Best of LessWrong: November 2014

1. You have a set amount of "weirdness points". Spend them wisely.
2. First(?) Rationalist elected to state government
3. The Hostile Arguer
4. Breaking the vicious cycle
5. MIRI Research Guide
6. Bayes Academy: Development report 1
7. A List of Nuances
8. Productivity through self-loyalty

Best of LessWrong: November 2014

You have a set amount of "weirdness points". Spend them wisely.

I've heard of the concept of "weirdness points" many times before, but after a bit of searching I can't find a definitive post describing the concept, so I've decided to make one. As a disclaimer, I don't think the evidence backing this post is all that strong and I am skeptical, but I do think it's strong enough to be worth considering, and I'm probably going to make some minor life changes based on it.

-

Chances are that if you're reading this post, you're probably a bit weird in some way.

No offense, of course. In fact, I actually mean it as a compliment. Weirdness is *incredibly important*. If people weren't willing to deviate from society and hold weird beliefs, we wouldn't have had the important social movements that ended slavery and pushed back against racism, that created democracy, that expanded social roles for women, and that made the world a better place in numerous other ways.

Many things we take for granted now as why our current society as great were once... *weird*.

Joseph Overton theorized that policy develops through six stages: *unthinkable*, then *radical*, then *acceptable*, then *sensible*, then *popular*, then *actual policy.* We could see this happen with many policies -- currently same-sex marriage is making its way from popular to actual policy, but not to long ago it was merely acceptable, and not too long before that it was pretty radical.

Some good ideas are currently in the *radical* range. Effective altruism itself is such a collection of beliefs typical people would consider pretty radical. Many people think donating 3% of their income is a lot, let alone the 10% demand that Giving What We Can places, or the 50%+ that some people in the community do.

And that's not all. Others would suggest that everyone become vegetarian, advocating for open borders and/or universal basic income, theabolishment of gendered language, having more resources into mitigating existential risk, focusing on research into Friendly AI, cryonicsand curing death, etc.

While many of these ideas might make the world a better place if made into policy, all of these ideas are pretty *weird*.

Weirdness, of course, is a drawback. People take weird opinions less seriously.

The absurdity heuristic is a real bias that people -- even you -- have. If an idea sounds *weird* to you, you're less likely to try and believe it,even if there's overwhelming evidence. And social proof matters -- if less people believe something,

people will be less likely to believe it. Lastly, don't forget the halo effect -- if one part of you seems weird, the rest of you will seem weird too!

(**Update:** apparently this concept is, itself, already known to social psychology as idiosyncrasy credits. Thanks, Mr. Commenter!)

...But we can use this knowledge to our advantage. The halo effect can work in reverse -- if we're normal in many ways, our weird beliefs will seem more normal too. If we have a notion of weirdness as a kind of currency that we have a limited supply of, we can spend it wisely, without looking like a crank.

All of this leads to the following actionable principles:

Recognize you only have a few "weirdness points" to spend. Trying to convince all your friends to donate 50% of their income to MIRI, become a vegan, get a cryonics plan, and demand open borders will be met with a lot of resistance. But -- I hypothesize -- that if you pick one of these ideas and push it, you'll have a lot more success.

Spend your weirdness points effectively. Perhaps it's really important that people advocate for open borders. But, perhaps, getting people to donate to developing world health would overall do more good. In that case, I'd focus on moving donations to the developing world and leave open borders alone, even though it is really important. You should triage your weirdness effectively the same way you would triage your donations.

Clean up and look good. Lookism is a problem in society, and I wish people could look "weird" and still be socially acceptable. But if you're a guy wearing a dress in public, or some punk rocker vegan advocate, recognize that you're spending your weirdness points fighting lookism, which means less weirdness points to spend promoting veganism or something else.

Advocate for more "normal" policies that are almost as good. Of course, allocating your "weirdness points" on a few issues doesn't mean you have to stop advocating for other important issues -- just consider being less weird about it. Perhaps universal basic income truly would be a very effective policy to help the poor in the United States. But reforming the earned income tax credit and relaxing zoning laws would also both do a lot to help the poor in the US, and such suggestions aren't weird.

Use the foot-in-door technique and the door-in-face technique. The foot-in-door technique involves starting with a small ask and gradually building up the ask, such as suggesting people donate a little bit effectively, and then gradually get them to take the Giving What We Can Pledge. The door-in-face technique involves making a big ask (e.g., join Giving What We Can) and then substituting it for a smaller ask, like the Life You Can Save pledge or Try Out Giving.

Reconsider effective altruism's clustering of beliefs. Right now, effective altruism is associated strongly with donating a lot of money and donating effectively, less strongly with impact in career choice, veganism, and existential risk. Of course, I'm not saying that we should drop some of these memes completely. But maybe EA should disconnect a bit more and compartmentalize -- for example, leaving AI risk to MIRI, for example, and not talk about it much, say, on 80,000 Hours. And maybe

instead of asking people to both give more AND give more effectively, we could focus more exclusively on asking people to donate what they already do more effectively.

Evaluate the above with more research. While I think the evidence base behind this is decent, it's not great and I haven't spent that much time developing it. I think we should look into this more with a review of the relevant literature and some careful, targeted, market research on the individual beliefs within effective altruism (how weird are they?) and how they should be connected or left disconnected. Maybe this has already been done some?

-

Also discussed on the EA Forum and EA Facebook group.

First(?) Rationalist elected to state government

Has no one else mentioned this on LW yet?

Elizabeth Edwards has been elected as a New Hampshire State Rep, self-identifies as a Rationalist and explicitly mentions Less Wrong in her first post-election blog post.

Sorry if this is a repost

The Hostile Arguer

“Your instinct is to talk your way out of the situation, but that is an instinct born of prior interactions with reasonable people of good faith, and inapplicable to this interaction…”

– Ken White

One of the Less Wrong Study Hall denizens has been having a bit of an issue recently. He became an atheist some time ago. His family was in denial about it for a while, but in recent days they have 1. stopped with the denial bit, and 2. been less than understanding about it. In the course of discussing the issue during break, this line jumped out at me:

“I can defend my views fine enough, just not to my parents.”

And I thought: Well, of course you can’t, because they’re not interested in your views. At all.

I never had to deal with the Religion Argument with my parents, but I did spend my fair share of time failing to argumentatively defend myself. I think I have some useful things to say to those younger and less the-hell-out-of-the-house than me.

A clever arguer is someone that has already decided on their conclusion and is making the best case they possibly can for it. A clever arguer is not necessarily interested in what you currently believe; they are arguing for proposition A and against proposition B. But there is a specific sort of clever arguer, one that I have difficulty defining explicitly but can characterize fairly easily. I call it, as of today, the Hostile Arguer.

It looks something like this:

When your theist parents ask you, “What? Why would you believe that?! We should talk about this,” they do not *actually* want to know why you believe anything, despite the form of the question. There is no genuine curiosity there. They are instead looking for *ammunition*. Which, if they are cleverer arguers than you, you are likely to provide. Unless you are epistemically *perfect*, you believe things that you cannot, on demand, come up with an explicit defense for. Even important things.

In accepting that the onus is solely on you to defend your position – which is what you are implicitly doing, in engaging the question – you are putting yourself at a disadvantage. That is the real point of the question: to bait you into an argument that your interlocutor knows you will lose, whereupon they will expect you to acknowledge defeat and toe the line they define.

Someone in the chat compared this to politics, which makes sense, but I don’t think it’s the best comparison. Politicians usually meet each other as equals. So do debate teams. This is more like a cop asking a suspect where they were on the night of X, or an employer asking a job candidate how much they made at their last job. Answering can hurt you, but can never help you. The question is inherently a trap.

The central characteristic of a hostile arguer is the insincere question. “Why do you believe there is/isn’t a God?” may be genuine curiosity from an impartial friend, or righteous fury from a zealous authority, even though the words themselves are the

same. What separates them is the response to answers. The curious friend updates their model of you with your answers; the Hostile Arguer instead updates their battle plan.[1]

So, what do you do about it?

Advice often fails to generalize, so take this with a grain of salt. It seems to me that argument in this sense has at least some of the characteristics of the Prisoner's Dilemma. Cooperation represents the pursuit of mutual understanding; defection represents the pursuit of victory in debate. Once you are aware that they are defecting, cooperating in return is highly non-optimal. On the other hand, mutual defection – a flamewar online, perhaps, or a big fight in real life in which neither party learns much of anything except how to be pissed off – kind of sucks, too. Especially if you have reason to care, on a personal level, about your opponent. If they're family, you probably do.

It seems to me that getting out of the game is the way to go, if you can do it.

Never try to defend a proposition against a hostile arguer.[2] They *do not care*. Your best arguments will fall on deaf ears. Your worst will be picked apart by people who are much better at this than you. Your insecurities will be exploited. If they have direct power over you, it will be abused.

This is especially true for parents, where obstinate disagreement can be viewed as disrespect, and where their power over you is close to absolute. I'm sort of of the opinion that all parents should be considered epistemically hostile until one moves out, as a practical application of the SNAFU Principle. If you find yourself wanting to acknowledge defeat in order to avoid imminent punishment, this is what is going on.

If you have some disagreement important enough for this advice to be relevant, you probably genuinely care about what you believe, and you probably genuinely want to be understood. On some level, you want the other party to "see things your way." So my second piece of advice is this: Accept that they *won't*, and especially accept that it will not happen as a result of anything you say in an argument. If you must explain yourself, write a blog or something and point them to it a few years later. If it's a religious argument, maybe write the Atheist Sequences. Or the Theist Sequences, if that's your bent. But don't let them make you defend yourself on the spot.

The previous point, incidentally, was my personal failure through most of my teenage years (although my difficulties stemmed from school, not religion). I *really* want to be understood, and I really approach discussion as a search for mutual understanding rather than an attempt at persuasion, by default. I expect most here do the same, which is one reason I feel so at home here. The failure mode I'm warning against is adopting this approach with people who will not respect it and will, in fact, punish your use of it.[3]

It takes two to have an argument, so don't be the second party, ever, and they will eventually get tired of talking to a wall. You are not morally obliged to justify yourself to people who have pre-judged your justifications. You are not morally obliged to convince the unconvinceable. Silence is always an option. "No comment" also works well, if repeated enough times.

There is the possibility that the other party is able and willing to punish you for refusing to engage. Aside from promoting them from "treat as Hostile Arguer" to "treat as hostile, period", I'm not sure what to do about this. Someone in the Hall

suggested supplying random, irrelevant justifications, as requiring minimal cognitive load while still subverting the argument. I'm not certain how well that will work. It sounds plausible, but I suspect that if someone is running the algorithm "punish all responses that are not 'yes, I agree and I am sorry and I will do or believe as you say'", then you're probably screwed (and should get out sooner rather than later if at all possible).

None of the above advice implies that you are right and they are wrong. You may still be incorrect on whatever factual matter the argument is about. The point I'm trying to make is that, in arguments of this form, *the argument is not really about correctness*. So if you care about correctness, don't have it.

Above all, remember this: Tapping out is not *just* for Less Wrong.

(thanks to all LWSH people who offered suggestions on this post)

After reading the comments and thinking some more about this, I think I need to revise my position a bit. I'm really talking about three different characteristics here:

1. People who have already made up their mind.
2. People who are personally invested in making you believe as they do.
3. People who have power over you.

For all three together, I think my advice still holds. MrMind puts it very concisely in the comments. In the absence of 3, though, JoshuaZ notes some good reasons one might argue anyway; to which I think one ought to add everything mentioned under the Fifth Virtue of Argument.

But one thing that ought not to be added to it is the hope of convincing the other party – either of your position, or of the proposition that you are not stupid or insane for holding it. These are cases where *you* are personally invested in what *they* believe, and all I can really say is "don't do that; it will hurt." Even if you are correct, you will fail for the reasons given above and more besides. It's very much a case of Just Lose Hope Already.

---

1. I'm using religious authorities harshing on atheists as the example here because that was the immediate cause of this post, but atheists take caution: If you're asking someone "why do you believe in God?" with the primary intent of cutting their answer down, you're guilty of this, too. ↩

2. Someone commenting on a draft of this post asked how to tell when you're dealing with a Hostile Arguer. This is the sort of micro-social question that I'm not very good at and probably shouldn't opine on. Suggestions requested in the comments. ↩

3. It occurs to me that the Gay Talk might have a lot in common with this as well. For those who've been on the wrong side of that: Did that also feel like a mismatched battle, with you trying to be understood, and them trying to break you down? ↩

Breaking the vicious cycle

You may know me as the guy who posts a lot of controversial stuff about LW and MIRI. I don't enjoy doing this and do not want to continue with it. One reason being that the debate is turning into a flame war. Another reason is that I noticed that it does affect my health negatively (e.g. my high blood pressure (I actually had a single-sided hearing loss over this xkcd comic on Friday)).

This all started in 2010 when I encountered something I perceived to be wrong. But the specifics are irrelevant for this post. The problem is that ever since that time there have been various reasons that made me feel forced to continue the controversy. Sometimes it was the urge to clarify what I wrote, other times I thought it was necessary to respond to a reply I got. What matters is that I couldn't stop. But I believe that this is now possible, given my health concerns.

One problem is that I don't want to leave possible misrepresentations behind. And there very likely exist misrepresentations. There are many reasons for this, but I can assure you that I never deliberately lied and that I never deliberately tried to misrepresent anyone. The main reason might be that I feel very easily overwhelmed and never had the ability to force myself to invest the time that is necessary to do something correctly if I don't really enjoy doing it (for the same reason I probably failed school). Which means that most comments and posts are written in a tearing hurry, akin to a reflexive retraction from the painful stimulus.

<tldr>

I hate this fight and want to end it once and for all. I don't expect you to take my word for it. So instead, here is an offer:

I am willing to post counterstatements, endorsed by MIRI, of any length and content[1] at the top of any of my blog posts. You can either post them in the comments below or send me an email (da [at] kruel.co).

</tldr>

I have no idea if MIRI believes this to be worthwhile. But I couldn't think of a better way to solve this dilemma in a way that everyone can live with happily. But I am open to suggestions that don't stress me too much (also about how to prove that I am trying to be honest).

You obviously don't need to read all my posts. It can also be a general statement.

I am also aware that LW and MIRI are bothered by RationalWiki. As you can easily check from the fossil record, I have at points tried to correct specific problems. But, for the reasons given above, I have problems investing the time to go through every sentence to find possible errors and attempt to correct it in such a way that the edit is not reverted and that people who feel offended are satisfied.

[1] There are obviously some caveats regarding the content, such as no nude photos of Yudkowsky ;-)

MIRI Research Guide

We've recently published a guide to MIRI's research on MIRI's website. It overviews some of the major open problems in FAI research, and provides reading lists for those who want to get familiar with MIRI's technical agenda.

This guide updates and replaces the MIRI course list that started me on the path of becoming a MIRI researcher over a year ago. Many thanks to Louie Helm, who wrote the previous version.

This guide is a bit more focused than the old course list, and points you not only towards prerequisite textbooks but also towards a number of relevant papers and technical reports in something approximating the "appropriate order." By following this guide, you can get yourself pretty close to the cutting edge of our technical research (barring some results that we haven't written up yet). If you intend to embark on that quest, you are invited to let me know; I can provide both guidance and encouragement along the way.

I've reproduced the guide below. The canonical version is at intelligence.org/research-guide, and I intend to keep that version up to date. This post will not be kept current.

Finally, a note on content: the guide below discusses a number of FAI research subfields. The goal is to overview, rather than motivate, those subfields. These sketches are not intended to carry any arguments. Rather, they attempt to convey our current conclusions to readers who are already extending us significant charity. We're hard at work producing a number of documents describing why we think these particular subfields are important. (The first was released a few weeks ago, the rest should be published over the next two months.) In the meantime, please understand that the research guide is not able nor intended to provide strong motivation for these particular problems.

Friendly AI theory currently isn't about implementation, it's about figuring out how to ask the right questions. Even if we had unlimited finite computing resources *and* a solid understanding of general intelligence, we still wouldn't know how to specify a system that would reliably have a positive impact during and after an intelligence explosion. Such is the state of our ignorance.

For now, MIRI's research program aims to develop solutions that assume access to unbounded finite computing power, not because unbounded solutions are feasible, but in the hope that these solutions will help us understand which questions need to be answered in order to the lay the groundwork for the eventual specification of a Friendly AI. Hence, our current research is primarily in mathematics (as opposed to software engineering or machine learning, as many expect).

This guide outlines the topics that one can study to become able to contribute to one or more of MIRI's active research areas.

Table of Contents

How to use this guide

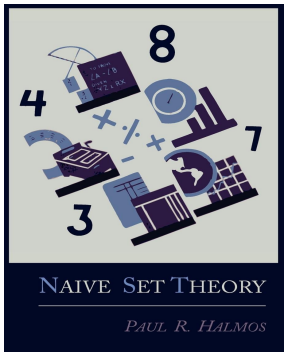
Perhaps the shortest path to being hired as a MIRI researcher is to study the materials below, then attend the nearest MIRIx workshop a few times, then attend a MIRI workshop or two and show an ability to contribute at the cutting edge. The same path (read these materials, then work your way through some workshops) will also help if you want to research these topics at some other institution.

You can learn most of the requisite material by simply reading all of the textbooks and papers below. However, with all of the material in this guide, please do not grind away for the sake of grinding away. If you already know the material, skip ahead. If one of the active research areas fails to capture your interest, switch to a different one. If you don't like one of the recommended textbooks, find a better one or skip it entirely. The goal is to get yourself to the front lines with a solid understanding of what our research says. Hopefully, this guide can help you achieve that goal, but don't let it hinder you!

The basics

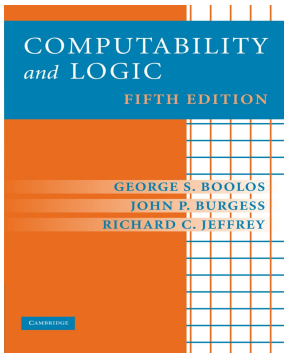
It's important to have some basic mathematical understanding before jumping directly into the active research topics. All of our research areas are well-served by a basic understanding of computation, logic, and probability theory. Below are some introductory resources to get you started.

You don't need to go through this section chronologically. Pick up whatever is interesting, and don't hesitate to skip back and forth between the research areas and the basics as necessary.



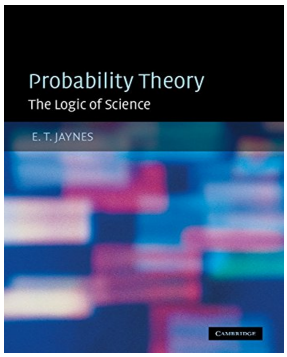
Set Theory

Most of modern mathematics is formalized in set theory, and a familiarity with set theory is a great place to start.



Computation and Logic

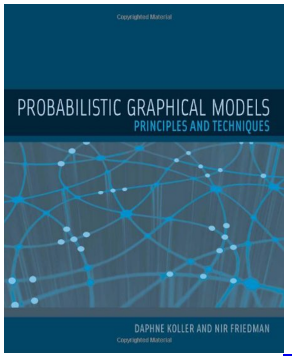
Computability (and the limits posed by diagonalization) are foundational to understanding what can and can't be done by computation.



chapters 1-4

Probability Theory

Probability theory is central to an understanding of rational agency, and an understanding of probability theory is critical in all of our active research areas.



chapters 3-5 and 9

Probabilistic graphical models

This book will help flesh out an understanding of how inference can be done using probabilistic world-models.

It's also very important to understand the concept of VNM rationality, which I recommend learning from the Wikipedia article but which can also be picked up from the original book. Von Neumann and Morgenstern showed that any agent obeying a few simple consistency axioms acts with preferences characterizable by a utility function. While many expect that we may ultimately need to abandon VNM rationality in order to construct Friendly agents, the VNM framework remains the most expressive framework we have for characterizing the behavior of sufficiently powerful agents. (For example, see the *orthogonality thesis* and the *instrumental convergence thesis* from Bostrom's "The Superintelligent Will.") The concept of VNM rationality is used throughout all our active research areas.

Corrigibility

As artificially intelligent systems grow in intelligence and capability, some of their available options may allow them to resist intervention by their programmers. We call an AI system "corrigible" if it cooperates with what its creators regard as a corrective intervention, despite default incentives for rational agents to resist attempts to shut them down or modify their preferences.

This field of research is basically brand-new, so all it takes in order to get up to speed is to read a paper or two:

1. Soares et al.'s "Corrigibility" introduces the field at large, along with a few open problems.

2. Armstrong's "Utility indifference" discusses one potential approach for making agents indifferent between which utility function they maximize, which is a small step towards agents that allow themselves to be modified.

Some early work in corrigibility was done in discussions on the web forum LessWrong. Most of the relevant results are captured in the above papers. However, additional historical work in this area can be read in the following blog posts:

1. Cake or Death outlines an example of the "motivated value selection" problem. In this example, an agent with uncertainty about its utility function benefits from avoiding information that reduces its uncertainty.

2. Proper value learning through indifference outlines the original utility indifference idea. It is largely interesting for historical reasons, and is subsumed by Armstrong's Utility Indifference paper linked above.

Our current work on corrigibility focuses mainly on a small subproblem knows as the "shutdown problem:" how do you construct an agent that shuts down upon the press of a shutdown button, and which does not have incentives to cause or prevent the pressing of the button? Within that subproblem, we currently focus on the utility indifference problem: how could you construct an agent which allows you to switch which utility function it maximizes, without giving it incentives to affect whether the switch occurs? Even if we had a satisfactory solution to the utility indifference problem, this would not yield a satisfactory solution to the shutdown problem, as it still seems difficult to adequately specify "shutdown behavior" in a manner that is immune to perverse instantiation. Stuart Armstrong has written a short series of blog posts about the specification of "reduced impact" AGIs:

1. The mathematics of reduced impact: help needed
2. Domesticating reduced impact AIs
3. Reduced impact AI: no back channels
4. Reduced impact in practice: randomly sampling the future

A satisfactory understanding of how to specify "reduced impact" AI systems has not yet been attained, but these blog posts will get you up to speed on the state of the question.

Tiling agents

An agent undergoing an intelligence explosion may need to execute many self-modifications to its core algorithms and agent architecture, one after the next. Even if the agent at the beginning of this process functioned exactly as planned, if it made a single crucial mistake in choosing one of these rewrites, the end result might be very far from the intended one.

In order to prevent this, we expect that a Friendly system would need some way to limit itself to executing self-modifications only after it has gained extremely high confidence that the resulting system would still be Friendly. Self-confidence of this form, done naively, runs afoul of mathematical problems of self-reference, and it currently seems that formal systems which can gain high self-confidence must walk a fine line between self-trust and unsoundness.

(What do we mean by "high confidence", "self-confidence", "self-trust", and "formal systems"? We don't quite know yet. Part of the problem is figuring out how to formalize these intuitive concepts in a way that avoids Gödelian pitfalls.)

The study of tiling agents is the study of agents which are able to self-modify in a highly reliable way, specifically via the study of formal systems that can gain some

form of confidence in similar systems.

Recommended reading:

1. Chang & Keisler, *Model Theory*, chs. 1-2.
    - MIRI's existing toy models for studying tiling agents are largely based on first order logic. Understanding this logic and its nuances is crucial in order to understand the existing tools we have developed for studying formal systems capable of something approaching confidence in similar systems.
2. Yudkowsky & Herreshoff, "Tiling Agents for Self-Modifying AI"
    - This paper introduces the field of tiling agents, and some of the toy formalisms and partial solutions that MIRI has developed thus far. The paper is a little choppy, but my walkthrough might help make it go down easier.
3. Yudkowsky, "The procrastination paradox"
    - The Löbian obstacle (a problem stemming from too little "self trust") described in the tiling agents paper turns out to be only half the problem: many solutions to the Löbian obstacle run afoul of unsoundnesses that come from too *much* self-trust. Satisfactory solutions will need to walk a fine line between these two problems.
4. Christiano et al., "Definability of Truth in Probabilistic Logic"
    - This describes an early attempt to create a formal system that can reason about itself while avoiding paradoxes of self-reference. It succeeds, but has ultimately been shown to be unsound. My walkthrough for this paper may help put it into a bit more context.
5. Fallenstein & Soares, "Problems of self-reference in self-improving space-time embedded intelligence"
    - This describes our simple suggester-verifier model used for studying tiling agents, and demonstrates a toy scenario in which sound agents can successfully tile to (e.g. gain high confidence in) other similar agents.

If you're really excited about this research topic, there are a number of other relevant tech reports. Unfortunately, most of them don't explain their motivations well, and have not yet been put into the greater context.

1. Fallenstein's "Procrastination in Probabilistic Logic" illustrates how Christiano et al's probabilistic reasoning system is unsound and vulnerable to the procrastination paradox.
2. Fallenstein's "Decreasing mathematical strength…" describes one unsatisfactory property of Parametric Polymorphism, a partial solution to the Lobian obstacle.
3. Soares' "Fallenstein's monster" describes a hackish formal system which avoids the above problem. It also showcases a mechanism for restricting an agent's goal predicate which can also be used by Parametric Polymorphism to create a less restrictive version of PP than the one explored in the tiling agents paper.

4. Fallenstein's "An infinitely descending sequence of sound theories…" describes a more elegant partial solution to the Lobian obstacle, which is now among our favored partial solutions.

5. Yudkowsky's "Distributions allowing tiling…" takes some early steps towards probabilistic tiling settings.

An understanding of recursive ordinals provides a useful context from which to understand these results, and can be gained by reading Franzén's "Transfinite progressions: a second look at completeness."

Logical Uncertainty

Imagine a black box, with one input chute and two output chutes. A ball can be put into the input chute, and it will come out of one of the two output chutes. Inside the black box is a Rube Goldberg machine which takes the ball from the input chute to one of the output chutes. A perfect probabilistic reasoner can be uncertain about which output chute will take the ball, but only insofar as they are uncertain about which machine is inside the black box: it is assumed that if they knew the machine (and how it worked) then they would know which chute the ball would come out. It is assumed that probabilistic reasoners are *logically omniscient*, that they know all logical consequences of the things they know.

In reality, we are not logically omniscient: we can know precisely which machine the box implements and precisely how the machine works, and just not have the time to deduce where the ball comes out. We reason under *logical uncertainty*. A formal understanding of reasoning under logical uncertainty does not yet exist, but seems necessary in the construction of a safe artificial intelligence. (Self modification involves reasoning about the unknown output of two known programs; it seems difficult to gain confidence in any reasoning system intended to do this sort of reasoning under logical uncertainty before gaining a formal understanding of idealized reasoning under logical uncertainty.)

Unfortunately, the field of logical uncertainty is not yet well-understood, and I am not aware of good textbooks introducing the material. A solid understanding of probability theory is a must; consider augmenting the first few chapters of Jaynes with Feller, chapters 1, 5, 6, and 9.

An overview of the subject can be gained by reading the following papers.

1. Gaifman, "Concerning measures in first-order calculi." Gaifman started looking at this problem many years ago, and has largely focused on a relevant subproblem, which is the assignment of probabilities to different models of a formal system (assuming that once the model is known, all consequences of that model are known). We are now attempting to expand this approach to a more complete notion of logical uncertainty (where a reasoner can know what the model is but not know the implications of that model), but work by Gaifman is still useful to gain a historical context and an understanding of the difficulties surrounding logical uncertainty. See also

    - Gaifman & Snir, "Probabilities over rich languages…"

    - Gaifman, "Reasoning with limited resources and assigning probabilities to arithmetic statements"

2. Hutter et al., "Probabilities on sentences in an expressive logic" largely looks at the problem of logical uncertainty assuming access to infinite computing power (and many levels of halting oracles). Again, we take a slightly different approach, asking how an idealized reasoner should handle logical uncertainty given unlimited but finite amounts of computing power. Nevertheless, understanding Hutter's approach (and what can be done with infinite computing power) helps flesh out one's understanding of where the difficult questions lie.

3. Demski, "Logical prior probability" provides an approximately computable logical prior. Following Demski, our work largely focuses on the creation of a prior probability distribution over logical sentences, in the hopes that understanding the creation of logical priors will lead us to a better understanding of how they could be updated, and from there a better understanding of logical uncertainty more generally.

4. Christiano, "Non-omniscience, probabilistic inference, and metamathematics" largely follows this approach. This paper provides some early practical considerations about the generation of logical priors, and highlights a few open problems.

Decision theory

We do not yet understand a decision algorithm which would, given access to unlimited finite computing power and an arbitrarily accurate world-model, always take the best available action. Intuitively, specifying such an algorithm (with respect to some VNM-rational set of preferences) may seem easy: simply loop through available actions and evaluate the expected utility achieved by taking that action, and then choose the action that yields the highest utility. In practice, however, this is quite difficult: the algorithm is in fact going to choose only one of the available actions, and in order to evaluate what "would have" happened if the agent instead took a different action that it "could have" taken requires a formalization of "would" and "could": what does it mean to say that a deterministic algorithm "could have had" a different output, and how is this circumstance (which runs counter to the laws of logic and/or physics) evaluated?

Solving this problem will require a better understanding of counterfactual reasoning; this is the domain of decision theory. Modern decision theories do not provide satisfactory methods for counterfactual reasoning, and are insufficient for use in a superintelligence. Existing methods of counterfactual reasoning turn out to be unsatisfactory both in the short term (in the sense that they fail systematically on certain classes of problems) and in the long term (in the sense that self-modifying agents reasoning using bad counterfactuals would, according to those broken counterfactuals, decide that they should not in fact fix all of their flaws): see my talk "Why aint you rich?"

We are currently in the process of writing up an introduction to decision theory as an FAI problem. In the interim, I suggest the following resources in order to understand MIRI's decision theory research:

1. Peterson's *An Introduction to Decision Theory* or Muehlhauser's "Decision Theory FAQ" explain the field of normative decision theory in broad strokes.

2. Hintze's "Problem class dominance in predictive dilemmas" contrasts four different normative decision theories: CDT, EDT, TDT, and UDT, and argues that UDT dominates the others on a certain class of decision problems.

3. Several posts by Yudkowsky and Soares explain why causal counterfactual reasoning is not sufficient for use in an intelligent agent: "Newcomb's problem and the regret of rationality," "Causal decision theory is unsatisfactory," "An introduction to Newcomblike problems," "Newcomblike problems are the norm."

Alternative decision theories have been developed which are by no means solutions, but which constitute progress. The most promising of these is Updateless Decision Theory, developed by Wei Dai and Vladimir Slepnev among others:

1. Dai's "Towards a New Decision Theory" introduces UDT.

2. Slepnev's "A model of UDT with a halting oracle" provides an early first formalization.

3. Fallenstein's alternative formalization provides a probabilistic formalization.

UDT has a number of problems of its own. Unfortunately, satisfactory write-ups detailing these problems do not yet exist. Two of the open problems have been outlined in blog posts by Vladimir Slepnev:

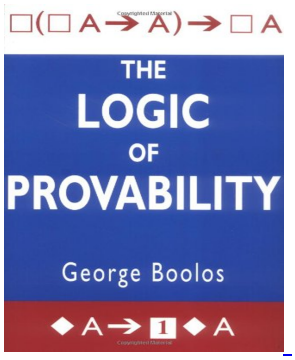
1. "An example of self-fulfilling spurious proofs in UDT" explains how UDT can achieve sub-optimal results due to spurious proofs.

2. "Agent simulates predictor" describes a strange problem wherein it seems as if agents are rewarded for having less intelligence.

A somewhat unsatisfactory solution is discussed in the following write-up by Tsvi Benson-Tilsen:

1. "UDT with known search order" contains a formalization of UDT with known proof-search order and demonstrates the necessity of playing a technique known as “playing chicken with the universe” in order to avoid spurious proofs.

In order to study multi-agent settings, Patrick LaVictoire has developed a modal agents framework, which has also allowed us to make some novel progress in the field of decision theory. To understand this, you’ll first need to understand provability logic:



chapters 1-5
(+6-9 if enthusiastic)

Provability Logic

In logical toy models of agents reflecting upon systems similar to themselves, the central question is what the parent system can prove about the child system. Our Tiling Agent research makes heavy use of provability logic, which can elegantly express these problems.

This should be sufficient to help you understand the modal agents framework:

1. Barasz et al's "Robust cooperation in the Prisoner's dilemma": roughly, this allows us to consider agents which decide whether or not to cooperate with each other based only upon what they can *prove* about each other's behavior. This prevents infinite regress, and in fact, the behavior of two agents which act only according to what they can prove about the behavior of the other can be determined in quadratic time using results from provability logic.

Many open problems in decision theory involve multi-agent settings, and in order to contribute to cutting-edge research it is also important to understand game theory. I have heard good things about the following textbook, but have not read it myself:

1. Tadelis' *Game Theory: An Introduction*.

You also may have luck with Yvain's Game Theory sequence on LessWrong.

Value learning

Perhaps the most promising approach for loading values into a powerful AI is to specify a criterion for *learning* what to value. While this problem dominates the public mindspace with regards to Friendly AI problems (if you could build an FAI, what would you have it do?), we actually find that it is somewhat less approachable than many other important problems (how do you build something stable, how do you verify its decision-making behavior, etc.). That said, a number of papers on value learning exist, and can be used to understand the current state of value learning:

1. Dewey's "Learning what to value" discusses the difficulty of the problem.

2. The orthogonality thesis further motivates why the problem will not be solved by default.

3. One approach to value learning is Bostrom & Ord's "parliamentary model," which suggests that value learning is somewhat equivalent to a voter aggregation

problem, and that many value learning systems can be modeled as parliamentary voting systems (where the voters are possible utility functions).

4. MacAskill's "Normative Uncertainty" provides a framework for discussing normative uncertainty. Be warned, the full work, while containing many insights, is very long. You can get away with skimming parts and/or skipping around some, especially if you're more excited about other areas of active research.

5. Fallenstein & Stiennon's "Loudness" discusses a concern with aggregating utility functions stemming from the fact that the preferences encoded by utility functions are preserved under positive affine transformation (e.g. as the utility function is scaled or shifted). This implies that special care is required in order to normalize the set of possible utility functions.

6. Owen Cotton-Barratt's "Geometric reasons for normalising…" discusses the normalization of utility functions.

7. De Blanc's "Ontological crises in artificial agents' value systems" discusses a separate problem in the space of value learning: how are values retained as the system's model of reality changes drastically? It seems likely that explicit resolution mechanisms will be required, but it is not yet clear how to have an agent learn values in a manner that is robust to ontological shifts.

Naturalized induction

How should an agent treat itself as if it is a part of the world? How should it learn as if it (and its sensors and its memory) are embedded in the environment, rather than sitting outside the environment? How can an agent make choices when its world-model stops modeling its own action as a fundamentally basic causal node and starts modelling it as a deterministic process resulting from a collection of transistors following physics? Many narrow AI systems assume an agent/environment separation, and we still have some confusion surrounding the nature of learners and actors that treat themselves as part of their environment.

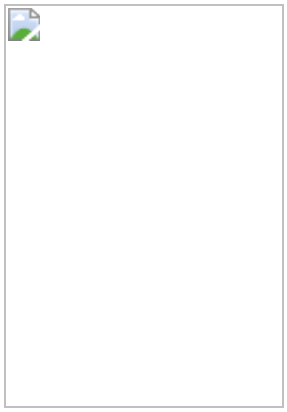
We've been referring to this as the problem of "naturalized induction,". While there has been little research done in this space, here is some reading that can help you better understand the problems:

1. Bensinger, "Naturalized induction"

Other tools

Mastery in any subject can be a very powerful tool, especially in the realm of mathematics, where seemingly disjoint topics are actually deeply connected. Many fields of mathematics have the property that if you understand them very very well, then that understanding is useful no matter where you go. With that in mind, while the subjects listed below are not necessary in order to understand MIRI's active research,

an understanding of each of these subjects constitutes an additional tool in the mathematical toolbox that will often prove quite useful when doing new research.

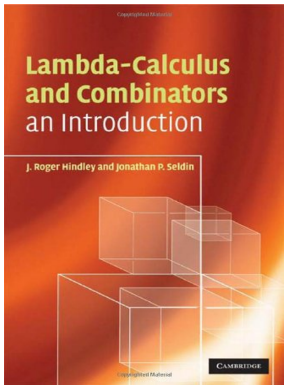


Discrete Math

Most math studies either continuous or discrete structures. Many people find discrete mathematics more intuitive, and a solid understanding of discrete mathematics will help you gain a quick handle on the discrete versions of many other mathematical tools such as group theory, topology, and information theory.

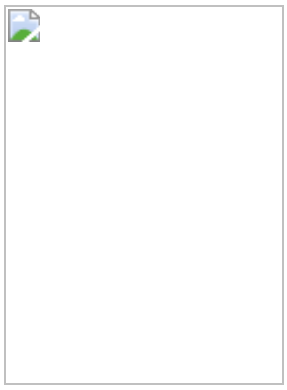
Linear Algebra

Linear algebra is one of those tools that shows up almost everywhere in mathematics. A solid understanding of linear algebra will be helpful in many domains.



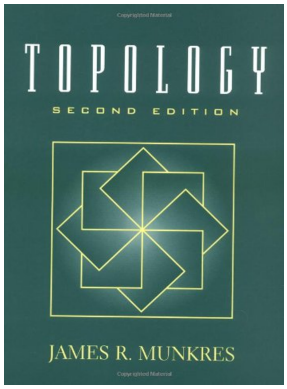
Type Theory

Set theory commonly serves as the foundation for modern mathematics, but it's not the only available foundations. Type theory can also serve as a foundation for mathematics, and in many cases, type theory is a better fit for the problems at hand. Type theory also bridges much of the theoretical gap between computer programs and mathematical proofs, and is therefore often relevant to certain types of AI research.



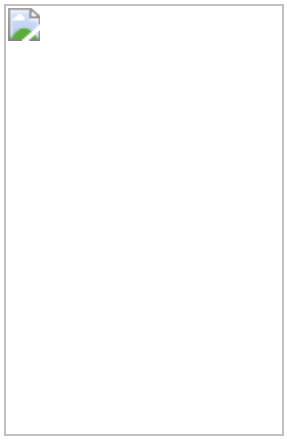
Category Theory

Category theory studies many mathematical structures at a very high level of abstraction. This can help you notice patterns in disparate branches of mathematics, and makes it much easier to transfer your mathematical tools from one domain to another.



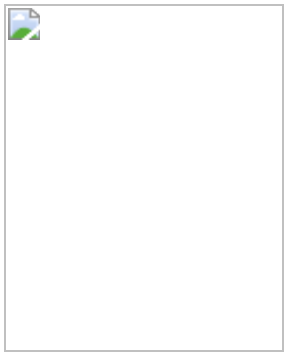
Topology

Topology is another one of those subjects that shows up pretty much everywhere in mathematics. A solid understanding of topology turns out to be helpful in many unexpected places.



Computability and Complexity

MIRI's math research is working towards solutions that will eventually be relevant to computer programs. A good intuition for what computers are capable of is often essential.

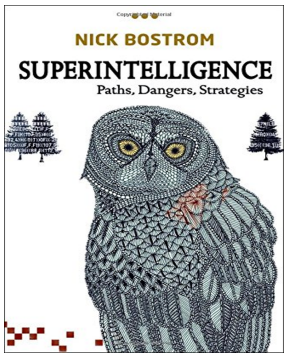


Program Verification

Program verification techniques allow programmers to become confident that a specific program will actually act according to some specification. (It is, of course, still difficult to validate that the specification describes the intended behavior.) While MIRI's work is not currently concerned with verifying real-world programs, it is quite useful to understand what modern program verification techniques can and cannot do.

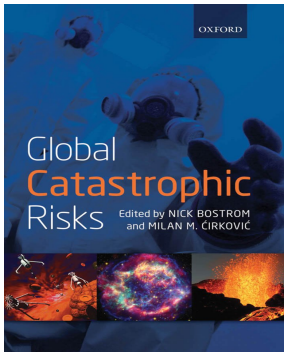
Understanding the mission

Why do this kind of research in the first place? (The first book below is the most important.)



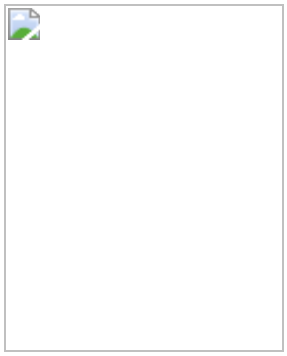
Superintelligence

This guide largely assumes that you're already on board with MIRI's mission, but if you're wondering why so many people think this is an important and urgent area of research in the first place, *Superintelligence* provides a nice overview.



Global Catastrophic Risks

But what about other global risks? How does AI compare to them? This book provides an introductory overview of the global risk landscape.



The Sequences

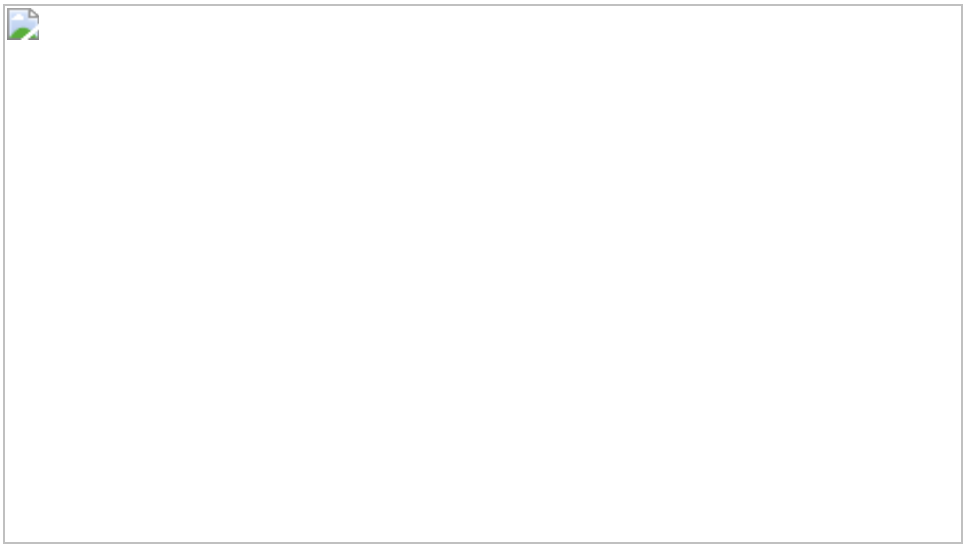
This long series of blog posts, soon to be compiled as a book, explains much of the philosophy and cognitive science motivating MIRI's paradigm for Friendly AI research.

Bayes Academy: Development report 1

Some of you may remember me proposing a game idea that went by the name of The Fundamental Question. Some of you may also remember me talking a lot about developing an educational game about Bayesian Networks for my MSc thesis, but not actually showing you much in the way of results.

Insert the usual excuses here. But thanks to SSRIs and mytomatoes.com and all kinds of other stuff, I'm now finally on track towards actually accomplishing something. Here's a report on a very early prototype.

This game has basically two goals: to teach its players something about Bayesian networks and probabilistic reasoning, and to be fun. (And third, to let me graduate by giving me material for my Master's thesis.)



We start with the main character stating that she is nervous. Hitting any key, the player proceeds through a number of lines of internal monologue:

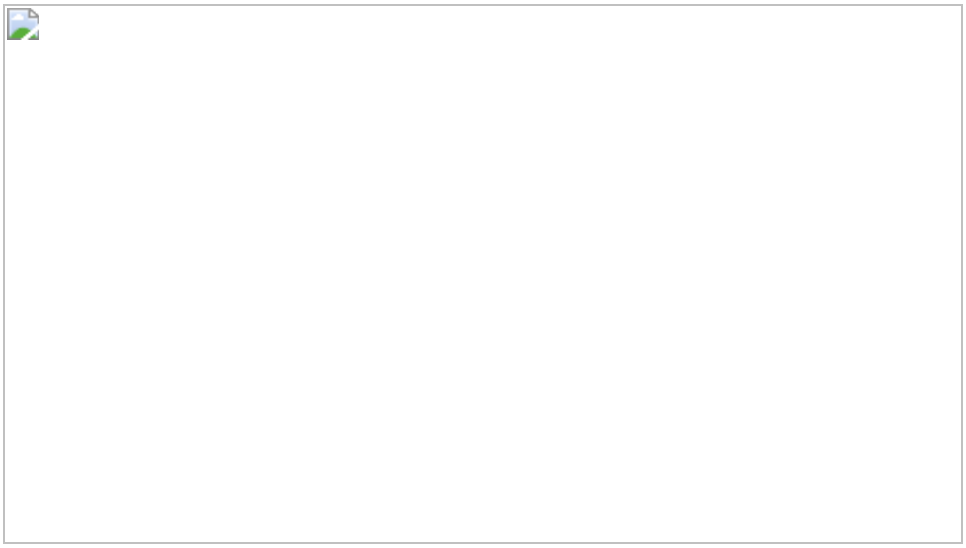
> I am nervous.
>
> I'm standing at the gates of the Academy, the school where my brother Opin was studying when he disappeared. When we asked the school to investigate, they were oddly reluctant, and told us to drop the issue.
>
> The police were more helpful at first, until they got in contact with the school. Then they actually started threatening us, and told us that we would get thrown in prison if we didn't forget about Opin.
>
> That was three years ago. Ever since it happened, I've been studying hard to make sure that I could join the Academy once I was old enough, to find out what exactly happened to Opin. The answer lies somewhere inside the Academy gates, I'm sure of it.

Now I’m finally 16, and facing the Academy entrance exams. I have to do everything I can to pass them, and I have to keep my relation to Opin a secret, too.

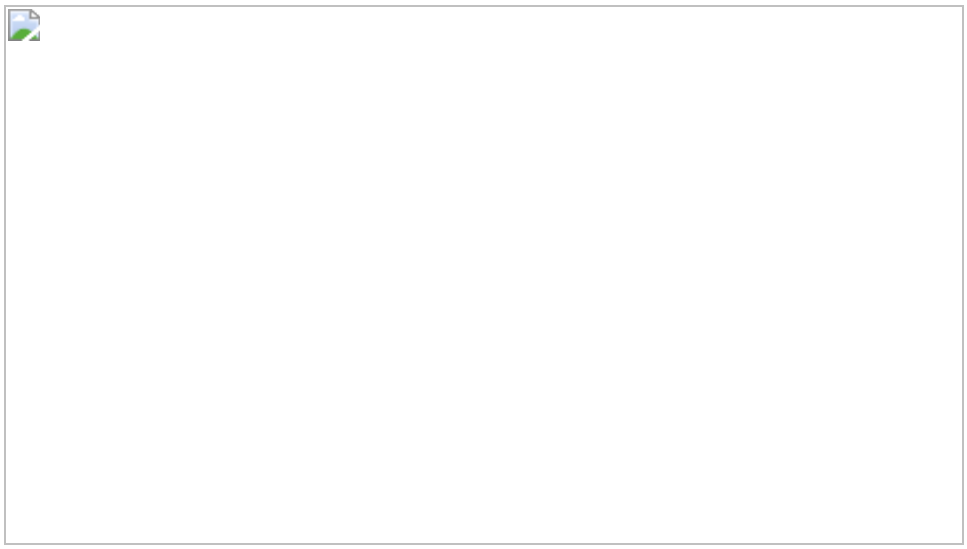
???: “Hey there.”

Eep! Someone is talking to me! Is he another applicant, or a staff member? Wait, let me think… I’m guessing that applicant would look a lot younger than staff members! So, to find that out… I should look at him!

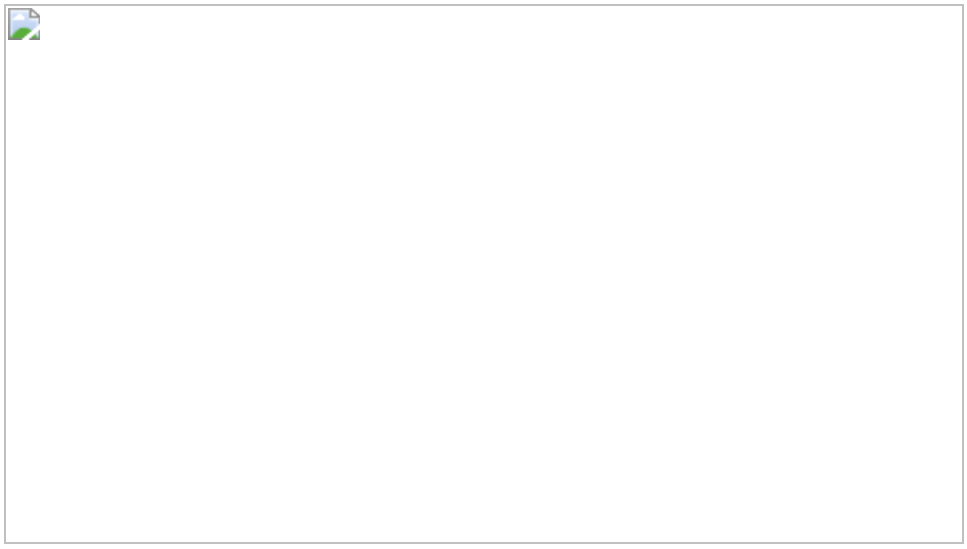
[You are trying to figure out whether the voice you heard is a staff member or another applicant. While you can't directly observe his staff-nature, you believe that he'll look young if he's an applicant, and like an adult if he's a staff member. You can look at him, and therefore reveal his staff-nature, by right-clicking on the node representing his apperance.]



Here is our very first Bayesian Network! Well, it's not really much of a network: I'm starting with the simplest possible case in order to provide an easy start for the player. We have one node that cannot be observed ("Student", its hidden nature represented by showing it in greyscale), and an observable node ("Young-looking") whose truth value is equal to that of the Student node. All nodes are binary random variables, either true or false.

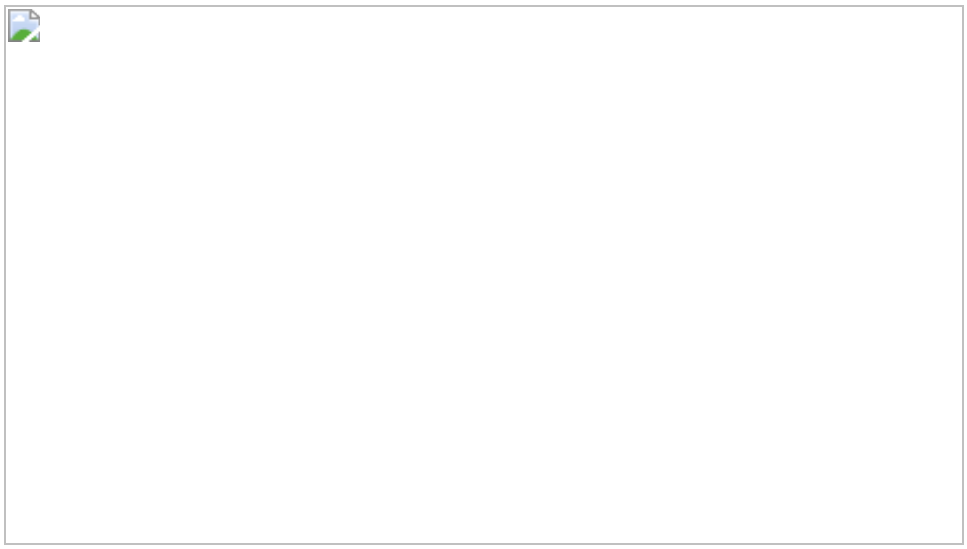


According to our current model of the world, "Student" has a 50% chance of being true, so it's half-colored in white (representing the probability of it being true) and half-colored in black (representing the probability of it being false). "Young-looking" inherits its probability directly. The player can get a bit of information about the two nodes by left-clicking on them.

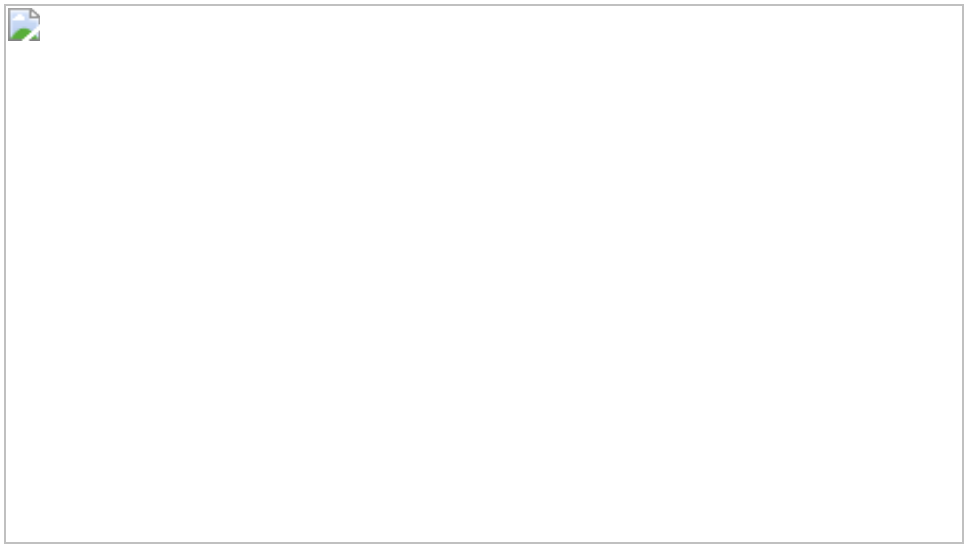


The game also offers alternate color schemes for colorblind people who may have difficulties distinguishing red and green.

Now we want to examine the person who spoke to us. Let's look at him, by right-clicking on the "Young-looking" node.



Not too many options here, because we're just getting started. Let's click on "Look at him", and find out that he is indeed young, and thus a student.



This was the simplest type of minigame offered within the game. You are given a set of hidden nodes whose values you're tasked with discovering by choosing which observable nodes to observe. Here the player had no way to fail, but later on, the minigames will involve a time limit and too many observable nodes to inspect within that time limit. It then becomes crucial to understand how probability flows within a Bayesian network, and which nodes will actually let you know the values of the hidden nodes.

The story continues!

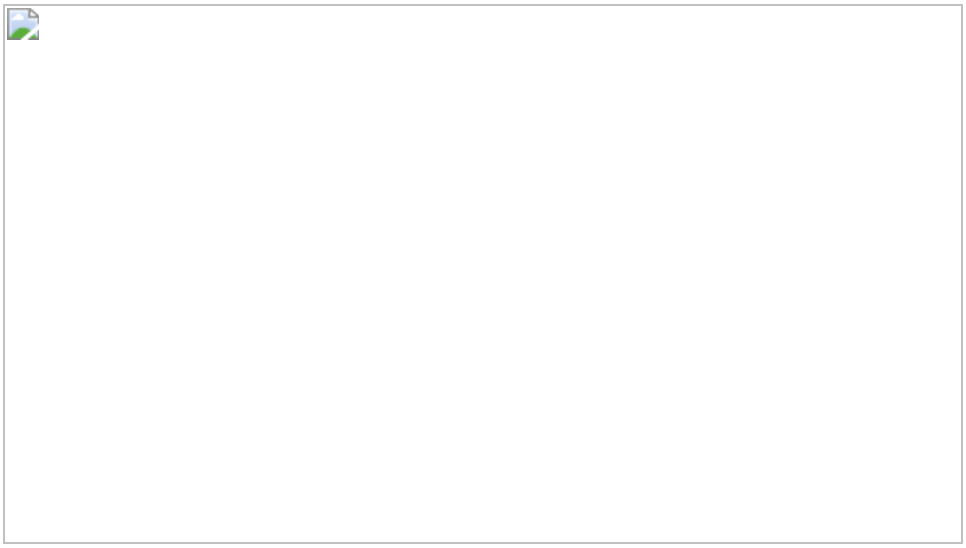
> Short for an adult, face has boyish look, teenagerish clothes... yeah, he looks young!
>
> He's a student!
>
> ...I feel like I’m overthinking things now.

...he’s looking at me.

I’m guessing he’s either waiting for me to respond, or there’s something to see behind me, and he’s actually looking past me. If there isn’t anything behind me, then I know that he must be waiting for me to respond.

Maybe there's a monster behind me, and he's paralyzed with fear! I should check that possibility before it eats me!

[You want to find out whether the boy is waiting for your reply or staring at a monster behind you. You know that he's looking at you, and your model of the world suggests that he will only look in your direction if he's waiting for you to reply, or if there's a monster behind you. So if there's no monster behind you, you know that he's waiting for you to reply!]



Slightly more complicated network, but still, there's only one option here. Oops, apparently the "Looks at you" node says it's an observable variable that you can right-click to observe, despite the fact that it's already been observed. I need to fix that.

Anyway, right-clicking on "Attacking monster" brings up a "Look behind you" option, which we'll choose.

You see nothing there. Besides trees, that is.

Boy: “Um, are you okay?”

“Yeah, sorry. I just… you were looking in my direction, and I wasn’t sure of whether you were expecting me to reply, or whether there was a monster behind me.”

He blinks.

Boy: “You thought that there was a reasonable chance for a monster to be behind you?”

I’m embarrassed to admit it, but I’m not really sure of what the probability of a monster having snuck up behind me really should have been.

My studies have entirely focused on getting into this school, and Monsterology isn’t one of the subjects on the entrance exam!

I just went with a 50-50 chance since I didn’t know any better.

'Okay, look. Monsterology is my favorite subject. Monsters avoid the Academy, since it’s surrounded by a mystical protective field. There’s no chance of them getting even near! 0 percent chance.'

'Oh. Okay.'

[Your model of the world has been updated! The prior of the variable 'Monster Near The Academy' is now 0%.]

Then stuff happens and they go stand in line for the entrance exam or something. I haven't written this part. Anyway, then things get more exciting, for a wild monster appears!

Stuff happens

AAAAAAH! A MONSTER BEHIND ME!

Huh, the monster is carrying a sword.

Well, I may not have studied Monsterology, but I sure did study fencing!

[You draw your sword. Seeing this, the monster rushes at you.]

He looks like he's going to strike. But is it really a strike, or is it a feint?

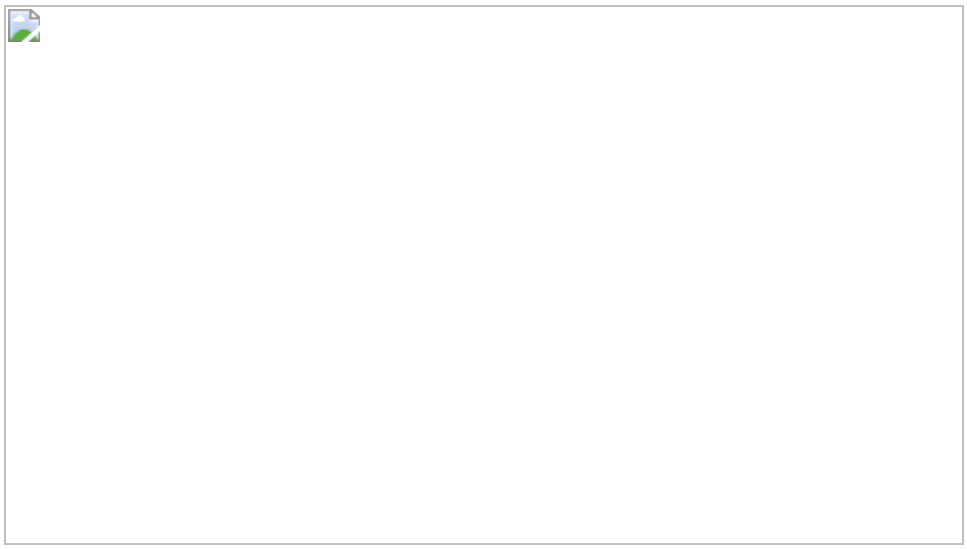
If it's a strike, I want to block and counter-attack. But if it's a feint, that leaves him vulnerable to my attack.

I have to choose wisely. If I make the wrong choice, I may be dead.

What did my master say? If the opponent has at least two of dancing legs, an accelerating midbody, and ferocious eyes, then it's an attack!

Otherwise it's a feint! Quick, I need to read his body language before it's too late!

Now get to the second type of minigame! Here, you again need to discover the values of some number of hidden variables within a time limit, but here it is in order to find out the consequences of your decision. In this one, the consequence is simple - either you live or you die. I'll let the screenshot and tutorial text speak for themselves:

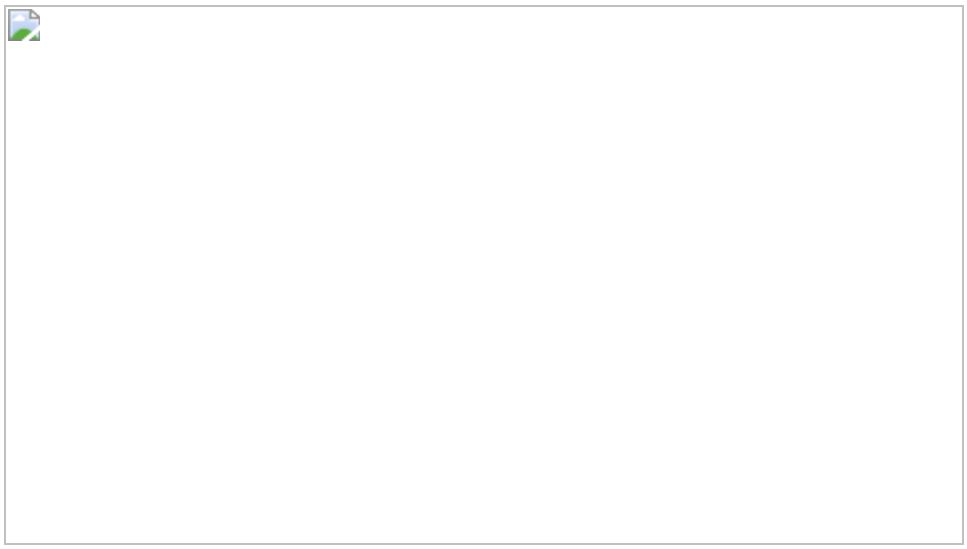


> [Now for some actual decision-making! The node in the middle represents the monster's intention to attack (or to feint, if it's false). Again, you cannot directly observe his intention, but on the top row, there are things about his body language that signal his intention. If at least two of them are true, then he intends to attack.]
>
> [Your possible actions are on the bottom row. If he intends to attack, then you want to block, and if he intends to feint, you want to attack. You need to inspect his body language and then choose an action based on his intentions. But hurry up! Your third decision must be an action, or he'll slice you in two!]

In reality, the top three variables are not really independent of each other. We want to make sure that the player can always win this battle despite only having three actions. That's two actions for inspecting variables, and one action for actually making a decision. So this battle is rigged: either the top three variables are all true, or they're all false.

...actually, now that I think of it, the order of the variables is wrong. Logically, the body language should be caused by the intention to attack, and not vice versa, so the arrows should point from the intention to body language. I'll need to change that. I got these mixed up because the prototypical exemplar of a decision minigame is one where you need to predict someone's reaction from their personality traits, and there the personality traits do cause the reaction. Anyway, I want to get this post written before I go to bed, so I won't change that now.

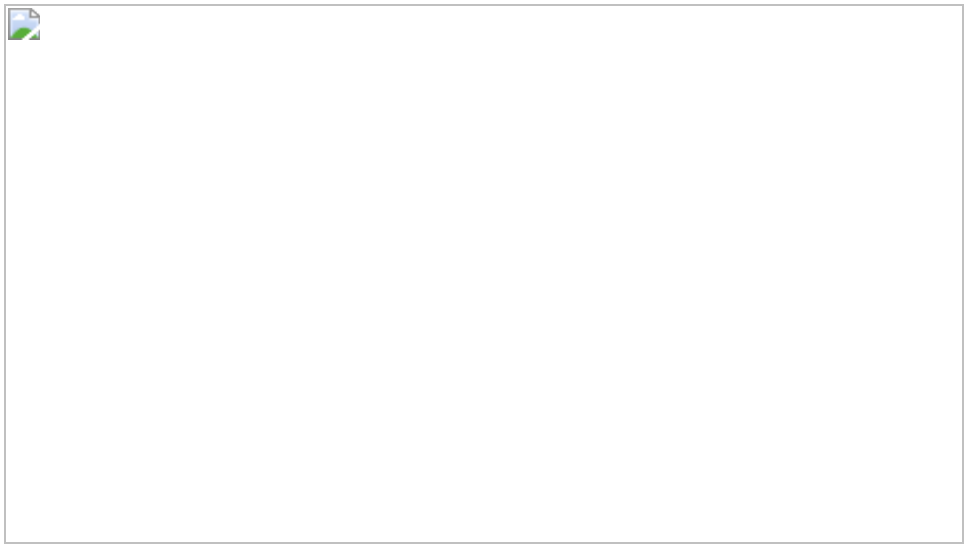
Right-clicking "Dancing legs", we now see two options besides "Never mind"!



We can find out the dancingness of the enemy's legs by thinking about our own legs - we are well-trained, so our legs are instinctively mirroring our opponent's actions to prevent them from getting an advantage over us - or by just instinctively feeling where they are, without the need to think about them! Feeling them would allow us to observe this node without spending an action.

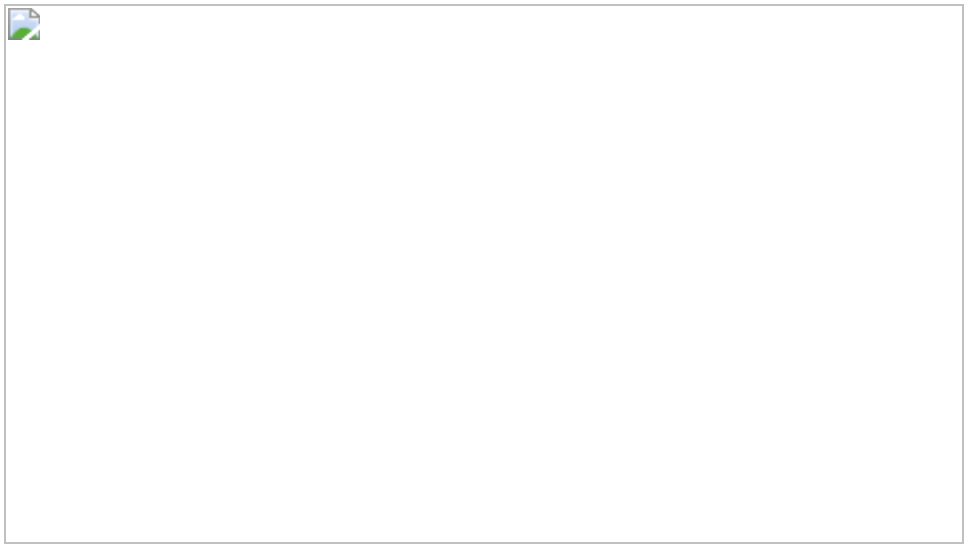
Unfortunately, feeling them has "Fencing 2" as a prerequisite skill, and we don't have that. Neither could we have them, in this point of the game. The option is just there to let the player know that there are skills to be gained in this game, and make them look forward to the moment when they can actually gain that skill. As well as giving them an idea of how the skill can be used.

Anyway, we take a moment to think of our legs, and even though our opponent gets closer to us in that time, we realize that our legs our dancing! So his legs must be dancing as well!

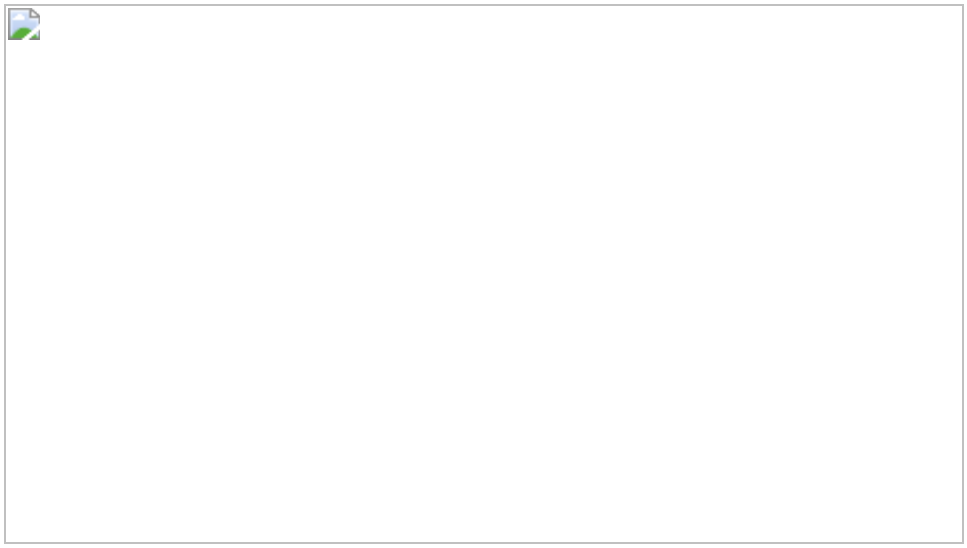


With our insider knowledge, we now know that he's attacking, and we could pick "Block" right away. But let's play this through. The network has automatically

recalculated the probabilities to reflect our increased knowledge, and is now predicting a 75% chance for our enemy to be attacking, and for "Blocking" to thus be the right decision to make.



Next we decide to find out what his eyes say, by matching our gaze with his. Again, there would be a special option that cost us no time - this time around, one enabled by Empathy 1 - but we again don't have that option.



Except that his gaze is so ferocious that we are forced to look away! While we are momentarily distracted, he closes the distance, ready to make his move. But now we know what to do... block!

Success!

Now the only thing that remains to do is to ask our new-found friend for an explanation.

"You told me there was a 0% chance of a monster near the academy!"

Boy: “Ehh… yeah. I guess I misremembered. I only read like half of our course book anyway, it was really boring.”

“Didn’t you say that Monsterology was your favorite subject?”

Boy: “Hey, that only means that all the other subjects were even more boring!”

“. . .”

I guess I shouldn’t put too much faith on what he says.

[Your model of the world has been updated! The prior of the variable 'Monster Near The Academy' is now 50%.]

[Your model of the world has been updated! You have a new conditional probability variable: 'True Given That The Boy Says It's True', 25%]

And that's all for now. Now that the basic building blocks are in place, future progress ought to be much faster.

Notes:

As you might have noticed, my "graphics" suck. A few of my friends have promised to draw art, but besides that, the whole generic Java look could go. This is where I was originally planning to put in the sentence "and if you're a Java graphics whiz and want to help fix that, the current source code is conveniently available at GitHub", ~~but then getting things to his point took longer than I expected and I didn't have the time to actually figure out how the whole Eclipse-GitHub integration works. I'll get to that soon.~~ Github link here!

I also want to make the nodes more informative - right now they only show their marginal probability. Ideally, clicking on them would expand them to a representation where you could visually see what components their probability composed of. I've got some scribbled sketches of what this should look like for various node types, but none of that is implemented yet.

I expect some of you to also note that the actual Bayes theorem hasn't shown up yet, at least in no form resembling the classic mammography problem. (It is used implicitly in the network belief updates, though.) That's intentional - there will be a third minigame involving that form of the theorem, but somehow it felt more natural to start this way, to give the player a rough feeling of how probability flows through Bayesian networks. Admittedly I'm not sure of how well that's happening so far, but hopefully more minigames should help the player figure it out better.

What's next? Once the main character (who needs a name) manages to get into the Academy, there will be a lot of social scheming, and many mysteries to solve in order for her to find out just what did happen to her brother... also, I don't mind people suggesting things, such as what could happen next, and what kinds of network configurations the character might face in different minigames.

(Also, everything that you've seen might get thrown out and rewritten if I decide it's no good. Let me know what you think of the stuff so far!)

A List of Nuances

Abram Demski and Grognor

Much of rationality is pattern-matching. An article on lesswrong might point out a thing to look for. Noticing this thing changes your reasoning in some way. This essay is a list of things to look for. These things are all associated, but the reader should take care not to lump them together. Each dichotomy is distinct, and although the brain will tend to abstract them into some sort of yin/yang correlated mush, in reality they have a more complicated structure; some things may be similar, but if possible, try to focus on the complex interrelationships.

1. Map vs. Territory
    1. Eliezer's sequences use this as a jump-off point for discussion of rationality.
    2. Many thinking mistakes are map vs. territory confusions.
        1. A map and territory mistake is a mix-up of seeming vs being.
        2. Humans need frequent reminders that we are not omniscient.
2. Cached Thoughts vs. Thinking
    1. This document is a list of cached thoughts.
3. Clusters vs. Properties
    1. These words could be used in different ways, but the distinction I want to point at is that of labels we put on things vs actual differences in things.
    2. The mind projection fallacy is the fallacy of thinking a mental category (a "cluster") is an actual property things have.
        1. If we see something as good for one reason, we are likely to attribute other good properties to it, as if it had inherent goodness. This is called the halo effect. (If we see something as bad and infer other bad properties as a result, it is referred to as the reverse-halo effect.)
    3. Categories are inference applicability heuristics; ruling X an instance of Y without expecting novel inferences is cargo cult classification.
4. Syntax vs. Semantics
    1. The syntax is the physical instantiation of the map. The semantics is the way we are meant to read the map; that is, the intended relationship to the territory.
5. Semantics vs. Pragmatics
    1. The semantics is the literal contents of a message, whereas the pragmatics is the intended result of conveying the message.
        1. An example of a message with no semantics and only pragmatics is a command, such as "Stop!".
        2. Almost no messages lack pragmatics, and for good reason. However, if you seek truth in a discussion, it is important to foster a willingness to say things with less pragmatic baggage.
        3. Usually when we say things, we do so with some "point" which is beyond the semantics of our statement. The point is usually to build up or knock down some larger item of discussion. This is not inherently a bad thing, but has a failure mode where arguments are battles and statements are weapons, and the cleverer arguer wins.
    2. The meaning of a thing is the way you should be influenced by it.
6. Object-level vs. Meta-level
    1. The difference between making a map and writing a book about map-making.

    2. A good meta-level theory helps get things right at the object level, but it is usually impossible to get things right at the meta level before before you've made significant progress at the object level.
7. Seeming vs. Being
    1. We can only deal with how things seem, not how they are. Yet, we must strive to deal with things as they are, not as they seem.
        1. This is yet another reminder that we are not omniscient.
    2. If we optimize too hard for things which seem good rather than things which are good, we will get things which seem *very* good but which may only be somewhat good, or even bad.
    3. The dangerous cases are the cases where you do not notice there is a distinction.
        1. This is why humans need constant reminders that we are not omniscient.
    4. We must take care to notice the difference between how things seem to seem, and how they actually seem.
8. Signal vs. Noise
    1. Not all information is equal. It is often the case that we desire certain sorts of information and desire to ignore other sorts.
    2. In a technical setting, this has to do with the error rate present in a communication channel; imperfections in the channel will corrupt some bits, making a need for redundancy in the message being sent.
    3. In a social setting, this is often used to refer to the amount of good information vs irrelevant information in a discussion. For example, letting a mediocre writer add material to a group blog might increase the absolute amount of good information, yet worsen the signal-to-noise ratio.
    4. Attention is a scarce resource; yes everyone has something to teach you, but many people are much more efficient sources of wisdom than others.
9. Selection Effects
    1. Filtered evidence.
        1. In many situations, if we can present evidence to a Bayesian agent without the agent knowing that we are being selective, we can convince the agent of anything we like. For example, if I want to convince you that smoking causes obesity, I could find many people who became obese after they started smoking.
        2. The solution to this is for the Bayesian agent to model where the information is coming from. If you know I am selecting people based on this criteria, then you will not take it as evidence of anything, because the evidence has been cherry-picked.
        3. Most of the information you receive is intensely filtered. Nothing comes to your attention with a good conscience.
    2. The silent evidence problem.
        1. Selection bias need not be the result of purposeful interference as in cherry-picking. Often, an unrelated process may hide some of the evidence needed. For example, we hear far more about successful people than unsuccessful. It is tempting to look at successful people and attempt to draw conclusion about what it takes to be successful. This approach suffers from the silent evidence problem: we also need to look at the unsuccessful people and examine what is *different* about the two groups.
    3. Observer selection effects.
10. What You Mean vs. What You Think You Mean
    1. Very often, people will say something and then that thing will be refuted. The common response to this is to claim you meant something slightly

different, which is more easily defended.
        1. We often do this without noticing, making it dangerous for thinking. It is an automatic response generated by our brains, not a conscious decision to defend ourselves from being discredited. You do this far more often than you notice. The brain fills in a false memory of what you meant without asking for permission.
11. What You Mean vs. What the Others Think You Mean
    1. The illusion of transparency.
    2. The double illusion of transparency.
    3. Wiio's Laws
12. What You Optimize vs. What You Think You Optimize
    1. Evolution optimizes for reproduction but in doing so creates animals with a variety of goals which are correlated with reproduction.
    2. Extrinsic motivation is weaker than intrinsic motivation.
    3. The people who value practice for its own sake do better than the people who only value being good at what they're practicing.
    4. "Consequentialism is true, but virtue ethics is what works."
13. Stated Preferences vs. Revealed Preferences
    1. Revealed preferences are the preferences we can infer from your actions. These are usually different from your stated preferences.
        1. X is not about Y:
            1. Food isn't about nutrition.
            2. Clothes aren't about comfort.
            3. Bedrooms aren't about sleep.
            4. Marriage isn't about love.
            5. Talk isn't about information.
            6. Laughter isn't about humour.
            7. Charity isn't about helping.
            8. Church isn't about God.
            9. Art isn't about insight.
            10. Medicine isn't about health.
            11. Consulting isn't about advice.
            12. School isn't about learning.
            13. Research isn't about progress.
            14. Politics isn't about policy.
            15. Going meta isn't about the object level.
            16. Language isn't about communication.
            17. The rationality movement isn't about epistemology.
        2. Everything is actually about signalling.
    2. Humans Are Not Automatically Strategic
        1. Never attribute to malice that which can be adequately explained by stupidity. The difference between stated preferences and revealed preferences does not indicate dishonest intent. We should expect the two to differ in the absence of a mechanism to align them.
        2. Hidden Motives vs. Innocent Failure
    3. People, ideas, and organizations respond to incentives.
        1. Evolution selects humans who have reproductively selfish behavioral tendencies, but prosocial and idealistic stated preferences.
            1. Near vs. Far
        2. Social forces select ideas for virality and comprehensibility as opposed to truth or even usefulness.
            1. Motte-and-bailey fallacy
        3. Organizations are by default bad at being strategic about their own survival, but the ones that survive are the ones you see.

14. What You Achieve vs. What You Think You Achieve
    1. Most of the consequences of our actions are totally unknown to us.
    2. It is impossible to optimize without proper feedback.
15. What You Optimize vs. What You Actually Achieve
    1. Consequentialism is more about expected consequences than actual consequences.
16. What You Seem Like vs. What You Are
    1. You can try to imagine yourself from the outside, but no one has the full picture.
17. What Other People Seem Like vs. What They Are
    1. When people assume that they understand others, they are wrong.
18. What People Look Like vs. What They Think They Look Like
    1. People underestimate the gap between stated preferences and revealed preferences.
19. What Your Brain Does vs. What You Think It Does
    1. You are running on corrupted hardware.
        1. The brain's machinations are fundamentally social; it automatically does things like signal, save face, etc., which distort the truth.
    2. The reverse of stupidity is not intelligence.
        1. Knowing that you are running on corrupted hardware should cause skepticism about the outputs of your thought-processes. Yet, too much skepticism will cause you to stumble, particularly when fast thinking is needed.
            1. Producing a correct result plus justification is harder than producing only the correct result.
            2. Justifications are important, but the correct result is more important.
            3. Much of our apparent self-reflection is confabulation, generating plausible explanations after the brain spits out an answer.
            4. Example: doing quick mental math. If you are good at this, attempting to explicitly justify every step as you go would likely slow you down.
            5. Example: impressions formed over a long period of time. Wrong or right, it is unlikely that you can explicitly give all your reasons for the impression. Requiring your own beliefs to be justifiable would preempt impressions that require lots of experience and/or many non-obvious chains of subconscious inference.
            6. Impressions are not beliefs and they are always useful data.
20. Clever Argument vs. Truth-seeking; The Bottom Line
    1. People believe what they want to believe.
        1. Believing X for some reason unrelated to X being true is referred to as motivated cognition.
        2. Giving a smart person more information and more methods of argument may actually make their beliefs less accurate, because you are giving them more tools to construct clever arguments for what they want to believe.
    2. Your actual reason for believing X determines how well your belief correlates with the truth.
        1. If you believe X because you want to, any arguments you make for X no matter how strong they sound are devoid of informational context about X and should properly be ignored by a truth-seeker.

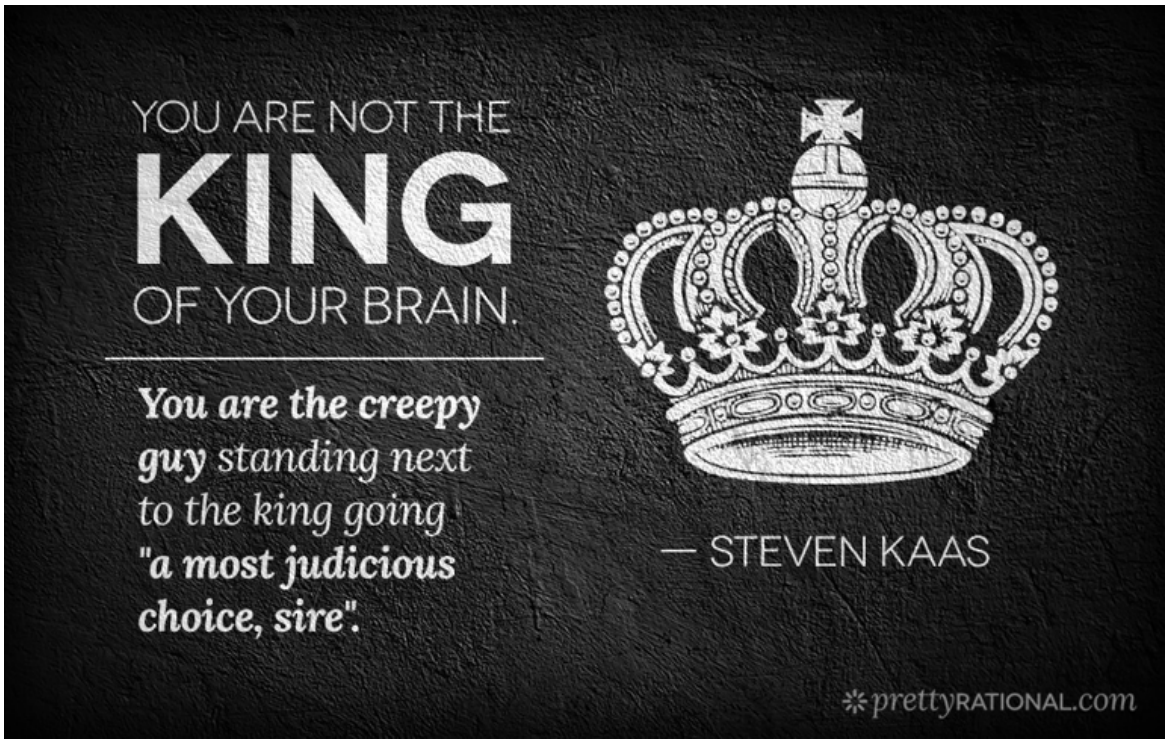
    3. If you believe true things when doing so improves your life, that is no credit to you at all. Everyone does that.
21. Lumpers vs. Splitters
    1. A lumper is a thinker who attempts to fit things into overarching patterns. A splitter is a thinker who makes as many distinctions as possible, recognizing the importance of being specific and getting the details right.
    2. Specifically, some people want big Wikipedia and TVTropes articles that discuss many things, and others want smaller articles that discuss fewer things.
    3. This list of nuances is a lumper attempting to think more like a splitter.
22. Fox vs. Hedgehog
    1. “A fox knows many things, but a hedgehog knows One Big Thing.” Closely related to a splitter, a fox is a thinker whose strength is in a broad array of knowledge. A hedgehog is a thinker who, in contrast, has one big idea and applies it everywhere.
    2. The fox mindset is better for making accurate judgements, according to Tetlock.
23. Traps vs. Gardens
    1. Well-kept gardens die by pacifism.
        1. Conversations tend to slide toward contentious and useless topics.
        2. Societies tend to decay.
        3. Systems in general work poorly or not at all.
        4. Thermodynamic equilibrium is entropic.
        5. Without proper institutions being already in place, it takes large amounts of constant effort and vigilance to stay out of traps.
    2. From the outside of a broken Molochian system it is easy to see how to fix. But it cannot be fixed from the inside.

Productivity through self-loyalty

1

I can be pretty dang productive when I put my mind to it.

Many people have a generic mind-model which runs roughly as follows: a person's reported desires are but one voice among the thronging mob of forces that govern the brain, and it takes significant effort and force of will to align the mob for long enough for people to get something done. Many of us have experienced a desire to stop procrastinating, and then have watched helplessly as we continue to surf the internet. Many of us have resolved to do something difficult, only to watch the opportunity flit by us as we stand motionless at the sidelines.



People use something like this model when they speak of akrasia, the tendency to act against your own better judgement. Haidt analogizes the brain to someone riding an elephant, where the conscious mind is a rider struggling to steer. Kahneman writes of a dichotomy between "fast," emotional, immediate processes that govern most of our thinking, and "slow," deliberate, conscious processes that occasionally assume command. I have found that the "spoon theory" model of energy reserves resonates for many people, even those who aren't chronically ill or otherwise disabled.

In all these models, there is a tendency to separate the voice from the mob. Insofar as the voice has the ability to direct the mob (steer the elephant, convince system 1, etc.), we get to do what we want. But when the mob loses interest or focus or motivation, we are at it's leisure.

I find a lot of truth in these models, and so do many others. Thus, many people, upon seeing my high levels of productivity, expect that I must be very very good at keeping tight control

over the mental mob, and forcing them to do things that they would rather not do. It's not uncommon for people to remark that I need to be careful about strong-arming the mob (as eventually they rebel, leading to burn-out), or for people to tell me that I must have some sort of iron will (which they cannot replicate).

I don't think this is the case. As I said last time:

> *A problem isn't solved until it's solved automatically, without need for attention or willpower.*

It's *possible* to force the mob to do something, and this is why willpower is often useful in the short term. But it's seldom a good idea to try to force yourself to do things the mob doesn't want to do in the long term. Ultimately, the mob is the one actually managing your motivation systems, and any plan that relies upon a permanent use of mental force is unlikely to succeed.

It is much better to have the mob on the same side as the voice of reason.

But this is something of a catch 22: many people have mind-mobs that just want to sit around all day and watch TV shows or surf the internet. If your mind-mob just wants to rest and I'm cautioning against force, then how does one ever attain high levels of productivity?

My answer is complex, and relies upon many tools. I've discussed a few of them in the past, and today I'll discuss another.

2

First, a word of warning: remember the law of equal and opposite advice. For every piece of advice useful to one person, there is some other person who needs exactly the opposite advice.

I am going to discuss a technique that I use for productivity which results in a sense of austerity through compassion/camaraderie: the parts of me that need rest take as much rest as they need, but also try to take as little as they need out of awareness of the scarcity of resources and compassion for the other parts of me.

This has proved a powerful technique for me, but it may be exactly the wrong tool for many others. The goal is *not* to guilt-trip the parts of you that need extra rest, and the goal is *not* to give yourself over to self-indulgent whims. I personally find a lot of power somewhere in between, at "compasionate austerity," but many others may react poorly to any internal narrative of scarce resources and mental frugality. *Remember the law of equal and opposite advice.*

3

Imagine a student who has been assigned a very important bit of homework with a deadline looming ever closer. Let's say they're trying to kick themselves into high productivity mode. How can they do this? Well, they can pull out the whips and cattle prods and *force* their mind-mob to be productive (with gritted teeth and building malcontent), or they can use their most desperate voice and plead with themselves, promising rewards for good behavior (that the mob might just take anyway, if it suits them), or they can wait until the deadline is so close that even the short-sighted mob can see it, at which point they'll go into panic mode (which is kinda like high productivity mode, if you squint).

But there's also a fourth option, which is something like "gain the trust of the mob, and build rapport." If the student gets the mob onto *their* side, then the paper will be done

automatically, no willpower or pleading or panic necessary. This obviously sounds nice, but how is it done?

I do this, at least in part, by *showing the mob that I am on their side first.* This involves self-signalling, as discussed in last week's post. Specifically, it involves signaling to yourself that *you are loyal to the mob.*

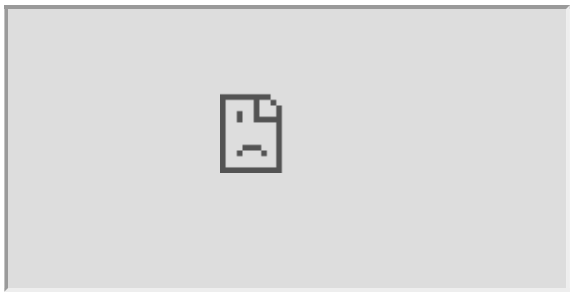
Sometimes, the mob in you will make demands that sound unreasonable, such as "cancel everything today, I need a break." In these situations, it's easy to try to force or plea or bargain with yourself. I take a different tactic: I ask myself if this is really what I need, and if it is, *then I do it.*

I show the mob that I respect its demands, and that I'm on its side. After all, we have the same goals; and furthermore, I am not the king in my mind. I do not desire a fight (and if I did, I wouldn't win it).

There are some really bad ways to do this (remember the law of equal and opposite advice!), and if you do this incorrectly it may lead to destructive self-indulgence. If your voice of reason signals helplessness in the face of the mob's whims, if it *gives itself up* to the mob, then you might end up unhappily pursuing short-sighted whims. The trick is to signal *respect* for the mob instead: what my mind reports it needs, it gets. This—an unflinching willingness to get the mob what it wants—*tempers* the mob's demands.

The appropriate sentiment can perhaps best be described by this clip from the film *It's a Wonderful Life*:

(start at 2:58, watch through 6:26)



4

This scene portrays a bank run during the beginning of the great depression. It features the protagonist, George Bailey, trying to calm down a worried mob by reminding them that they're all in this crisis together. The mob doesn't really go for it, and he ends up using his honeymoon money to keep the bank alive.

The first member of the mob to get his money out of the bank demands the full value of his account, $242. George pleads for austerity, reminding him that they're all in this together, but Tom still demands all his money. George doesn't protest or argue, he just nods and pays out Tom's entire account (and then extends the man a little extra compassion, to boot). The next two members of the mob say they can get away with $20, and are starting to express some concern for George using his own money for this. Then Mrs. Davis bids *lower*, asking for only $17.50. Overcome, George gives her a kiss on the cheek.

This is the sort of relationship—between George Bailey and the mob—that I have the "voice of reason" cultivate with the varied and disparate parts of my mind. When some part of me demands that I pay its full account, I'll ask it once how much it *needs,* but if it still demands

its full account I'll pay up without hesitation (and extend some additional compassion). This is done not in an appeasing way, but it a respectful way: we're all in this together.

The mob understands that the voice of reason is responsible for many of the good outcomes that I've achieved, and the mob understands that things like "rest" and "relaxation" and "procrastination" are expensive in terms of ability to achieve good outcomes—I'm "paying out of the honeymoon money."

But the voice of reason, in turn, is *willing* to pay out of its honeymoon money. It knows that everyone is going to need some resources to make it through, and does not begrudge any part of me for that.

There are two important components to this sort of self-relationship: First, the mob must respect the voice of reason, by understanding that the voice of reason achieves many nice things, such as food and roofs and clever schemes and so on. Second, the mob must know that the voice of reason is *loyal to them.* When some mind part *does* demand something ostentatious, such as "a few days of doing nothing," then the voice of reason is willing to acquire it.

My loyalty is not to any individual appointment or task. My own mental health is among my top priorities.

Once the mob sees this, once the mob *knows* that I will move the heavens and the earth in order to meet its needs, it doesn't tend to demand the full account. Because, in fact, the mob *respects* the scarcity of scarce resources, it *wants* the voice of reason to have enough flexibility to keep on achieving good outcomes. Done right, the mob enters a sort of camaraderie where it takes as little as it can out of *compassion,* because we all know that life can be hard.

When Mrs. Davis leaves that bank with $17.50, she isn't feeling resentment or smugness. She knows that she's going to have to struggle a bit to live on only $17.50 until the bank re-opens, but she isn't dreading the struggle or muttering curses. No, she goes home filled with compassion, with respect for George Bailey who is taking great pains to get everyone through this crisis together, and with a tighter feeling of community and closeness to those around her enduring similar austerity. She goes home happy and warm.

5

This is the mindstate in which I attain high productivity: various parts of the mob of my mind occasionally need rest, recuperation, and procrastination. Parts of me ask for these things. When they do, I ask them how much they really need, how much they can get by with. Do I actually need to take four days off? Because I will, but it's expensive.

Often, when a part of me really needs a break, and throws up its hands feeling overwhelmed, its initial demands are unrealistic—"two weeks with no responsibilities!" So then I ask it again, with the demeanor of George Bailey, what it really needs to get by. And that part of me quickly remembers that all of me is in this together, and that I'm trying to do some very difficult things, and that all parts of me are constrained by scarce resources. Then the part that protested searches for what it really needs, the bare minimum, and it usually answers something like "I can get the rest I need in fifteen minutes."

And this sacrifice can leave me feeling stronger, feeling warmth and compassion and self-camaraderie, the same feeling that spurs George Bailey to kiss Mrs. Davis' cheek in the video clip above.

6

There is only so much time and attention that we have in this world, and we're trying to do many amazing and wonderful things. If you want to be able to do more than you're currently doing, I don't suggest trying to force yourself. Instead, I suggest showing yourself that you really are willing to move the heaven and earth *for yourself,* in order to satisfy your needs. This, in turn, can help you build up the mental camaraderie (and resulting austerity) that comes from all the parts of you understanding that you're all in this together.