll not quite have reached Uberman in the space of a week--that is, you'll still be hanging on to some extra naps on July 30th. Then you'll be walloped by a new bout of sleepiness. This flavor of tired will be different from the last. If it's is tolerable, drop straight to full Uberman and try to hold out for at least 24hrs, then convert to the Everyman 3. If the new flavor of tired is intolerable, convert to E3 as soon as the new tired hits, and expect the next week or so to be tougher on you than on the lucky ones.



Why are we doing this weird naptation adaptation plan thing instead of just going straight for the Everyman 3? Mostly because Matthew Fallshaw said to. If you know Matt, that's enough. In case you don't: It takes people about a month to adapt to the Everyman 3, but only about a week to adapt to the Uberman. The Uberman forces your body to learn to get its REM and SWS in those tiny 20 minute naps. If you're still giving it core sleep time, your body won't take the fullest possible advantage of naps right away.

If you think you can keep the Uberman schedule indefinitely, go for it! But keep me informed about it so I know what's up with my data.

*A clarification from Matt: "Drop naps as quickly as you can while remaining functional. The most important part of this period is that you don't sleep for longer than 20 minutes at a time, but the earlier you can get to a pure Uberman schedule the better. Take naps as you need them (with at least 40 minutes awake and moving around between naps) while pushing towards Uberman. The longer you can maintain pure Uberman before introducing a longer core sleep block the further along you'll be to a full adaptation."

ETA: I chose the psychomotor vigilance task (which you'll find if you check out the form I linked to) because the specific thing I'm trying to do here is distinguish polyphasic from chronic partial sleep restriction. If people on polyphasic return to their monophasic PVT baseline after a couple of weeks, especially if they stay there for a long time, that's clear evidence that the polyphasers are not experiencing the same physiological phenomenon as people suffering from chronic partial sleep restriction, which is what I'm actually concerned about. One of the only really well established facts in the literature on partial sleep restriction is that people who are deprived of a couple of hours of sleep a night get worse at the PVT as a function of time. If it's the case that the polyphasers will end up with memory problems, attention problems, and related issues, the simplest explanation is that they're suffering from a kind of chronic partial sleep restriction. I hope that clears some things up.

Why I'm Skeptical About Unproven Causes (And You Should Be Too)

Since living in Oxford, one of the centers of the "effective altruism" movement, I've been spending a lot of time discussing the classic “effective altruism” topic -- where it would be best to focus our time and money.

Some people here seem to think that the most important thing we should be focusing our time and money on are speculative projects, or projects that promise a very high impact, but involve a lot of uncertainty. One such very common example is "existential risk reduction", or attempts to make a long-term far future for humans more likely, say by reducing the chance of things that would cause human extinction.

I do agree that the far future is the most important thing to consider, by far (see papers by Nick Bostrom and Nick Beckstead). And I do think we can influence the far future. I just don't think we can do it in a *reliable* way. All we have are guesses about what the far future will be like and guesses about how we can affect it. All of these ideas are unproven, speculative projects, and I don't think they deserve the main focus of our funding.

While I waffled in cause indecision for a while, I'm now going to resume donating to GiveWell's top charities, except when I have an opportunity to use a donation to learn more about impact. Why? My case is that speculative causes, or any cause with high uncertainty (reducing nonhuman animal suffering, reducing existential risk, etc.) requires that we rely on our commonsense to evaluate them with naīve cost-effectiveness calculations, and this is (1) demonstrably unreliable with a bad track record, (2) plays right into common biases, and (3) doesn’t make sense based on how we ideally make decisions. While it’s unclear what long-term impact a donation to a GiveWell top charity will have, the near-term benefit is quite clear and worth investing in.

Focusing on Speculative Causes Requires Unreliable Commonsense

How can we reduce the chance of human extinction? It just makes sense that if we fund cultural exchange programs between the US and China, there will be more goodwill for the other within each country, and therefore the countries will be less likely to nuke each other. Since nuclear war would likely be very bad, it's of high value to fund cultural exchange programs, right?

Let's try another. The Machine Intelligence Research Institute (MIRI) thinks that someday artificial intelligent agents will become better than humans at making AIs. At this point, AI will build a smarter AI which will build an even smarter AI, and -- FOOM! -- we have a superintelligence. It's important that this superintelligence be programmed to be benevolent, or things will likely be very bad. And we can stop this bad event by funding MIRI to write more papers about AI, right?

Or how about this one? It seems like there will be challenges in the far future that will be very daunting, and if humanity handles them wrong, things will be very bad. But if people were better educated and had more resources, surely they'd be better at handling those problems, whatever they may be. Therefore we should focus on speeding up economic development, right?

These three examples are very common appeals to commonsense. But commonsense hasn't worked very well in the domain of finding optimal causes.

Can You Pick the Winning Social Program?

Benjamin Todd makes this point well in "Social Interventions Gone Wrong", where he provides a quiz with eight social programs and asks readers to guess whether they succeeded or failed.

I'll wait for you to take the quiz first... doo doo doo... la la la...

Ok, welcome back. I don't know how well you did, but success on this quiz is very rare, and this poses problems for commonsense. Sure, I'll grant you that Scared Straight sounds pretty suspicious. But the Even Start Family Literacy Program? It just makes sense that providing education to boost literacy skills and promote parent-child literacy activities should boost literacy rates, right? Unfortunately, it was wrong. Wrong in a very counter-intuitive way. There wasn't an effect.

GiveWell and Commonsense's Track Record of Failure

Commonsense actually has a track record of failure. GiveWell has been talking about this for ages. Every time GiveWell has found an intervention hyped by commonsense notions of high-impact and they've looked at it further, they've ended up disappointed.

The first was the Fred Hollows Foundation. A lot of people had been repeating the figure that the Fred Hollows Foundation could cure blindness for $50. But GiveWell found that number suspect.

The second was VillageReach. GiveWell originally put them as their top charity and estimated them as saving a life for under $1000. But further investigation kept leading them to revise their estimate until ultimately they weren't even sure if VillageReach had an impact at all.

Third, there is deworming. Originally, deworming was announced as saving a year of healthy life (DALY) for every $3.41 spent. But when GiveWell dove into the spreadsheets that resulted in that number, they found five errors. When the dust settled, the $3.41 figure was found to actually be off by a factor of 100. It was revised to $326.43.

Why shouldn't we expect this trend to not be the case in other areas where calculations are even looser and numbers are even less settled, like efforts devoted to speculative causes? Our only recourse is to fall back on interventions that are actually studied.

People Are Notoriously Bad At Predicting the (Far) Future

Cost-effectiveness estimates also frequently require making predictions about the future. Existential risk reduction, for example, requires making predictions about what will happen in the far future, and how your actions are likely to effect events hundreds of years down the road. Yet, experts are notoriously bad at making these kinds of predictions.

James Shanteau found in "Competence in Experts: The Role of Task Characteristics" (see also Kahneman and Klein's "Conditions for Intuitive Expertise: A Failure to Disagree") that experts perform well when thinking about static stimuli, thinking about things, and when there is feedback and objective analysis available. Furthermore, experts perform pretty badly when thinking about dynamic stimuli, thinking about behavior, and feedback and objective analysis are unavailable.

Predictions about existential risk reduction and the far future are firmly in the second category. So how can we trust our predictions about our impact on the far future? Our only recourse is to fall back on interventions that we can reliably predict, until we get better at prediction (or invest money in getting better at making predictions).

Even Broad Effects Require Specific Attempts

One potential resolution to this problem is to argue for “broad effects” rather than “specific attempts”. Perhaps it’s difficult to know whether a particular intervention will go well or mistaken to focus entirely on Friendly AI, but surely if we improved incentives and norms in academic work to better advance human knowledge (meta-research), improved education, or advocated for effective altruism, the far future would be much better equipped to handle threats.

I agree that these broad effects would make the far future better and I agree that it’s possible to implement these broad effects and change the far future. The problem, however, is it can’t be done in an easy or well understood way. Any attempt to implement a broad effect would require a specific action that has an unknown expectation of success and unknown cost-effectiveness. It’s definitely beneficial to advocate for effective altruism, but could this be done in a cost-effective way? A way that’s more cost-effective at producing welfare than AMF? How would you know?

In order to accomplish these broad effects, you’d need specific organizations and interventions to channel your time and money into. And by picking these specific organizations and interventions, you’re losing the advantage of broad effects and tying yourself to particular things with poorly understood impact and no track record to evaluate.

Focusing on Speculative Causes Plays Into Our Biases

We've now known for quite a long time that people are not all that rational. Instead, human thinking fails in very predictable and systematic ways. Some of these ways make us less likely to take speculative causes seriously, such as ambiguity aversion, the absurdity heuristic, scope neglect, and overconfidence bias.

But there’s also a different side of the coin, with biases that might make people think badly about existential risk:

Optimism bias. People generally think things will turn out better than they actually will. This could lead people to think that their projects will have a higher impact than they actually will, which would lead to higher estimates of cost-effectiveness than is reasonable.

Control bias. People like to think they have more control over things than they actually do. This plausibly also includes control over the far future. Therefore, people are probably biased into thinking they have more control over the far future than they actually do, leading to higher estimates of ability to influence the future than is reasonable.

"Wow factor" bias. People seem attracted to more impressive claims. Saving a life for $2500 through a malaria bed net seems much more boring compared to the chance of saving the entire world by averting a global catastrophe. Within the Effective Altruist / LessWrong community, existential risk reduction is cool and high status, whereas averting global poverty is not. This might lead to more endorsement of existential risk reduction than is reasonable.

Conjunction fallacy. People have a problem assessing probability properly when there are many steps involved, each of which has a chance of not happening. Ten steps, each with an independent 90% success rate, has only a 35% chance of success. Focusing on the far future seems to involve that a lot of largely independent events happen the way that is predicted. This would mean people are worse at estimating their chances of helping the far future, creating higher cost-effectiveness estimates than is reasonable.

Selection bias. When trying to find trends in history that are favorable for affecting the far future, some examples can be provided. However, this is because we usually hear about the interventions that end up working, whereas all the failed attempts to influence the far future are never heard of again. This creates a very skewed sample that can negatively bias our thinking about our success of influencing the far future.

It’s concerning there are numerous biases both weighted in favor and weighted against speculative causes, and this means we must tread carefully when assessing their merits. However, I would strongly expect biases to be even worse in favor of speculative causes rather than against them, because speculative causes lack the available feedback and objective evidence needed to help insulate against bias, whereas a focus on global health does not.

Focusing on Speculative Causes Uses Bad Decision Theory

Furthermore, not only is the case for speculative causes undermined by a bad track record and possible cognitive biases, but the underlying decision theory seems suspect in a way that's difficult to place.

Would you play a lottery with no stated odds?

Imagine another thought experiment -- you're asked to play a lottery. You have to pay $2 to play, but you have a chance at winning $100. Do you play?

Of course, you don't know, because you're not given odds. Rationally, it makes sense to play any lottery where you expect to come out ahead more often than not. If the lottery is a coin flip, it makes sense to pay $2 to have a 50/50 shot to win $100, since you'd expect to win $50 on average, and come ahead $48 each time. With a sufficiently high reward, even a one in a million chance is worth it. Pay $2 for a 1/1M chance of winning $1B, and you'd expect to come out ahead by $998 each time.

But $2 for the chance to win $100, without knowing what the chance is? Even if you had some sort of bounds, like you knew the odds had to be at least 1/150 and at most 1/10, though you could be off by a little bit. Would you accept that bet?

Such a bet seems intuitively uninviting to me, yet this is the bet that speculative causes offer me.

"Conservative Orders of Magnitude" Arguments

In response to these considerations, I've seen people endorsing speculative causes look at their calculations and remark that even if their estimate were off by 1000x, or three orders of magnitude, they still would be on solid ground for high impact, and there's no way they're actually off by three orders of magnitude. However, Nate Silver's **The Signal and the Noise: Why So Many Predictions Fail — but Some Don't** offers a cautionary tale:

> Moody’s, for instance, went through a period of making ad hoc adjustments to its model in which it increased the default probability assigned to AAA-rated securities by 50 percent. That might seem like a very prudent attitude: surely a 50 percent buffer will suffice to account for any slack in one’s assumptions? It might have been fine had the potential for error in their forecasts been linear and arithmetic. But leverage, or investments financed by debt, can make the error in a forecast compound many times over, and introduces the potential of highly geometric and nonlinear mistakes.
>
> Moody’s 50 percent adjustment was like applying sunscreen and claiming it protected you from a nuclear meltdown—wholly inadequate to the scale of the problem. It wasn’t just a possibility that their estimates of default risk could be 50 percent too low: they might just as easily have underestimated it by 500 percent or 5,000 percent. In practice, defaults were two hundred times more likely than the ratings agencies claimed, meaning that their model was off by a mere 20,000 percent.

Silver points out that when estimating how safe mortgage backed securities were, the difference between assuming defaults are perfectly uncorrelated and defaults are perfectly correlated is a difference of 160,000x in your risk estimate -- or five orders of magnitude.

If these kinds of five-orders-of-magnitude errors are possible in a realm that has actual feedback and is moderately understood, how do we know the estimates for cost-effectiveness are safe for speculative causes that are poorly understood and offer no feedback? Again, our only recourse is to fall back on interventions that we can reliably predict, until we get better at prediction.

Value of Information, Exploring, and Exploiting

Of course, there still is one important aspect of this problem that has not been discussed -- value of information -- or the idea that sometimes it's worth doing something just to learn more about how the world works. This is important in effective altruism too, where we focus specifically on "giving to learn", or using our resources to figure out more about the impact of various causes.

I think this is actually really important and is not a victim to any of my previous arguments, because we're not talking about impact, but rather learning value. Perhaps one could look to an "explore-exploit model", or the idea that we achieve the best outcome when we spend a lot of time exploring first (learning more about how to achieve better outcomes) before exploiting (focusing resources on achieving the best outcome we can). Therefore, whenever we have an opportunity to "explore" further or learn more about what causes have high impact, we should take it.

Learning in Practice

Unfortunately, in practice, I think these opportunities are very rare. Many organizations that I think are "promising" and worth funding further to see what their impact looks like do not have sufficiently good self-measurement in place to actually assess their impact or sufficient transparency to provide that information, therefore making it difficult to actually learn from them. And on the other side of things, many very promising opportunities to learn more are already fully funded. One must be careful to ensure that it's actually one's marginal dollar that is getting marginal information.

The Typical Donor

Additionally, I don't think the typical donor is in a very good position to assess where there is high value of information or have the time and knowledge to act upon this information once it is acquired. I think there's a good argument for people in the "effective altruist" movement to perhaps make small investments in EA organizations and encourage transparency and good measurement in their operations to see if they're successfully doing what they claim (or potentially create an EA startup themselves to see if it would work, though this carries large risks of further splitting the resources of the movement).

But even that would take a very savvy and involved effective altruist to pull off. Assessing the value of information on more massive investments like large-scale research or innovation efforts would be significantly more difficult, beyond the talent and resources of nearly all effective altruists, and are probably left to full-time foundations or subject-matter experts.

GiveWell’s Top Charities Also Have High Value of Information

As Luke Muehlhauser mentions in "Start Under the Streetlight, Then Push Into the Shadows", lots of lessons can be learned only by focusing on the easiest causes first, even if we have strong theoretical reasons to expect that they won’t end up being the highest impact causes once we have more complete knowledge.

We can use global health cost-effectiveness considerations as practice for slowly and carefully moving into the more complex and less understood domains. There even are some very natural transitions, such as beginning to look at "flow through effects" of reducing disease in the third-world and beginning to look at how more esoteric things affect the disease burden, like climate change. Therefore, even additional funding for GiveWell’s top charities has high value of information. And notably, GiveWell is beginning this "push" through GiveWell Labs.

Conclusion

The bottom line is that sometimes things look too good to be true. Therefore, I should expect that the actual impact of speculative causes that make large promises, upon a thorough investigation, will be much lower.

And this has been true in other domains. People are notoriously bad at estimating the effects of causes in both the developed world and developing world, and those are the causes that are near to us, provide us with feedback, and are easy to predict. Yet, from the Even Start Family Literacy Program to deworming estimates, our commonsense has failed us.

Add to that the fact that we should expect ourselves to perform even worse at predicting the far future. Add to that optimism bias, control bias, "wow factor" bias, and the conjunction fallacy, which make it difficult for us to think realistically about speculative causes. And then add to that considerations in decision theory, and whether we would bet on a lottery with no stated odds.

When all is said and done, I'm very skeptical of speculative projects. Therefore, I think we should be focused on exploring and exploiting. We should do whatever we can to fund projects aimed at learning more, when those are available, but be careful to make sure they actually have learning value. And when exploring isn’t available, we should exploit what opportunities we have and fund proven interventions.

But don’t confuse these two concepts and fund causes intended for learning because of their actual impact value. I’m skeptical about these causes actually being high impact, though I’m open to the idea that they might be and look forward to funding them in the future when they become better proven.

-

Followed up in: "What Would It Take To 'Prove' A Skeptical Cause" and "Where I've Changed My Mind on My Approach to Speculative Causes".

This was also cross-posted to my blog and to effective-altruism.com.

I'd like to thank Nick Beckstead, Joey Savoie, Xio Kikauka, Carl Shulman, Ryan Carey, Tom Ash, Pablo Stafforini, Eliezer Yudkowsky, and Ben Hoskin for providing feedback on this essay, even if some of them might strongly disagree with it's conclusion.

Prisoner's dilemma tournament results

The prisoner's dilemma tournament is over. There were a total of 21 entries. The winner is Margaret Sy, with a total of 39 points. 2nd and 3rd place go to rpglover64 and THE BLACK KNIGHT, with scores of 38 and 36 points respectively. There were some fairly intricate strategies in the tournament, but all three of these top scorers submitted programs that completely ignored the source code of the other player and acted randomly, with the winner having a bias towards defecting.

You can download a chart describing the outcomes here, and the source codes for the entries can be downloaded here.

I represented each submission with a single letter while running the tournament. Here is a directory of the entries, along with their scores: (some people gave me a term to refer to the player by, while others gave me a term to refer to the program. I went with whatever they gave me, and if they gave me both, I put the player first and then the program)

A: rpglover64 (38)
B: Watson Ladd (27)
c: THE BLACK KNIGHT (36)
D: skepsci (24)
E: Devin Bayer (30)
F: Billy, Mimic-- (27)
G: itaibn (34)
H: CooperateBot (24)
I: Sean Nolan (28)
J: oaz (26)
K: selbram (34)
L: Alexei (25)
M: LEmma (25)
N: BloodyShrimp (34)
O: caa (32)
P: nshepperd (25)
Q: Margaret Sy (39)
R: So8res, NateBot (33)
S: Quinn (33)
T: HonoreDB (23)
U: SlappedTogetherAtTheLastMinuteBot (20)