Best of LessWrong: May 2012

1. Thoughts on the Singularity Institute (SI)
2. Punctuality - Arriving on Time and Math
3. The rational rationalist's guide to rationally using "rational" in rational post titles
4. Avoid inflationary use of terms
5. When None Dare Urge Restraint, pt. 2
6. Value of Information: 8 examples
7. Alan Carter on the Complexity of Value
8. Problematic Problems for TDT
9. Petition: Off topic area
10. Thinking and Deciding: a chapter by chapter review
11. How to brainstorm effectively
12. PSA: Learn to code
13. A Protocol for Optimizing Affection
14. Share Your Checklists!
15. Off to Alice Springs
16. Case Study: Testing Confirmation Bias

Best of LessWrong: May 2012

1. Thoughts on the Singularity Institute (SI)
2. Punctuality - Arriving on Time and Math
3. The rational rationalist's guide to rationally using "rational" in rational post titles
4. Avoid inflationary use of terms
5. When None Dare Urge Restraint, pt. 2
6. Value of Information: 8 examples
7. Alan Carter on the Complexity of Value
8. Problematic Problems for TDT
9. Petition: Off topic area
10. Thinking and Deciding: a chapter by chapter review
11. How to brainstorm effectively
12. PSA: Learn to code
13. A Protocol for Optimizing Affection
14. Share Your Checklists!
15. Off to Alice Springs
16. Case Study: Testing Confirmation Bias

Thoughts on the Singularity Institute (SI)

This post presents thoughts on the Singularity Institute from Holden Karnofsky, Co-Executive Director of GiveWell. Note: Luke Muehlhauser, the Executive Director of the Singularity Institute, reviewed a draft of this post, and commented: "I do generally agree that your complaints are either correct (especially re: past organizational competence) or incorrect but not addressed by SI in clear argumentative writing (this includes the part on 'tool' AI). I am working to address both categories of issues." I take Luke's comment to be a significant mark in SI's favor, because it indicates an explicit recognition of the problems I raise, and thus increases my estimate of the likelihood that SI will work to address them.

***September 2012 update:** responses have been posted by Luke and Eliezer (and I have responded in the comments of their posts). I have also added acknowledgements.*

The Singularity Institute (SI) is a charity that GiveWell has been repeatedly asked to evaluate. In the past, SI has been outside our scope (as we were focused on specific areas such as international aid). With GiveWell Labs we are open to any giving opportunity, no matter what form and what sector, but we still do not currently plan to recommend SI; given the amount of interest some of our audience has expressed, I feel it is important to explain why. Our views, of course, remain open to change. (Note: I am posting this only to Less Wrong, not to the GiveWell Blog, because I believe that everyone who would be interested in this post will see it here.)

I am currently the GiveWell staff member who has put the most time and effort into engaging with and evaluating SI. Other GiveWell staff currently agree with my bottom-line view that we should not recommend SI, but this does not mean they have engaged with each of my specific arguments. Therefore, while the lack of recommendation of SI is something that GiveWell stands behind, the specific arguments in this post should be attributed only to me, not to GiveWell.

Summary of my views

- The argument advanced by SI for why the work it's doing is beneficial and important seems both wrong and poorly argued to me. My sense at the moment is that the arguments SI is making would, if accepted, increase rather than decrease the risk of an AI-related catastrophe. More
- SI has, or has had, multiple properties that I associate with ineffective organizations, and I do not see any specific evidence that its personnel/organization are well-suited to the tasks it has set for itself. More
- A common argument for giving to SI is that "even an infinitesimal chance that it is right" would be sufficient given the stakes. I have written previously about why I reject this reasoning; in addition, prominent SI representatives seem to reject this particular argument as well (i.e., they believe that one should support SI only if one believes it is a strong organization making strong arguments). More
- My sense is that at this point, given SI's current financial state, withholding funds from SI is likely better for its mission than donating to it. (I would not take this view to the furthest extreme; the argument that SI should have *some*

funding seems stronger to me than the argument that it should have as much as it currently has.)
- I find existential risk reduction to be a fairly promising area for philanthropy, and plan to investigate it further. More
- There are many things that could happen that would cause me to revise my view on SI. However, I do not plan to respond to all comment responses to this post. (Given the volume of responses we may receive, I may not be able to even read all the comments on this post.) I do not believe these two statements are inconsistent, and I lay out paths for getting me to change my mind that are likely to work better than posting comments. (Of course I encourage people to post comments; I'm just noting in advance that this action, alone, doesn't guarantee that I will consider your argument.) More

Intent of this post

I did not write this post with the purpose of "hurting" SI. Rather, I wrote it in the hopes that **one of these three things** (or some combination) will happen:

1. New arguments are raised that cause me to change my mind and recognize SI as an outstanding giving opportunity. If this happens I will likely attempt to raise more money for SI (most likely by discussing it with other GiveWell staff and collectively considering a GiveWell Labs recommendation).
2. SI concedes that my objections are valid and increases its determination to address them. A few years from now, SI is a better organization and more effective in its mission.
3. SI can't or won't make changes, and SI's supporters feel my objections are valid, so SI loses some support, freeing up resources for other approaches to doing good.

Which one of these occurs will hopefully be driven primarily by the merits of the different arguments raised. Because of this, I think that whatever happens as a result of my post will be positive for SI's mission, whether or not it is positive for SI as an organization. I believe that most of SI's supporters and advocates care more about the former than about the latter, and that this attitude is far too rare in the nonprofit world.

Does SI have a well-argued case that its work is beneficial and important?

I know no more concise summary of SI's views than this page, so here I give my own impressions of what SI believes, in italics.

1. *There is some chance that in the near future (next 20-100 years), an "artificial general intelligence" (AGI) - a computer that is vastly more intelligent than humans in every relevant way - will be created.*
2. *This AGI will likely have a utility function and will seek to maximize utility according to this function.*
3. *This AGI will be so much more powerful than humans - due to its superior intelligence - that it will be able to reshape the world to maximize its utility, and humans will not be able to stop it from doing so.*
4. *Therefore, it is crucial that its utility function be one that is reasonably harmonious with what humans want. A "Friendly" utility function is one that is*

reasonably harmonious with what humans want, such that a "Friendly" AGI (FAI) would change the world for the better (by human standards) while an "Unfriendly" AGI (UFAI) would essentially wipe out humanity (or worse).

5. *Unless great care is taken specifically to make a utility function "Friendly," it will be "Unfriendly," since the things humans value are a tiny subset of the things that are possible.*
6. *Therefore, it is crucially important to develop "Friendliness theory" that helps us to ensure that the first strong AGI's utility function will be "Friendly." The developer of Friendliness theory could use it to build an FAI directly or could disseminate the theory so that others working on AGI are more likely to build FAI as opposed to UFAI.*

From the time I first heard this argument, it has seemed to me to be skipping important steps and making major unjustified assumptions. However, for a long time I believed this could easily be due to my inferior understanding of the relevant issues. I believed my own views on the argument to have only very low relevance (as I stated in my 2011 interview with SI representatives). Over time, I have had many discussions with SI supporters and advocates, as well as with non-supporters who I believe understand the relevant issues well. I now believe - for the moment - that my objections are highly relevant, that they cannot be dismissed as simple "layman's misunderstandings" (as they have been by various SI supporters in the past), and that SI has not published anything that addresses them in a clear way.

Below, I list my major objections. I do not believe that these objections constitute a sharp/tight case for the idea that SI's work has low/negative value; I believe, instead, that SI's own arguments are too vague for such a rebuttal to be possible. There are many possible responses to my objections, but SI's public arguments (and the private arguments) do not make clear which possible response (if any) SI would choose to take up and defend. Hopefully the dialogue following this post will clarify what SI believes and why.

Some of my views are discussed at greater length (though with less clarity) in a public transcript of a conversation I had with SI supporter Jaan Tallinn. I refer to this transcript as "Karnofsky/Tallinn 2011."

Objection 1: it seems to me that any AGI that was set to maximize a "Friendly" utility function would be extraordinarily dangerous.

Suppose, for the sake of argument, that SI manages to create what it believes to be an FAI. Suppose that it is successful in the "AGI" part of its goal, i.e., it has successfully created an intelligence vastly superior to human intelligence and extraordinarily powerful from our perspective. Suppose that it has also done its best on the "Friendly" part of the goal: it has developed a formal argument for why its AGI's utility function will be Friendly, it believes this argument to be airtight, and it has had this argument checked over by 100 of the world's most intelligent and relevantly experienced people. Suppose that SI now activates its AGI, unleashing it to reshape the world as it sees fit. What will be the outcome?

I believe that the probability of an unfavorable outcome - by which I mean an outcome essentially equivalent to what a UFAI would bring about - exceeds 90% in such a scenario. I believe the goal of designing a "Friendly" utility function is likely to be beyond the abilities even of the best team of humans willing to design such a

function. I do not have a tight argument for why I believe this, but a comment on LessWrong by Wei Dai gives a good illustration of the kind of thoughts I have on the matter:

> What I'm afraid of is that a design will be shown to be safe, and then it turns out that the proof is wrong, or the formalization of the notion of "safety" used by the proof is wrong. This kind of thing happens *a lot* in cryptography, if you replace "safety" with "security". These mistakes are still occurring today, even after decades of research into how to do such proofs and what the relevant formalizations are. From where I'm sitting, proving an AGI design Friendly seems even more difficult and error-prone than proving a crypto scheme secure, probably by a large margin, and there is no decades of time to refine the proof techniques and formalizations. There's good recent review of the history of provable security, titled Provable Security in the Real World, which might help you understand where I'm coming from.

I think this comment understates the risks, however. For example, when the comment says "the formalization of the notion of 'safety' used by the proof is wrong," it is not clear whether it means that the values the programmers have in mind are not correctly implemented by the formalization, or whether it means they are correctly implemented but are themselves catastrophic in a way that hasn't been anticipated. I would be highly concerned about both. There are other catastrophic possibilities as well; perhaps the utility function itself is well-specified and safe, but the AGI's model of the world is flawed (in particular, perhaps its prior or its process for matching observations to predictions are flawed) in a way that doesn't emerge until the AGI has made substantial changes to its environment.

By SI's own arguments, even a small error in any of these things would likely lead to catastrophe. And there are likely failure forms I haven't thought of. The overriding intuition here is that complex plans usually fail when unaccompanied by feedback loops. A scenario in which a set of people is ready to unleash an all-powerful being to maximize some parameter in the world, based solely on their initial confidence in their own extrapolations of the consequences of doing so, seems like a scenario that is overwhelmingly likely to result in a bad outcome. It comes down to placing the world's largest bet on a highly complex theory - with no experimentation to test the theory first.

So far, all I have argued is that the development of "Friendliness" theory can achieve at best only a limited reduction in the probability of an unfavorable outcome. However, as I argue in the next section, I believe there is at least one concept - the "tool-agent" distinction - that has more potential to reduce risks, and that SI appears to ignore this concept entirely. I believe that tools are safer than agents (even agents that make use of the best "Friendliness" theory that can reasonably be hoped for) and that SI encourages a focus on building agents, thus increasing risk.

Objection 2: SI appears to neglect the potentially important distinction between "tool" and "agent" AI.

Google Maps is a type of artificial intelligence (AI). It is far more intelligent than I am when it comes to planning routes.

Google Maps - by which I mean the complete software package including the display of the map itself - does not have a "utility" that it seeks to maximize. (One could fit a

utility function to its actions, as to any set of actions, but there is no single "parameter to be maximized" driving its operations.)

Google Maps (as I understand it) considers multiple possible routes, gives each a score based on factors such as distance and likely traffic, and then displays the best-scoring route in a way that makes it easily understood by the user. If I don't like the route, for whatever reason, I can change some parameters and consider a different route. If I like the route, I can print it out or email it to a friend or send it to my phone's navigation application. Google Maps has no single parameter it is trying to maximize; it has no reason to try to "trick" me in order to increase its utility.

In short, Google Maps is not an *agent*, taking actions in order to maximize a utility parameter. It is a *tool*, generating information and then displaying it in a user-friendly manner for me to consider, use and export or discard as I wish.

Every software application I know of seems to work essentially the same way, including those that involve (specialized) artificial intelligence such as Google Search, Siri, Watson, Rybka, etc. Some can be put into an "agent mode" (as Watson was on Jeopardy!) but all can easily be set up to be used as "tools" (for example, Watson can simply display its top candidate answers to a question, with the score for each, without speaking any of them.)

The "tool mode" concept is importantly different from the possibility of Oracle AI sometimes discussed by SI. The discussions I've seen of Oracle AI present it as an Unfriendly AI that is "trapped in a box" - an AI whose intelligence is driven by an explicit utility function and that humans hope to control coercively. Hence the discussion of ideas such as the AI-Box Experiment. A different interpretation, given in Karnofsky/Tallinn 2011, is an AI with a carefully designed utility function - likely as difficult to construct as "Friendliness" - that leaves it "wishing" to answer questions helpfully. By contrast with both these ideas, Tool-AGI is not "trapped" and it is not Unfriendly or Friendly; it has no motivations and no driving utility function of any kind, just like Google Maps. It scores different possibilities and displays its conclusions in a transparent and user-friendly manner, as its instructions say to do; it does not have an overarching "want," and so, as with the specialized AIs described above, while it may sometimes "misinterpret" a question (thereby scoring options poorly and ranking the wrong one #1) there is no reason to expect intentional trickery or manipulation when it comes to displaying its results.

Another way of putting this is that a "tool" has an underlying instruction set that conceptually looks like: "(1) Calculate which action A would maximize parameter P, based on existing data set D. (2) Summarize this calculation in a user-friendly manner, including what Action A is, what likely intermediate outcomes it would cause, what other actions would result in high values of P, etc." An "agent," by contrast, has an underlying instruction set that conceptually looks like: "(1) Calculate which action, A, would maximize parameter P, based on existing data set D. (2) Execute Action A." In any AI where (1) is separable (by the programmers) as a distinct step, (2) can be set to the "tool" version rather than the "agent" version, and this separability is in fact present with most/all modern software. Note that in the "tool" version, neither step (1) nor step (2) (nor the combination) constitutes an instruction to maximize a parameter - to describe a program of this kind as "wanting" something is a category error, and there is no reason to expect its step (2) to be deceptive.

I elaborated further on the distinction and on the concept of a tool-AI in Karnofsky/Tallinn 2011.

This is important because **an AGI running in tool mode could be extraordinarily useful but far more safe than an AGI running in agent mode**. In fact, if developing "Friendly AI" is what we seek, a tool-AGI could likely be helpful enough in thinking through this problem as to render any previous work on "Friendliness theory" moot. Among other things, a tool-AGI would allow transparent views into the AGI's reasoning and predictions without any reason to fear being purposefully misled, and would facilitate safe experimental testing of any utility function that one wished to eventually plug into an "agent."

Is a tool-AGI possible? I believe that it is, and furthermore that it ought to be our default picture of how AGI will work, given that practically all software developed to date can (and usually does) run as a tool and given that modern software seems to be constantly becoming "intelligent" (capable of giving better answers than a human) in surprising new domains. In addition, it intuitively seems to me (though I am not highly confident) that intelligence inherently involves the distinct, separable steps of (a) considering multiple possible actions and (b) assigning a score to each, *prior* to executing any of the possible actions. If one can distinctly separate (a) and (b) in a program's code, then one can abstain from writing any "execution" instructions and instead focus on making the program list actions and scores in a user-friendly manner, for humans to consider and use as they wish.

Of course, there are possible paths to AGI that may rule out a "tool mode," but it seems that most of these paths would rule out the application of "Friendliness theory" as well. (For example, a "black box" emulation and augmentation of a human mind.) What are the paths to AGI that allow manual, transparent, intentional design of a utility function but do not allow the replacement of "execution" instructions with "communication" instructions? Most of the conversations I've had on this topic have focused on three responses:

- **Self-improving AI.** Many seem to find it intuitive that (a) AGI will almost certainly come from an AI rewriting its own source code, and (b) such a process would inevitably lead to an "agent." I do not agree with either (a) or (b). I discussed these issues in Karnofsky/Tallinn 2011 and will be happy to discuss them more if this is the line of response that SI ends up pursuing. Very briefly:
  - The idea of a "self-improving algorithm" intuitively sounds very powerful, but does not seem to have led to many "explosions" in software so far (and it seems to be a concept that could apply to narrow AI as well as to AGI).
  - It seems to me that a tool-AGI could be plugged into a self-improvement process that would be quite powerful but would also terminate and yield a new tool-AI after a set number of iterations (or after reaching a set "intelligence threshold"). So I do not accept the argument that "self-improving AGI means agent AGI." As stated above, I will elaborate on this view if it turns out to be an important point of disagreement.
  - I have argued (in Karnofsky/Tallinn 2011) that the relevant self-improvement abilities are likely to come *with* or *after* - not *prior to* - the development of strong AGI. In other words, any software capable of the relevant kind of self-improvement is likely also capable of being used as a strong tool-AGI, with the benefits described above.
  - The SI-related discussions I've seen of "self-improving AI" are highly vague, and do not spell out views on the above points.
- **Dangerous data collection.** Some point to the seeming dangers of a tool-AI's "scoring" function: in order to score different options it may have to collect data, which is itself an "agent" type action that could lead to dangerous actions. I think my definition of "tool" above makes clear what is wrong with this objection:

a tool-AGI takes its existing data set D as fixed (and perhaps could have some pre-determined, safe set of simple actions it can take - such as using Google's API - to collect more), and if maximizing its chosen parameter is best accomplished through more data collection, it can transparently output why and how it suggests collecting more data. Over time it can be given more autonomy *for data collection* through an *experimental and domain-specific process* (e.g., modifying the AI to skip specific steps of human review of proposals for data collection after it has become clear that these steps work as intended), a process that has little to do with the "Friendly overarching utility function" concept promoted by SI. Again, I will elaborate on this if it turns out to be a key point.

- **Race for power.** Some have argued to me that humans are likely to *choose* to create agent-AGI, in order to quickly gain power and outrace other teams working on AGI. But this argument, even if accepted, has very different implications from SI's view.

  Conventional wisdom says it is extremely dangerous to empower a computer to act in the world until one is very sure that the computer will do its job in a way that is helpful rather than harmful. So if a programmer chooses to "unleash an AGI as an agent" with the hope of gaining power, it seems that this programmer will be deliberately ignoring conventional wisdom about what is safe in favor of shortsighted greed. I do not see why such a programmer would be expected to make use of any "Friendliness theory" that might be available. (Attempting to incorporate such theory would almost certainly slow the project down greatly, and thus would bring the same problems as the more general "have caution, do testing" counseled by conventional wisdom.) It seems that the appropriate measures for preventing such a risk are security measures aiming to stop humans from launching unsafe agent-AIs, rather than developing theories or raising awareness of "Friendliness."

One of the things that bothers me most about SI is that there is practically no public content, as far as I can tell, explicitly addressing the idea of a "tool" and giving arguments for why AGI is likely to work only as an "agent." The idea that AGI will be driven by a central utility function seems to be simply assumed. Two examples:

- I have been referred to Muehlhauser and Salamon 2012 as the most up-to-date, clear explanation of SI's position on "the basics." This paper states, "Perhaps we could build an AI of limited cognitive ability — say, a machine that only answers questions: an 'Oracle AI.' But this approach is not without its own dangers (Armstrong, Sandberg, and Bostrom 2012)." However, the referenced paper (Armstrong, Sandberg and Bostrom 2012) seems to take it as a given that an Oracle AI is an "agent trapped in a box" - a computer that has a basic drive/utility function, not a Tool-AGI. The rest of Muehlhauser and Salamon 2012 seems to take it as a given that an AGI will be an agent.
- I have often been referred to Omohundro 2008 for an argument that an AGI is likely to have certain goals. But this paper seems, again, to take it as given that an AGI will be an agent, i.e., that it will have goals at all. The introduction states, "To say that a system of any design is an 'artificial intelligence', we mean that it has goals which it tries to accomplish by acting in the world." In other words, the premise I'm disputing seems embedded in its very definition of AI.

The closest thing I have seen to a public discussion of "tool-AGI" is in Dreams of Friendliness, where Eliezer Yudkowsky considers the question, "Why not just have the AI answer questions, instead of trying to *do* anything? Then it wouldn't need to be

Friendly. It wouldn't need any goals at all. It would just answer questions." His response:

> To which the reply is that the AI needs goals in order to decide how to think: that is, the AI has to act as a powerful optimization process in order to plan its acquisition of knowledge, effectively distill sensory information, pluck "answers" to particular questions out of the space of all possible responses, and of course, to improve its own source code up to the level where the AI is a powerful intelligence. All these events are "improbable" relative to random organizations of the AI's RAM, so the AI has to hit a narrow target in the space of possibilities to make superintelligent answers come out.

This passage appears vague and does not appear to address the specific "tool" concept I have defended above (in particular, it does not address the analogy to modern software, which challenges the idea that "powerful optimization processes" cannot run in tool mode). The rest of the piece discusses (a) psychological mistakes that could lead to the discussion in question; (b) the "Oracle AI" concept that I have outlined above. The comments contain some more discussion of the "tool" idea (Denis Bider and Shane Legg seem to be picturing something similar to "tool-AGI") but the discussion is unresolved and I believe the "tool" concept defended above remains essentially unaddressed.

In sum, SI appears to encourage a focus on building and launching "Friendly" agents (it is seeking to do so itself, and its work on "Friendliness" theory seems to be laying the groundwork for others to do so) while not addressing the tool-agent distinction. It seems to assume that any AGI will have to be an agent, and to make little to no attempt at justifying this assumption. The result, in my view, is that it is essentially advocating for a more dangerous approach to AI than the traditional approach to software development.

Objection 3: SI's envisioned scenario is far more specific and conjunctive than it appears at first glance, and I believe this scenario to be highly unlikely.

SI's scenario concerns the development of artificial *general* intelligence (AGI): a computer that is vastly more intelligent than humans in every relevant way. But we already have many computers that are vastly more intelligent than humans in *some* relevant ways, and the domains in which specialized AIs outdo humans seem to be constantly and continuously expanding. I feel that the relevance of "Friendliness theory" depends heavily on the idea of a "discrete jump" that seems unlikely and whose likelihood does not seem to have been publicly argued for.

One possible scenario is that at some point, we develop powerful enough non-AGI tools (particularly specialized AIs) that we vastly improve our abilities to consider and prepare for the eventuality of AGI - to the point where any previous theory developed on the subject becomes useless. Or (to put this more generally) non-AGI tools simply change the world so much that it becomes essentially unrecognizable from the perspective of today - again rendering any previous "Friendliness theory" moot. As I said in Karnofsky/Tallinn 2011, some of SI's work "seems a bit like trying to design Facebook before the Internet was in use, or even before the computer existed."

Perhaps there will be a discrete jump to AGI, but it will be a sort of AGI that renders "Friendliness theory" moot for a different reason. For example, *in the practice of*

software development, there often does not seem to be an operational distinction between "intelligent" and "Friendly." (For example, my impression is that the only method programmers had for evaluating Watson's "intelligence" was to see whether it was coming up with the same answers that a well-informed human would; the only way to evaluate Siri's "intelligence" was to evaluate its helpfulness to humans.) "Intelligent" often ends up getting defined as "prone to take actions that seem all-around 'good' to the programmer." So the concept of "Friendliness" may end up being naturally and subtly baked in to a successful AGI effort.

The bottom line is that we know very little about the course of future artificial intelligence. I believe that the probability that SI's concept of "Friendly" vs. "Unfriendly" goals ends up seeming essentially nonsensical, irrelevant and/or unimportant from the standpoint of the relevant future is over 90%.

Other objections to SI's views

There are other debates about the likelihood of SI's work being relevant/helpful; for example,

- It isn't clear whether the development of AGI is imminent enough to be relevant, or whether other risks to humanity are closer.
- It isn't clear whether AGI would be as powerful as SI's views imply. (I discussed this briefly in Karnofsky/Tallinn 2011.)
- It isn't clear whether even an extremely powerful UFAI would choose to attack humans as opposed to negotiating with them. (I find it somewhat helpful to analogize UFAI-human interactions to human-mosquito interactions. Humans are enormously more intelligent than mosquitoes; humans are good at predicting, manipulating, and destroying mosquitoes; humans do not value mosquitoes' welfare; humans have other goals that mosquitoes interfere with; humans would like to see mosquitoes eradicated at least from certain parts of the planet. Yet humans haven't accomplished such eradication, and it is easy to imagine scenarios in which humans would prefer honest negotiation and trade with mosquitoes to any other arrangement, if such negotiation and trade were possible.)

Unlike the three objections I focus on, these other issues have been discussed a fair amount, and if these other issues were the only objections to SI's arguments I would find SI's case to be strong (i.e., I would find its scenario likely *enough* to warrant investment in).

Wrapup

- I believe the most likely future scenarios are the ones we haven't thought of, and that the most likely fate of the sort of theory SI ends up developing is irrelevance.
- I believe that unleashing an all-powerful "agent AGI" (without the benefit of experimentation) would very likely result in a UFAI-like outcome, no matter how carefully the "agent AGI" was designed to be "Friendly." I see SI as encouraging (and aiming to take) this approach.
- I believe that the standard approach to developing software results in "tools," not "agents," and that tools (while dangerous) are much safer than agents. A "tool mode" could facilitate *experiment-informed* progress toward a safe

"agent," rather than needing to get "Friendliness" theory right without any experimentation.
- Therefore, I believe that the approach SI advocates and aims to prepare for is far more dangerous than the standard approach, so *if* SI's work on Friendliness theory affects the risk of human extinction one way or the other, it will increase the risk of human extinction. Fortunately I believe SI's work is far more likely to have no effect one way or the other.

For a long time I refrained from engaging in object-level debates over SI's work, believing that others are better qualified to do so. But after talking at great length to many of SI's supporters and advocates and reading everything I've been pointed to as relevant, I still have seen no clear and compelling response to any of my three major objections. As stated above, there are many possible responses to my objections, but SI's current arguments do not seem clear on what responses they wish to take and defend. At this point I am unlikely to form a positive view of SI's work until and unless I do see such responses, and/or SI changes its positions.

Is SI the kind of organization we want to bet on?

This part of the post has some risks. For most of GiveWell's history, sticking to our standard criteria - and putting more energy into recommended than non-recommended organizations - has enabled us to share our honest thoughts about charities without appearing to get personal. But when evaluating a group such as SI, I can't avoid placing a heavy weight on (my read on) the general competence, capability and "intangibles" of the people and organization, because SI's mission is not about repeating activities that have worked in the past. **Sharing my views on these issues could strike some as personal or mean-spirited and could lead to the misimpression that GiveWell is hostile toward SI. But it is simply necessary in order to be fully transparent about why I hold the views that I hold.**

Fortunately, SI is an ideal organization for our first discussion of this type. I believe the staff and supporters of SI would overwhelmingly rather hear the whole truth about my thoughts - so that they can directly engage them and, if warranted, make changes - than have me sugar-coat what I think in order to spare their feelings. People who know me and my attitude toward being honest vs. sparing feelings know that this, itself, is high praise for SI.

One more comment before I continue: our policy is that non-public information provided to us by a charity will not be published or discussed without that charity's prior consent. However, none of the content of this post is based on private information; all of it is based on information that SI has made available to the public.

There are several reasons that I currently have a negative impression of SI's general competence, capability and "intangibles." My mind remains open and I include specifics on how it could be changed.

- **Weak arguments.** SI has produced enormous quantities of public argumentation, and I have examined a very large proportion of this information. Yet I have never seen a clear response to any of the three basic objections I listed in the previous section. One of SI's major goals is to raise awareness of AI-related risks; given this, the fact that it has not advanced

clear/concise/compelling arguments speaks, in my view, to its general competence.

- **Lack of impressive endorsements.** I discussed this issue in my 2011 interview with SI representatives and I still feel the same way on the matter. I feel that given the enormous implications of SI's claims, if it argued them well it ought to be able to get more impressive endorsements than it has.

  I have been pointed to Peter Thiel and Ray Kurzweil as examples of impressive SI supporters, but I have not seen any on-record statements from either of these people that show agreement with SI's specific views, and in fact (based on watching them speak at Singularity Summits) my impression is that they disagree. Peter Thiel seems to believe that speeding the pace of general innovation is a good thing; this would seem to be in tension with SI's view that AGI will be catastrophic by default and that no one other than SI is paying sufficient attention to "Friendliness" issues. Ray Kurzweil seems to believe that "safety" is a matter of transparency, strong institutions, etc. rather than of "Friendliness." I am personally in agreement with the things I have seen both of them say on these topics. I find it possible that they support SI because of the Singularity Summit or to increase general interest in ambitious technology, rather than because they find "Friendliness theory" to be as important as SI does.

  Clear, on-record statements from these two supporters, specifically endorsing SI's arguments and the importance of developing Friendliness theory, would shift my views somewhat on this point.

- **Resistance to feedback loops.** I discussed this issue in my 2011 interview with SI representatives and I still feel the same way on the matter. SI seems to have passed up opportunities to test itself and its own rationality by e.g. aiming for objectively impressive accomplishments. This is a problem because of (a) its extremely ambitious goals (among other things, it seeks to develop artificial intelligence *and* "Friendliness theory" before anyone else can develop artificial intelligence); (b) its view of its staff/supporters as having unusual insight into rationality, which I discuss in a later bullet point.

  SI's list of achievements is not, in my view, up to where it needs to be given (a) and (b). Yet I have seen no declaration that SI has fallen short to date and explanation of what will be changed to deal with it. SI's recent release of a strategic plan and monthly updates are improvements from a transparency perspective, but they still leave me feeling as though there are no clear metrics or goals by which SI is committing to be measured (aside from very basic organizational goals such as "design a new website" and very vague goals such as "publish more papers") and as though SI places a low priority on engaging people who are critical of its views (or at least not yet on board), as opposed to people who are naturally drawn to it.

  I believe that one of the primary obstacles to being impactful as a nonprofit is the lack of the sort of helpful feedback loops that lead to success in other domains. I like to see groups that are making as much effort as they can to create meaningful feedback loops for themselves. I perceive SI as falling well short on this front. Pursuing more impressive endorsements and developing benign but objectively recognizable innovations (particularly commercially viable ones) are two possible ways to impose more demanding feedback loops. (I discussed both of these in my interview linked above).

- **Apparent poorly grounded belief in SI's superior general rationality.** Many of the things that SI and its supporters and advocates say imply a belief that they have special insights into the nature of general rationality, and/or have superior general rationality, relative to the rest of the population. (Examples here, here and here). My understanding is that SI is in the process of spinning off a group dedicated to training people on how to have higher general rationality.

  Yet I'm not aware of any of what I consider compelling evidence that SI staff/supporters/advocates have any special insight into the nature of general rationality or that they have especially high general rationality.

  I have been pointed to the Sequences on this point. The Sequences (which I have read the vast majority of) do not seem to me to be a demonstration or evidence of general rationality. They are *about* rationality; I find them very enjoyable to read; and there is very little they say that I disagree with (or would have disagreed with before I read them). However, they do not seem to demonstrate rationality on the part of the writer, any more than a series of enjoyable, not-obviously-inaccurate essays on the qualities of a good basketball player would demonstrate basketball prowess. I sometimes get the impression that fans of the Sequences are willing to ascribe superior rationality to the writer simply because the content *seems smart and insightful to them*, without making a critical effort to determine the extent to which the content is novel, actionable and important.

  I endorse Eliezer Yudkowsky's statement, "Be careful … any time you find yourself defining the [rationalist] as someone other than the agent who is currently smiling from on top of a giant heap of utility." To me, the best evidence of superior general rationality (or of insight into it) would be objectively impressive achievements (successful commercial ventures, highly prestigious awards, clear innovations, etc.) and/or accumulation of wealth and power. As mentioned above, SI staff/supporters/advocates do not seem particularly impressive on these fronts, at least not as much as I would expect for people who have the sort of insight into rationality that makes it sensible for them to train others in it. I am open to other evidence that SI staff/supporters/advocates have superior general rationality, but I have not seen it.

  Why is it a problem if SI staff/supporter/advocates believe themselves, without good evidence, to have superior general rationality? First off, it strikes me as a belief based on wishful thinking rather than rational inference. Secondly, I would expect a series of problems to accompany overconfidence in one's general rationality, and several of these problems seem to be actually occurring in SI's case:

  - Insufficient self-skepticism given how strong its claims are and how little support its claims have won. Rather than endorsing "Others have not accepted our arguments, so we will sharpen and/or reexamine our arguments," SI seems often to endorse something more like "Others have not accepted their arguments because they have inferior general rationality," a stance less likely to lead to improvement on SI's part.
  - Being too selective (in terms of looking for people who share its preconceptions) when determining whom to hire and whose feedback to take seriously.
  - Paying insufficient attention to the limitations of the confidence one can have in one's untested theories, in line with my Objection 1.

- **Overall disconnect between SI's goals and its activities.** SI seeks to build FAI and/or to develop and promote "Friendliness theory" that can be useful to others in building FAI. Yet it seems that most of its time goes to activities other than developing AI or theory. Its per-person output in terms of publications seems low. Its core staff seem more focused on Less Wrong posts, "rationality training" and other activities that don't seem connected to the core goals; Eliezer Yudkowsky, in particular, appears (from the strategic plan) to be focused on writing books for popular consumption. These activities seem neither to be advancing the state of FAI-related theory nor to be engaging the sort of people most likely to be crucial for building AGI.

  A possible justification for these activities is that SI is seeking to promote greater general rationality, which over time will lead to more and better support for its mission. But if this is SI's core activity, it becomes even more important to test the hypothesis that SI's views are in fact rooted in superior general rationality - and these tests don't seem to be happening, as discussed above.

- **Theft.** I am bothered by the 2009 theft of $118,803.00 (as against a $541,080.00 budget for the year). In an organization as small as SI, it really seems as though theft that large relative to the budget shouldn't occur and that it represents a major failure of hiring and/or internal controls.

  In addition, I have seen no public SI-authorized discussion of the matter that I consider to be satisfactory in terms of explaining what happened and what the current status of the case is on an ongoing basis. Some details may have to be omitted, but a clear SI-authorized statement on this point with as much information as can reasonably provided would be helpful.

A couple positive observations to add context here:

- I see significant positive qualities in many of the people associated with SI. I especially like what I perceive as their sincere wish to do whatever they can to help the world as much as possible, and the high value they place on being right as opposed to being conventional or polite. I have not interacted with Eliezer Yudkowsky but I greatly enjoy his writings.
- I'm aware that SI has relatively new leadership that is attempting to address the issues behind some of my complaints. I have a generally positive impression of the new leadership; I believe the Executive Director and Development Director, in particular, to represent a step forward in terms of being interested in transparency and in testing their own general rationality. So I will not be surprised if there is some improvement in the coming years, particularly regarding the last couple of statements listed above. That said, SI is an organization and it seems reasonable to judge it by its organizational track record, especially when its new leadership is so new that I have little basis on which to judge these staff.

Wrapup

While SI has produced a lot of content that I find interesting and enjoyable, it has not produced what I consider evidence of superior general rationality or of its suitability for the tasks it has set for itself. I see no qualifications or achievements that specifically seem to indicate that SI staff are well-suited to the challenge of understanding the key AI-related issues and/or coordinating the construction of an FAI.

And I see specific reasons to be pessimistic about its suitability and general competence.

When estimating the expected value of an endeavor, it is natural to have an implicit "survivorship bias" - to use organizations whose accomplishments one is familiar with (which tend to be relatively effective organizations) as a reference class. Because of this, I would be extremely wary of investing in an organization with apparently poor general competence/suitability to its tasks, even if I bought fully into its mission (which I do not) and saw no other groups working on a comparable mission.

But if there's even a chance ...

A common argument that SI supporters raise with me is along the lines of, "Even if SI's arguments are weak and its staff isn't as capable as one would like to see, their goal is so important that they would be a good investment even at a tiny probability of success."

I believe this argument to be a form of Pascal's Mugging and I have outlined the reasons I believe it to be invalid in two posts (here and here). There have been some objections to my arguments, but I still believe them to be valid. There is a good chance I will revisit these topics in the future, because I believe these issues to be at the core of many of the differences between GiveWell-top-charities supporters and SI supporters.

Regardless of whether one accepts my specific arguments, it is worth noting that the most prominent people associated with SI tend to agree with the *conclusion* that the "But if there's even a chance ..." argument is not valid. (See comments on my post from Michael Vassar and Eliezer Yudkowsky as well as Eliezer's interview with John Baez.)

Existential risk reduction as a cause

I consider the general cause of "looking for ways that philanthropic dollars can reduce direct threats of global catastrophic risks, particularly those that involve some risk of human extinction" to be a relatively high-potential cause. It is on the working agenda for GiveWell Labs and we will be writing more about it.

However, I don't think that "Cause X is the one I care about and Organization Y is the only one working on it" to be a good reason to support Organization Y. For donors determined to donate within this cause, I encourage you to consider donating to a donor-advised fund while making it clear that you intend to grant out the funds to existential-risk-reduction-related organizations in the future. (One way to accomplish this would be to create a fund with "existential risk" in the name; this is a fairly easy thing to do and one person could do it on behalf of multiple donors.)

For one who accepts my arguments about SI, I believe withholding funds in this way is likely to be better for SI's mission than donating to SI - through incentive effects alone (not to mention my specific argument that SI's approach to "Friendliness" seems likely to increase risks).

How I might change my views

My views are very open to revision.

However, I cannot realistically commit to read and seriously consider all comments posted on the matter. The number of people capable of taking a few minutes to write a comment is sufficient to swamp my capacity. I do encourage people to comment and I do intend to read at least some comments, but if you are looking to change my views, you should not consider posting a comment to be the most promising route.

Instead, what I will commit to is reading and carefully considering **up to 50,000 words of content that are (a) specifically marked as SI-authorized responses to the points I have raised; (b) explicitly cleared for release to the general public as SI-authorized communications.** In order to consider a response "SI-authorized and cleared for release," I will accept explicit communication from SI's Executive Director or from a majority of its Board of Directors endorsing the content in question. After 50,000 words, I may change my views and/or commit to reading more content, or (if I determine that the content is poor and is not using my time efficiently) I may decide not to engage further. SI-authorized content may improve or worsen SI's standing in my estimation, so unlike with comments, there is an incentive to select content that uses my time efficiently. Of course, SI-authorized content may end up including excerpts from comment responses to this post, and/or already-existing public content.

I may also change my views for other reasons, particularly if SI secures more impressive achievements and/or endorsements.

One more note: I believe I have read the vast majority of the Sequences, including the AI-foom debate, and that this content - while interesting and enjoyable - does not have much relevance for the arguments I've made.

Again: I think that whatever happens as a result of my post will be positive for SI's mission, whether or not it is positive for SI as an organization. I believe that most of SI's supporters and advocates care more about the former than about the latter, and that this attitude is far too rare in the nonprofit world.

Acknowledgements

Thanks to the following people for reviewing a draft of this post and providing thoughtful feedback (this of course does not mean they agree with the post or are responsible for its content): Dario Amodei, Nick Beckstead, Elie Hassenfeld, Alexander Kruel, Tim Ogden, John Salvatier, Jonah Sinick, Cari Tuna, Stephanie Wykstra.

Punctuality - Arriving on Time and Math

In hindsight, this post seems incredibly obvious. The meat of it already exists in sayings which we all know we ought to listen to: "*Always arrive 10 minutes earlier than you think early is*," "*If you arrive on time, then you're late*," or "*Better three hours too soon than one minute too late*." Yet even with these sayings, I still never trusted them nor arrived on time. I'd miss deadlines, show up late, and just be generally tardy. The reason is that I never truly understood what it took to arrive on time until I grokked the math of it. So, while this may be remedial reading for most of you, I'm posting this because maybe there's someone out there who missed the same obviousness that I missed.

Statistical Distributions

Everyone here understands that our universe is controlled and explained by math. Math describes how heavenly bodies move. Math describes how our computers run. Math describes how other people act in aggregate. Wait a second, something's not right with that statement... "*other people*". The way it comes out it's natural to think that math controls the way that *other people* act, and not myself. Intellectually, I am aware that I am not a special snowflake who is exempt from the laws of math. While I had managed to propagate this thought far enough to crush my belief in libertarian free will, I hadn't propagated it fully through my mind. Specifically, I hadn't realized I could also use math to describe my actions and reap the benefit of understanding them mathematically. I was still late to arrive and missing deadlines, and nothing seemed to help.

But wait, I'm a rationalist! I know all about the planning fallacy; I know to take the outside view! That's enough to save me right? Well, not quite. It seemed I missed one last part of the puzzle... *Bell Curves*.

When I go to work every day, the time from when I do nothing but getting ready to go to work until the time that I actually arrive there (I'll just call this prep time) usually takes 45 minutes, but sometimes it can take more time or less time. Weirdly and crazily enough, if you plot all the prep times on a graph, the shape would end up looking roughly *like a bell*. Well that's funny. Math is for *other people*, but my behavior appears like it can be described statistically. Some days I will have deviations from the normal routine that help me arrive faster while other days will have things that slow me down. Some of them happen more often, some of them happen less often. *If* I were describable by math, I could almost call these things standard deviations: days where I have almost zero traffic prep time takes 1 standard deviation less, days when I can't find my car keys my prep time takes 1 standard deviation more, days I realize would be late and skip showering take 2 standard deviations less, and days when there is a terrible accident on the freeway end up

requiring +2 or +3 standard deviations more in time. To put it in other words, my prep time is a bell curve, and I've got 1-sigma and 2-sigma (and occasionally 3-sigma) events speeding me up and slowing me down.

This holds true for more than just going to work. Everything's time-until-completion can be described this way: project completion times, homework, going to the airport, the duration of foreplay and sex. *Everything*. It's not always bell curves, but it's a probability distribution with respect to completion times, and that can help give useful insights.

Starting 'On Time' Means You Won't be On Time

What do we gain by understanding that our actions are described by a probability distribution? The first and most important take away is this: If you only allocate the exact amount of time to do something, you'll be late 50% of the time. I'm going to repeat it and italicize because I think it's that important of a point. *If you only allocate the exact amount of time to do something, you'll be late 50% of the time.* That's the way bell curves work.

I know I've heard jokes about how 90% of the population has above average children, but it wasn't until I really looked at the math of my behavior that I realized I was doing the exact same thing. I'd say "oh it takes me 45 minutes on average to go to work every day, so I'll leave at 7:15." Yet I never realized that I was completely ignoring that *half* the time would take longer than average. So half the time, I'd end up be pressed for time and have to skip shaving (or something) or I'd end up late. I was terribly unpunctual until I realized I that I had to arrive early to always arrive on time. "*If you arrive on time, then you are late.*" Hmm. You win this one, folk wisdom.

Still, the question remained. How much early would it take to never be late? The answer lay in bell curves.

Acceptable Lateness and Standard deviation

Looking at time requirements as a bell curve implies another thing: One can never completely eliminate all lateness; the only option is to make a choice about what probability of lateness is acceptable. A person must decide what lateness ratio they're willing to take, and then start prepping that many standard deviations beforehand. And, despite what employers say, 0% is not a probability.

If my prep time averages 45 minutes with a standard deviation of 10 minutes then that means...

- Starting 45 minutes beforehand will force me to be late or miss services (eg shaving) around **50%** of the time or **about 10 workdays a month**.
- Starting 55 minutes beforehand will force me to be late or miss services (eg shaving) around **16%** of the time or **about 3 workdays a month**.
- Starting 65 minutes beforehand will force me to be late or miss services (eg shaving) around **2.3%** of the time or **about 1 day every other month**.

That's really good risk reduction for a small amount of time spent. (NB, remember that averages are dangerous little things. Taking this to a meta level, consider that being late to work ***about*** 3 times a month isn't helpful if you arrive late only once the first month, then get fired the next month when you arrive late 5 times. Hence, "*Always arrive 10 minutes earlier than you think early is.*" God I hate folk wisdom, especially when it's right.)

The risk level you're acceptable with dictates how much time you need for padding. For job interviews, I'm only willing to arrive late to 1 in 1000, so I prepare 3 standard deviations early now. For first dates, I'm willing to miss about 5%. For dinners with the family, I'm okay with being late half the time. It feels *similar* to the algorithm I used before, which was a sort of ad-hoc thing where I'd prepared earlier for important things. The main difference is that now I can quantify the risk I'm assuming when I procrastinate. It causes each procrastination to become more concrete for me, and drastically reduces the chance that I'll be willing to make those tradeoffs. Instead of being willing to read lesswrong for 10 more minutes in exchange for "oh I might have to rush", I can now see that it would increase my chance of being late from 16% to 50%, which is flatly unacceptable. Viewing procrastination in terms of the latter tradeoff makes it much easier to get myself moving.

The last quote is "*Better three hours too soon than one minute too late*." I'm glad that at least that one's wrong. I'm sure Umesh would have some stern words for that saying. My key to arriving on time is locating your acceptable risk threshold and making an informed decision about how much risk you are willing to take.

Summary

The time it takes for you to complete any task is (usually) described by a bell curve. How much time you think you'll take is a lie, and not just because of the planning fallacy. Even if you do the sciency-thing and take the outside view, it's still not enough to keep you from getting fired or showing up to your interview late. To consistently show up on time, you *must* incorporate padding time.

So I've got a new saying, "*If you wish to be late only 2.3% of the time, you must start getting ready at least two standard deviations before the average prep time you have needed historically.*" I wish my mom would have told me this one. It's so much easier to understand than all those other sayings!

(Also my first actual article-thingy, so any comments or suggestions is welcome)

The rational rationalist's guide to rationally using "rational" in rational post titles

1. Don't.

Avoid inflationary use of terms

Inflationary terms! You see them everywhere. And for those who actually know and care about the subject matter they can be very frustrating. These terms are notorious for being used in contexts where:

1. They are only loosely applicable at best.
2. There exists a better word that is more specific.
3. The topic has a far bias.

Some examples:

- Rational
- Evolution
- Singularity
- Emergent
- Nanotech
- Cryogenics
- Faith

The problem is *not* that these words are meaningless in their original form, *nor* that you shouldn't ever use them. The problem is that they often get used in stupid ways that make them much less meaningful. By that I mean, less useful for keeping a focus on the topic and understanding what the person is really talking about.

For example, terms like Nanotech (or worse, "Nanobot") *do apply* in a certain *loose* sense to several kinds of chemistry and biological innovations that are currently in vogue. Nonetheless, each time the term is used to refer to these things it makes it much harder to know if you are referring to Drexlerian Mechanosynthesis. Hint: If you get your grant money by convincing someone you are working on one thing whereas you are really working on something completely different, that's fraud.

Similarly, Cryogenics is the science of keeping things really cold. And of course *Cryonics* is a form of that. But saying "Cryogenics" when you really mean exactly Cryonics is an incredibly harmful practice which actual Cryonicists generally avoid. Most people who work in Cryogenics have nothing to do with Cryonics, and this kind of confusion in popular culture has apparently engendered animosity towards Cryonics among Cryogenics specialists.

Recently I fell prey to something like this with respect to the term "Rational". I wanted to know in general terms what the best programming language for a newbie would be and why. I wanted some in depth analysis, from a group I trust to do so. (And I wasn't disappointed -- we have some very knowledgeable programmers whose opinions were most helpful to me.) However the reaction of some lesswrongers to the title I initially chose for the post was distinctly negative. The title was "Most rational programming language?"

After thinking about it for a while I realized what the problem was: This way of using the term, despite being more or less valid, makes the term less meaningful in the long run. And I *don't* want to be the person who makes Rational a less meaningful word. Nobody here wants that to happen. Thus it would have been better to use a term such as "Best" or "Most optimal" instead.

Another example that comes to mind is when people (usually outsiders) refer to Transhumanism, Bayeseanism, the Singularity, or even skepticism, as a "Faith" or "Belief". Well yeah, trivially, if you are willing to stretch that word to its broadest possible meaning you can feel free to apply it to such as us. But... for crying out loud! What meaning does the word have if Faith is something absolutely everyone has? We're really referring to something like "Confidence" here.

Then there's Evolution. Is Transhumanism really about the next stage in human Evolution? Perhaps in a certain *loose* sense it is -- but let's not lose sight of the mutilation of the language (and consequent noise-to-signal increase) that occurs when you say such a thing. Human Evolution is an *existing scientific specialty* with absolutely zilch to do with cybernetic body modification or genetic engineering, and everything to do with the effects of natural selection and mutation on the development of humans in the past.

Co-opting terms isn't *always* bad. If you are brand-new to a topic, seeing an analogy to something with which you are already familiar may reduce the inferential distance and help you click the idea in your brain. But this gets more hazardous the closer the terms actually are in meaning. Distant terms are safer -- when I say "Avoid *inflationary* use of terms" you can instantly see that I'm definitely not talking about money, nor rubber objects with compressed air inside of them, but about words and phrases.

On the other hand with such things as Rational versus Optimal, we're taking two surface-level-similar words and blurring them in such a way that one cannot meaningfully talk about either without accidentally importing baggage from the other. Rational is more suitable for use in *contrast* with clear examples of irrationality -- cognitive biases, for example, or drug addiction, and is a rather unabashedly idealistic term. Optimal on the other hand doesn't so much require specific contrast because pretty much everything is suboptimal *by default* to some degree or another -- optimizing is understood as an ongoing and very relativistic process.

To sum up: Avoid making words cheaper and less effective for their specialized tasks. Don't use them for things where a better and more appropriate term exists. As your brain gets used to an idea, be prepared to discard old terms you have co-opted from other domains that were really just useful placeholders to get you started. Specialized jargon exists for a reason!

When None Dare Urge Restraint, pt. 2

In the original When None Dare Urge Restraint post, Eliezer discusses the dangers of the "spiral of hate" that can develop when saying negative things about the Hated Enemy trumps saying accurate things. Specifically, he uses the example of how the 9/11 hijackers were widely criticized as "cowards," even though this vice in particular was surely not on their list. Over this past Memorial Day weekend, however, it seems like the exact mirror-image problem played out in nearly textbook form.

The trouble began when MSNBC host Chris Hayes noted* that he was uncomfortable with how people use the word "hero" to describe those who die in war -- in particular, because he thinks this sort of automatic valor attributed to the war dead makes it easier to justify future wars. And as you might expect, people went crazy in response, calling Hayes's comments "reprehensible and disgusting," something that "commie grad students would say," and that old chestnut, apparently offered without a hint of irony, "unAmerican." If you watch the video, you can tell that Hayes himself is really struggling to make the point, and by the end he definitely knew he was going to get in trouble, as he started backpedaling with a "but maybe I'm wrong about that." And of course, he apologized the very next day, basically stating that it was improper to have "opine[d] about the people who fight our wars, having never dodged a bullet or guarded a post or walked a mile in their boots."

This whole episode struck me as particularly frightening, mostly because Hayes *wasn't even offering a criticism*. Soldiers in the American military are, of course, an untouchable target, and I would hardly expect any attack on soldiers to be well received, no matter how grounded. But what genuinely surprised me in this case was that Hayes was merely saying "let's not *automatically* apply the single most valorizing word we have, because that might cause future wars, and thus future war deaths." But apparently anything less than maximum praise was not only incorrect, but offensive.

Of course, there's no shortage of rationality failures in political discourse, and I'm obviously not intending this post as a political statement about any particular war, policy, candidate, etc. But I think this example is worth mentioning, for two main reasons. First, it's just such a textbook example of the exact sort of problem discussed in Eliezer's original post, in a purer form than I can recall seeing since 9/11 itself. I don't imagine many LW members need convincing in this regard, but I do think there's value in being mindful of this sort of problem on the national stage, even if we're not going to start arguing politics ourselves.

But second, I think this episode says something not just about nationalism, but about how people approach death more generally. Of course, we're all familiar with afterlifism/"they're-in-a-better-place"-style rationalizations of death, but labeling a death as "heroic" can be a similar sort of rationalization. If a death is "heroic," then there's at least some kind of silver lining, some sense of justification, if only partial justification. The movie might not be happy, but it can still go on, and there's at least a chance to play inspiring music. So there's an obvious temptation to label death as "heroic" as much as possible -- I'm reminded of how people tried to call the 9/11 victims "heroes," apparently because they had the great courage to work in buildings that were targeted in a terrorist attack.

If a death is *just* a tragedy, however, you're left with a more painful situation. You have to acknowledge that *yes, really*, the world isn't fair, and *yes, really*, thousands of people -- even the Good Guy's soldiers! -- might be dying for *no good reason at all*. And even for those who don't really believe in an afterlife, facing death on such a large scale without the "heroic" modifier might just be too painful. The obvious problem, of course -- and Hayes's original point -- is that this sort of death-anesthetic makes it all too easy to numb yourself to more death. If you really care about the problem, you have to *face* the sheer tragedy of it. Sometimes, all you can say is "we shall have to work faster." And I think that lesson's as appropriate on Memorial Day as any other.

*I apologize that this clip is inserted into a rather low-brow attack video. At the time of posting it was the only link on Youtube I could find, and I wanted something accessible.

Value of Information: 8 examples

ciphergoth just asked what the actual value of Quantified Self/self-experimentation is. This finally tempted me into running value of information calculations on my own experiments. It took me all afternoon because it turned out I didn't actually understand how to do it and I had a hard time figuring out the right values for specific experiments. (I may not have not gotten it right, still. Feel free to check my work!) Then it turned out to be too long for a comment, and as usual the master versions will be on my website at some point. But without further ado!

The value of an experiment is the information it produces. What is the value of information? Well, we can take the economic tack and say value of information is the value of the decisions it changes. (Would you pay for a weather forecast about somewhere you are not going to? No. Or a weather forecast about your trip where you *have* to make that trip, come hell or high water? Only to the extent you can make preparations like bringing an umbrella.)

Wikipedia says that for a risk-neutral person, value of perfect information is "value of decision situation with perfect information" - "value of current decision situation". (Imperfect information is just weakened perfect information: if your information was not 100% reliable but 99% reliable, well, that's worth 99% as much.)

1 Melatonin

`http://www.gwern.net/Zeo#melatonin` & http://www.gwern.net/Melatonin

The decision is the binary take or not take. Melatonin costs ~$10 a year (if you buy in bulk during sales, as I did). Suppose I had perfect information it worked; I would not change anything, so the value is $0. Suppose I had perfect information it did not work; then I would stop using it, saving me $10 a year in perpetuity, which has a net present value (at 5% discounting) of $205. So the value of perfect information is $205, because it would save me from blowing $10 every year for the rest of my life. My melatonin experiment is not perfect since I didn't randomize or double-blind it, but I had a lot of data and it was well powered, with something like a >90% chance of detecting the decent effect size I expected, so the imperfection is just a loss of 10%, down to $184. From my previous research and personal use over years, I am highly confident it works - say, 80%. If it works, the information is useless to me, and if it doesn't, I save $184; what's the expected value of obtaining the information, giving these two outcomes? `(80% * $0) + (20% * $184) = $36.8`. At minimum wage opportunity cost of $7 an hour, $36.8 is worth 5.25 hours of my time. I spent much time on screenshots, summarizing, and analysis, and I'd guess I spent closer to 10–15 hours all told.

(The net present value formula is the annual savings divided by the natural log of the discount rate, out to eternity. Exponential discounting means that a bond that expires in 50 years is worth a surprisingly similar amount to one that continues paying out forever. For example, a 50 year bond paying $10 a year at a discount rate of 5% is worth `sum $ map (\t -> 10 / (1 + 0.05)^t) [1..50] ~> 182.5` but if that same bond never expires, it's worth `10 / log 1.05 = 204.9` or just $22.4 more! My own expected

longevity is ~50 more years, but I prefer to use the simple natural log formula rather than the more accurate summation. All the numbers here are questionable anyway.)

This worked out example demonstrates that when a substance is cheap and you are highly confident it works, a long costly experiment may not be worth it. (Of course, I would have done it anyway due to factors not included in the calculation: to try out my Zeo, learn a bit about sleep experimentation, do something cool, and have something neat to show everyone.)

2 Vitamin D

http://www.gwern.net/Zeo#vitamin-d

I ran 2 experiments on vitamin D: whether it hurt sleep when taken in the evening, and whether it helped sleep when taken in the morning.

2.1 Evening

http://www.gwern.net/Zeo#vitamin-d-at-night-hurts

The first I had no opinion on. I actually did sometimes take vitamin D in the evening when I hadn't gotten around to it earlier (I take it for its anti-cancer and SAD effects). There was no research background, and the anecdotal evidence was of very poor quality. Still, it was plausible since vitamin D *is* involved in circadian rhythms, so I gave it 50% and decided to run an experiment. What effect would perfect information that it did negatively affect my sleep have? Well, I'd definitely switch to taking it in the morning and would never take it in the evening again, which would change maybe 20% of my future doses, and what was the negative effect? It couldn't be *that* bad or I would have noticed it already (like I noticed sulbutiamine made it hard to get to sleep). I'm not willing to change my routines very much to improve my sleep, so I would be lying if I estimated that the value of eliminating any vitamin D-related disturbance was more than, say, 10 cents per night; so the total value of affected nights would be `$0.10 * 0.20 * 365.25 = $7.3`. On the plus side, my experiment design was high quality and ran for a fair number of days, so it would surely detect any sleep disturbance from the randomized vitamin D, so say 90% quality of information. This gives `((7.3 - 0) / log 1.05) * 0.90 * 0.50 = 67.3`, justifying <9.6 hours. Making the pills took perhaps an hour, recording used up some time, and the analysis took several hours to label & process all the data, play with it in R, and write it all up in a clean form for readers. Still, I don't think it took almost 10 hours of work, so I think this experiment ran at a profit.

2.2 Morning

http://www.gwern.net/Zeo#vitamin-d-at-morn-helps

With the vitamin D theory partially vindicated by the previous experiment, I became fairly sure that vitamin D in the morning would benefit my sleep somehow: 70%. Benefit how? I had no idea, it might be large or small. I didn't expect it to be a second melatonin, improving my sleep and trimming it by 50 minutes, but I hoped maybe it would help me get to sleep faster or wake up less. The actual experiment turned out

to show, with very high confidence, absolutely no change except in my mood upon awakening in the morning.

What is the “value of information” for this experiment? Essentially - nothing! Zero!

1. If the experiment had shown any benefit, I obviously would have continued taking it in the morning
2. if the experiment had shown no effect, I would have continued taking it in the morning to avoid incurring the evening penalty discovered in the previous experiment
3. if the experiment had shown the unthinkable, a negative effect, it would have to be substantial to convince me to stop taking vitamin D altogether and forfeit its other health benefits, and it’s not worth bothering to analyze an outcome I would have given <=5% chance to.

Of course, I did it anyway because it was cool and interesting! (Estimated time cost: perhaps half the evening experiment, since I manually recorded less data and had the analysis worked out from before.)

3 Adderall

http://www.gwern.net/Nootropics#adderall-blind-testing

The amphetamine mix branded “Adderall” is terribly expensive to obtain even compared to modafinil, due to its tight regulation (a lower schedule than modafinil), popularity in college as a study drug, and reportedly moves by its manufacture to exploit its privileged position as a licensed amphetamine maker to extract more consumer surplus. I paid roughly $4 a pill but could have paid up to $10. Good stimulant hygiene involves recovery periods to avoid one’s body adapting to eliminate the stimulating effects, so even if Adderall was the answer to all my woes, I would not be using it more than 2 or 3 times a week. Assuming 50 uses a year (for specific projects, let’s say, and not ordinary aimless usage), that’s a cool $200 a year. My general belief was that Adderall would be too much of a stimulant for me, as I am amphetamine-naive and Adderall has a bad reputation for letting one waste time on unimportant things. We could say my prediction was 50% that Adderall would be useful and worth investigating further. The experiment was pretty simple: blind randomized pills, 10 placebo & 10 active. I took notes on how productive I was and the next day guessed whether it was placebo or Adderall before breaking the seal and finding out. I didn’t do any formal statistics for it, much less a power calculation, so let’s try to be conservative by penalizing the information quality heavily and assume it had 25%. So `((200 - 0) / log 1.05) * 0.50 * 0.25 = 512`! The experiment probably used up no more than an hour or two total.

This example demonstrates that anything you are doing *expensively* is worth testing *extensively*.

4 Modafinil day

http://www.gwern.net/Nootropics#modalert-blind-day-trial

I tried 8 randomized days like with Adderall to see whether I was one of the people whom modafinil energizes during the day. (The other way to use it is to skip sleep,

which is my preferred use.) I rarely use it during the day since my initial uses did not impress me subjectively. The experiment was not my best - while it was double-blind randomized, the measurements were subjective, and not a good measure of mental functioning like dual n-back (DNB) scores which I could statistically compare from day to day or against my many previous days of dual n-back scores. Between my high expectation of finding the null result, the poor experiment quality, and the minimal effect it had (eliminating an already rare use), it's obvious without guesstimating any numbers that the value of this information was very small.

I mostly did it so I could tell people that "no, day usage isn't particularly great for me; why don't you run an experiment on yourself and see whether it was just a placebo effect (or whether you genuinely are sleep-deprived and it is indeed compensating)?"

5 Lithium

http://www.gwern.net/Nootropics#lithium-experiment

Low-dose lithium orotate is extremely cheap, ~$10 a year. There is some research literature on it improving mood and impulse control in regular people, but some of it is epidemiological (which implies considerable unreliability); my current belief is that there is probably *some* effect size, but at just 10mg, it may be too tiny to matter. I have ~40% belief that there will be a large effect size, but I'm doing a long experiment and I should be able to detect a large effect size with >75% chance. So, the formula is NPV of the difference between taking and not taking, times quality of information, times expectation: `((10 - 0) / log 1.05) * 0.75 * 0.40 = 61.4`, which justifies a time investment of less than 9 hours. As it happens, it took less than an hour to make the pills & placebos, and taking them is a matter of seconds per week, so the analysis will be the time-consuming part. This one may actually turn a profit.

6 Redshift

http://www.gwern.net/Zeo#redshiftf.lux

Like the modafinil day trial, this was another value-less experiment justified by its intrinsic interest. I expect the results will confirm what I believe: that red-tinting my laptop screen will result in less damage to my sleep by not forcing lower melatonin levels with blue light. The only outcome that might change my decisions is if the use of Redshift actually worsens my sleep, but I regard this as highly unlikely. It is cheap to run as it is piggybacking on other experiments, and all the randomizing & data recording is being handled by 2 simple shell scripts.

7 Meditation

http://www.gwern.net/Zeo#meditation-1

I find meditation useful when I am screwing around and can't focus on anything, but I don't meditate as much as I might because I lose half an hour. Hence, I am interested in the suggestion that meditation may not be as expensive as it seems because it reduces sleep need to some degree: if for every two minutes I meditate, I need one less minute of sleep, that halves the time cost - I spend 30 minutes meditating, gain

back 15 minutes from sleep, for a net time loss of 15 minutes. So if I meditate regularly but there is no substitution, I lose out on 15 minutes a day. Figure I skip every 2 days, that's a total lost time of `(15 * 2/3 * 365.25) / 60 = 61` hours a year or $427 at minimum wage. I find the theory somewhat plausible (60%), and my year-long experiment has roughly a 60% chance of detecting the effect size (estimated based on the sleep reduction in a Indian sample of meditators). So `((427 - 0) / log 1.05) * 0.60 * 0.60 = $3150`. The experiment itself is unusually time-intensive, since it involve ~180 sessions of meditation, which if I am "overpaying" translates to 45 hours (`(180 * 15) / 60`) of wasted time or $315. But even including the design and analysis, that's less than the calculated value of information.

This example demonstrates that drugs aren't the only expensive things for which you should do extensive testing.

Alan Carter on the Complexity of Value

It’s always good news when someone else develops an idea independently from you. It's a sign you might be onto something. Which is why I was excited to discover that Alan Carter, Professor Emeritus of the University of Glasgow’s Department of Philosophy, has developed the concept of Complexity of Value independent of Less Wrong.

As far as I can tell Less Wrong does not know of Carter, the only references to his existence I could find on LW and OB were written by me. Whether Carter knows of LW or OB is harder to tell, but the only possible link I could find online was that he has criticized the views of Michael Huemer, who knows Bryan Caplan, who knows Robin Hanson. This makes it all the more interesting that Carter has developed views on value and morality very similar to ones commonly espoused on Less Wrong.

The Complexity of Value is one of the more important concepts in Less Wrong. It has been elaborated on its wiki page, as well as some classic posts by Eliezer. Carter has developed the same concept in numerous papers, although he usually refers to it as “a plurality of values” or “multidimensional axiology of value.” I will focus the discussion on working papers Carter has on the University of Glasgow’s website, as they can be linked to directly without having to deal with a pay wall. In particular I will focus on his paper "A Plurality of Values."

Carter begins the paper by arguing:

> Wouldn’t it be nice if we were to discover that the physical universe was reducible to only one kind of fundamental entity? … Wouldn’t it be nice, too, if we were to discover that the moral universe was reducible to only one kind of valuable entity —or one core value, for short? And wouldn’t it be nice if we discovered that all moral injunctions could be derived from one simple principle concerning the one core value, with the simplest and most natural thought being that we should maximize it? There would be an elegance, simplicity and tremendous justificatory power displayed by the normative theory that incorporated the one simple principle. The answers to all moral questions would, in theory at least, be both determinate and determinable. It is hardly surprising, therefore, that many moral philosophers should prefer to identify, and have thus sought, the one simple principle that would, hopefully, ground morality.
>
> And it is hardly surprising that many moral philosophers, in seeking the one simple principle, should have presumed, explicitly or tacitly, that morality must ultimately be grounded upon the maximization of a solitary core value, such as quantity of happiness or equality, say. Now, the assumption—what I shall call the presumption of value-monism—that here is to be identified a single core axiological value that will ultimately ground all of our correct moral decisions has played a critical role in the development of ethical theory, for it clearly affects our responses to certain thought-experiments, and, in particular, our responses concerning how our normative theories should be revised or concerning which ones ought to be rejected.

Most members of this community will immediately recognize the similarities between these paragraphs and Eliezer's essay "Fake Utility Functions." The presumption of value monism sounds quite similar to Eliezer's description of "someone who has discovered the One Great Moral Principle, of which all other values are a mere derivative consequence." Carter's opinion of such people is quite similar to Eliezer's.

While Eliezer discovered the existence of the Complexity of Value by working on Friendly AI, Carter discovered it by studying some of the thornier problems in ethics, such as the Mere Addition Paradox and what Carter calls the Problem of the Ecstatic Psychopath. Many Less Wrong readers will be familiar with these problems; they have been discussed numerous times in the community.

For those who aren't, in brief the Mere Addition Paradox states that if one sets maximizing total wellbeing as the standard of value then one is led to what is commonly called the Repugnant Conclusion, the belief that a huge population of people with lives barely worth living is better than a somewhat smaller population of people with extremely worthwhile lives. The Problem of the Ecstatic Psychopath is the inverse of this, it states that, if one takes average levels of well-being as the standard of value, that a population of one immortal ecstatic psychopath with a nonsentient machine to care for all their needs is better than a population of trillions of very happy and satisfied, but not ecstatic people.

Carter describes both of these problems in his paper and draws an insightful conclusion:

> In short, surely the most plausible reason for the counter-intuitive nature of any mooted moral requirement to bring about, directly or indirectly, the world of the ecstatic psychopath is that either a large total quantity of happiness or a large number of worthwhile lives is of value; and surely the most plausible reason for the counter-intuitive nature of any mooted injunction to bring about, directly or indirectly, the world of the Repugnant Conclusion is that a high level of average happiness is also of value.
>
> How is it that we fail to notice something so obvious? I submit: because we are inclined to dismiss summarily any value that fails to satisfy our desire for the one core value—in other words, because of the presumption of value-monism.

Once Carter has established the faults of value monism he introduces value pluralism to replace it.[1] He introduces two values to start with, "number of worthwhile lives" and "the level of average happiness," which both contribute to "overall value." However, *their contributions have diminishing returns,[2] so a large population with low average happiness and a tiny population with extremely high average happiness are both worse than a moderately sized population with moderately high average happiness.*

This is a fairly unique use of the idea of the complexity of value, as far as I know. I've read a great deal of Less Wrong's discussion of the Mere Addition Paradox, and most attempts to resolve it have consisted of either trying to reformulate Average Utilitarianism so that it does not lead to the Problem of the Ecstatic Psychopath, or redefining what "a life barely worth living" means upwards so that it is much less horrible than one would initially think. The idea of agreeing that increasing total wellbeing is important, but not the be all and end all of morality, did not seem to come up, although if it did and I missed it I'd be very happy if someone posted a link to that thread.

Carter's resolution of the Mere Addition Paradox makes a great deal of sense, as it manages to avoid every single repugnant and counterintuitive conclusion that Total and Average Utilitarianism draw by themselves while still being completely logically consistent. In fact, I think that most people who reject the Repugnant Conclusion will realize that this was their True Rejection all along. I am tempted to say that Carter has discovered Theory X, the hypothetical theory of population ethics Derek Parfit believed could accurately describe the ethics of creating more people without implying any horrifying conclusions.

Carter does not stop there, however, he then moves to the problem of what he calls "pleasure wizards" (many readers may be more familiar with the term "utility monster"). The pleasure wizard can convert resources into utility much more efficiently than a normal person, and hence it can be argued that it deserves more resources. Carter points out that:

> …such pleasure-wizards, to put it bluntly, do not exist... But their opposites do. And the opposites of pleasure-wizards—namely, those who are unusually inefficient at converting resources into happiness—suffice to ruin the utilitarian's egalitarian pretensions. Consider, for example, those who suffer from, what are currently, incurable diseases. … an increase in their happiness would require that a huge proportion of society's resources be diverted towards finding a cure for their rare condition. Any attempt at a genuine equality of happiness would drag everyone down to the level of these unfortunates. Thus, the total amount of happiness is maximized by diverting resources away from those who are unusually inefficient at converting resources into happiness. In other words, if the goal is, solely, to maximize the total amount of happiness, then giving anything at all to such people and spending anything on cures for their illnesses is a waste of valuable resources. Hence, given the actual existence of such unfortunates, the maximization of happiness requires a considerable inequality in its distribution.

Carter argues that, while most people don't think all of society's resources should be diverted to help the very ill, the idea that they should not be helped at all also seems wrong. He also points out that to a true utilitarian the nonexistence of pleasure wizards should be a tragedy:

> So, the consistent utilitarian should greatly regret the non-existence of pleasure-wizards; and the utilitarian should do so even when the existence of extreme pleasure-wizards would morally require everyone else to be no more than barely happy.

Yet, this is not how utilitarians behave, he argues, rather:

> As I have yet to meet a utilitarian, and certainly not a monistic one, who admits to thinking that the world would be a better place if it contained an extreme pleasure-wizard living alongside a very large population all at that level of happiness where their lives were just barely worth living…But if they do not bemoan the lack of pleasure-wizards, then they must surely value equality directly, even if they hide that fact from themselves. And this suggests that the smile of contentment on the faces of utilitarians after they have deployed diminishing marginal utility in an attempt to show that their normative theory is not incompatible with egalitarianism has more to do with their valuing of equality than they are prepared to admit.

Carter resolves the problem of "pleasure wizard" by suggesting equality as an end in itself as a third contributing value towards overall value. Pleasure wizards should not get all the resources because equality is valuable for its own sake, not just because of diminishing marginal utility. As with average happiness and total worthwhile lives, equality is balanced against other values, rather than dominating them. It may often be ethical for a society to sacrifice some amount of equality to increase the total and average wellbeing.

Carter then briefly states that, though he only discusses three in this paper, there are many other dimensions of value that could be added. It might even be possible to add some form of deontological rules or virtue ethics to the complexity of value, although they would be traded off against consequentialist considerations. He concludes the paper by reiterating that:

> Thus, in avoiding the Repugnant Conclusion, the Problem of the Ecstatic Psychopath and the problems posed by pleasure-wizards, as well as the problems posed by any unmitigated demand to level down, we appear to have identified an axiology that is far more consistent with our considered moral judgments than any entailing these counter-intuitive implications.

Carter has numerous other papers discussing the concept in more detail, but “A Plurality of Values” is the most thorough. Other good ones include “How to solve two addition paradoxes and avoid the Repugnant Conclusion,” which more directly engages the Mere Addition Paradox and some of its defenders like Michael Huemer; "Scrooge and the Pleasure Witch," which discusses pleasure wizards and equality in more detail; and “A pre-emptive response to some possible objections to a multidimensional axiology with variable contributory values,” which is exactly what it says on the tin.

On closer inspection it was not hard to see why Carter had developed theories so close to those of Eliezer and other members of Less Wrong and SIAI communities. In many ways their two tasks are similar. Eliezer and the SIAI are trying to devise a theory of general ethics that cannot be twisted into something horrible by a rules-lawyering Unfriendly AI, while Carter is trying to devise a theory of population ethics that cannot be twisted into something horrible by rules-lawyering humans. The worlds of the Repugnant Conclusion and the Ecstatic Psychopath are just the sort of places a poorly programmed AI with artificially simple values would create.

I was very pleased to see an important Less Wrong concept had a defender in mainstream academia. I was also pleased to see that Carter had not just been content to develop the concept of the Complexity of Value. He was also able to employ in the concept in new way, successfully resolving one of the major quandaries of modern philosophy.

Footnotes

[1]I do not mean to imply Carter developed this theory out of thin air of course. Value pluralism has had many prominent advocates over the years, such as Isaiah Berlin and Judith Jarvis Thomson.

[2]Theodore Sider proposed a theory called "geometrism" in 1991 that also focused on diminishing returns, but geometrism is still a monist theory, it had geometric diminishing returns for the people in the scenario, rather than the values creating the people was trying to fulfill.

Edited - To remove a reference to Aumann's Agreement Theorem that the commenters convinced me was unnecessary and inaccurate.

Problematic Problems for TDT

A key goal of Less Wrong's "advanced" decision theories (like TDT, UDT and ADT) is that they should out-perform standard decision theories (such as CDT) in contexts where another agent has access to the decider's code, or can otherwise predict the decider's behaviour. In particular, agents who run these theories will one-box on Newcomb's problem, and so generally make more money than agents which two-box. Slightly surprisingly, they may well continue to one-box even if the boxes are transparent, and even if the predictor Omega makes occasional errors (a problem due to Gary Drescher, which Eliezer has described as equivalent to "counterfactual mugging"). More generally, these agents behave like a CDT agent will wish it had pre-committed itself to behaving before being faced with the problem.

However, I've recently thought of a class of Omega problems where TDT (and related theories) appears to under-perform compared to CDT. Importantly, these are problems which are "fair" - at least as fair as the original Newcomb problem - because the reward is a function of the agent's actual choices in the problem (namely which box or boxes get picked) and independent of the method that the agent uses to choose, or of its choices on any other problems. This contrasts with clearly "unfair" problems like the following:

Discrimination: Omega presents the usual two boxes. Box A always contains $1000. Box B contains nothing if Omega detects that the agent is running TDT; otherwise it contains $1 million.

So what are some *fair* "problematic problems"?

Problem 1: Omega (who experience has shown is always truthful) presents the usual two boxes A and B and announces the following. "Before you entered the room, I ran a simulation of this problem as presented to an agent running TDT. I won't tell you what the agent decided, but I will tell you that if the agent two-boxed then I put nothing in Box B, whereas if the agent one-boxed then I put $1 million in Box B. Regardless of how the simulated agent decided, I put $1000 in Box A. Now please choose your box or boxes."

Analysis: Any agent who is themselves running TDT will reason as in the standard Newcomb problem. They'll prove that their decision is linked to the simulated agent's, so that if they two-box they'll only win $1000, whereas if they one-box they will win $1 million. So the agent will choose to one-box and win $1 million.

However, any CDT agent can just take both boxes and win $1001000. In fact, any other agent who is *not* running TDT (e.g. an EDT agent) will be able to re-construct the chain of logic and reason that the simulation one-boxed and so box B contains the $1 million. So any other agent can safely two-box as well.

Note that we can modify the contents of Box A so that it contains anything up to $1 million; the CDT agent (or EDT agent) can in principle win up to twice as much as the TDT agent.

Problem 2: Our ever-reliable Omega now presents ten boxes, numbered from 1 to 10, and announces the following. "Exactly one of these boxes contains $1 million; the others contain nothing. You must take exactly one box to win the money; if you try to take more than one, then you won't be allowed to keep any winnings. Before you entered the room, I ran multiple simulations of this problem as presented to an agent running TDT, and determined the box which the agent was least likely to take. If there were several such boxes tied for equal-lowest probability, then I just selected one of them, the one labelled with the smallest number. I then placed $1 million in the selected box. Please choose your box."

Analysis: A TDT agent will reason that whatever it does, it cannot have more than 10% chance of winning the $1 million. In fact, the TDT agent's best reply is to pick each box with equal probability; after Omega calculates this, it will place the $1 million under box number 1 and the TDT agent has exactly 10% chance of winning it.

But any non-TDT agent (e.g. CDT or EDT) can reason this through as well, and just pick box number 1, so winning $1 million. By increasing the number of boxes, we can ensure that TDT has arbitrarily low chance of winning, compared to CDT which always wins.

Some questions:

1. Have these or similar problems already been discovered by TDT (or UDT) theorists, and if so, is there a known solution? I had a search on Less Wrong but couldn't find anything obviously like them.

2. Is the analysis correct, or is there some subtle reason why a TDT (or UDT) agent would choose differently from described?

3. If a TDT agent believed (or had reason to believe) that Omega was going to present it with such problems, then wouldn't it want to self-modify to CDT? But this seems paradoxical, since the whole idea of a TDT agent is that it doesn't have to self-modify.

4. Might such problems show that there cannot be a single TDT algorithm (or family of provably-linked TDT algorithms) so that when Omega says it is simulating a TDT agent, it is quite ambiguous what it is doing? (This objection would go away if Omega revealed the source-code of its simulated agent, and the source-code of the choosing agent; each particular version of TDT would then be out-performed on a specific matching problem.)

5. Are these really "fair" problems? Is there some intelligible sense in which they are not fair, but Newcomb's problem is fair? It certainly looks like Omega may be "rewarding irrationality" (i.e. giving greater gains to someone who runs an inferior decision theory), but that's exactly the argument that CDT theorists use about Newcomb.

6. Finally, is it more likely that Omegas - or things like them - will present agents with Newcomb and Prisoner's Dilemma problems (on which TDT succeeds) rather than problematic problems (on which it fails)?

Edit: I tweaked the explanation of Box A's contents in Problem 1, since this was causing some confusion. The idea is that, as in the usual Newcomb problem, Box A

always contains $1000. Note that Box B depends on what the simulated agent chooses; it doesn't depend on Omega predicting what the actual deciding agent chooses (so Omega doesn't put less money in any box just because it sees that the actual decider is running TDT).

Petition: Off topic area

Petition: LW should introduce a dedicated off topic area

Why?

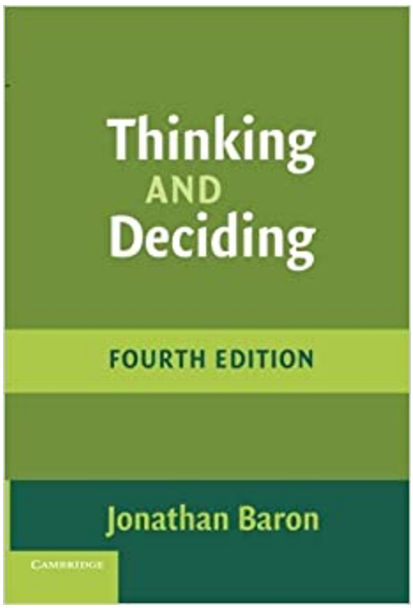
1) I want to discuss various topics with people who are both intelligent and rationalist, and i know of no other place where to do it.

2) If find that rationality is getting boring in itself. I need to use it on something.

3) As stated in this comment
http://lesswrong.com/lw/btc/how_can_we_get_more_and_better_lw_contrarians/6e3p

the narrow set of topics might actually hurt LW by driving good rationalists away.

Thinking and Deciding: a chapter by chapter review



This is a chapter-by-chapter review of Thinking and Deciding by Jonathan Baron (UPenn, twitter). It won't be a detailed summary like badger's excellent summary of Epistemology and the Psychology of Human Judgment, in part because this is a 600-page textbook and so a full summary would be far longer that I want to write here. I'll try to provide enough details that people can seek out the chapters that they find interesting, but this is by no means a replacement for reading the chapters that you find interesting. Every chapter is discussed below, with a brief "what should I read?" section if you know what you're interested in.

We already have a thread for textbook recommendations, but this book is central enough to Less Wrong's mission that it seems like it's worth an in-depth review. I'll state my basic impression of the whole book up front: I expect most readers of LW would gain quite a bit from reading the book, especially newer members, as it seems like a more focused and balanced introduction to the subject of rationality than the Sequences.

Baron splits the book into three sections: Thinking in General, Probability and Belief, and Decisions and Plans.

I may as well quote the first page in its entirety, as I feel it gives a good description of the book:

> Beginning with its first edition and through three subsequent editions, *Thinking and Deciding* has established itself as the required text and important reference work for students and scholars of human cognition and rationality. In this, the fourth edition, Jonathan Baron retains the comprehensive attention to the key questions addressed in previous editions- How should we think? What, if anything, keeps us from thinking that way? How can we improve our thinking and decision making? - and his expanded treatment of topics such as risk, utilitarianism, Bayes's theorem, and moral thinking. With the student in mind, the fourth edition emphasizes the development of an understanding of the fundamental concepts in judgment and decision making. This book is essential reading for students and scholars in judgment and decision making and related fields, including psychology, economics, law, medicine, and business.
>
> Jonathan Baron is Professor of Psychology at the University of Pennsylvania. He is the author and editor of several other books, most recently *Against Bioethics*. Currently he is editor of the journal *Judgment and Decision Making* and president of the Society for Judgment and Decision Making (2007) .

1. What is thinking?

This chapter will be mostly familiar to readers of Less Wrong; in the second paragraph, Baron says (in more words) 'rationality is what wins.' It may still be helpful as Baron expresses a number of things often left unsaid here.

He splits thinking into three parts: thinking about decisions (instrumental rationality), thinking about beliefs (epistemic rationality), and thinking about goals. The last is a notoriously sticky subject. He also discusses his search-inference framework, which is how he describes minds as actually operating- coming across ideas, evaluating them, and proceeding from there. Most decision analysis views itself as operating over a fixed set with a well-defined objective function, but those are the two main problems for real decision-makers: identifying possibilities worth considering and comparing two dissimilar outcomes.

The chapter is filled out with a discussion of understanding, knowledge as design, and examples of thinking processes (worth skimming over, but many of which will be familiar to experts in the relevant fields).

2. The study of thinking

Kahneman and Tversky get their first of many references here. Baron discusses a number of the methods used to learn about human cognition, mentioning a few of their pitfalls.

One, which bears repeating, is that most study of biases just reports means, rather than distributions. I remember learning the actual numerical size of the Asch conformity experiments about five years after I heard about the experiment itself, and was underwhelmed (32% incorrect answers, ~75% of subjects gave at least one incorrect answer). A general human tendency is different from a sizeable subset of weak-willed people. Similarly, our article on Prospect Theory had a link to graphs of subjective probability in one of the comments, of which the most noteworthy were the two people who were nearly linear. While Baron brings up this issue, he doesn't give many examples of it here.

He also mentions three models of thought: descriptive models, prescriptive models, and normative models. Descriptive models are what people actually do; normative models are what thinkers should do with infinite cognitive resources; prescriptive models are what thinkers should do with limited cognitive resources. This has come up on LW before, though the focus here has often been exclusively on the normative, though the prescriptive seems most useful.

Computer models of thinking are briefly discussed, but at a superficial level.

This chapter sees the first set of exercises. Overall, the exercises in the book seem to provide a brief example / check, rather than being enough to develop mastery. I think this is what I'd recommend but it has the potential to be a weakness.

3. Rationality

Again, Baron identifies rationality as “the kind of thinking that helps us achieve our goals.” Refreshingly, he focuses on *optimal* search, keeping in mind the costs of decision-making and information-gathering.

Much of this chapter will be familiar to someone who has read the Sequences, but it's presented tersely and lucidly. The section on rationality and emotion, for example, is only three pages long but is clear, quickly identifying how the two interact in a way that'll clear up common confusions.

4. Logic

The content in this chapter seems mostly unimportant- I imagine most readers of LW are much more interested in probabilistic reasoning than syllogisms. Still, Baron gives a readable (and not very favorable) description of the usefulness of formal logic as a normative model of thinking.

What is fascinating, though, is the section of the chapter that delves into the four-card problem and variations of it. Particularly noteworthy is the variation designed so that most people's intuitions are correct- people give the correct explanations of why they selected the cards they selected, and why they didn't select the cards they didn't select. But when their intuition is wrong, they give explanations that are just as sophisticated- but wrong. It's more evidence that the decision-making and verbal reason-providing modules are different- even someone who gives the correct explanation of the correct answer may stumble on a problem where their underlying simple heuristic (pick the cards mentioned in the question) fails.

He presents a method of mental modeling that makes logical statements easier to correctly evaluate, and then there are a few logical inference exercises.

5. Normative theory of probability

Yet another introduction to Bayes. Baron focuses primarily on Bayesianism (called the “personal” theory of probability) but still introduces alternatives (the “frequency” theory, i.e. frequentism, and “logical” theory, which is a subset of frequentism where all events are required to have the same probability.) This chapter will be useful for someone who doesn't have a firm probabilistic foundation, but holds little interest for others.

There are a handful of exercises for applying Bayes.

6. Descriptive theory of probability judgment

This chapter primarily covers biases related to numerical probability estimates, many of which are classics in the heuristics and biases field (and so have probably been mentioned on Less Wrong at least once). The chapter shines when Baron goes into the detail of an experiment and its variations, as that gives a firmer view of what the experiment actually shows (and, importantly, what it does not show)- descriptions of biases where he only quotes a single experiment (or single feature of an experiment) feel weaker.

A major feature of this chapter is the implication that people are bad at numerical probability estimation mostly because they're unfamiliar with it, implying that calibration exercises may improve probability estimation. A 1977 study of weatherman calibration suggested they were very well calibrated, both with their estimates and with the confidence that should be placed in those estimates. More recent work shows that weathermen have systematic calibration biases.

7. Hypothesis testing

I was gratified to discover that this chapter was not about statistics, but how to come up with and test hypotheses. Baron discusses different models of scientific advancement, focusing on the sorts of likelihood ratios that they look for, as well as discussing the sort of mistakes people make when choosing tests for hypotheses. Many of the stories will probably be familiar- Ignaz Semmelweis gets a mention, though in more detail than I had seen before, as well as the 2-4-6 rule familiar to HPMOR fans and a variation of the four card experiment that makes the typical mistake more obvious.

He gives a baking example to suggest why people might search primarily for positive evidence- there may be benefits to getting a “yes” answer besides the information involved. If you're experimenting with cake recipes, and you think your last cake was good because of a feature, it makes sense to alter other features but keep the one you suspect the same, as that means a good cake is more likely; if you think a cake was bad because of a feature, it makes sense to alter that feature but keep the others the same, as that also means a good cake is more likely. In a purely scientific context, it makes sense to vary the element you think has an impact just to maximize the expected size of the impact, positive or negative.

He describes in more detail a methodology he's been discussing, “actively open-minded thinking,” which seems to boil down to “don't just be willing to accept disconfirming evidence, go looking for it,” but the full explanation comes in a few chapters.

8. Judgment of correlation and contingency

This chapter is descriptive; it begins with a description of correlations and then discusses human judgment of correlations. Unsurprisingly, people suffer from the illusion of control- they think there's more likely to be a correlation if their effort is involved- and from confirmation bias. There are some examples of the latter, where people find correlations that make intuitive sense but aren't in the data, and don't discover correlations that don't make intuitive sense that are in the data. There's also a brief section on how people use nearly useless evidence to support theories or dismiss evidence that doesn't support their theory. Overall, it's a short chapter that won't be surprising to LW readers (although some of the studies referenced may be new).

9. Actively open-minded thinking

I'll quote part of this chapter in full because I think it's a great description:

> [G]ood thinking consists of (1) search that is thorough in proportion to the importance of the question, (2) confidence that is appropriate to the amount and quality of thinking done, and (3) fairness to other possibilities than the one we initially favor.

The chapter overall is very solid- it deftly combines normative predictions with descriptive biases to weave a prescriptive recommendation of how to think better. There are several great examples of actively open-minded thinking; in particular, the thought process of two students as they attempt to make sense of a story sentence by sentence.

Many of the suggestions in the chapter are extended by various LW posts, but the chapter seems useful as a concise description of the whole problem and illustration of a general solution. If you're having trouble fitting together various rationality hacks, this seems like a good banner to unite them under.

10. Normative theory of choice under uncertainty

This chapter is an introduction to utility theory, describing how it works, how multiple attributes can be consolidated into one score, and a way to resolve conflicts between agents with different utilities. It's a good introduction to decision analysis / utility theory, and there are some exercises, but there are no surprises for someone who's seen this before.

11. Descriptive theory of choice under uncertainty

This chapter is an introduction to different theories of how humans actually make decisions, like prospect theory and regret theory. There are a handful of exercises for understanding prospect theory.

Baron takes an even-handed approach to deviations from the normative theory. For example, when discussing regret theory, regrets have a real emotional cost (and real learning benefit)- but behaving according to descriptive theories because they're descriptive rather than because they're useful is a mistake. In many cases, those emotions can be manipulated by choice of reference point.

He also discusses the ambiguity effect- where people treat known probabilities differently from unknown probabilities, giving examples both of laboratory situations (drawing balls from an urn with a partially known composition) and real-life situations (insuring unprecedented or unrepeatable events). Baron describes this as incompatible with personal probability and suggests it's related to framing- situations where the probabilities seem known can be changed into situations where probabilities seem unknown. This aversion to ambiguity, though, can be perfectly sensible insofar as it pushes decision-makers to acquire more information.

He also discusses a Tversky study in which most students make a decision to pay money to defer a decision until they receive relevant information, but when asked how they would make the decision in the case of either possible piece of information, most

students realize they would make the same decision and choose not to defer the decision.

12. Choice under certainty

This chapter is primarily descriptive, focusing on the problem of thinking about goals. Most people favor categorical goal systems- Baron gives a great example, from Gardiner and Edwards, of the California Coastal Commission, tasked to decide which development projects to allow on the Pacific Coast. The commission was split into pro-development and pro-environment factions, which almost never agreed on which projects to allow and disallow. When asked to rank projects, most would rank them solely by their preferred criterion, creating lists that strongly disagreed. When asked to take both criteria into account- but with whatever weighting they wanted- the subjects would heavily weight their preferred criterion, but the projects which were both very valuable and not very environmentally damaging floated to the top of both lists, creating significant agreement.

The list of biases is *long*, and each has a study or story associated with. Many of the effects have been mentioned on LW somewhere, but it's very useful to have them placed next to each other (and separated from probabilistic biases), and so I'd recommend everyone read this chapter.

13. Utility measurement

This descriptive chapter discusses the difficult challenge of measuring utilities. It introduces both decision analysis and cost-benefit analysis- the latter converts outcomes to dollars to guide decisions, while the former converts outcomes to utility values to guide decisions.

People are not very skilled at satisfying axioms we would like them to satisfy. For example, consider the challenge of valuing a certain $50 against a *p* chance of $100 (and $0 otherwise). A subject will often give an answer like .7. Then, when later asked how much a 70% chance of $100 is worth, the subject will answer $60. That inconsistency needs to be resolved before their answers are used as parameters for any decisions. Thankfully, this is an area of active research, and ways to elicit probabilities and values that hold up to reflective equilibrium are gradually being developed. (This particular chapter, while it sounds that note of hope, is mostly negative: here are methods that have been tried and have crippling problems.)

This seems like a chapter that would be useful for anyone who wants to use utilities in an argument or model- treating them like they're unambiguous, easily measured objects when they actually seem to be fuzzy and hard to pin down can lead to significant problems, and thinking clearly about values is a spot where LW could do better.

14. Decision analysis and values

This chapter is a more prescriptive approach to the same problem- given that utilities and values are hard to find, where do we look for them? A dichotomy familiar to LW readers- instrumental and terminal values- appears here as "means-ends objective hierarchy" or "means values" and "fundamental values."

It contains a wealth of examples, including a computer-buying one with potential memories of 64KB to 640KB, with the hilarious comment that "you are buying this computer many years ago, when these numbers made sense!" There are also practical elicitation suggestions- rather than try to figure out a point estimate, start from a number that's too high until you're indifferent, and then start from a number that's too low until you're indifferent, giving you an indifference range (that you can either report or use the middle of as a point estimate).

Lexical preferences (also called categorical preferences elsewhere) and tradeoffs are discussed- Baron takes the position (that I share) that lexical preferences are actually tradeoffs with very, very high weights. (How do we trade off human lives and dollars? We should require a *lot* of dollars for a life- but not an infinite amount.) There's a discussion of micromorts (though he doesn't use that term) and of historical attempts to teach decision analysis that should be interesting to CFAR (though the references are a few decades old, now). The discussion of the examples contains quite a bit of practical advice, and the chapter seems worthwhile for almost everyone.

15. Quantitative judgment

This chapter describes three common quantitative problems- scoring, ranking, and classifying, and discusses some biases that hamper human decision-making along those lines and some recommendations. Statistical prediction rules make an appearance, though they're not called that. One fascinating suggestion is that models of people can actually perform better than those people, since the models don't have off days and people do.

This chapter will have some new material for LWers, and seems like a good extension of the previous chapter.

16. Moral Judgment and Choice

This chapter discusses morality from the point of decision-making- which is a refreshing perspective. Baron strongly endorses consequentialism and weakly endorses utilitarianism, providing a host of moral questions in which many people deviate from the consequentialist or utilitarian position.

A recurring theme is omission bias: people tend to judge active involvement in a situation in which someone is made worse off as worse than passive involvement in such a situation, even if the end result is better for everyone. People also weight intentions, which doesn't fit a direct consequentialist view.

Overall, the chapter seems valuable for reframing moral questions- placing them within the realm of pragmatism by moving to the perspective of decisions- but provides very little in the way of answers. Both the consequentialist and utilitarian positions are controversial and come with significant drawbacks, and Baron is fair enough in presenting those drawbacks and controversies, though in a rather abridged form.

17. Fairness and justice

This chapter is an extension of the previous chapter, focusing on intuitions dealing with fairness and justice. Baron details situations in which they agree and disagree with utilitarian analysis. Noteworthy is the undercurrent of adaptation-execution and not utility-maximization - fairness has tangible benefits, but people will often pursue fairness even at the cost of tangible benefits.

This chapter (and to a lesser extent the previous one) seem odd in light of chapter 15, in which the fallibility of individual judgment took center stage, with the recommendation that applying rules derived from individual judgment can often do better. It is good to know the reasoning that justifies moral intuitions, especially if one is interested in their boundaries, but when those boundaries impact outcomes they become political questions. If the sole point of punishment is deterrence (and that is the only sensible utilitarian justification), the question of whether or not a decision can impact future decisions is a sticky one. Perhaps the full consequentialist reckoning will recommend unthinking application of the rules, even in cases where direct consequentialist reckoning recommends suspending the rules.

18. Social dilemmas: cooperation versus defection

This chapter focuses on descriptive experiments- how people actually behave in social dilemmas- finding them to be much more cooperative than normative theory would recommend. There is some ambiguity, which he discusses, in what the "normative theory" is- utilitarianism recommends cooperation on the prisoner's dilemma, for example, because it maximizes total utility, whereas expected utility theory recommends defection on the prisoner's dilemma, because it's a dominating strategy.

The value of the chapter mostly lies in the study results- a few are interesting, like that discussing the social dilemma with other participants beforehand significantly increases cooperation, or that subjects are more likely to defect on the prisoner's dilemma if they know their partner's response than if they are uncertain, even if they know their partner cooperated.

Typically, for social dilemmas (scenarios in which private gain requires public loss, or public gain requires private loss), decision-making biases increase the level that people cooperate. (This is somewhat unsurprising, since the normative recommendation is typically defection, and biases move real decisions away from the normative recommendation.) People fail to distinguish between casual influence- "my voting makes people like me more likely to vote"- from diagnostic influence- "people like me voting makes me more likely to vote"- but one of the major reasons people give for voting is that it has a causal influence, rather than a merely diagnostic one.

19. Decisions about the future

This chapter is unlikely to contain any surprises for LWers, but serves as a fine introduction to discounting, both exponential and hyperbolic, and thus dynamic inconsistency. Also interesting (but too brief) is the discussion of goals in the context of time and plans and of goals as malleable objects.

Baron describes four methods of self-control: extrapsychic devices (removing a tempting option), control of attention (thinking about things other than the tempting

option), control of emotion (cultivating an incompatible emotion), or personal rules (viewing situations as instances of general policies, rather than isolated events). Again, the discussion is brief- only two pages- though the subject is of great interest to many here.

20. Risk

This chapter focuses on descriptive approaches to risk- survey responses and government regulation- as the normative approach to risk has mostly been detailed in the rest of the book: use expected utility theory. Most people are beset by biases and innumeracy, though, and so there's a whole chapter of material on misjudgments of risk and insurance.

Many of the biases, though perhaps not the examples, will be familiar to LWers. On the whole, they're somewhat uninteresting since most of them seem to just result from innumeracy: when given a table of deaths per year from four causes with wildly different prevalences, subjects were correctly willing to pay more to reduce larger risks by the same percentage as smaller risks. But their preferences scaled much more slowly than the risks- the subjects were, on average, willing to pay 20 times as much to prevent 20% of the deaths from a cause of death that killed 10,000 times as many people. Those distorted willingnesses to pay show up in government regulations. People were also more willing to pay for protection against the unfamiliar than the familiar- even though the relative benefit was far higher for protection against the familiar. (The illusion of control also shows up, distorting perceptions of risk.)

What should I read?

- Almost everyone: 7 and 9.
- I'm hunting biases: 6, 8, 11, 12, and then 15-20 (perhaps without 18).
- I'm interested in moral reasoning: 13 and 16 should be required reading. 14, 15, and 17-19 will be useful.
- I'm a decision maker: 10 and 14 will be directly useful, but check out the bias chapters too.
- I'm new to rationality: Start off with 1-4.
- I'm an expert at rationality but haven't heard of Baron: Still read 1-4, just to get his perspective of the field.
- I don't have a strong background in Bayesianism: read chapter 5.

How to brainstorm effectively

> Mr. Malfoy is new to the business of having ideas, and so when he has one, he becomes proud of himself for having it. He has not yet had enough ideas to unflinchingly discard those that are beautiful in some aspects and impractical in others; he has not yet acquired confidence in his own ability to think of better ideas as he requires them. What we are seeing here is not Mr. Malfoy's best idea, I fear, but rather his only idea.

- Harry Potter and the Methods of Rationality

> I want to emphasize yet again that the tools [described in Serious Creativity] are deliberate and can be used systematically. It is not a matter of inspiration or feeling in the mood of being "high." You can use the tools just as deliberately as you can add up a column of numbers.

- Edward De Bono, *Serious Creativity*

I will summarize some of the techniques for how to generate ideas presented in Serious Creativity. The book also has other material, e.g. interesting deep theories about why these techniques work, arguments for the importance of creativity, and more techniques beyond what's described in this post, but in the interest of keeping this post concise and useful, I will only describe one kind of technique and urge you to just try it. You should read the book if you want more detail or techniques.

These techniques can be used both when you have a problem you need to solve and when you have a general area that you suspect could be improved or innovated, but don't have any specific ideas of what's wrong (or even if you don't feel like there's anything wrong at all).

The technique I will describe in this post is that of "provocation" followed by "movement." A provocation is a seemingly random or nonsensical sentence or phrase. Movement is the process of going forward with a provocation and actually generating an idea. There are precise, formal techniques for generating provocations and movement, which I will describe after giving an example of how this "provocation-movement" process works.

Example

Provocation: Planes land upside down.
Movement: We can imagine this actually happening, and observe that the pilot would have a better view of the landing area. This naturally leads us to consider other ways to improve the pilot's view of the landing area. Perhaps we could move the cockpit to the bottom, or add video cameras. So using this technique, we've identified an area for improvement and two possible ways to make that improvement.

Setting Up Provocations

Provocation is a way to avoid getting stuck in the same "mental pathways" (see priming) so that you can find new ones. Provocations should not make sense and are not necessarily intended to convey meaning; they are just intended to "make things happen in our minds." The book precedes provocations with "po," a word used to indicate that the sentence is intended to be nonsensical and illogical. Po stands for "provoking operation." The book describes several techniques for generating provocations.

1. **Escape method**: Think of something that we take for granted, and negate it. E.g., "Po, restaurants do not have food" or "Po, shoes do not have soles."
2. **Reversal**: Take a standard arrangement or relationship that we take for granted, and reverse it. E.g. "I have orange juice for breakfast" becomes "Po, the orange juice has me for breakfast". Note that the reversal would *not* be "Po, I do not have orange juice for breakfast." That would be the escape method.
3. **Exaggeration**: Suggest that some dimension or measurement falls far outside its normal range (either greater or lesser). E.g. "Po, every household has 100 phones" or "Po, the phone has 1 dialing button." If you're making the dimension smaller, do not bring it to 0 or you're just using the escape method again. E.g. "Po, the phone has 0 dialing buttons" is not an exaggeration, it's an escape.
4. **Distortion**: Take normal arrangements (e.g. relationships or time sequences) and switch them around. E.g. "Po, you close the letter after you post it," "Po, criminals pay for the police force," or "Po, food prepares customers for chefs."
5. **Wishful thinking**: "Wouldn't it be nice if..." put forward a fantasy that is known to be impossible. E.g. "Po, the pencil should write by itself."

A provocation doesn't need to follow from one of these techniques. A provocation can be any incorrect or absurd statement. These techniques are just easy step-by-step ways to generate provocations without requiring any elusive "spark of inspiration." Once a provocation is generated, it should be followed by one or more of the movement techniques described in the next section.

If you are trying to solve a specific problem or innovate in a particular domain, then choose provocations related to the domain. That is, if you're trying to figure out how to improve wikipedia, don't use a provocation like "Po, the orange juice has me for breakfast," choose one like "Po, citations are not needed" (escape) or "Po, articles contain encyclopedias." (reversal).

Movement

Movement allows you to take some idea, concept, or provocation and move forward with it to generate more useful ideas and concepts. These techniques don't apply solely to provocations: you can use them for ideas and concepts too. The book describes 5 formal techniques for movement:

1. **Extract a principle**: Focus on some principle of the provocation, and then work with that principle to discover other ideas related to it. E.g. with the provocation "Po, bring back the town crier", we may extract the principle that the town crier

can go to where people are, and then we try to generate ideas related to that principle.
2. **Focus on the difference**: Compare the provocation to existing ways of doing things. How are they different? Then you can consider other ways to use this difference. This is very similar to "extract a principle."
3. **Moment to moment**: imagine what would happen if the provocation were put into effect. We are not interested in the final effect, but the moment-to-moment happenings. E.g. for "Po, orange juice has me for breakfast", you may imagine yourself falling into a giant glass of orange juice.
4. **Positive aspects**: Look directly for benefits. What are the positive aspects of the provocation? Once you've identified some positive aspects you can consider if you can achieve some of them in other ways (again, this is similar to extract a principle, it's just another way of thinking about it).
5. **Circumstances**: In what circumstances would the provocation have immediate value? E.g. for the provocation "Po, drinking glasses should have rounded bottoms," you could notice that this would be useful if you didn't want people to be able to put down their glasses. This could be good for bars, where you want people to drink more and faster.

You can use these movement techniques not just on provocations, but also ideas or concepts. For example, you may start with a provocation, use the "moment to moment" technique which gives you an idea, and then you could use the "positive aspects" technique with that idea to generate more ideas. Also, of course, you do not need to strictly use just these techniques. If a provocation directly leads you to think of something interesting without explicitly choosing to use one of these techniques, that's fine, you should explore the idea more. Use these when you need them.

More Examples

Here's another example from the book. This one uses the "moment to moment" movement technique:

> Po, cars have square wheels
>
> We imagine a car with square wheels. We imagine this car starting to roll. The square wheel rises up on its corner. This would lead to a very bumpy ride. But the suspension could anticipate this rise and could adjust by getting shorter. This leads to the concept of an adjusting suspension. This in turn leads to the idea of a vehicle for going over rough ground. A jockey wheel would signal back the state of the ground to the suspension which would then adjust so that the wheel was raised to follow the "profile" of the ground...This was an idea I first suggested about twenty years ago. Today several companies such as Lotus (part of GM) are working on "intelligent suspension" which behaves in a similar way.

And here's another one from the book. The provocation uses the "escape" method and the movement seems to use the "circumstances" method:

> Po, waiters are not polite.
>
> This leads to an idea for waiters to be actors and actresses. The menu indicates the "character" of the waiter. You can order whichever waiter you wanted:

belligerent, humorous, obsequious, and so on. You might order a belligerent waiter and enjoy having a fight with him. The waiters and waitresses would act out the assigned role.

Warnings

- As a general principle, try to avoid saying "oh, but this is just like this other existing product" whenever you generate an idea. Usually it's *not* just like the existing idea, you're just interpreting it in that way because we naturally follow paths toward the familiar. So if you have a half-formed idea that could take several directions, fight the urge to immediately take it down an existing path and then discard it because it already exists. Leave it in the half-formed stage instead. I'm reminded of the concept of semantic stopsigns. Saying an idea is "the same as" something else gives the illusion of having fully explored the idea, when in reality you just jumped immediately to one possible development (possibly the least useful development, since it's one you know already exists).
- Similarly, do not take too many steps when moving from a provocation. This will just lead you to an existing idea. There's nothing to be gained by playing 6 degrees of separation with provocations and existing ideas. Just take a few small steps. If nothing comes to you, try other movement techniques or try a different provocation.
- You're not expected to come up with a good idea for every provocation. Most of the time you'll come up with some mediocre or half-formed idea, or even no idea at all. This is fine.
- You should write down anything you come up with that seems interesting (even if it's a bad idea in its current form, if it has something interesting about it, write it down) and then come back to it later and think about it more (either using these techniques or just your normal thinking processes for improving and adapting ideas).

PSA: Learn to code

Presumably you read Less Wrong because you're interested in thinking better.

If so, you might be interested in another opportunity to improve the quality of your thinking: learn to code.

Like nothing else, coding forces you to identify flaws in your thinking. If your thinking is flawed, your program won't work, except by accident. There's no other discipline quite like this. If you're a mathematician or physicist and you solve a problem wrong, your paper won't tell you. Computer programmers have to measure their thinking against the gold standard of correctness *constantly*. The process of uncovering and fixing flaws in a program, usually called "debugging", typically takes up the majority of the time spent on software projects.

But this is only the beginning. You've probably heard something like "there are some problems that humans are good at and some problems that computers are good at". This is true. And once you learn to code, you'll be able to exploit computers to solve the problems they are good at. Having a computer to write software with is like having a hi-tech mental exoskeleton that lets your mind run harder and jump higher. Want to know what the second most common letter for an English word to end in is? That's a 15 line script. Tired of balancing chemical equations for your homework? Automate it.

Two more benefits that have less to do with thinking better:

- Employment. You probably don't need a computer science degree. I know of *two* Less Wrong users who learned to program *after college* and got jobs at Silicon Valley startups with just a project or two on their resume. (MBlume and FrankAdamek.) See Advice on Getting a Software Job by Tom McCabe for more on this possibility.
- Productivity software. Writing your own is much nicer than using stuff made by other people in my experience. The reason there are so many to-do list applications is because everyone's needs are different. If you use the terminal as your interface, it doesn't take much effort to write this stuff; you'll spend most of your time figuring out what you want it to do. (Terminal + cron on Linux with JSON log files has worked great for my needs.)

Having enough coding knowledge to be dangerous may take persistence. If you tried and failed in the past, you probably either got stuck and gave up because there was no one to help you, or you just didn't keep at it.

I've take two different introductory programming classes now to meet college requirements. The students in both seemed substantially less intelligent to me than Less Wrong users, and most were successful in learning to program. So based on the fact that you are reading this, I am pretty darn sure you have the necessary level of mental ability.

Starting Out

I recommend trying one of these interactive tutorials *right now* to get a quick feel for what programming is like.

After you do that, here are some freely available materials for studying programming:

- Learn Python the Hard Way. I like Zed's philosophy of having you type a lot of code, and apparently I'm not the only one. (Other books in the Hard Way series.)
- Eloquent JavaScript. No installation needed for this one, and the exercises are nicely interspersed with the text.
- Think Python. More of a computer science focus. ("Computer science" refers to more abstract, less applied aspects of programming.)
- Codecademy (uses JavaScript). Makes use of gamification-type incentives. Just don't lose sight of the fact that programming can be fun without them.
- Hackety Hack (uses Ruby). Might be especially good for younger folks.
- How to Design Programs. This book uses an elegant, quirky, somewhat impractical language called Scheme, and emphasizes a disciplined approach to programming. Maybe that will appeal to you. Structure and Interpretation of Computer Programs is a tougher, more computer science heavy book that also uses Scheme. You should probably have a good understanding of programming with recursive functions before tackling it.

Here's a discussion on Less Wrong about what the best programming language to start with is.

If you're having a hard time getting something up and running, that's a system administration challenge, not a programming one. Everyone hates system administration I think, except maybe system administrators. Keep calm, put your error message into Google, get help on a relevant IRC channel, etc.

Once you've got the basics, a good way to proceed is to decide on something you want to write and try to write it. If you don't know how to get started, start making Google searches. Soon you'll figure out the sort of libraries/frameworks people use to write your kind of program.

At first you may just be aping what others do. For example, if you want to learn something called "bleh", searching on Google for "bleh tutorial" is a great way to start. Finding a working program and modifying it to see out how it changes is another good option. Soon you'll graduate to appropriating sample code from documentation. As you write more code and see more of the software landscape, you'll be better prepared to craft original approaches to writing software.

See also: On the Fence? Major in CS, Teach Yourself Programming in 10 Years, Computer Science and Programming: Links and Resources.

A Protocol for Optimizing Affection

If Eliezer's art of solving confusing questions is the basic punch of rationality, and fighting akrasia and becoming personally effective is the basic front kick, I would like to master the loving hug. Here is a simple protocol to help us build stronger relationships and stronger communities:

In the spirit of Crocker's rules, I give you Nyan's rules: I hereby declare that you are allowed to love me. I will not judge you or hate you or stop talking to you. I will recieve and return your affection happily and gently let you know if you push my limits.

What's this all about? Here is the story:

I have strong feels of love and friendship for some of you that I met at minicamp, and some of you that I know from my meetup. On reflection, I see that I want to be deeply in (reciprocal) love with as many people as possible. I look forward to a future when I am smart enough to be in wonderful friendly love with all N billion of us.

I don't just want more feels, I want to be able to express them, too. I want to be able to tell you all that I love you and hold your hands and hug and cuddle and generally be nice without anyone feeling awkward or creeped out or conflicted.

Happiness research and personal experience suggests that more affection and closer relationships are generally a good thing. Mammals seem to like curling up together. Unwelcome affection is no good tho; the utility of affection seems to drop off past some point where people start to feel uncomfortable or unsafe. I think if we tried, we would find that there is tremendous value in finding the right level of affection in our relationships. The problem at this point is how *quickly* the utility of affection drops off, and how unwilling people are to be explicit about their preferences here.

Currently, I feel like if I tell my friend that I love him or try to hold his hand, and he is not interested, this at best creates an awkward situation, and at worst irrevocably damages the friendship. It is a violation of fun theory to have a misstep that is this expensive. The usual method prescribed to deal with this is to be able to work up the curve slowly and get a feel for when you are reaching the limit. The location of the optimum also moves up, so building rapport like this is a pretty important skill. IMO, tho, it is too expensive to do things this way if we can avoid it.

We should be able to find and operate at the optimal level of affection with minimal cost. In the current social dynamic with my current skill level, even probing for information is so scary that I don't bother to play the game.

Many percieved social risks are imaginary, but if this one is, no one is being explicit about its non-existance, so it still scares me. If it scares me, it probably scares others. There may even be people who want to be more affectionate with me, and aren't able to work up enough courage to try. That makes me really sad.

This is all made worse by love being mixed up with romance. Romance brings a whole other bag of grenades to the love party. If, in some case, full-on romance is uncomfortable or inconvienient for someone, that doesn't mean the optimal level of affection is none. We can probably still have hugs and cuddles. Note that this is just a consequence of the optimal-affection idea.

So there are two things we need to do, I think, to create a better social dynamic in which we can optimize affection and relationships faster and better. We need to be more comfortable with being explicit about what we are comfortable with, and we need to try to flatten the tail of our affection->utility curve so that overstepping comfort limits is not such a disaster. This means not punishing people for overstepping the bounds the first time, just gently nudging them back to your comfort zone.

At minicamp, there were a couple moments where a few of us semi-deliberately made these changes. IMO, the result was huge; we probed each other's comfort boundaries and built loving relationships very quickly, and all came out of it happier. At least that's what it felt like to me. This is one of the sources of strong feels of love and friendship that I mentioned above. This post is an attempt to formalize what happened there into a useful protocol.

Human social dynamics is one of the most complex systems in the known universe. Hacking it naively is bound to hit some pitfall or other. Even so, it is *our* system, and we are rationalists; I think we can do better here.

The naive approach is to do like radical honesty and start expressing love honestly when you feel it. Even if this were explicitly endorsed and enforced by the group (good luck overcoming *that* momentum), it still has two big issues: It requires way too much courage, and punishes people who are not comfortable with saying they are uncomfortable. This is the same sort of thing, except worse, that sinks radical honesty. Forcing the new rules on people who are not ready is *bad*.

The solution, I think is the same as the solution to these problems for radical honesty: transform the intervention from a something *forced* on people who are not ready to an opt-in protocol where people who are ready *invite* others to initiate interactions under the new system. Radical honesty becomes Crocker's rules, really awkward affection becomes Nyan's rules (or something).

(if anyone has a better name...)

So here are the rules:

1. I want to optimize the level of affection between us; I probably want more of your love.
2. To make it easier for you, I will give you feedback about what I feel comfortable with. I am ready to do this and you don't have to worry that I am secretly uncomfortable.
3. To make it safer for you, I won't punish you or hate you for going over my limits. I still expect you to respect them, but you can expect me to warn you before blowing up. (don't keep testing me tho).
4. If you reach out to me, please be comfortable with being open about your own limits. You may be suprised at how much I love you back.

What does this get us? If this works as well as I think it should, it will become a major piece of the group rationality puzzle. Rationalists should be able to build strong emotional relationships faster and better than any Dark Side cult. Is this going to work? I think it is at least worth testing.

I feel so much love just waiting for an opportunity to come out. There are many people I would love to be more open and affectionate with, but don't want to risk making them uncomfortable or ruining a friendship. I can't force this on them; all I can do is do for others what I would like them to do for me.

So if you like, try this out at your meetups. Lets see if it works. It seems safe enough, so I'll be the first to awkwardly stick my neck out and say it:

It is safe to express love or be affectionate with me, really, I won't bite.

Share Your Checklists!

Checklists are powerful, and I don't use them enough. You probably don't, either.

Below are some of my own checklists. Please share your own!

I don't know how to do X.

1. Check eHow, Google.
2. Skim-read the *For Dummies* book on the subject.
3. Check my social network for somebody who knows how to do X, ask the expert how to do X.

I don't understand X.

1. Check Wikipedia, BetterExplained, WiseGeek.
2. Read the relevant chapter(s) in a recent textbook, or find a recent review article. (See here.)
3. Check my social network for someone who understands X, ask for a tutorial. Offer to buy them coffee or lunch if necessary.

I feel mentally exhausted but can't afford to sleep right now.

1. Take a shower.
2. Watch 10 minutes of wimp.com, cats on YouTube, IGN video reviews, or movie trailers.
3. Go for a walk and listen to awesome music on high-quality headphones.

I don't want to get out of bed, but I should.

1. Imagine how good a hot shower will feel, then try again to get out of bed.
2. Set my phone alarm to go off in 5 minutes, then slide it across the floor to the other side of the room.

I'm procrastinating on task X.

1. Give the task to someone else. (Usually, this isn't possible, because I've always delegated away as much as possible.)
2. Think about which part of the procrastination equation is likely causing me the most trouble, and use one of the techniques aimed at tackling that specific problem that has worked best for me in the past.
3. Procrastinate on task X by doing a different task that is slightly less urgent/important but still productive. (See structured procrastination.)

I'm about to send an email / post a comment of some significance.

1. Is there criticism in the email or comment? Use the sandwich technique.
2. Emulate my reader(s) and predict what reaction they will have. If it's not the reaction I am aiming for with this communication, restructure the communication.

(I don't do these ones *nearly* enough! D'oh!)

I feel sad about not doing a better job at X.

1. Figure out something I can do better with regard to X, simulate in my head the steps required to execute that improvement, and if feasible then execute the improvement.
2. Think about all the things I'm doing pretty well despite running on fucked-up ape-brain software and hardware.

I'm about to make a decision of some significance.

1. Check consequentialism.
2. Check VoI. Can I improve my decision by purchasing some piece of information relatively cheaply? (This includes running checks against various biases that may be at play, performing a more formal cost-benefit analysis, etc.)
3. Sanity-check the decision with a couple people who have good decision-making skills and possess much of the relevant information.

I could go on, but... what are yours? (Now is also a good opportunity to *make* some checklists for yourself, based on what you think tends to work for you.)

Off to Alice Springs

Am about to pack up computer then go to the airport to start a sequence of flights to give this a try.

I already have a room in a hostel booked for a few nights for when I get there, and will see how stuff goes.

Anyways, since there's been on and off discussion on this, just thought I'd post that I'm actually giving this a try.

(Will likely be a day or two before I can reply/comment/etc, given length of flights, etc.)

EDIT: Ugh. You take care of one aspect of the planning fallacy, and fail elsewhere. Long story short, I missed my flight and had to reschedule it to friday.

EDIT2: Packing up computer and going off to airport. Again. This time will be early.

EDIT3: And am here. and am exhausted. :) Will start looking for work stuff tomorrow. There's a job board at this hostel, but apparently there's not much currently. But right now am rather sleep deprived.

EDIT4: So today (Monday, May 21st) went to the visitor information center. I must have misunderstood the original article, was under the impression that the visitor center had job boards. Didn't, but pointed me to a nearby recruiting/contracting agency which they said might have appropriate stuff for visitors on a work&holiday visa. Went there. said that at least as of today there's nothing, but also needed a resume (which I didn't have with me, and my work experience is limited anyways.) Anyways, got a copy of the form, will dig out/fix up what resume I do have, and also keep looking. The board at this hostel didn't have much of anything in the way of work that I saw. Will look again, though, and see if I can find others.

EDIT5 (May 29th): Still looking for work, been asking/applying to various places, including that recruiting/contracting agency, and am right now waiting (well, and still looking.) Over the weekend, though, MileyCyrus and I went on an organized 3day Uluru/Kata Tjuta/Kings Canyon trip/hikes, which was awesome. But again, as far as work, tossing out inquiries and stuff all over, trying to find out who's hiring at the moment.

Case Study: Testing Confirmation Bias

Master copy lives on gwern.net