Best of LessWrong: June 2013

1. Robust Cooperation in the Prisoner's Dilemma
2. Tiling Agents for Self-Modifying AI (OPFAI #2)
3. Earning to Give vs. Altruistic Career Choice Revisited
4. For FAI: Is "Molecular Nanotechnology" putting our best foot forward?
5. After critical event W happens, they still won't believe you
6. How to Write Deep Characters
7. Start Under the Streetlight, then Push into the Shadows
8. Prisoner's Dilemma (with visible source code) Tournament
9. How probable is Molecular Nanotech?
10. Do Earths with slower economic growth have a better chance at FAI?
11. Public Service Announcement Collection
12. Useful Concepts Repository

Best of LessWrong: June 2013

1. Robust Cooperation in the Prisoner's Dilemma
2. Tiling Agents for Self-Modifying AI (OPFAI #2)
3. Earning to Give vs. Altruistic Career Choice Revisited
4. For FAI: Is "Molecular Nanotechnology" putting our best foot forward?
5. After critical event W happens, they still won't believe you
6. How to Write Deep Characters
7. Start Under the Streetlight, then Push into the Shadows
8. Prisoner's Dilemma (with visible source code) Tournament
9. How probable is Molecular Nanotech?
10. Do Earths with slower economic growth have a better chance at FAI?
11. Public Service Announcement Collection
12. Useful Concepts Repository

Robust Cooperation in the Prisoner's Dilemma

I'm proud to announce the preprint of Robust Cooperation in the Prisoner's Dilemma: Program Equilibrium via Provability Logic, a joint paper with Mihaly Barasz, Paul Christiano, Benja Fallenstein, Marcello Herreshoff, Patrick LaVictoire (me), and Eliezer Yudkowsky.

This paper was one of three projects to come out of the 2nd MIRI Workshop on Probability and Reflection in April 2013, and had its genesis in ideas about formalizations of decision theory that have appeared on LessWrong. (At the end of this post, I'll include links for further reading.)

Below, I'll briefly outline the problem we considered, the results we proved, and the (many) open questions that remain. Thanks in advance for your thoughts and suggestions!

Background: Writing programs to play the PD with source code swap

(If you're not familiar with the Prisoner's Dilemma, see here.)

The paper concerns the following setup, which has come up in academic research on game theory: say that you have the chance to write a computer program **X**, which takes in one input and returns either *Cooperate* or *Defect*. This program will face off against some other computer program **Y**, but with a twist: **X** will receive the source code of **Y** as input, and **Y** will receive the source code of **X** as input. And you will be given your program's winnings, so you should think carefully about what sort of program you'd write!

Of course, you could simply write a program that defects regardless of its input; we call this program **DefectBot**, and call the program that cooperates on all inputs **CooperateBot**. But with the wealth of information afforded by the setup, you might wonder if there's some program that might be able to achieve mutual cooperation in situations where **DefectBot** achieves mutual defection, without thereby risking a sucker's payoff. (Douglas Hofstadter would call this a perfect opportunity for superrationality...)

Previously known: CliqueBot and FairBot

And indeed, there's a way to do this that's been known since at least the 1980s. You can write a computer program that knows its own source code, compares it to the input, and returns *C* if and only if the two are identical (and *D* otherwise). Thus it achieves mutual cooperation in one important case where it intuitively ought to: when playing against itself! We call this program **CliqueBot**, since it cooperates only with the "clique" of agents identical to itself.

There's one particularly irksome issue with **CliqueBot**, and that's the fragility of its cooperation. If two people write functionally analogous but syntactically different versions of it, those programs will defect against one another! This problem can be patched somewhat, but not fully fixed. Moreover, mutual cooperation might be the best strategy against some agents that are not even functionally identical, and extending this approach requires you to explicitly delineate the list of programs that you're willing to cooperate with. Is there a more flexible and robust kind of program you could write instead?

As it turns out, there is: in a 2010 post on LessWrong, cousin_it introduced an algorithm that we now call **FairBot**. Given the source code of **Y**, **FairBot** searches for a proof (of less than some large fixed length) that **Y** returns *C* when given the source code of **FairBot**, and then returns *C* if and only if it discovers such a proof (otherwise it returns *D*). Clearly, if our proof system is consistent, **FairBot** only cooperates when that cooperation will be mutual. But the really fascinating thing is what happens when you play two versions of **FairBot** against each other. Intuitively, it seems that *either* mutual cooperation or mutual defection would be stable outcomes, but it turns out that if their limits on proof lengths are sufficiently high, they will achieve mutual cooperation!

The proof that they mutually cooperate follows from a bounded version of Löb's Theorem from mathematical logic. (If you're not familiar with this result, you might enjoy Eliezer's Cartoon Guide to Löb's Theorem, which is a correct formal proof written in much more intuitive notation.) Essentially, the asymmetry comes from the fact that both programs are searching for the same outcome, so that a short proof that one of them cooperates leads to a short proof that the other cooperates, and vice versa. (The opposite is not true, because the formal system can't know it won't find a contradiction. This is a subtle but essential feature of mathematical logic!)

Generalization: Modal Agents

Unfortunately, **FairBot** isn't what I'd consider an ideal program to write: it happily cooperates with **CooperateBot**, when it could do better by defecting. This is problematic because in real life, the world isn't separated into agents and non-agents, and any natural phenomenon that doesn't predict your actions can be thought of as a **CooperateBot** (or a **DefectBot**). You don't want your agent to be making concessions to rocks that happened not to fall on them. (There's an important caveat: some things have utility functions that you care about, but don't have sufficient ability to predicate their actions on yours. In that case, though, it wouldn't be a true Prisoner's Dilemma if your values actually prefer the outcome (*C*,*C*) to (*D*,*C*).)

However, **FairBot** belongs to a promising class of algorithms: those that decide on their action by looking for short proofs of logical statements that concern their opponent's actions. In fact, there's a really convenient mathematical structure that's analogous to the class of such algorithms: the modal logic of provability (known as GL, for Gödel-Löb).

So that's the subject of this preprint: **what can we achieve in decision theory by considering agents defined by formulas of provability logic?**

More formally *(skip the next two paragraphs if you're willing to trust me)*, we inductively define the class of "modal agents" as formulas using propositional variables and logical connectives and the modal operator (which represents provability in some

base-level formal system like Peano Arithmetic), of the form , where is fully modalized (i.e. all instances of variables are contained in an expression), and with each corresponding to a fixed modal agent of lower rank. For example, **FairBot** is represented by the modal formula .

When two modal agents play against each other, the outcome is given by the unique fixed point of the system of modal statements, where the variables are identified with each other so that represents the expression  , represents  , and the represent the actions of lower-rank modal agents against and vice-versa. (Modal rank is defined as a natural number, so this always bottoms out in a finite number of modal statements; also, we interpret outcomes as statements of provability in Peano Arithmetic, evaluated in the model where PA is consistent, PA+Con(PA) is consistent, and so on. See the paper for the actual details.)

The nice part about modal agents is that there are simple tools for finding the fixed points without having to search through proofs; in fact, Mihaly and Marcello wrote up a computer program to deduce the outcome of the source-code-swap Prisoner's Dilemma between any two (reasonably simple) modal agents. These tools also made it much easier to prove general theorems about such agents.

PrudentBot: The best of both worlds?

Can we find a modal agent that seems to improve on **FairBot**? In particular, we should want at least the following properties:

- It should be un-exploitable: if our axioms are consistent in the first place, then it had better only end up cooperating when it's mutual.
- It should cooperate with itself, and also mutually cooperate with **FairBot** (both are, common-sensically, the best actions in those cases).
- It should defect, however, against **CooperateBot** and lots of similarly exploitable modal agents.

It's nontrivial that such an agent exists: you may remember the post I wrote about the Masquerade agent, which is a modal agent that does *almost* all of those things (it doesn't cooperate with the original **FairBot**, though it does cooperate with some more complicated variants), and indeed we didn't find anything better until after we had Mihaly and Marcello's modal-agent-evaluator to help us.

But as it turns out, there is such an agent, and it's pretty elegant: we call it **PrudentBot**, and its modal version cooperates with another agent **Y** if and only if (there's a proof in Peano Arithmetic that **Y** cooperates with **PrudentBot** and there's a proof in PA+Con(PA) that **Y** defects against **DefectBot**). This agent can be seen to satisfy all of our criteria. But is it *optimal* among modal agents, by any reasonable criterion?

Results: Obstacles to Optimality

It turns out that, even within the class of modal agents, it's hard to formulate a definition of optimality that's actually true of something, and which meaningfully corresponds to our intuitions about the "right" decisions on decision-theoretic problems. (This intuition is not formally defined, so I'm using scare quotes.)

There are agents that give preferential treatment to **DefectBot**, **FairBot**, or even **CooperateBot**, compared to **PrudentBot**, though these agents are not ones you'd program in an attempt to win at the Prisoner's Dilemma. (For instance, one agent that rewards **CooperateBot** over **PrudentBot** is the agent that cooperates with **Y** iff PA proves that **Y** cooperates against **DefectBot**; we've taken to jokingly calling that agent **TrollBot**.) One might well suppose that a modal agent could still be optimal in the sense of making the "right" decision in every case, regardless of whether it's being punished for some other decision. However, this is not the only obstacle to a useful concept of optimality.

The second obstacle is that any modal agent only checks proofs at some finite number of levels on the hierarchy of formal systems, and agents that appear indistinguishable at all those levels may have obviously different "right" decisions. And thirdly, an agent might mimic another agent in such a way that the "right" decision is to treat the mimic differently from the agent it imitates, but in some cases one can prove that no modal agent can treat the two differently.

These three strikes appear to indicate that if we're looking to formalize more advanced decision theories, modal agents are too restrictive of a class to work with. We might instead allow things like quantifiers over agents, which would invalidate these specific obstacles, but may well introduce new ones (and certainly would make for more complicated proofs). But for a "good enough" algorithm on the original problem (assuming that the computer will have lots of computational resources), one could definitely do worse than submit a finite version of **PrudentBot**.

Why is this awesome, and what's next?

In my opinion, the result of Löbian cooperation deserves to be published for its illustration of Hofstadterian superrationality in action, apart from anything else! It's *really cool* that two agents reasoning about each other can in theory come to mutual cooperation for genuine reasons that don't have to involve being clones of each other (or other anthropic dodges). It's a far cry from a practical approach, of course, but it's a start: mathematicians always begin with a simplified and artificial model to see what happens, then add complications one at a time.

As for what's next: First, we don't *actually* know that there's no meaningful non-vacuous concept of optimality for modal agents; it would be nice to know that one way or another. Secondly, we'd like to see if some other class of agents contains a simple example with really nice properties (the way that classical game theory doesn't always have a pure Nash equilibrium, but always has a mixed one). Thirdly, we might hope that there's an actual implementation of a decision theory (TDT, UDT, etc) in the context of program equilibrium.

If we succeed in the positive direction on any of those, we'd next want to extend them in several important ways: using probabilistic information rather than certainty, considering more general games than the Prisoner's Dilemma (bargaining games have many further challenges, and games of more than two players could be more convoluted still), etc. I personally hope to work on such topics in future MIRI workshops.

Further Reading on LessWrong

Here are some LessWrong posts that have tackled similar material to the preprint:

- AI cooperation in practice, cousin_it, 2010
- Notion of Preference in Ambient Control, Vladimir_Nesov, 2010
- A model of UDT with a halting oracle, cousin_it, 2011
- Formulas of arithmetic that behave like decision agents, Nisan, 2012
- A model of UDT without proof limits, An example of self-fulfilling spurious proofs in UDT, Löbian cooperation, version 2, Bounded versions of Gödel's and Löb's theorems, cousin_it, 2012
- Predictability of decisions and the diagonal method, Consequentialist formal systems, Vladimir_Nesov, 2012
- Decision Theories: A Semi-Formal Analysis: Part 0 (A LessWrong Primer), Part 1 (The Problem with Naive Decision Theory), Part 2 (Causal Decision Theory and Substitution), Part 3 (Formalizing Timeless Decision Theory), Part 3.5 (Halt, Melt, and Catch Fire), Part 3.75 (Hang On, I Think This Works After All), orthonormal, 2012

Tiling Agents for Self-Modifying AI (OPFAI #2)

An early draft of publication #2 in the Open Problems in Friendly AI series is now available: Tiling Agents for Self-Modifying AI, and the Lobian Obstacle. ~20,000 words, aimed at mathematicians or the highly mathematically literate. The research reported on was conducted by Yudkowsky and Herreshoff, substantially refined at the November 2012 MIRI Workshop with Mihaly Barasz and Paul Christiano, and refined further at the April 2013 MIRI Workshop.

Abstract:

We model self-modication in AI by introducing 'tiling' agents whose decision systems will approve the construction of highly similar agents, creating a repeating pattern (including similarity of the offspring's goals). Constructing a formalism in the most straightforward way produces a G odelian difficulty, the Lobian obstacle. By technical methods we demonstrate the possibility of avoiding this obstacle, but the underlying puzzles of rational coherence are thus only partially addressed. We extend the formalism to partially unknown deterministic environments, and show a very crude extension to probabilistic environments and expected utility; but the problem of finding a fundamental decision criterion for self-modifying probabilistic agents remains open.

Commenting here is the preferred venue for discussion of the paper. This is an early draft and has not been reviewed, so it may contain mathematical errors, and reporting of these will be much appreciated.

The overall agenda of the paper is introduce the conceptual notion of a self-reproducing decision pattern which includes reproduction of the goal or utility function, by exposing a particular possible problem with a tiling logical decision pattern and coming up with some partial technical solutions. This then makes it conceptually much clearer to point out the even deeper problems with "We can't yet describe a probabilistic way to do this because of non-monotonicity" and "We don't have a good bounded way to do this because maximization is impossible, satisficing is too weak and Schmidhuber's swapping criterion is underspecified." The paper uses first-order logic (FOL) because FOL has a lot of useful standard machinery for reflection which we can then invoke; in real life, FOL is of course a poor representational fit to most real-world environments outside a human-constructed computer chip with thermodynamically expensive crisp variable states.

As further background, the idea that something-like-proof might be relevant to Friendly AI is not about achieving some chimera of absolute safety-feeling, but rather about the idea that the total probability of catastrophic failure should not have a significant conditionally independent component on each self-modification, and that self-modification will (at least in initial stages) take place within the highly deterministic environment of a computer chip. This means that statistical testing methods (e.g. an evolutionary algorithm's evaluation of average fitness on a set of test problems) are not suitable for self-modifications which can potentially induce catastrophic failure (e.g. of parts of code that can affect the representation or interpretation of the goals). Mathematical proofs have the property that they are as strong as their axioms and have no significant conditionally independent per-step

failure probability if their axioms are semantically true, which suggests that something like mathematical reasoning may be appropriate for certain particular types of self-modification during some developmental stages.

Thus the content of the paper is very far off from how a realistic AI would work, but conversely, if you can't even answer the kinds of simple problems posed within the paper (both those we partially solve and those we only pose) then you must be very far off from being able to build a stable self-modifying AI. Being able to say how to build a theoretical device that would play perfect chess given infinite computing power, is very far off from the ability to build Deep Blue. However, if you can't even say how to play perfect chess given infinite computing power, you are confused about the rules of the chess or the structure of chess-playing computation in a way that would make it entirely hopeless for you to figure out how to build a bounded chess-player. Thus "In real life we're always bounded" is no excuse for not being able to solve the much simpler unbounded form of the problem, and being able to describe the infinite chess-player would be substantial and useful conceptual progress compared to *not* being able to do that. We can't be absolutely certain that an analogous situation holds between solving the challenges posed in the paper, and realistic self-modifying AIs with stable goal systems, but every line of investigation has to start somewhere.

Parts of the paper will be easier to understand if you've read Highly Advanced Epistemology 101 For Beginners including the parts on correspondence theories of truth (relevant to section 6) and model-theoretic semantics of logic (relevant to 3, 4, and 6), and there are footnotes intended to make the paper somewhat more accessible than usual, but the paper is still essentially aimed at mathematically sophisticated readers.

Earning to Give vs. Altruistic Career Choice Revisited

A commonly voiced sentiment in the effective altruist community is that the best way to do the most good is generally to make as much money as possible, with a view toward donating to the most cost-effective charities. This is often referred to as "earning to give." In the article To save the world, don't get a job at a charity; go work on Wall Street William MacAskill wrote:

> *Top undergraduates who want to "make a difference" are encouraged to forgo the allure of Wall Street and work in the charity sector ... while researching ethical career choice, I concluded that it's in fact better to earn a lot of money and donate a good chunk of it to the most cost-effective charities, a path that I call "earning to give." ... In general, the charitable sector is people-rich but money-poor. Adding another person to the labor pool just isn't as valuable as providing more money, so that more workers can be hired.*

In private correspondence, MacAskill clarified that he wasn't arguing that "earning to give" is the *best* way to do good, only that it's often better than working at a given nonprofit. In a recent comment MacAskill wrote

> *I think there's too much emphasis on "earning to give" as the *best* option rather than as the *baseline* option*

and raises a number of counter-considerations against "earning to give." Despite this, the idea that "earning to give" is optimal has caught on in the effective altruist community, and so it's important to discuss it.

Over the past three years, I myself have shifted from the position that "earning to give" is philanthropically optimal, to the position that **it's generally the case that one can do more good by choosing a career with high direct social value than by choosing a lucrative career with a view toward donating as much as possible**.

In this post I'll outline some arguments in favor of this view.

Responses to MacAskill's Considerations

In the article To save the world, don't get a job at a charity; go work on Wall Street, MacAskill gives three considerations in favor of "earning to give." I respond to these considerations below. What I write should be read as a response to the article, rather than to MacAskill's views.

Variance in cost-effectiveness of charities

MacAskill wrote

> *... charities vary tremendously in the amount of good they do with the money they receive. For example, it costs about $40,000 to train and provide a guide dog for one person, but it costs less than $25 to cure one person of sight-destroying trachoma. For the cost of improving the life of one person with blindness, you can cure 1,000 people of it...it's unlikely that you can work for only*

> *the very best charities. In contrast, if you earn to give, you can donate anywhere, preferably to the most cost-effective charities, and change your donations as often as you like.*

GiveWell has spent about five years looking for the best giving opportunities in global health, and its current #1 ranked charity is Against Malaria Foundation (AMF). GiveWell estimates that AMF saves an infant's life for ~ $2,300, not counting other benefits. These other benefits not withstanding, AMF's cost per DALY saved is much higher than the implied cost per DALY saved associated with the figure cited for curing sight-destroying trachoma.

GiveWell may have missed giving opportunities in global health that are much more cost-effective than AMF is, but given the amount of time, energy and attention that GiveWell spent on its search, one should have a strong prior against the possibility that one can easily find a better giving opportunity in global health. So a plausible estimate of the cost-effectiveness of donating to the best charity that delivers direct global health interventions is much lower than the above quotation suggests.

Furthermore, the phenomenon of the optimizer's curse suggests that all charities with robust case for fairly high cost-effectiveness are closer in cost-effectiveness to AMF than explicit cost-effectiveness calculations indicate. This narrows the variance in cost-effectiveness amongst charities.

So the advantage of being able to choose a charity to support and change at any time is smaller than the above quotation suggests.

Discrepancy in earnings

MacAskill wrote:

> *Annual salaries in banking or investment start at $80,000 and grow to over $500,000 if you do well. A lifetime salary of over $10 million is typical. Careers in nonprofits start at about $40,000, and don't typically exceed $100,000, even for executive directors ... By entering finance and donating 50% of your lifetime earnings, you could pay for two nonprofit workers in your place—while still living on double what you would have if you'd chosen that route.*

The assumption "if you do well" is a very strong one. Only about 1% of Americans make ~$500k/year. There are some people who have a strong comparative advantage in finance, for whom "earning to give" to give may be especially compelling. But people who are able to make ~$500k/year in finance who **don't** have a large comparative advantage in finance have **very strong transferable skills**. Such people are significantly more capable than the average non-profit worker, and can plausibly have a bigger impact than 2 or 3 such workers by working directly on something with high social value.

Replaceability

MacAskill wrote:

> *..."making a difference" requires doing something that wouldn't have happened anyway...The competition for not-for-profit jobs is fierce, and if someone else takes the job instead of you, he or she likely won't be much worse at it than you would have been. So the difference you make by taking the job is*

> *only the difference between the good you would do, and the good that the other person would have done.*

I would guess that there are some highly cost-effective humanitarian interventions that are sufficiently easy to implement that the implementers are easily replaceable. I could easily imagine that this is the case for vaccination efforts.

But funding opportunities for these interventions can be thought of as "low hanging fruit." Broad market efficiency suggests that such interventions will be funded. And indeed, GiveWell has found that straightforward immunization efforts are already largely funded, to the point that GiveWell has been unable to find giving opportunities for individual donors in this area.

This suggests that at the margin, **very high value humanitarian efforts require highly skilled and highly motivated laborers**.

High skilled laborers are a relatively small subset of laborers, so there are fewer people available to do these sorts of jobs than other jobs. Doing a hard, non-routine job well requires high motivation. The collection of people who are sufficiently highly motivated to do a hard job with high social value that doesn't pay well, and who could otherwise be making much more money, largely consists of people who are trying to have a significant positive social impact.

So suppose that you're a highly skilled laborer deciding whether to "earn to give" or take a job with high social value that requires high skills and motivation. If you don't take the job with high social value, your counterfactual replacement is likely be one of the following:

1. Substantially less capable than you on account of having low skills, or low altruistic motivation.
2. A highly skilled person with high motivation, *who would be doing something else with high social value if you had taken the job, and who can't do this because they have to do the job that you would have done*.
3. Nonexistent.

So the replaceability consideration carries less weight than it might seem.

Admittedly there's a counterconsideration — broad market efficiency cuts both ways, and one could imagine that the low hanging fruit in *working directly on projects with high social valu*e is also plucked, and this counter-consideration pushes in favor of "earning to give." I have a fairly strong intuition that "if you don't fund it, somebody else will" is more true than "if you don't do it, somebody else will" so that this counter-consideration is outweighed. It's important to note that many projects of high social value are the first of their kind, and that finding somebody else to execute such a project is highly nontrivial. I think that it's also relevant that 114 billionaires have signed the Giving Pledge, committing to giving 50+% of their wealth away in their lifetimes.

In any case, there isn't a clear-cut, unconditional argument that favors "earning to give": whether "earning to give" is the best option very much depends on nuanced empirical considerations rather than a general abstract argument.

Other important considerations that favor an altruistic career

There are additional important considerations that favor pursuing a career with high social value over “earning to give”:

Asymmetric implications of the existence of small probability failure modes

In Robustness of Cost-Effectiveness Estimates and Philanthropy, I described how a large collection of small probability failure modes conspires to substantially reduce the expected value of a funding opportunity. The same issue applies to choosing a narrow career goal with a view toward directly having a high positive social impact. But **a worker has more capacity than a donor does to learn whether small probability failure modes prevail in practice, and can switch to a different job if he or she finds that such a failure mode prevails.**

Here’s an example. Suppose that you go to medical school with a view toward the possibility of performing cleft palate surgeries in the developing world. It’s probably the case that the opportunity isn’t as promising as it seems. But if you try it, then you’ll be able to see how effective the intervention is firsthand. If it’s highly effective, then you can keep doing it. If it’s not highly effective, then you can explore other possibilities, such as

- Starting your own surgery organization.
- Switching to doing a different kind of surgery in the developing world, such as cataract removal.
- Working in a poor community in the developed world (which could have a bigger impact than working in the developing world owing to flow-through effects).
- Working for a biotech company.
- Getting involved in clinical medical research.
- Other things that haven't occurred to me.

By experimenting, one can hope to hone in on a job that has both high ostensible cost-effectiveness, and and a relatively small mass of small probability failure modes.

Altruistic careers extend beyond the nonprofit world

Even on the assumption that “earning to give” is better than working at a nonprofit, it doesn’t follow that “earning to give” optimizes social impact. There are ways to have a positive social impact in the for-profit world, in scientific research, and in the government.

Historical Precedent

For the most part, the people who have had the biggest positive impact on the world haven’t had their impact by “earning to give.”

There are a few possible exceptions, such as Bill Gates and Warren Buffett, whose philanthropic activities could be having a huge impact (though it’s hard to tell from the outside) and could well outstrip the value that they contributed through their labor. But they appear to have an unusually high ratio of wealth to direct positive impact of their work, and so appear to be unrepresentative.

Steve Jobs' highest net worth was on the order of $10 billion, whereas Bill Gates' highest net worth was on the order of $100 billion. I don't think that Bill Gates contributed 10x as much as Steve Jobs to technology, and I don't think that Jobs could have had a bigger social impact by donating than through his work (which had massive positive flow-through effects). I acknowledge that Jobs is a cherry picked example, but I think that the general principle still holds.

Mainstream consensus

Few people think that "earning to give" is the best way to make the world a better place. This could be attributable to irrationality or to low altruism, but my experience is that there are many people who care about global welfare, or just welfare within a specific cause, and many people who are highly intelligent. In light of the existence of illusory superiority, one should be wary of holding an implicit view that one knows more about how to make the world a better place than the vast majority of the population.

Steelmanning wealth maximization

It's worth highlighting some factors that *favor* choosing a career with a view toward maximizing wealth in some situations:

- **Comparative advantage** — Some people are unusually good at making money relative to doing other things. Such people may do better to "earn to give" than to try to choose a job that has a direct positive impact (which they're relatively bad at).
- **The market mechanism** — In the for-profit world, maximizing wealth is often correlated with maximizing positive social impact, and so can be used as a proxy goal for maximizing positive social impact.

- **Connections and personal growth** — People with high earnings are generally more capable and more knowledgeable than people in other contexts, and tend to be well connected, so positioning oneself among such people can increase one's prospects of soaring to greater heights. Jeff Bezos started his career in finance, and later created Amazon, which has had massive positive social impact (both direct, and via flow-through effects).

- **Unusual values** — If one cares about causes that very few people care about, then it could be difficult to find funding for work on them, so "earning to give" could be necessary. I don't believe this to be the case, but it's a consideration that's been raised by others, and so is worth mentioning.

Closing summary

There are many arguments against the claim that "earning to give" is generally the best way to maximize one's positive social impact, and I believe that choosing a job where one can do as much good as possible through one's work is generally the best way to maximize one's positive social impact. However, for some people in unusual situations, "earning to give" may be the best way to have a positive social impact.

Note: I formerly worked as a research analyst at GiveWell. All views expressed here are my own.

Acknowledgements: I thank Nick Beckstead, ModusPonies and Will Crouch for helpful feedback on an earlier version of this article.

For FAI: Is "Molecular Nanotechnology" putting our best foot forward?

Molecular nanotechnology, or MNT for those of you who love acronyms, seems to be a fairly common trope on LW and related literature. It's not really clear to me why. In many of the examples of "How could AI's help us" or "How could AI's rise to power" phrases like "cracks protein folding" or "making a block of diamond is just as easy as making a block of coal" are thrown about in ways that make me very very uncomfortable. Maybe it's all true, maybe I'm just late to the transhumanist party and the obviousness of this information was with my invitation that got lost in the mail, but seeing all the physics swept under the rug like that sets off every crackpot alarm I have.

I must post the disclaimer that I have done a little bit of materials science, so maybe I'm just annoyed that you're making me obsolete, but I don't see why this particular possible future gets so much attention. Let us assume that a smarter than human AI will be very difficult to control and represents a large positive or negative utility for the entirety of the human race. Even given that assumption, it's still not clear to me that MNT is a likely element of the future. It isn't clear to me than MNT is physically practical. I don't doubt that it can be done. I don't doubt that very clever metastable arrangements of atoms with novel properties can be dreamed up. Indeed, that's my day job, but I have a hard time believing the only reason you can't make a nanoassembler capable of arbitrary manipulations out of a handful of bottles you ordered from Sigma-Aldrich is because we're just not smart enough. Manipulating individuals atoms means climbing huge binding energy curves, it's an enormously steep, enormously complicated energy landscape, and the Schrodinger Equation scales very very poorly as you add additional particles and degrees of freedom. Building molecular nanotechnology seems to me to be roughly equivalent to being able to make arbitrary lego structures by shaking a large bin of lego in a particular way while blindfolded. Maybe a super human intelligence is capable of doing so, but it's not at all clear to me that it's even possible.

I assume the reason than MNT is added to a discussion on AI is because we're trying to make the future sound more plausible via adding burdensome details. I understand that AI and MNT is less probable than AI or MNT alone, but that both is supposed to sound more plausible. This is precisely where I have difficulty. I would estimate the probability of molecular nanotechnology (in the form of programmable replicators, grey goo, and the like) as lower than the probability of human or super human level AI. I can think of all sorts of objection to the former, but very few objections to the latter. Including MNT as a consequence of AI, especially including it without addressing any of the fundamental difficulties of MNT, I would argue harms the credibility of AI researchers. It makes me nervous about sharing FAI literature with people I work with, and it continues to bother me.

I am particularly bothered by this because it seems irrelevant to FAI. I'm fully convinced that a smarter than human AI could take control of the Earth via less magical means, using time tested methods such as manipulating humans, rigging elections, making friends, killing its enemies, and generally only being a marginally more clever and motivated than a typical human leader. A smarter than human AI

could out-manipulate human institutions and out-plan human opponents with the sort of ruthless efficiency that modern computers beat humans in chess. I don't think convincing people that smarter than human AI's have enormous potential for good and evil is particularly difficult, once you can get them to concede that smarter than human AIs are possible. I do think that waving your hands and saying super-intelligence at things that may be physically impossible makes the whole endeavor seem less serious. If I had read the chain of reasoning smart computer->nanobots before I had built up a store of good-will from reading the Sequences, I would have almost immediately dismissed the whole FAI movement a bunch of soft science fiction, and it would have been very difficult to get me to take a second look.

Put in LW parlance, suggesting things not known to be possible by modern physics without detailed explanations puts you in the reference class "people on the internet who have their own ideas about physics". It didn't help, in my particular case, that one of my first interactions on LW was in fact with someone who appears to have their own view about a continuous version of quantum mechanics.

And maybe it's just me. Maybe this did not bother anyone else, and it's an incredible shortcut for getting people to realize just how different a future a greater than human intelligence makes possible and there is no better example. It does alarm me though, because I think that physicists and the kind of people who notice and get uncomfortable when you start invoking magic in your explanations may be the kind of people FAI is trying to attract.

After critical event W happens, they still won't believe you

In general and across all instances I can think of so far, I do not agree with the part of your futurological forecast in which you reason, "After event W happens, everyone will see the truth of proposition X, leading them to endorse Y and agree with me about policy decision Z."

Example 1: "After a 2-year-old mouse is rejuvenated to allow 3 years of additional life, society will realize that human rejuvenation is possible, turn against deathism as the prospect of lifespan / healthspan extension starts to seem real, and demand a huge Manhattan Project to get it done." (EDIT: This has not happened, and the hypothetical is mouse healthspan extension, not anything cryonic. It's being cited because this is Aubrey de Grey's reasoning behind the Methuselah Mouse Prize.)

Alternative projection: Some media brouhaha. Lots of bioethicists acting concerned. Discussion dies off after a week. Nobody thinks about it afterward. The rest of society does not reason the same way Aubrey de Grey does.

Example 2: "As AI gets more sophisticated, everyone will realize that real AI is on the way and then they'll start taking Friendly AI development seriously."

Alternative projection: As AI gets more sophisticated, the rest of society can't see any difference between the latest breakthrough reported in a press release and that business earlier with Watson beating Ken Jennings or Deep Blue beating Kasparov; it seems like the same sort of press release to them. The same people who were talking about robot overlords earlier continue to talk about robot overlords. The same people who were talking about human irreproducibility continue to talk about human specialness. Concern is expressed over technological unemployment the same as today or Keynes in 1930, and this is used to fuel someone's previous ideological commitment to a basic income guarantee, inequality reduction, or whatever. The same tiny segment of unusually consequentialist people are concerned about Friendly AI as before. If anyone in the science community does start thinking that superintelligent AI is on the way, they exhibit the same distribution of performance as modern scientists who think it's on the way, e.g. Hugo de Garis, Ben Goertzel, etc.

Consider the situation in macroeconomics. When the Federal Reserve dropped interest rates to nearly zero and started printing money via quantitative easing, we had some people loudly predicting hyperinflation just because the monetary base had, you know, gone up by a factor of 10 or whatever it was. Which is kind of understandable. But still, a lot of mainstream economists (such as the Fed) thought we would not get hyperinflation, the implied spread on inflation-protected Treasuries and numerous other indicators showed that the free market thought we were due for below-trend inflation, and then in actual reality we got below-trend inflation. It's one thing to disagree with economists, another thing to disagree with implied market forecasts (why aren't you betting, if you really believe?) but you can still do it sometimes; but when conventional economics, market forecasts, *and reality* all agree on something, it's time to shut up and ask the economists how they knew. I had some credence in inflationary worries before that experience, but not afterward... So what about the rest of the world? In the heavily scientific community you live in, or if you read econblogs, you will find that a number of people actually have started to worry

less about inflation and more about sub-trend nominal GDP growth. You will also find that right now these econblogs are having worry-fits about the Fed prematurely exiting QE and choking off the recovery because the elderly senior people with power have updated more slowly than the econblogs. And in larger society, if you look at what happens when Congresscritters question Bernanke, you will find that they are all terribly, terribly concerned about inflation. Still. The same as before. Some econblogs are very harsh on Bernanke because the Fed did not print enough money, but when I look at the kind of pressure Bernanke was getting from Congress, he starts to look to me like something of a hero just for following conventional macroeconomics as much as he did.

That issue is a hell of a lot more clear-cut than the medical science for human rejuvenation, which in turn is far more clear-cut ethically and policy-wise than issues in AI.

After event W happens, a few more relatively young scientists will see the truth of proposition X, and the larger society won't be able to tell a damn difference. This won't change the situation very much, there are probably already some scientists who endorse X, since X is probably pretty predictable even today if you're unbiased. The scientists who see the truth of X won't all rush to endorse Y, any more than current scientists who take X seriously all rush to endorse Y. As for people in power lining up behind your preferred policy option Z, forget it, they're old and set in their ways and Z is relatively novel without a large existing constituency favoring it. Expect W to be used as argument fodder to support conventional policy options that already have political force behind them, and for Z to not even be on the table.

How to Write Deep Characters

Triggered by: Future Story Status

A helpful key to understanding the art and technique of character in storytelling, is to consider the folk-psychological notion from Internal Family Systems of people being composed of different 'parts' embodying different drives or goals. A shallow character is a character with only one 'part'.

A good rule of thumb is that to create a 3D character, that person must contain at least two different 2D characters who come into conflict. Contrary to the first thought that crosses your mind, three-dimensional good people are constructed by combining at least two different good people with two different ideals, not by combining a good person and a bad person. Deep sympathetic characters have two sympathetic parts in conflict, not a sympathetic part in conflict with an unsympathetic part. Deep smart characters are created by combining at least two different people who are geniuses.

E.g. HPMOR!Hermione contains both a sensible young girl who tries to keep herself and her friends out of trouble, and a starry-eyed heroine, neither of whom are stupid. (Actually, since HPMOR!Hermione is also the one character who I created as close to her canon self as I could manage - she didn't *need* upgrading - I should credit this one to J. K. Rowling.) (Admittedly, I didn't actually follow that rule deliberately to construct Methods, I figured it out afterward when everyone was praising the characterization and I was like, "Wait, people are calling me a character author now? What the hell did I just do right?")

If instead you try to construct a genius character by having an emotionally impoverished 'genius' part in conflict with a warm nongenius part... ugh. Cliche. Don't write the first thing that pops into your head from watching Star Trek. This is not how real geniuses work. HPMOR!Harry, the primary protagonist, contains so many different people he has to give them names, and none of them are stupid, nor does any one of them contain his emotions set aside in a neat jar; they contain different mixtures of emotions and ideals. Combining two cliche characters won't be enough to build a deep character. Combining two different *realistic* people in that character's situation works much better. Two is not a limit, it's a minimum, but everyone involved still has to be recognizably the same person when combined.

Closely related is Orson Scott Card's observation that a conflict between Good and Evil can be interesting, but it's often not half as interesting as a conflict between Good and Good. All standard rules about cliches still apply, and a conflict between good and good which you've previously read about and to which the reader can already guess your correct approved answer, cannot carry the story. A good rule of thumb is that if you have a conflict between good and good which you feel unsure about yourself, or which you can remember feeling unsure about, or you're not sure where exactly to draw the line, you can build a story around it. I consider the most successful moral conflict in HPMOR to be the argument between Harry and Dumbledore in Ch. 77 because it almost perfectly divided the readers on who was in the right *and* about whose side the author was taking. (*This* was done by deliberately following Orson Scott Card's rule, not by accident. Likewise _Three Worlds Collide_, though it was only afterward that I realized how much of the praise for that story, which I hadn't dreamed would be considered literarily meritful by serious SF writers, stemmed from the sheer rarity of stories built around genuinely open moral arguments. Orson Scott Card:

"Propaganda only works when the reader feels like you've been absolutely fair to other side", and writing about a moral dilemma where *you're* still trying to figure out the answer is an excellent way to achieve this.)

Character shallowness can be a symptom of moral shallowness if it reflects a conflict between Good and Evil drawn along lines too clear to bring two good parts of a good character into conflict. This is why it would've been hard for Lord of the Rings to contain conflicted characters without becoming an entirely different story, though as Robin Hanson has just remarked, LotR is a Mileu story, not a Character story. Conflicts between evil and evil are even shallower than conflicts between good and evil, which is why what passes for 'maturity' in some literature is so uninteresting. There's nothing to choose there, no decision to await with bated breath, just an author showing off their disillusionment as a claim of sophistication.

Start Under the Streetlight, then Push into the Shadows

See also: Hack Away at the Edges.

The streetlight effect

You've heard the joke before:

> Late at night, a police officer finds a drunk man crawling around on his hands and knees under a streetlight. The drunk man tells the officer he's looking for his wallet. When the officer asks if he's sure this is where he dropped the wallet, the man replies that he thinks he more likely dropped it across the street. Then why are you looking over here? the befuddled officer asks. Because the light's better here, explains the drunk man.

The joke illustrates the streetlight effect: we "tend to look for answers where the looking is good, rather than where the answers are likely to be hiding."

Freedman (2010) documents at length some harms caused by the streetlight effect. For example:

> A bolt of excitement ran through the field of cardiology in the early 1980s when anti-arrhythmia drugs burst onto the scene. Researchers knew that heart-attack victims with steady heartbeats had the best odds of survival, so a medication that could tamp down irregularities seemed like a no-brainer. The drugs became the standard of care for heart-attack patients and were soon smoothing out heartbeats in intensive care wards across the United States.
>
> But in the early 1990s, cardiologists realized that the drugs were also doing something else: killing about 56,000 heart-attack patients a year. Yes, hearts were beating more regularly on the drugs than off, but their owners were, on average, one-third as likely to pull through. Cardiologists had been so focused on immediately measurable arrhythmias that they had overlooked the longer-term but far more important variable of *death*.

Start under the streetlight

Of course, there are good reasons to search under the streetlight:

> It is often extremely difficult or even impossible to cleanly measure what is really important, so scientists instead cleanly measure what they can, hoping it turns out to be relevant.

In retrospect, we might wish cardiologists had done a decade-long longitudinal study measuring the long-term effects of the new anti-arrhythmia drugs of the 1980s. But it's easy to understand why they didn't. Decades-long longitudinal studies are

expensive, and resources are limited. It was more efficient to rely on an easily-measurable proxy variable like arrhythmias.

We must remember, however, that the analogy to the streetlight joke isn't exact. Searching under the streetlight gives the drunkard virtually *no* information about where his wallet might be. But in science and other disciplines, searching under the streetlight can reveal helpful clues about the puzzle you're investigating. Given limited resources, it's often best to start searching under the streetlight and then, initial clues in hand, push into the shadows.[1]

The problem with streetlight science isn't that it relies on easily-measurable proxy variables. If you want to figure out how some psychological trait works, start with a small study and use free undergraduates at your home university — that's a good way to test hypotheses cheaply. The problem comes in when researchers don't appropriately *flag* the fact their subjects were WEIRD and that a larger study needs to be done on a more representative population before we start drawing conclusions. (Another problem is that despite some researcher's cautions against overgeneralizing from a study of WEIRD subjects, the media will write splashy, universalizing headlines anyway.)

But money and time aren't the only resources that might be limited. Another is *human reasoning ability*. Human brains were built for hunting and gathering in the savannah, not for unlocking the mysteries of fundamental physics or intelligence or consciousness. So even if time and money aren't limiting factors, it's often best to break a complex problem into pieces and think through the simplest pieces, or the pieces for which our data are most robust, before trying to answer the questions you *most* want to solve.

As Pólya advises in his hugely popular *How to Solve It*, "If you cannot solve the proposed problem, try to solve first some related [but easier] problem." In physics, this related but easier problem is often called a toy model. In other fields, it is sometimes called a toy problem. Animal models are often used as toy models in biology and medicine.

Or, as Scott Aaronson put it:

> ...I don’t spend my life thinking about P versus NP [because] there are vastly easier prerequisite questions that we already don’t know how to answer. In a field like [theoretical computer science], you very quickly get used to being able to state a problem with perfect clarity, knowing exactly what would constitute a solution, and still not having any clue how to solve it... And at least in my experience, being pounded with this situation again and again slowly reorients your worldview... Faced with a [very difficult question,] you learn to respond: “What’s another question that’s easier to answer, and that probably has to be answered anyway before we have any chance on the original one?”

I'll close with two examples: GiveWell on effective altruism and MIRI on stability under self-modification.

GiveWell on effective altruism

GiveWell's mission is "to find outstanding giving opportunities and publish the full details of our analysis to help donors decide where to give."

But finding and verifying outstanding giving opportunities is *hard*. Consider the case of one straightforward-seeming intervention: deworming.

Nearly 2 billion people (mostly in poor countries) are infected by parasitic worms that hinder their cognitive development and overall health. This is also producing barriers to economic development where parasitic worms are common. Luckily, deworming pills are cheap, and early studies indicated that they improved educational outcomes. The DCP2, produced by over 300 contributors and in collaboration with the World Health Organization, estimated that a particular deworming treatment was one of the most cost-effective treatments in global health, at just $3.41 per DALY.

Unfortunately, things are not so simple. A careful review of the evidence in 2008 by The Cochrane Collaboration concluded that, due to weaknesses in some studies' designs and other factors, "No effect [of deworming drugs] on cognition or school performance has been demonstrated." And in 2011, GiveWell found that a spreadsheet used to produce the DCP2's estimates contained *5 separate errors* that, when corrected, increased the cost estimate for deworming by roughly *a factor of 100*. In 2012, another Cochrane review was even more damning for the effectiveness of deworming, concluding that "Routine deworming drugs given to school children... has not shown benefit on weight in most studies... For haemoglobin and cognition, community deworming seems to have little or no effect, and the evidence in relation to school attendance, and school performance is generally poor, with no obvious or consistent effect."

On the other hand, Innovations for Poverty Action critiqued the 2012 Cochrane review, and GiveWell said the review did not fully undermine the case for its #3 recommended charity, which focuses on deworming.

What are we to make of this? Thousands of hours of data collection and synthesis went into producing the initial case for deworming as a cost-effective intervention, and thousands of additional hours were required to discover flaws in those initial analyses. In the end, GiveWell recommends one deworming charity, the Schistosomiasis Control Initiative, but their page on SCI is littered with qualifications and concerns and "We don't know"s.

GiveWell had to wrestle with these complications despite the fact that it *chose* to search under the streetlight. Global health interventions are among the *easiest* interventions to analyze, and have often been subjected to multiple randomized controlled trials and dozens of experimental studies. Such high-quality evidence usually isn't available when trying to estimate the cost-effectiveness of, say, certain forms of political activism.

GiveWell co-founder Holden Karnofsky suspects that the best giving opportunities are *not* in the domain of global health, but GiveWell began their search in global health — under the spotlight — (in part) because the evidence was clearer there.[2]

It's difficult to do counterfactual history, but I suspect GiveWell made the right choice. While investigating global health, GiveWell has learned many important lessons about effective altruism — lessons it would have been more difficult to learn with the same clarity if they had begun with investigations of even-more-challenging domains like meta-research and pollitical activism. But now that they've learned those lessons,

they're beginning to push into the shadows where the evidence is less clear, via GiveWell Labs.

MIRI on stability under self-modification

MIRI's mission is "to ensure that the creation of smarter-than-human intelligence has a positive impact."

Many different interventions have been proposed as methods for increasing the odds that smarter-than-human intelligence has a positive impact, but for several reasons MIRI decided to focus its efforts on "Friendly AI research" during 2013.

The FAI research program decomposes into a wide variety of technical research questions. One of those questions is the question of *stability under self-modification*:

> How can we ensure that an AI will serve its intended purpose even after repeated self-modification?

This is a challenging and ill-defined question. How might we make progress on such a puzzle?

For puzzles such as this one, Scott Aaronson recommends a strategy he calls "bait and switch":

> [Philosophical] progress has almost always involved a [kind of] “bait-and-switch.” In other words: one replaces an unanswerable philosophical riddle Q by a “merely” scientific or mathematical question Q′, which captures part of what people have wanted to know when they’ve asked Q. Then, with luck, one solves Q′… this process of “breaking off” answerable parts of unanswerable riddles, then trying to answer those parts, is the closest thing to philosophical progress that there is.
>
> Successful examples of this breaking-off process fill intellectual history. The use of calculus to treat infinite series, the link between mental activity and nerve impulses, natural selection, set theory and first-order logic, special relativity, Gödel’s theorem, game theory, information theory, computability and complexity theory, the Bell inequality, the theory of common knowledge, Bayesian causal networks — each of these advances addressed questions that could rightly have been called “philosophical” before the advance was made.

The recent MIRI report on Tiling Agents performs one such "bait and switch." It replaces the philosophical puzzle of "How can we ensure that an AI will serve its intended purpose even after repeated self-modification?" (Q) with a better-specified *formal* puzzle on which it is possible to make measurable progress: "How can an agent perform perfectly tiling self-modifications despite Löb's Theorem?" (Q')

This allows us to state at least three crisp technical problems: Löb and coherent quantified belief (sec. 3 of 'Tiling Agents'), nonmonotonicity of probabilistic reasoning (secs. 5.2 & 7), and maximizing/satisficing not being satisfactory for bounded agents (sec. 8). It also allows us to identify progress: formal results that mankind had not previously uncovered (sec. 4).

Of course, even if Q' is eventually solved, we'll need to check whether there are other pieces of Q we need to solve. Or perhaps Q will have been *dissolved* by our efforts to solve Q', similar to how the question "What force distinguishes living matter from non-living matter?" was dissolved by 20th century biology.

Notes

[1] Karnofsky (2011) suggests that it may often be best to start under the streetlight *and stay there*, at least in the context of effective altruism. Karnofsky asks, "What does it look like when we build knowledge only where we're best at building knowledge, rather than building knowledge on the 'most important problems?'" His reply is: "Researching topics we're good at researching can have a lot of benefits, some unexpected, some pertaining to problems we never expected such research to address. Researching topics we're bad at researching doesn't seem like a good idea no matter how important the topics are. Of course I'm in favor of thinking about how to develop new research methods to make research good at what it was formerly bad at, but I'm against applying current problematic research methods to current projects just because they're the best methods available." Here's one example: "what has done more for political engagement in the U.S.: studying how to improve political engagement, or studying the technology that led to the development of the Internet, the World Wide Web, and ultimately to sites like Change.org...?" I am sympathetic with Karnofsky's view in many cases, but I will give two points of reply with respect to my post above. First, in the above post I wanted to focus on the question of how to tackle difficult questions, not the question of whether difficult questions should be tackled in the first place. And conditional on one's choice to tackle a difficult question, I recommend one start under the streetlight and push into the shadows. Second, my guess is that I'm talking about a broader notion of the streetlight effect than Karnofsky is. For example, I doubt Karnofsky would object to the process of tackling a problem in theoretical computer science or math by trying to solve easier, related problems first.

[2] In GiveWell's January 24th, 2013 board meeting (starting at 6:35 in the MP3 recording), GiveWell co-founder Holden Karnofsky said that interventions outside global health are "where we would bet today that we'll find... the best giving opportunities... that best fulfill GiveWell's mission as originally [outlined] in the mission statement." This doesn't appear to be a recently acquired view of things, either. Starting at 22:47 in the same recording, Karnofsky says "There were reasons that we focused on [robustly evidence-backed] interventions for GiveWell initially, but... the [vision] I've been pointing to [of finding giving opportunities outside global health, where less evidence is available]... has [to me] been the vision all along." In personal communication with me, Karnofsky wrote that "We sought to start 'under the streetlight,' as you say, and so focused on finding opportunities to fund things with strong documented evidence of being 'proven, cost-effective and scalable.' Initially we looked at both U.S. and global interventions, and within developing-world interventions we looked at health but also economic empowerment. We ended up focusing on global health because it performed best by these criteria."

Prisoner's Dilemma (with visible source code) Tournament

After the iterated prisoner's dilemma tournament organized by prase two years ago, there was discussion of running tournaments for several variants, including one in which two players submit programs, each of which are given the source code of the other player's program, and outputs either “cooperate” or “defect”. However, as far as I know, no such tournament has been run until now.

Here's how it's going to work: Each player will submit a file containing a single Scheme lambda-function. The function should take one input. Your program will play exactly one round against each other program submitted (not including itself). In each round, two programs will be run, each given the source code of the other as input, and will be expected to return either of the symbols “C” or “D” (for "cooperate" and "defect", respectively). The programs will receive points based on the following payoff matrix:

“Other” includes any result other than returning “C” or “D”, including failing to terminate, throwing an exception, and even returning the string “Cooperate”. Notice that “Other” results in a worst-of-both-worlds scenario where you get the same payoff as you would have if you cooperated, but the other player gets the same payoff as if you had defected. This is an attempt to ensure that no one ever has incentive for their program to fail to run properly, or to trick another program into doing so.

Your score is the sum of the number of points you earn in each round. The player with the highest score wins the tournament. **Edit: There is a 0.5 bitcoin prize being offered for the winner. Thanks, VincentYu!**

Details:
All submissions must be emailed to wardenPD@gmail.com by July 5, at noon PDT (Edit: that's 19:00 UTC). Your email should also say how you would like to be identified when I announce the tournament results.
Each program will be allowed to run for 10 seconds. If it has not returned either “C” or “D” by then, it will be stopped, and treated as returning “Other”. For consistency, I will have Scheme collect garbage right before each run.
One submission per person or team. No person may contribute to more than one entry. **Edit: This also means no copying from each others' source code. Describing the behavior of your program to others is okay.**
I will be running the submissions in Racket. You may be interested in how Racket handles time (especially the (current-milliseconds) function), threads (in particular, “thread”, “kill-thread”, “sleep”, and “thread-dead?”), and possibly randomness.
Don't try to open the file you wrote your program in (or any other file, for that matter). I'll add code to the file before running it, so if you want your program to use a copy of your source code, you will need to use a quine. **Edit: No I/O of any sort.**
Unless you tell me otherwise, I assume I have permission to publish your code after

the contest.
You are encouraged to discuss strategies for achieving mutual cooperation in the comments thread.
I'm hoping to get as many entries as possible. If you know someone who might be interested in this, please tell them.
It's possible that I've said something stupid that I'll have to change or clarify, so you might want to come back to this page again occasionally to look for changes to the rules. Any edits will be bolded, and I'll try not to change anything too drastically, or make any edits late in the contest.

Here is an example of a correct entry, which cooperates with you if and only if you would cooperate with a program that always cooperates (actually, if and only if you would cooperate with one particular program that always cooperates):

```
(lambda (x)
   (if (eq? ((eval x) '(lambda (y) 'C)) 'C)
       'C
       'D))
```

How probable is Molecular Nanotech?

Circa a week ago I posted asking whether bringing up molecular nanotechnology(MNT) as a possible threat avenue for an unfriendly artificial intelligence made FAI research seem less credible because MNT seemed to me to be not obviously possible. I was told to some extent, to put up and address the science of MNT or shut up. A couple of people also expressed an interest in seeing a more fact and less PR oriented discussion, so I got the ball rolling and you all have no one to blame but yourselves. I should note before starting, that I do not personally have a strong opinion on whether Drexler-style MNT is possible. This isn't something I've researched previously, and I'm open to being convinced one way or the other. If MNT turns out to be likely at the end of this investigation, then hopefully this discussion can provide a good resource for LW/FAI on the topic for people like myself not yet convinced that MNT is the way of future. As far as I'm concerned, at this point all paths lead to victory.

While *Nanosystems* was the canonical reference mentioned in the last conversation. I purchased it, then about 2/3rds of the way through this I figured Engines of Creation was giving me enough to work with and cancelled my order. If the science in Nanosystems is really much better than in EoC I can reorder it, but I figured we'd get started for free. 50 bucks is a lot of money to spend on an internet argument.

Before I begin I would like to post the following disclaimers.

1. I am not an expert in many of the claims that border on MNT. I did work at a Nanotechnology center for a year, but that experience was essentially nothing like what Drexler describes. More relevantly I am in the process of completing a Ph.D. in Physics, and my thesis work is on computational modeling of novel materials. I don't really like squishy things, so I'm very much out of my depth when it comes to discussions as to what ribosomes can and cannot accomplish, and I'll happily defer to other authorities on the more biological subjects. With that being said, several of my colleagues run MD simulations of protein folding all day every day, and if a biology issue is particularly important, I can shoot some emails around the department and try and get a more expert opinion.

2. There are several difficulties in precisely addressing Drexler's arguments, because it's not always clear to me at least exactly what his arguments are. I've been going through Engines of Creation and several of his other works, and I'll present my best guess outline here. If other people would like to contribute specific claims about molecular nanotech, I'll be happy to add them to the list and do my best to address them.

3. This discussion is intended to be scientific. As was pointed out previously, Drexler et al. have made many claims about time tables of when things might be invented. Judging the accuracy of these claims is difficult because of issues with definitions as mentioned in the previous paragraph. I'm not interested in having this discussion encompass Drexler's general prediction accuracy. Nature is the only authority I'm interested in consulting in this thread. If someone wants to make a Drexler's prediction accuracy thread, they're welcome to do so.

4. If you have any questions about the science underlying anything I say, don't hesitate to ask. This is a fairly technical topic, and I'm happy to bring anyone up to

speed on basic physics/chemistry terms and concepts.

Discussion

I'll begin by providing some background and highlighting why exactly I am not already convinced that MNT, and especially AI-assisted rapid MNT is the future, and then I'll try and address some specific claims made by Drexler in various publications.

Conservation of energy:

Feynman, and to some extent Drexler, spends an enormous amount of time addressing issues that we are familiar with from dealing with macroscopic pieces of equipment, such as how much space it takes to store things, how parts can wear out, etc. What is not mentioned in how we plan to power these Engines of Creation. Assembling nanotechnology is more than just getting atoms into the individual places you want them, it's a matter of very precise energetic control. The high resolution energy problem is equally as difficult as fine-grain control of atom positions, and this is further complicated by the fact that any energy delivery system you contrive for a nano-assembler is also going to impart momentum. In the macroscale world, your factory doesn't start sliding when you hook it up to the grid. At smaller sizes, that may not be true. It's very unclear in most of the discussions I read about these Nanofactories what's going to power them. What synthetic equivalent of ATP is going to allow us to out-compete the ribosome? What novel energy source is grey-goo going to have access to that will allow it break and reassemble the bonds necessary for nanofabrication?

Modelling is hard:

Solving the Schrodinger equation is essentially impossible. We can solve it more or less exactly for the Hydrogen atom, but things get very very difficult from there. This is because we don't have a simple solution for the three-body problem, much less the n-body problem. Approximately, the difficulty is that because each electron interacts with every other electron, you have a system where to determine the forces on electron 1, you need to know the position of electrons 2 through N, but the position of each of those electrons depends somewhat on electron 1. We have some tricks and approximations to get around this problem, but they're only justified empirically. The only way we know what approximations are good approximations is by testing them in experiments. Experiments are difficult and expensive, and if the AI is using MNT to gain infrastructure, then we can assume it doesn't already have the infrastructure to run its own physics lab.

A factory isn't the right analogy:

The discussion of nanotechnology seems to me to have an enormous emphasis on Assemblers, or nanofactories, but a factory doesn't run unless it has a steady supply of raw materials and energy resources both arriving at the correct time. The evocation of a factory calls to mind the rigid regularity of an assembly line, but the factory only works because it's situated in the larger, more chaotic world of the economy. Designing new nanofactories isn't a problem of building the factory, but a problem of designing an entire economy. There has to be a source of raw material, an energy source, and means of transporting material and energy from place to place. And, with a microscopic factory, Brownian motion may have moved the factory by the time the

delivery van gets there. This fact makes the modelling problem orders of magnitude more difficult. Drexler makes a big deal about how his rigid positional world isn't like the chaotic world of the chemists, but it seems like the chaos is still there; building a factory doesn't get rid of the logistics issue.

Chaos

The reason we can't solve the n-body problem, and lots of other problems such as the double pendulum and the weather is because it turns out to be a rather unfortunate fact of nature that many systems have a very sensitive dependence on initial conditions. This means that ANY error, any unaccounted for variable, can perturb a system in dramatic ways. Since there will always be some error (at the bare minimum $h/4\pi$) this means that our AI is going to have to do Monte Carlo simulations like the rest of us smucks and try to eliminate as many degrees of freedom as possible.

The laws of physics hold

I didn't think it would be necessary to mention this, but I believe that the laws of physics are pretty much the laws of physics we know right now. I would direct anyone who suggests that an AI has a shot at powering MNT with cold fusion, tachyons, or other physical phenomena not predicted by the standard model to this post. I am not saying there is no new no physics, but we understand quantum mechanics really well, and the Standard Model has been confirmed to enough decimal places that anyone who suggests something the Standard Model says can't happen is almost certainly wrong. Even if they have experimental evidence that is supposed to 99.9999% percent correct.

Specific Claims

Drexler's claims about what we can do now with respect to materials science in general are true. This should be unsurprising. It is not particularly difficult to predict the past. Here are 6 claims he makes that we can't currently accomplish which I'll try and evaluate:

1. Building "gear-like" nanostructures is possible (Toward Integrated Nanosystems)
2. Predicting crystal structures from first principles is possible (Toward Integrated Nanosystems)
3. Genetic engineering is a superior form of chemical synthesis to traditional chemical plants. (EoC 6)
4. "*Biochemical engineers, then, will construct new enzymes to assemble new patterns of atoms. For example, they might make an enzyme-like machine which will add carbon atoms to a small spot, layer on layer. If bonded correctly, the atoms will build up to form a fine, flexible diamond fiber having over fifty times as much strength as the same weight of aluminum*." (EoC 10)
5. Proteins can make and break diamond bonds (EoC 11)
6. Proteins are "programmable" (EoC 11)

1. Maybe. This depends on definitions. We can build molecules that rotate, and indeed they occur naturally, but those are a long way from Drexler's proposals. I haven't run any simulations as to whether specific designs such as the molecular planetary gear he exhibits are actually stable. If anyone has an xyz file for one of those doodads I'll be happy to run a simulation. You might look at the state of the art and imagine that if

we can make atomic flip books that molecular gears can't be too far off, but it's not really true. That video is more like molecular feet than molecular hands. We can push a molecule around on the floor, but we can't really do anything useful with it.

2. True. This isn't true yet, but should be possible. I might even work on this after I graduate, if don't go hedge fund or into AI research.

3. Not wrong, but misleading. The statement "*Genetic engineers have now programmed bacteria to make proteins ranging from human growth hormone to rennin, an enzyme used in making cheese*." is true in the same sense that copying and pasting someone else's code constitutes programming. Splicing a gene into a plasmid is sweet, but genetic programming implies more control than we have. Similarly, the statement: "*Whereas engineers running a chemical plant must work with vats of reacting chemicals (which often misarrange atoms and make noxious byproducts), engineers working with bacteria can make them absorb chemicals, carefully rearrange the atoms, and store a product or release it into the fluid around them*." implies that bacterial synthesis leads to better yields (false), that bacteria are careful(meaningless), and implies greater control over genetically modified E.Coli than we have.

4a. False. Flexible diamond doesn't make any sense. Diamond is sp3 bonded carbon and those bonds are highly directional. They're not going to flex.. Metals are flexible because metallic bonds, unlike covalent bonds, don't confine the electrons in space. Whatever this purported carbon fiber is, it either won't be flexible, or it won't be diamond.

4b. False. It isn't clear that this is even remotely possible. Enzymes don't work like this. Enzymes are catalysts for existing reactions. There is no existing reaction that results in a single carbon atom. That's an enormously energetically unfavorable state. Breaking a single carbon carbon double bond requires something like 636 kJ/mol (6.5eV) of energy. That's roughly equivalent to burning 30 units of ATP at the same time. How? How do you get all that energy into the right place at the right time? How does your enzyme manage to hold on to the carbons strongly enough to pull them apart?

5. "*A flexible, programmable protein machine will grasp a large molecule (the workpiece) while bringing a small molecule up against it in just the right place. Like an enzyme, it will then bond the molecules together. By bonding molecule after molecule to the workpiece, the machine will assemble a larger and larger structure while keeping complete control of how its atoms are arranged. This is the key ability that chemists have lacked.*" I'm no biologist, but this isn't how proteins work. Proteins aren't Turing machines. You don't set the state and ignore them. The conformation of a protein depends intimately on its environment. The really difficult part here is that the thing it's holding, the nanopart you're trying to assemble is a big part of the protein's environment. Drexler complains around how proteins are no good because they're soft and squishy, but then he claims they're strong enough to assemble diamond and metal parts. But if the stiff nanopart that you're assembling has a dangling carbon bond waiting to filled then it's just going to cannibalize the squishy protein that's holding it. What can a protein held together by Van der Waals bonds do to a diamond? How can it control the shape it takes well enough to build a fiber?

6. All of these tiny machines are repeatedly described as programmable, but that doesn't make any sense. What programs are they capable of accepting or executing? What set of instructions can a collection of 50 carbon atoms accept and execute? How

are these instructions being delivered? This gets back to my factory vs. economy complaint. If nothing else, this seems an enormously sloppy use of language.

Some things that are possible

I think we have or will have the technology to build some interesting artificial inorganic structures in very small quantities, primarily using ultra-cold, ultra-high-vacuum laser traps. It's even possible that eventually we could create some functional objects this way, though I can't see any practical way to scale that production up.

"Nanorobots" will be small pieces of metal or dieletric material that we manipulate with lasers or sophisticated magnetic fields, possibly attached to some sort of organic ligand. This isn't much of a prediction, we pretty much do this already. The nanoworld will continue to be statistical and messy.

We will gain some inorganic control over organics like protein and DNA (though not organic over inorganic). This hasn't really been done yet that I'm aware of, but stronger bonds>weaker bonds makes sense. I think there are people trying to read DNA/proteins by pushing the strands through tiny silicon windows. I feel like I heard a seminar along those lines, though I'm pretty sure I slept through it.

That brings me through the first 12 pages of EoC or so. More to follow. Let me know if the links don't work or the formatting is terrible or I said something confusing. Also, please contribute any specific MNT claims you'd like evaluated, and any resources or publications you think are relevant. Thank you.

Bibliography

Engines of Creation

Toward Integrated Nanosystems

Molecular Devices and Machines)

Do Earths with slower economic growth have a better chance at FAI?

I was raised as a good and proper child of the Enlightenment who grew up reading *The Incredible Bread Machine* and *A Step Farther Out,* taking for granted that economic growth was a huge in-practice component of human utility (plausibly the majority component if you asked yourself what was the major difference between the 21st century and the Middle Ages) and that the "Small is Beautiful" / "Sustainable Growth" crowds were living in impossible dreamworlds that rejected quantitative thinking in favor of protesting against nuclear power plants.

And so far as I know, such a view would still be an excellent first-order approximation if we were going to carry on into the future by steady technological progress: Economic growth = good.

But suppose my main-line projection is correct and the "probability of an OK outcome" / "astronomical benefit" scenario essentially comes down to a race between Friendly AI and unFriendly AI. So far as I can tell, the most likely reason we wouldn't get Friendly AI is the total *serial* research depth required to develop and implement a strong-enough theory of stable self-improvement with a possible side order of failing to solve the goal transfer problem. Relative to UFAI, FAI work seems like it would be mathier and more insight-based, where UFAI can more easily cobble together lots of pieces. This means that UFAI parallelizes better than FAI. UFAI also probably benefits from brute-force computing power more than FAI. Both of these imply, so far as I can tell, that slower economic growth is good news for FAI; it lengthens the deadline to UFAI and gives us more time to get the job done. I have sometimes thought half-jokingly and half-anthropically that I ought to try to find investment scenarios based on a continued Great Stagnation and an indefinite Great Recession where the whole developed world slowly goes the way of Spain, because these scenarios would account for a majority of surviving Everett branches.

Roughly, it seems to me like higher economic growth *speeds up time* and this is not a good thing. I wish I had more time, not less, in which to work on FAI; I would prefer worlds in which this research can proceed at a relatively less frenzied pace and still succeed, worlds in which the default timelines to UFAI terminate in 2055 instead of 2035.

I have various cute ideas for things which could improve a country's economic growth. The chance of these things eventuating seems small, the chance that they eventuate because I write about them seems tiny, and they would be good mainly for entertainment, links from econblogs, and possibly marginally impressing some people. I was thinking about collecting them into a post called "The Nice Things We Can't Have" based on my prediction that various forces will block, e.g., the all-robotic all-electric car grid which could be relatively trivial to build using present-day technology - that we are too far into the Great Stagnation and the bureaucratic maturity of developed countries to get nice things anymore. However I have a certain inhibition against trying things that would make everyone worse off if they actually succeeded, even if the probability of success is tiny. And it's not completely impossible that we'll see some actual experiments with small nation-states in the next few decades, that some of the people doing those experiments will have read Less Wrong, or that successful experiments will spread (if the US ever legalizes robotic cars or tries a city

with an all-robotic fleet, it'll be because China or Dubai or New Zealand tried it first). Other EAs (effective altruists) care much more strongly about economic growth directly and are trying to increase it directly. (An extremely understandable position which would typically be taken by good and virtuous people).

Throwing out remote, contrived scenarios where something accomplishes the opposite of its intended effect is cheap and meaningless (vide "But what if MIRI accomplishes the opposite of its purpose due to blah") but in this case I feel impelled to ask because my *mainline* visualization has the Great Stagnation being good news. I certainly *wish* that economic growth would align with FAI because then my virtues would align and my optimal policies have fewer downsides, but I am also aware that wishing does not make something more likely (or less likely) in reality.

To head off some obvious types of bad reasoning in advance: Yes, higher economic growth frees up resources for effective altruism and thereby increases resources going to FAI, but it also increases resources going to the AI field generally which is mostly pushing UFAI, and the problem *arguendo* is that UFAI parallelizes more easily.

Similarly, a planet with generally higher economic growth might develop intelligence amplification (IA) technology earlier. But this general advancement of science will also accelerate UFAI, so you might just be decreasing the amount of FAI research that gets done before IA and decreasing the amount of time available after IA before UFAI. Similarly to the more mundane idea that increased economic growth will produce more geniuses some of whom can work on FAI; there'd also be more geniuses working on UFAI, and UFAI probably parallelizes better and requires less serial depth of research. If you concentrate on some single good effect on *blah* and neglect the corresponding speeding-up of UFAI timelines, you will obviously be able to generate spurious arguments for economic growth having a positive effect on the balance.

So I pose the question: "Is slower economic growth good news?" or "Do you think Everett branches with 4% or 1% RGDP growth have a better chance of getting FAI before UFAI"? So far as I can tell, my current mainline guesses imply, "Everett branches with slower economic growth contain more serial depth of cognitive causality and have more effective time left on the clock before they end due to UFAI, which favors FAI research over UFAI research".

This seems like a good parameter to have a grasp on for any number of reasons, and I can't recall it previously being debated in the x-risk / EA community.

EDIT: To be clear, the idea is not that trying to *deliberately slow* world economic growth would be a maximally effective use of EA resources and better than current top targets; this seems likely to have very small marginal effects, and many such courses are risky. The question is whether a good and virtuous person ought to avoid, or alternatively seize, any opportunities which come their way to help out on world economic growth.

EDIT 2: Carl Shulman's opinion can be found on the Facebook discussion here.

Public Service Announcement Collection

P/S/A: There are single sentences which can create life-changing amounts of difference.

- P/S/A: If you're not sure whether or not you've ever had an orgasm, it means you haven't had one, a condition known as primary anorgasmia which is 90% treatable by cognitive-behavioral therapy.
- P/S/A: The people telling you to expect above-trend inflation when the Federal Reserve started printing money a few years back, disagreed with the market forecasts, disagreed with standard economics, turned out to be actually wrong in reality, and were wrong for reasonably fundamental reasons so don't buy gold when they tell you to.
- P/S/A: There are many many more submissive/masochistic men in the world than there are dominant/sadistic women, so if you are a woman who feels a strong temptation to command men and inflict pain on them, and you want a large harem of men serving your every need, it will suffice to state this fact anywhere on the Internet and you will have fifty applications by the next morning.
- P/S/A: Most of the personal-finance-advice industry is parasitic and/or self-deluded, and it's generally agreed on by economic theory and experimental measurement that an index fund will deliver the best returns you can get without huge amounts of effort.
- P/S/A: If you are smart and underemployed, you can very quickly check to see if you are a natural computer programmer by pulling up a page of Python source code and seeing whether it looks like it makes natural sense, and if this is the case you can teach yourself to program very quickly and get a much higher-paying job even without formal credentials.

Useful Concepts Repository

See also: Boring Advice Repository, Solved Problems Repository, Grad Student Advice Repository

I often find that my understanding of the world is strongly informed by a few key concepts. For example, I've repeatedly found the concept of opportunity cost to be a useful frame. My previous post on privileging the question is in some sense about the opportunity cost of paying attention to certain kinds of questions (namely that you don't get to use that attention on other kinds of questions). Efficient charity can also be thought of in terms of the opportunity cost of donating inefficiently to charity. I've also found the concept of incentive structure very useful for thinking about the behavior of groups of people in aggregate (see perverse incentive).

I'd like people to use this thread to post examples of concepts they've found particularly useful for understanding the world. I'm personally more interested in concepts that don't come from the Sequences, but comments describing a concept from the Sequences and explaining why you've found it useful may help people new to the Sequences. ("Useful" should be interpreted broadly: a concept specific to a particular field might be useful more generally as a metaphor.)