# Best of LessWrong: March 2015

# Best of LessWrong: March 2015

# Preface

You hold in your hands a compilation of two years of daily blog posts. In retrospect, I look back on that project and see a large number of things I did completely wrong. I'm fine with that. Looking back and *not* seeing a huge number of things I did wrong would mean that neither my writing nor my understanding had improved since 2009. *Oops* is the sound we make when we improve our beliefs and strategies; so to look back at a time and not see anything you did wrong means that you haven't learned anything or changed your mind since then.

It was a mistake that I didn't write my two years of blog posts with the intention of helping people do better in their everyday lives. I wrote it with the intention of helping people solve big, difficult, important problems, and I chose impressive-sounding, abstract problems as my examples.

In retrospect, this was the second-largest mistake in my approach. It ties in to the *first*-largest mistake in my writing, which was that I didn't realize that the big problem in learning this valuable way of thinking was figuring out how to practice it, not knowing the theory. I didn't realize that part was the priority; and regarding this I can only say "Oops" and "Duh."

Yes, sometimes those big issues really are big and really are important; but that doesn't change the basic truth that to master skills you need to practice them and it's harder to practice on things that are further away. (Today the Center for Applied Rationality is working on repairing this huge mistake of mine in a more systematic fashion.)

A third huge mistake I made was to focus too much on rational belief, too little on rational action.

The fourth-largest mistake I made was that I should have better organized the content I was presenting in the sequences. In particular, I should have created a wiki much earlier, and made it easier to read the posts in sequence.

*That* mistake at least is correctable. In the present work Rob Bensinger has reordered the posts and reorganized them as much as he can without trying to rewrite all the actual material (though he's rewritten a bit of it).

My fifth huge mistake was that I—as I saw it—tried to speak plainly about the stupidity of what appeared to me to be stupid ideas. I did try to avoid the fallacy known as Bulverism, which is where you *open* your discussion by talking about how stupid people are for believing something; I would always discuss the issue first, and only afterwards say, "And so this is stupid." But in 2009 it was an open question in my mind whether it might be important to have some people around who expressed contempt for homeopathy. I thought, and still do think, that there is an unfortunate problem wherein treating ideas courteously is processed by many people on some level as "Nothing bad will happen to me if I say I believe this; I won't lose status if I say I believe in homeopathy," and that derisive laughter by comedians can help people wake up from the dream.

Today I would write more courteously, I think. The discourtesy did serve a function, and I think there were people who were helped by reading it; but I now take more

seriously the risk of building communities where the normal and expected reaction to low-status outsider views is open mockery and contempt.

Despite my mistake, I am happy to say that my readership has so far been amazingly good about *not* using my rhetoric as an excuse to bully or belittle others. (I want to single out Scott Alexander in particular here, who is a nicer person than I am and an increasingly amazing writer on these topics, and may deserve part of the credit for making the culture of *Less Wrong* a healthy one.)

To be able to look backwards and say that you've "failed" implies that you had goals. So what was it that I was trying to do?

There is a certain valuable way of thinking, which is not yet taught in schools, in this present day. This certain way of thinking is not taught systematically at all. It is just absorbed by people who grow up reading books like *Surely You're Joking, Mr. Feynman* or who have an unusually great teacher in high school.

Most famously, this certain way of thinking has to do with science, and with the experimental method. The part of science where you go out and look at the universe instead of just making things up. The part where you say "Oops" and give up on a bad theory when the experiments don't support it.

But this certain way of thinking extends beyond that. It is deeper and more universal than a pair of goggles you put on when you enter a laboratory and take off when you leave. It applies to daily life, though this part is subtler and more difficult. But if you can't say "Oops" and give up when it looks like something isn't working, you have no choice but to keep shooting yourself in the foot. You have to keep reloading the shotgun and you have to keep pulling the trigger. You know people like this. And somewhere, someplace in your life you'd rather not think about, you *are* people like this. It would be nice if there was a certain way of thinking that could help us stop doing that.

In spite of how large my mistakes were, those two years of blog posting appeared to help a surprising number of people a surprising amount. It didn't work reliably, but it worked sometimes.

In modern society so little is taught of the skills of rational belief and decision-making, so little of the mathematics and sciences underlying them . . . that it turns out that just reading through a massive brain-dump full of problems in philosophy and science can, yes, be surprisingly good for you. Walking through all of that, from a dozen different angles, can sometimes convey a glimpse of the central rhythm.

Because it is all, in the end, one thing. I talked about big important distant problems and neglected immediate life, but the laws governing them aren't actually different. There are huge gaps in which parts I focused on, and I picked all the wrong examples; but it is all in the end one thing. I am proud to look back and say that, even after all the mistakes I made, and all the other times I said "Oops" . . .

Even five years later, it still appears to me that this is better than nothing.

—Eliezer Yudkowsky, February 2015

# Biases: An Introduction

Imagine reaching into an urn that contains seventy white balls and thirty red ones, and plucking out ten mystery balls.

Perhaps three of the ten balls will be red, and you'll correctly guess how many red balls total were in the urn. Or perhaps you'll happen to grab four red balls, or some other number. Then you'll probably get the total number wrong.

This random error is the cost of incomplete knowledge, and as errors go, it's not so bad. Your estimates won't be incorrect *on average*, and the more you learn, the smaller your error will tend to be.

On the other hand, suppose that the white balls are heavier, and sink to the bottom of the urn. Then your sample may be unrepresentative in a consistent direction.

*That* kind of error is called "statistical bias." When your method of learning about the world is biased, learning more may not help. Acquiring more data can even consistently *worsen* a biased prediction.

If you're used to holding knowledge and inquiry in high esteem, this is a scary prospect. If we want to be sure that learning more will help us, rather than making us worse off than we were before, we need to discover and correct for biases in our data.

The idea of *cognitive bias* in psychology works in an analogous way. A cognitive bias is a systematic error in *how we think*, as opposed to a random error or one that's merely caused by our ignorance. Whereas statistical bias skews a sample so that it less closely resembles a larger population, cognitive biases skew our thinking so that it less accurately tracks the truth (or less reliably serves our other goals).

Maybe you have an optimism bias, and you find out that the red balls can be used to treat a rare tropical disease besetting your brother, and you end up overestimating how many red balls the urn contains because you *wish* the balls were mostly red.

Like statistical biases, cognitive biases can distort our view of reality, they can't always be fixed by just gathering more data, and their effects can add up over time. But when the miscalibrated measuring instrument you're trying to fix is *you*, debiasing is a unique challenge.

Still, this is an obvious place to start. For if you can't trust your brain, how can you trust anything else?

## Noticing Bias

Imagine meeting someone for the first time, and knowing nothing about them except that they're shy.

Question: Is it more likely that this person is a librarian, or a salesperson?

Most people answer "librarian." Which is a mistake: shy salespeople are much more common than shy librarians, because salespeople in general are much more common than librarians—seventy-five times as common, in the United States.[1]

This is *base rate neglect*: grounding one's judgments in how well sets of characteristics feel like they fit together, and neglecting how common each characteristic is in the population at large.[2] Another example of a cognitive bias is the *sunk cost fallacy*—people's tendency to feel committed to things they've spent resources on in the past, when they should be cutting their losses and moving on.

Knowing about these biases, unfortunately, doesn't make you immune to them. It doesn't even mean you'll be able to notice them in action.

In a study of *bias blindness*, experimental subjects predicted that they would have a harder time neutrally evaluating the quality of paintings if they knew the paintings were by famous artists. And indeed, these subjects exhibited the very bias they had predicted when the experimenters later tested their prediction. When asked *afterward*, however, the very same subjects claimed that their assessments of the paintings had been objective and unaffected by the bias.[3]

Even when we correctly identify others' biases, we exhibit a *bias blind spot* when it comes to our own flaws.[4] Failing to detect any "biased-feeling thoughts" when we introspect, we draw the conclusion that we must just be less biased than everyone else.[5]

Yet it *is* possible to recognize and overcome biases. It's just not trivial. It's known that subjects can reduce base rate neglect, for example, by thinking of probabilities as frequencies of objects or events.

The approach to debiasing in this book is to communicate a systematic understanding of *why good reasoning works*, and of how the brain falls short of it. To the extent this volume does its job, its approach can be compared to the one described in Serfas (2010), who notes that "years of financially related work experience" didn't affect people's susceptibility to the sunk cost bias, whereas "the number of accounting courses attended" did help.

> As a consequence, it might be necessary to distinguish between experience and expertise, with expertise meaning "the development of a schematic principle that involves conceptual understanding of the problem," which in turn enables the decision maker to recognize particular biases. However, using expertise as countermeasure requires more than just being familiar with the situational content or being an expert in a particular domain. It requires that one fully understand the underlying rationale of the respective bias, is able to spot it in the particular setting, and also has the appropriate tools at hand to counteract the bias.[6]

The goal of this book is to lay the groundwork for creating rationality "expertise." That means acquiring a deep understanding of the structure of a very general problem: human bias, self-deception, and the thousand paths by which sophisticated thought can defeat itself.

## A Word About This Text

*Map and Territory* began its life as a series of essays by decision theorist Eliezer Yudkowsky, published between 2006 and 2009 on the economics blog *Overcoming Bias* and its spin-off community blog Less Wrong. Thematically linked essays were grouped together in "sequences," and thematically linked sequences were grouped into books. *Map and Territory* is the first of six such books, with the series as a whole going by the name *Rationality: From AI to Zombies*.[7]

In style, this series run the gamut from “lively textbook” to “compendium of vignettes” to “riotous manifesto,” and the content is correspondingly varied. The resultant rationality primer is frequently personal and irreverent—drawing, for example, from Yudkowsky’s experiences with his Orthodox Jewish mother (a psychiatrist) and father (a physicist), and from conversations on chat rooms and mailing lists. Readers who are familiar with Yudkowsky from *Harry Potter and the Methods of Rationality*, his science-oriented take-off of J.K. Rowling’s *Harry Potter* books, will recognize the same iconoclasm, and many of the same themes.

The philosopher Alfred Korzybski once wrote: “A map *is not* the territory it represents, but, if correct, it has a *similar structure* to the territory, which accounts for its usefulness.” And what can be said of maps here, as Korzybski noted, can also be said of beliefs, and assertions, and words.

“The map is not the territory.” This deceptively simple claim is the organizing idea behind this book, and behind the four sequences of essays collected here: **Predictably Wrong**, which concerns the systematic ways our beliefs fail to map the real world; **Fake Beliefs**, on what makes a belief a “map” in the first place; **Noticing Confusion**, on how this world-mapping thing our brains do actually works; and **Mysterious Answers**, which collides these points together. The book then concludes with “The Simple Truth,” a stand-alone dialogue on the idea of truth itself.

Humans aren’t rational; but, as behavioral economist Dan Ariely notes, we’re *predictably* irrational. There are patterns to how we screw up. And there are patterns to how we behave when we *don’t* screw up. Both admit of fuller understanding, and with it, the hope of leaning on that understanding to build a better future for ourselves.

---

[1] Wayne Weiten, *Psychology: Themes and Variations, Briefer Version, Eighth Edition* (Cengage Learning, 2010).

[2] Richards J. Heuer, *Psychology of Intelligence Analysis* (Center for the Study of Intelligence, Central Intelligence Agency, 1999) .

[3] Katherine Hansen et al., “People Claim Objectivity After Knowingly Using Biased Strategies,” *Personality and Social Psychology Bulletin* 40, no. 6 (2014): 691–699 .

[4] Emily Pronin, Daniel Y. Lin, and Lee Ross, “The Bias Blind Spot: Perceptions of Bias in Self versus Others,” *Personality and Social Psychology Bulletin* 28, no. 3 (2002): 369–381 .

[5] Joyce Ehrlinger, Thomas Gilovich, and Lee Ross, “Peering Into the Bias Blind Spot: People’s Assessments of Bias in Themselves and Others,” *Personality and Social Psychology Bulletin* 31, no. 5 (2005): 680–692.

[6] Sebastian Serfas, *Cognitive Biases in the Capital Investment Context: Theoretical Considerations and Empirical Experiments on Violations of Normative Rationality* (Springer, 2010).

[7] The first edition of *Rationality: From AI to Zombies* was released as a single sprawling ebook, before the series was edited and split up into separate volumes. The full book can also be found on http://lesswrong.com/rationality.

# Calibration Test with database of 150,000+ questions

Hi all,

I put this calibration test together this morning. It pulls from a trivia API of over 150,000 questions so you should be able to take this many, many times before you start seeing repeats.

http://www.2pih.com/caltest.php

A few notes:

1. The questions are "Jeopardy" style questions so the wording may be strange, and some of them might be impossible to answer without further context. On these just assign 0% confidence.

2. As the questions are open-ended, there is no answer-checking mechanism. You have to be honest with yourself as to whether or not you got the right answer. Because what would be the point of cheating at a calibration test?

I can't think of anything else. Please let me know if there are any features you would want to see added, or if there are any bugs, issues, etc.

****EDIT****

As per suggestion I have moved this to the main section. Here are the changes I'll be making soon:

- Label the axes and include an explanation of calibration curves.
- Make it so you can reverse your last selection in the event of a misclick.

Here are changes I'll make eventually:

- Create an account system so you can store your results online.
- Move trivia DB over to my own server to allow for flagging of bad/unanswerable questions.

**Here are the changes that are done:**

- *Change 0% to 0.1% and 99% to 99.9%*
- *Added second graph which shows the frequency of your confidence selections.*
- *Color code the "right" and "wrong" buttons and make them farther apart to prevent misclicks.*
- *Store your results locally so that you can see your calibration over time.*
- *Check to see if a question is blank and skip if so.*

# A map of LWers - find members of the community living near you.

There seems to be a lot of enthusiasm around LessWrong meetups, so I thought something like this might be interesting too. There is no need to register - just add your marker and keep an eye out for someone living near you.

Here's the link: https://www.zeemaps.com/map?group=1323143

I posted this on an Open Thread first. Below are some observations based on the previous discussion:

When creating a new marker you will be given a special URL you can use to edit it later. If you lose it, you can create a new one and ask me to delete the old marker. Try not to lose it though.

If someone you tried to contact is unreachable, notify me and I'll delete the marker in order to keep the map tidy. Also, try to keep your own marker updated.

It was suggested that it would be a good idea to circulate the map around survey time. I'll try to remind everyone to update their markers around that time. Any major changes (e.g. changing admin, switching services, remaking the map to eliminate dead markers) will also happen then.

The map data can be exported by anyone, so there's no need to start over if I disappear or whatever.

Edit: Please, you have to make it possible to contact you. If you choose to use a name that doesn't match your LW account, you have to add an email address or equivalent. If you don't do that, it is assumed that the name on the marker is your username here, but if it isn't you are essentially unreachable and will be removed.

# Rationality: From AI to Zombies online reading group

Update: When I posted this announcement I remarkably failed to make the connection that the April 15th is tax day here in the US, and as a prime example of the planning fallacy (a topic of the first sequence!), I failed to anticipate just how complicated my taxes would be this year. The first post of the reading group is basically done but a little rushed, and I want to take an extra day to get it right. Expect it to post on the next day, the 16th

On **Thursday, 16 April 2015**, just under a month out from this posting, I will hold the first session of an online reading group for the ebook Rationality: From AI to Zombies, a compilation of the LessWrong sequences by our own Eliezer Yudkowsky. I would like to model this on the very successful Superintelligence reading group led by KatjaGrace. This is an advanced warning, so that you can have a chance to get the ebook, make a donation to MIRI, and read the first sequence.

The point of this online reading group is to join with others to ask questions, discuss ideas, and probe the arguments more deeply. It is intended to add to the experience of reading the sequences in their new format or for the first time. It is intended to supplement discussion that has already occurred the original postings and the sequence reruns.

The reading group will 'meet' on a semi-monthly post on the LessWrong discussion forum. For each 'meeting' we will read one sequence from the the *Rationality* book, which contains a total of 26 lettered sequences. A few of the sequences are unusually long, and these might be split into two sessions. If so, advance warning will be given.

In each posting I will briefly summarize the salient points of the essays comprising the sequence, link to the original articles and discussion when possible, attempt to find, link to, and quote one or more related materials or opposing viewpoints from outside the text, and present a half-dozen or so question prompts to get the conversation rolling. Discussion will take place in the comments. Others are encouraged to provide their own question prompts or unprompted commentary as well.

We welcome both newcomers and veterans on the topic. If you've never read the sequences, this is a great opportunity to do so. If you are an old timer from the Overcoming Bias days then this is a chance to share your wisdom and perhaps revisit the material with fresh eyes. All levels of time commitment are welcome.

If this sounds like something you want to participate in, then please grab a copy of the book and get started reading the preface, introduction, and the 10 essays / 42 pages which comprise *Part A: Predictably Wrong*. The first virtual meeting (forum post) covering this material will go live before 6pm Thursday PDT (1am Friday UTC), 16 April 2015. Successive meetings will start no later than 6pm PDT on the first and third Wednesdays of a month.

Following this schedule it is expected that it will take just over a year to complete the entire book. If you prefer flexibility, come by any time! And if you are coming upon

this post from the future, please feel free leave your opinions as well. The discussion period never closes.

Topic for the first week is the preface by Eliezer Yudkowsky, the introduction by Rob Bensinger, and *Part A: Predictably Wrong*, a sequence covering rationality, the search for truth, and a handful of biases.

# New forum for MIRI research: Intelligent Agent Foundations Forum

Today, the Machine Intelligence Research Institute is launching a new forum for research discussion: the Intelligent Agent Foundations Forum! It's already been seeded with a bunch of new work on MIRI topics from the last few months.

We've covered most of the (what, why, how) subjects on the forum's new welcome post and the How to Contribute page, but this post is an easy place to comment if you have further questions (or if, maths forbid, there are technical issues *with* the forum instead of *on* it).

But before that, go ahead and check it out!

(Major thanks to Benja Fallenstein, Alice Monday, and Elliott Jin for their work on the forum code, and to all the contributors so far!)

**EDIT 3/22:** Jessica Taylor, Benja Fallenstein, and I wrote forum digest posts summarizing and linking to recent work (on the IAFF and elsewhere) on reflective oracle machines, on corrigibility, utility indifference, and related control ideas, and on updateless decision theory and the logic of provability, respectively! These are pretty excellent resources for reading up on those topics, in my biased opinion.

# Chapter 1: A Day of Very Low Probability

Disclaimer: J. K. Rowling owns Harry Potter, and no one owns the methods of rationality.

This fic is widely considered to have really hit its stride starting at around Chapter 5. If you still don't like it after Chapter 10, give up.

This is *not* a strict single-point-of-departure fic - there exists a primary point of departure, at some point in the past, but also other alterations. The best term I've heard for this fic is "parallel universe".

The text contains many clues: obvious clues, not-so-obvious clues, truly obscure hints which I was shocked to see some readers successfully decode, and massive evidence left out in plain sight. This is a rationalist story; its mysteries are solvable, and meant to be solved.

The pacing of the story is that of serial fiction, i.e., that of a TV show running for a predetermined number of seasons, whose episodes are individually plotted but with an overall arc building to a final conclusion.

All science mentioned is real science. But please keep in mind that, beyond the realm of science, the views of the characters may not be those of the author. Not everything the protagonist does is a lesson in wisdom, and advice offered by darker characters may be untrustworthy or dangerously double-edged.

---

*Beneath the moonlight glints a tiny fragment of silver, a fraction of a line...*

*(black robes, falling)*

*...blood spills out in litres, and someone screams a word.*

---

Every inch of wall space is covered by a bookcase. Each bookcase has six shelves, going almost to the ceiling. Some bookshelves are stacked to the brim with hardback books: science, maths, history, and everything else. Other shelves have two layers of paperback science fiction, with the back layer of books propped up on old tissue boxes or lengths of wood, so that you can see the back layer of books above the books in front. And it still isn't enough. Books are overflowing onto the tables and the sofas and making little heaps under the windows.

This is the living-room of the house occupied by the eminent Professor Michael Verres-Evans, and his wife, Mrs. Petunia Evans-Verres, and their adopted son, Harry James Potter-Evans-Verres.

There is a letter lying on the living-room table, and an unstamped envelope of yellowish parchment, addressed to *Mr. H. Potter* in emerald-green ink.

The Professor and his wife are speaking sharply at each other, but they are not shouting. The Professor considers shouting to be uncivilised.

"You're joking," Michael said to Petunia. His tone indicated that he was very much afraid that she was serious.

"My sister was a witch," Petunia repeated. She looked frightened, but stood her ground. "Her husband was a wizard."

"This is absurd!" Michael said sharply. "They were at our wedding - they visited for Christmas -"

"I told them you weren't to know," Petunia whispered. "But it's true. I've seen things -"

The Professor rolled his eyes. "Dear, I understand that you're not familiar with the sceptical literature. You may not realise how easy it is for a trained magician to fake the seemingly impossible. Remember how I taught Harry to bend spoons? If it seemed like they could always guess what you were thinking, that's called cold reading -"

"It wasn't bending spoons -"

"What was it, then?"

Petunia bit her lip. "I can't just tell you. You'll think I'm -" She swallowed. "Listen. Michael. I wasn't - always like this -" She gestured at herself, as though to indicate her lithe form. "Lily did this. Because I - because I *begged* her. For years, I begged her. Lily had *always* been prettier than me, and I'd... been mean to her, because of that, and then she got *magic,* can you imagine how I felt? And I *begged* her to use some of that magic on me so that I could be pretty too, even if I couldn't have her magic, at least I could be pretty."

Tears were gathering in Petunia's eyes.

"And Lily would tell me no, and make up the most ridiculous excuses, like the world would end if she were nice to her sister, or a centaur told her not to - the most ridiculous things, and I hated her for it. And when I had just graduated from university, I was going out with this boy, Vernon Dursley, he was fat and he was the only boy who would talk to me. And he said he wanted children, and that his first son would be named Dudley. And I thought to myself, *what kind of parent names their child Dudley Dursley?* It was like I saw my whole future life stretching out in front of me, and I couldn't stand it. And I wrote to my sister and told her that if she didn't help me I'd rather just -"

Petunia stopped.

"Anyway," Petunia said, her voice small, "she gave in. She told me it was dangerous, and I said I didn't care any more, and I drank this potion and I was sick for weeks, but when I got better my skin cleared up and I finally filled out and... I was beautiful, people were *nice* to me," her voice broke, "and after that I couldn't hate my sister any more, especially when I learned what her magic brought her in the end -"

"Darling," Michael said gently, "you got sick, you gained some weight while resting in bed, and your skin cleared up on its own. Or being sick made you change your diet -"

"She was a witch," Petunia repeated. "I saw it."

"Petunia," Michael said. The annoyance was creeping into his voice. "You *know* that can't be true. Do I really have to explain why?"

Petunia wrung her hands. She seemed to be on the verge of tears. "My love, I know I can't win arguments with you, but please, you have to trust me on this -"

"*Dad! Mum!*"

The two of them stopped and looked at Harry as though they'd forgotten there was a third person in the room.

Harry took a deep breath. "Mum, *your* parents didn't have magic, did they?"

"No," Petunia said, looking puzzled.

"Then no one in your family knew about magic when Lily got her letter. How did *they* get convinced?"

"Ah..." Petunia said. "They didn't just send a letter. They sent a professor from Hogwarts. He -" Petunia's eyes flicked to Michael. "He showed us some magic."

"Then you don't have to fight over this," Harry said firmly. Hoping against hope that this time, just this once, they would listen to him. "If it's true, we can just get a Hogwarts professor here and see the magic for ourselves, and Dad will admit that it's true. And if not, then Mum will admit that it's false. That's what the experimental method is for, so that we don't have to resolve things just by arguing."

The Professor turned and looked down at him, dismissive as usual. "Oh, come now, Harry. Really, *magic?* I thought *you'd* know better than to take this seriously, son, even if you're only ten. Magic is just about the most unscientific thing there is!"

Harry's mouth twisted bitterly. He was treated well, probably better than most genetic fathers treated their own children. Harry had been sent to the best primary schools - and when that didn't work out, he was provided with tutors from the endless pool of starving students. Always Harry had been encouraged to study whatever caught his attention, bought all the books that caught his fancy, sponsored in whatever maths or science competitions he entered. He was given anything reasonable that he wanted, except, maybe, the slightest shred of respect. A Doctor teaching biochemistry at Oxford could hardly be expected to listen to the advice of a little boy. You would listen to Show Interest, of course; that's what a Good Parent would do, and so, if you conceived of yourself as a Good Parent, you would do it. But take a ten-year-old *seriously?* Hardly.

Sometimes Harry wanted to scream at his father.

"Mum," Harry said. "If you want to win this argument with Dad, look in chapter two of the first book of the Feynman Lectures on Physics. There's a quote there about how philosophers say a great deal about what science absolutely requires, and it is all wrong, because the only rule in science is that the final arbiter is observation - that you just have to look at the world and report what you see. Um... off the top of my head I can't think of where to find something about how it's an ideal of science to settle things by experiment instead of arguments -"

His mother looked down at him and smiled. "Thank you, Harry. But -" her head rose back up to stare at her husband. "I don't want to win an argument with your father. I want my husband to, to listen to his wife who loves him, and trust her just this once -"

Harry closed his eyes briefly. *Hopeless.* Both of his parents were just hopeless.

Now his parents were getting into one of *those* arguments again, one where his mother tried to make his father feel guilty, and his father tried to make his mother feel stupid.

"I'm going to go to my room," Harry announced. His voice trembled a little. "Please try not to fight too much about this, Mum, Dad, we'll know soon enough how it comes out, right?"

"Of course, Harry," said his father, and his mother gave him a reassuring kiss, and then they went on fighting while Harry climbed the stairs to his bedroom.

He shut the door behind him and tried to think.

The funny thing was, he *should* have agreed with Dad. No one had ever seen any evidence of magic, and according to Mum, there was a whole magical world out there. How could anyone keep something like that a secret? More magic? That seemed like a rather suspicious sort of excuse.

It should have been a clean case for Mum joking, lying or being insane, in ascending order of awfulness. If Mum had sent the letter herself, that would explain how it arrived at the letterbox without a stamp. A little insanity was far, far less improbable than the universe really working like that.

Except that some part of Harry was utterly convinced that magic was real, and had been since the instant he saw the putative letter from the Hogwarts School of Witchcraft and Wizardry.

Harry rubbed his forehead, grimacing. *Don't believe everything you think,* one of his books had said.

But this bizarre certainty... Harry was finding himself just *expecting* that, yes, a Hogwarts professor would show up and wave a wand and magic would come out. The strange certainty was making no effort to guard itself against falsification - wasn't making excuses in advance for why there wouldn't be a professor, or the professor would only be able to bend spoons.

*Where do you come from, strange little prediction?* Harry directed the thought at his brain. *Why do I believe what I believe?*

Usually Harry was pretty good at answering that question, but in this particular case, he had no *clue* what his brain was thinking.

Harry mentally shrugged. A flat metal plate on a door affords pushing, and a handle on a door affords pulling, and the thing to do with a testable hypothesis is to go and test it.

He took a piece of lined paper from his desk, and started writing.

*Dear Deputy Headmistress*

Harry paused, reflecting; then discarded the paper for another, tapping another millimetre of graphite from his mechanical pencil. This called for careful calligraphy.

*Dear Deputy Headmistress Minerva McGonagall,*

*Or Whomsoever It May Concern:*

*I recently received your letter of acceptance to Hogwarts, addressed to Mr. H. Potter. You may not be aware that my genetic parents, James Potter and Lily Potter (formerly Lily Evans) are dead. I was adopted by Lily's sister, Petunia Evans-Verres, and her husband, Michael Verres-Evans.*

*I am extremely interested in attending Hogwarts, conditional on such a place actually existing. Only my mother Petunia says she knows about magic, and she can't use it herself. My father is highly sceptical. I myself am uncertain. I also don't know where to obtain any of the books or equipment listed in your acceptance letter.*

*Mother mentioned that you sent a Hogwarts representative to Lily Potter (then Lily Evans) in order to demonstrate to her family that magic was real, and, I presume, help Lily obtain her school materials. If you could do this for my own family it would be extremely helpful.*

*Sincerely,*

*Harry James Potter-Evans-Verres.*

Harry added their current address, then folded up the letter and put it in an envelope, which he addressed to Hogwarts. Further consideration led him to obtain a candle and drip wax onto the flap of the envelope, into which, using a penknife's tip, he impressed the initials H.J.P.E.V. If he was going to descend into this madness, he was going to do it with style.

Then he opened his door and went back downstairs. His father was sitting in the living-room and reading a book of higher maths to show how smart he was; and his mother was in the kitchen preparing one of his father's favourite meals to show how loving she was. It didn't look like they were talking to one another at all. As scary as arguments could be, *not arguing* was somehow much worse.

"Mum," Harry said into the unnerving silence, "I'm going to test the hypothesis. According to your theory, how do I send an owl to Hogwarts?"

His mother turned from the kitchen sink to stare at him, looking shocked. "I - I don't know, I think you just have to own a magic owl."

That should've sounded highly suspicious, *oh, so there's no way to test your theory then,* but the peculiar certainty in Harry seemed willing to stick its neck out even further.

"Well, the letter got here somehow," Harry said, "so I'll just wave it around outside and call 'letter for Hogwarts!' and see if an owl picks it up. Dad, do you want to come and watch?"

His father shook his head minutely and kept on reading. *Of course,* Harry thought to himself. Magic was a disgraceful thing that only stupid people believed in; if his father went so far as to *test* the hypothesis, or even *watch* it being tested, that would feel like *associating* himself with that...

Only as Harry stumped out the back door, into the back garden, did it occur to him that if an owl *did* come down and snatch the letter, he was going to have some trouble telling Dad about it.

*But - well - that can't* really *happen, can it? No matter what my brain seems to believe. If an owl really comes down and grabs this envelope, I'm going to have worries a lot more important than what Dad thinks.*

Harry took a deep breath, and raised the envelope into the air.

He swallowed.

Calling out *Letter for Hogwarts!* while holding an envelope high in the air in the middle of your own back garden was... actually pretty embarrassing, now that he thought about it.

*No. I'm better than Dad. I will use the scientific method even if it makes me feel stupid.*

"Letter -" Harry said, but it actually came out as more of a whispered croak.

Harry steeled his will, and shouted into the empty sky, "*Letter for Hogwarts! Can I get an owl?*"

"Harry?" asked a bemused woman's voice, one of the neighbours.

Harry pulled down his hand like it was on fire and hid the envelope behind his back like it was drug money. His whole face was hot with shame.

An old woman's face peered out from above the neighbouring fence, grizzled grey hair escaping from her hairnet. Mrs. Figg, the occasional babysitter. "What are you doing, Harry?"

"Nothing," Harry said in a strangled voice. "Just - testing a really silly theory -"

"Did you get your acceptance letter from Hogwarts?"

Harry froze in place.

"Yes," Harry's lips said a little while later. "I got a letter from Hogwarts. They say they want my owl by the 31st of July, but -"

"But you don't *have* an owl. Poor dear! I can't imagine *what* someone must have been thinking, sending you just the standard letter."

A wrinkled arm stretched out over the fence, and opened an expectant hand. Hardly even thinking at this point, Harry gave over his envelope.

"Just leave it to me, dear," said Mrs. Figg, "and in a jiffy or two I'll have someone over."

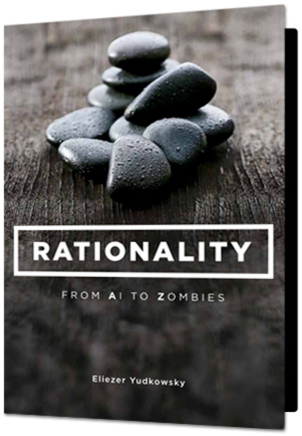
And her face disappeared from over the fence.

There was a long silence in the garden.

Then a boy's voice said, calmly and quietly, "What."

# Rationality: From AI to Zombies

Eliezer Yudkowsky's original Sequences have been edited, reordered, and converted into an ebook!




*Rationality: From AI to Zombies* is now available in PDF, EPUB, and MOBI versions on intelligence.org (**link**). You can choose your own price to pay for it (minimum $0.00), or buy it for $4.99 from Amazon (**link**). The contents are:

- 333 essays from Eliezer's 2006-2009 writings on *Overcoming Bias* and *Less Wrong*, including 58 posts that were not originally included in a named sequence.
- 5 supplemental essays from yudkowsky.net, written between 2003 and 2008.
- 6 new introductions by me, spaced throughout the book, plus a short preface by Eliezer.

The ebook's release has been timed to coincide with the end of Eliezer's other well-known introduction to rationality, *Harry Potter and the Methods of Rationality*. The two share many similar themes, and although *Rationality: From AI to Zombies* is (mostly) nonfiction, it is decidedly unconventional nonfiction, freely drifting in style from cryptic allegory to personal vignette to impassioned manifesto.

The 333 posts have been reorganized into twenty-six sequences, lettered A through Z. In order, these are titled:

- A — Predictably Wrong
- B — Fake Beliefs
- C — Noticing Confusion
- D — Mysterious Answers
- E — Overly Convenient Excuses
- F — Politics and Rationality
- G — Against Rationalization
- H — Against Doublethink
- I — Seeing with Fresh Eyes
- J — Death Spirals
- K — Letting Go
- L — The Simple Math of Evolution
- M — **Fragile Purposes**
- N — A Human's Guide to Words
- O — **Lawful Truth**
- P — Reductionism 101
- Q — Joy in the Merely Real
- R — Physicalism 201
- S — Quantum Physics and Many Worlds
- T — Science and Rationality
- U — **Fake Preferences**

- V — Value Theory
- W — **Quantified Humanism**
- X — Yudkowsky's Coming of Age
- Y — Challenging the Difficult
- Z — The Craft and the Community

Several sequences and posts have been renamed, so you'll need to consult the ebook's table of contents to spot all the correspondences. Four of these sequences (marked in **bold**) are almost completely new. They were written at the same time as Eliezer's other *Overcoming Bias* posts, but were never ordered or grouped together. Some of the others (A, C, L, S, V, Y, Z) have been substantially expanded, shrunk, or rearranged, but are still based largely on old content from the Sequences.

One of the most common complaints about the old Sequences was that there was no canonical default order, especially for people who didn't want to read the entire blog archive chronologically. Despite being called "sequences," their structure looked more like a complicated, looping web than like a line. With *Rationality: From AI to Zombies*, it will still be possible to hop back and forth between different parts of the book, but this will no longer be *required* for basic comprehension. The contents have been reviewed for consistency and in-context continuity, so that they can genuinely be read *in sequence*. You can simply read the book as a book.

I have also created a community-edited Glossary for *Rationality: From AI to Zombies*. You're invited to improve on the definitions and explanations there, and add new ones if you think of any while reading. When we release print versions of the ebook (as a six-volume set), a future version of the Glossary will probably be included.

# Announcing the Complice Less Wrong Study Hall

(If you're familiar with the backstory of the LWSH, you can skip to paragraph 5. If you just want the link to the chat, click here: LWSH on Complice)

The Less Wrong Study Hall was created as a tinychat room in March 2013, following Mqrius and ShannonFriedman's desire to create a virtual context for productivity. In retrospect, I think it's hilarious that a bunch of the comments ended up being a discussion of whether LW had the numbers to get a room that consistently had someone in it. The funny part is that they were based around the assumption that people would spend about 1h/day in it.

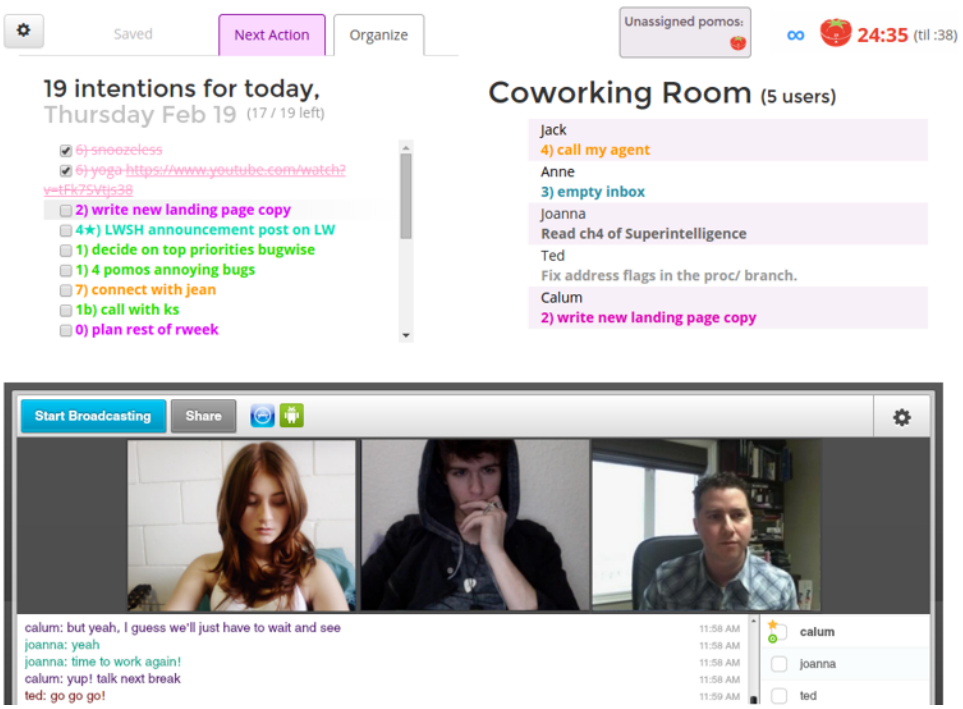
Once it was created, it was so effective that people started spending **their entire day** doing pomodoros (with 32minsWork+8minsBreak) in the LWSH and now often even stay logged in while doing chores away from their computers, just for cadence of focus and the sense of company. So there's almost always someone there, and often 5-10 people.

A week in, a call was put out for volunteers to program a replacement for the much-maligned tinychat. As it turns out though, *video chat is a hard problem*.

So nearly 2 years later, people are still using the tinychat.

But a few weeks ago, I discovered that you can embed the tinychat applet into an arbitrary page. I immediately set out to integrate LWSH into Complice, the productivity app I've been building for over a year, which counts many rationalists among its alpha & beta users.

The focal point of Complice is its today page, which consists of a list of everything you're planning to accomplish that day, colorized by goal. Plus a pomodoro timer. My habit for a long time has been to have this open next to LWSH. So what I basically did was integrate these two pages. On the left, you have a list of your own tasks. On the right, a list of other users in the room, with whatever task they're doing next. Then below all of that, the chatroom.




(Something important to note: I'm not planning to point existing Complice users, who may not be LWers, at the LW Study Hall. Any Complice user can create their own coworking room by going to complice.co/createroom)

With this integration, I've solved many of the core problems that people wanted addressed for the study hall:

- an actual ding sound beyond people typing in the chat
- synchronized pomodoro time visibility
- pomos that automatically start, so breaks don't run over
- Intentions — what am I working on this pomo?
- a list of what other users are working on
- the ability to show off how many pomos you've done
- better welcoming & explanation of group norms

There are a couple other requested features that I can definitely solve but decided could come after this launch:

- rooms with different pomodoro durations
- member profiles
- the ability to precommit to showing up at a certain time (maybe through Beeminder?!)

The following points were brought up in the Programming the LW Study Hall post or on the List of desired features on the github/nnmm/lwsh wiki, but can't be fixed without replacing tinychat:

- efficient with respect to bandwidth and CPU
- page layout with videos lined up down the left for use on the side of monitors
- chat history
- encryption
- everything else that generally sucks about tinychat

It's also worth noting that if you were to think of the entirety of Complice as an addition to LWSH... well, it would definitely look like feature creep, but at any rate there would be several other notable improvements:

- daily emails prompting you to decide what you're going to do that day
- a historical record of what you've done, with guided weekly, monthly, and yearly reviews
- optional accountability partner who gets emails with what you've done every day (the LWSH might be a great place to find partners!)

So, if you haven't clicked the link already, check out: complice.co/room/lesswrong

(This article posted to Main because that's where the rest of the LWSH posts are, and this represents a substantial update.)

# Don't Be Afraid of Asking Personally Important Questions of Less Wrong

**Related:** LessWrong as a social catalyst

I primarily used my prior user profile asked questions of Less Wrong. When I had an inkling for a query, but I didn't have a fully formed hypothesis, I wouldn't know how to search for answers to questions on the Internet myself, so I asked them on Less Wrong.

The reception I have received has been mostly positive. Here are some examples:

- I asked for a cost-benefit analysis of deliberately using nicotine for its nootropic effects.
- Back when I was trying to figure out which college major to pursue, I queried Less Wrong about which one was worth my effort. I followed this up with a discussion about whether it was worthwhile for me to personally, and for someone in general, to pursue graduate studies.

Other student users of Less Wrong benefit from the insight of their careered peers:

- A friend of mine was considering pursuing medicine to earn to give. In the same vein as my own discussion, I suggested he pose the question to Less Wrong. He didn't feel like it at first, so I posed the query on his behalf. In a few days, he received feedback which returned the conclusion that pursuing medical school through the avenues he was aiming for wasn't his best option relative to his other considerations. He showed up in the thread, and expressed his gratitude. The entirely of the online rationalist community was willing to respond provided valuable information for an important question. It might have taken him lots of time, attention, and effort to look for the answers to this question by himself.
- My friends, users Peter Hurford and Arkanj3l, have had similar experiences for choosing a career and switching majors, respectively.

In engaging with Less Wrong, with the rest of you, my experience has been that Less Wrong isn't just useful as an archive of blog posts, but is actively useful as a community of people. As weird as it may seem, you can generate positive externalities that improve the lives of others by merely writing a blog post. This extends to responding in the comments section too. Stupid Questions Threads are a great example of this; you can ask questions about your procedural knowledge gaps without fear of reprisal.  People have gotten great responses about getting more value out of conversations, to being more socially successful, to learning and appreciating music as an adult. Less Wrong may be one of few online communities for which even the comments sections are useful, by default.

For the above examples, even though they weren't the most popular discussions ever started, and likely didn't get as much traffic, it's because of the feedback they received that made them more personally valuable to one individual than several others.

At the CFAR workshop I attended, I was taught two relevant skills:

* **Value of Information Calculations**: formulating a question well, and performing a Fermi estimate, or back-of-the-envelope question, in an attempt to answer it, generates quantified insight you wouldn't have otherwise anticipated.

* **Social Comfort Zone Expansion**: humans tend to have a greater aversion to trying new things socially than is maximally effective, and one way of viscerally teaching System 1 this lesson is by trial-and-error of taking small risks. Posting on Less Wrong, especially, e.g., in a special thread, is really a low-risk action. The pang of losing karma can feel real, but losing karma really is a valuable signal that one should try again differently. Also, it's not as bad as failing at taking risks in meatspace.

When I've received downvotes for a comment, I interpret that as useful information, try to model what I did wrong, and thank others for correcting my confused thinking. If you're worried about writing something embarrassing, that's understandable, but realize it's a fact about your untested anticipations, not a fact about everyone else using Less Wrong. There are dozens of brilliant people with valuable insights at the ready, reading Less Wrong for fun, and who like helping us answer our own personal questions. Users shminux and Carl Shulman are exemplars of this.

This isn't an issue for all users, but I feel as if not enough users are taking advantage of the personal value they can get by asking more questions. This post is intended to encourage them. User Gunnar Zarnacke suggested that if enough examples of experiences like this were accrued, it could be transformed into some sort of repository of personal value from Less Wrong

# Political topics attract participants inclined to use the norms of mainstream political debate, risking a tipping point to lower quality discussion

(I hope that is the least click-baity title ever.)

Political topics elicit lower quality participation, holding the set of participants fixed. This is the thesis of "politics is the mind-killer".

Here's a separate effect: Political topics attract mind-killed participants. This can happen even when the initial participants are not mind-killed by the topic.

Since outreach is important, this could be a good thing. Raise the sanity water line! But the sea of people eager to enter political discussions is vast, and the epistemic problems can run deep. Of course not everyone needs to come perfectly prealigned with community norms, but any community will be limited in how robustly it can handle an influx of participants expecting a different set of norms. If you look at other forums, it seems to take very little overt contemporary political discussion before the whole place is swamped, and politics becomes endemic. As appealing as "LW, but with slightly more contemporary politics" sounds, it's probably not even an option. You have "LW, with politics in every thread", and "LW, with as little politics as we can manage".

That said, most of the problems are avoided by just not saying anything that patterns matches too easily to current political issues. From what I can tell, LW has always had tons of meta-political content, which doesn't seem to cause problems, as well as standard political points presented in unusual ways, and contrarian political opinions that are too marginal to raise concern. Frankly, if you have a "no politics" norm, people will still talk about politics, but to a limited degree. But if you don't even half-heartedly (or even hypocritically) discourage politics, then a open-entry site that accepts general topics will risk spiraling too far in a political direction.

As an aside, I'm not apolitical. Although some people advance a more sweeping dismissal of the importance or utility of political debate, this isn't required to justify restricting politics in certain contexts. The sort of the argument I've sketched (I don't want LW to be swamped by the worse sorts of people who can be attracted to political debate) is enough. There's no hypocrisy in not wanting politics on LW, but accepting political talk (and the warts it entails) elsewhere. Of the top of my head, Yvain is one LW affiliate who now largely writes about more politically charged topics on their own blog (SlateStarCodex), and there are some other progressive blogs in that direction. There are libertarians and right-leaning (reactionary? NRx-lbgt?) connections. I would love a grand unification as much as anyone, (of course, provided we all realize that I've been right all along), but please let's not tell the generals to bring their armies here for the negotiations.

# HPMOR Q&A by Eliezer at Wrap Party in Berkeley [Transcription]

Transcribed from maxikov's posted videos.

Verbal filler removed for clarity.

Audience Laughter denoted with [L], Applause with [A]

---

**Eliezer:** So, any questions? Do we have a microphone for the audience?

**Guy Offscreen:** We don't have a microphone for the audience, have we?

**Some Other Guy:** We have this furry thing, wait, no that's not hooked up. Never mind.

**Eliezer:** Alright, come on over to the microphone.

**Guy with 'Berkeley Lab' shirt:** So, this question is sort of on behalf of the HPMOR subreddit. You say you don't give red herrings, but like... He's making faces at me like... [L] You say you don't give red herrings, but while he's sitting during in the Quidditch game thinking of who he can bring along, he stares at Cedric Diggory, and he's like, "He would be useful to have at my side!", and then he never shows up. Why was there not a Cedric Diggory?

**Eliezer:** The true Cedrics Diggory are inside all of our hearts. [L] And in the mirror. [L] And in Harry's glasses. [L] And, well, I mean the notion is, you're going to look at that and think, "Hey, he's going to bring along Cedric Diggory as a spare wand, and he's gonna die! Right?" And then, Lestath Lestrange shows up and it's supposed to be humorous, or something. I guess I can't do humor. [L]

**Guy Dressed as a Witch:** Does Quirrell's attitude towards reckless muggle scientists have anything to do with your attitude towards AI researchers that aren't you? [L]

**Eliezer:** That is unfair. There are at least a dozen safety conscious AI researchers on the face of the earth. [L] At least one of them is respected. [L] With that said, I mean if you have a version of Voldemort who is smart and seems to be going around killing muggleborns, and sort of pretty generally down on muggles... Like, why would anyone go around killing muggleborns? I mean, there's more than one rationalization you could apply to this situation, but the sort of obvious one is that you disapprove of their conduct with nuclear weapons. From Tom Riddle's perspective that is.

I do think I sort of try to never have leakage from that thing I spend all day talking about into a place it really didn't belong, and there's a saying that goes 'A fanatic is someone who cannot change his mind, and will not change the subject.' And I'm like ok, so if I'm not going to change my mind, I'll at least endeavor to be able to change the subject. [L] Like, towards the very end of the story we are getting into the realm where sort of the convergent attitude that any sort of carefully reasoning person will take towards global catastrophic risks, and the realization that you are in fact a complete crap rationalist, and you're going to have to start over and actually try this time. These things are sort of reflective of the story outside the story, but apart from 'there is only one king upon a chessboard', and 'I need to raise the level of my game or fail', and perhaps, one little thing that was said about the mirror of VEC, as some people called it.

Aside from those things I would say that I was treating it more as convergent evolution rather than any sort of attempted parable or Professor Quirrell speaking form me. He usually doesn't... [L] I wish more people would realize that... [L] I mean, you know the... How can I put this exactly. There are these people who are sort of to the right side of the political spectrum and occasionally they tell me that they wish I'd just let Professor Quirrell take over my brain and run my body. And they are literally Republicans for You Know Who. And there you have it basically. Next Question! ... No more questions, ok. [L] I see that no one has any questions left; Oh, there you are.

**Fidgety Guy:** One of the chapters you posted was the final exam chapter where you had everybody brainstorm solutions to the predicament that Harry was in. Did you have any favorite alternate solution besides the one that made it into the book.

**Eliezer:** So, not to give away the intended solution for anyone who hasn't reached that chapter yet, though really you're just going to have the living daylight spoiled out of you, there's no way to avoid that really. So, the most brilliant solution I had not thought of at all, was for Harry to precommit to transfigure something that would cause a large explosion visible from the Quidditch stands which had observed no such explosion, thereby unless help sent via Time-Turner showed up at that point, thereby insuring that the simplest timeline was not the one where he never reached the Time-Turner. And assuring that some self-consistent set of events would occur which caused him not to carry through on his precommitment. I, you know, I suspect that I might have ruled that that wouldn't work because of the Unbreakable Vow preventing Harry from actually doing that because it might, in effect, count as trying to destroy that timeline, or filter it, and thereby have that count as trying to destroy the world, or just risk destroying it, or something along those lines, but it was brilliant! [L] I was staring at the computer screen going, "I can't believe how brilliant these people are!" "That's not something I usually hear you say," Brienne said. "I'm not usually watching hundreds of peoples' collective intelligence coming up with solutions way better than anything I thought of!" I replied to her.

And the sort of most fun lateral thinking solution was to call 'Up!' to, or pull Quirinus Quirrell's body over using transfigured carbon nanotubes and some padding, and call 'Up!' and ride away on his broomstick bones. [L] That is definitely going in 'Omake files #5: Collective Intelligence'! Next question!

**Guy Wearing Black:** So in the chapter with the mirror, there was a point at which

Dumbledore had said something like, "I am on this side of the mirror and I always have been." That was never explained that I could tell. I'm wondering if you could clarify that.

**Eliezer:** It is a reference to the fanfic 'Seventh Horcrux' that *totally* ripped off HPMOR despite being written slightly earlier than it... [L] I was slapping my forehead pretty hard when that happened. Which contains the line "Perhaps Albus Dumbledore really was inside the mirror all along." Sort of arc words as it were. And I also figured that there was simply some by-location effect using one of the advanced settings of the mirror that Dumbledore was using so that the trap would always be springable as opposed to him having to know at what time Tom Riddle would appear before the mirror and be trapped. Next!

**Black Guy:** So, how did Moody and the rest of them retrieve the items Dumbledore threw in the mirror of VEC?

**Eliezer:** Dumbledore threw them outside the mirrors range, thereby causing those not to be sealed in the corresponding real world when the duplicate mode of Dumbledore inside the mirror was sealed. So wherever Dumbledore was at the time, probably investigating Nicolas Flamel's house, he suddenly popped away and the line of Merlin Unbroken and the Elder Wand just fell to the floor from where he was.

**Asian Guy:** In the 'Something to Protect: Severus Snape', you wrote that he laughed. And I was really curious, what exactly does Severus Snape sound like when he laughs. [L]

**Person in Audience:** Perform for us!

**Eliezer:** He He He. [L]

**Girl in Audience:** Do it again now, everybody together!

**Audience:** He He He. [L]

**Guy in Blue Shirt:** So I was curious about the motivation between making Sirius re-evil again and having Peter be a good guy again, their relationship. What was the motivation?

**Eliezer:** In character or out character?

**Guy in Blue Shirt:** Well, yes. [L]

**Eliezer:** All right, well, in character Peter can be pretty attractive when he wants to be, and Sirius was a teenager. Or, you were asking about the alignment shift part?

**Guy in Blue Shirt:** Yeah, the alignment and their relationship.

**Eliezer:** So, in the alignment, I'm just ruling it always was that way. The whole Sirius Black thing is a puzzle, is the way I'm looking at it. And the canon solution to that puzzle is perfectly fine for a children's book, which I say once again requires a higher level of skill than a grown-up book, but just did not make sense in context. So I was just looking at the puzzle and being like, ok, so what can be the actual solution to this puzzle? And also, a further important factor, this had to happen. There's a whole lot of fanfictions out there of Harry Potter. More than half a million, and that was years ago. And 'Methods of Rationality' is fundamentally set in the universe of Harry Potter fanfiction, more than canon. And in many many of these fanfictions someone goes back in time to redo the seven years, and they know that Scabbers is secretly Peter Pettigrew, and there's a scene where they stun Scabbers the rat and take him over to Dumbledore, and Head Auror, and the Minister of Magic and get them to check out this rat over here, and uncover Peter Pettigrew. And in all the times I had read that scene, at least a dozen times literally, it was never once played out the way it would in real life, where that is just a rat, and you're crazy. [L] And that was the sort of basic seed of, "Ok, we're going to play this straight, the sort of loonier conspiracies are false, but there is still a grain of conspiracy truth to it." And then I introduced the whole accounting of what happened with Sirius Black in the same chapter where Hermione just happens to mention that there's a Metamorphmagus in Hufflepuff, and exactly one person posted to the reviews in chapter 28, based on the clue that the Metamorphmagus had been mentioned in the same chapter, "Aha! I present you the tale of Peter Pettigrew, the unfortunate Metamorphmagus." [L] See! You could've solved it, you could've solved it, but you didn't! Someone solved it, you did not solve that. Next Question!

**Guy in White:** First, [pulls out wand] Avada Kedavra. How do you feel about your security? [L] Second, have you considered the next time you need a large group of very smart people to really work on a hard problem, presenting it to them in fiction?

**Eliezer:** So, of course I always keep my Patronus Charm going inside of me. [Aww/L] And if that fails, I do have my amulet that triggers my emergency kitten shield. [L] And indeed one of the higher, more attractive things I'm considering to potentially do for the next major project is 'Precisely Bound Djinn and their Behavior'. The theme of which is you have these people who can summon djinn, or command the djinn effect, and you can sort of negotiate with them in the language of djinn and they will always interpret your wish in the worst way possible, or you can give them mathematically precise orders; Which they can apparently carry out using unlimited computing power, which obviously ends the world in fairly short order, causing our protagonist to be caught in a groundhog day loop as they try over and over again to both maybe arrange for conditions outside to be such that they can get some research done for longer than a few months before the world ends again, and also try to figure out what to tell their djinn. And, you know, I figure that if anyone can give me an unboundedly computable specification of a value aligned advanced agent, the story ends, the characters win, hopefully that person gets a large monetary prize if I can swing it, the world is safer, and I can go onto my next fiction writing project, which will be the one

with the boundedly specified [L] value aligned advanced agents. [A]

**Guy with Purple Tie:** So, what is the source of magic?

**Eliezer:** Alright, so, there was a bit of literary miscommunication in HPMOR. I tried as hard as I could to signal that unraveling the true nature of magic and everything that adheres in it is actually this kind of this large project that they were not going to complete during Harry's first year of Hogwarts. [L] You know, 35 years, even if someone is helping you is a reasonable amount of time for a project like that to take. And if it's something really difficult, like AIs, you might need more that two people even. [L] At least if you want the value aligned version. Anyway, where was I?

So the only way I think that fundamentally to come up with a non-nitwit explanation of magic, you need to get started from the non-nitwit explanation, and then generate the laws of magic, so that when you reveal the answer behind the mystery, everything actually fits with it. You may have noticed this kind of philosophy showing up elsewhere in the literary theory of HPMOR at various points where it turns out that things fit with things you have already seen. But with magic, ultimately the source material was not designed as a hard science fiction story. The magic that we start with as a phenomenon is not designed to be solvable, and what did happen was that the characters thought of experiments, and I in my role of the universe thought of the answer to it, and if they had ever reached the point where there was only one explanation left, then the magic would have had rules, and they would have been arrived at in a fairly organic way that I could have felt good about; Not as a sudden, "Aha! I gotcha! I revealed this thing that you had no way of guessing."

Now I could speculate. And I even tried to write a little section where Harry runs into Dumbledore's writings that Dumbledore left behind, where Dumbledore writes some of his own speculation, but there was no good place to put that into the final chapter. But maybe I'll later be able... The final edits were kind of rushed honestly, sleep deprivation, 3am. But maybe in the second edit or something I'll be able to put that paragraph, that set of paragraphs in there. In Dumbledore's office, Dumbledore has speculated. He's mostly just taking the best of some of the other writers that he's read. That, look at the size of the universe, that seems to be mundane. Dumbledore was around during World War 2, he does know that muggles have telescopes. He has talked with muggle scientists a bit and those muggle scientists seem very confident that all the universe they can see looks like it's mundane. And Dumbledore wondered, why is there this sort of small magical section, and this much larger mundane section, or this much larger muggle section? And that seemed to Dumbledore to suggest that as a certain other magical philosopher had written, If you consider the question, what is the underlying nature of reality, is it that it was mundane to begin with, and then magic arises from mundanity, or is the universe magic to begin with, and then mundanity has been imposed above it? Now mundanity by itself will clearly never give rise to magic, yet magic permits mundanity to be imposed, and so, this other magical philosopher wrote, therefore he thinks that the universe is magical to begin with and the mundane sections are imposed above the magic. And Dumbledore himself had speculated, having been antiquated with the line of Merlin for much of his life, that just as the Interdict of Merlin was imposed to restrict the spread an the number of people who had sufficiently powerful magic, perhaps the mundane world itself, is an attempt to bring order to something that was on the verge of falling apart in Atlantis, or in whatever came before Atlantis. Perhaps the thing that happened with the Interdict of Merlin has happened over and over again. People trying to impose law

upon reality, and that law having flaws, and the flaws being more and more exploited until they reach a point of power that recons to destroy the world, and the most adapt wielders of that power try to once again impose mundanity.

And I will also observe, although Dumbledore had no way of figuring this out, and I think Harry might not have figured it out yet because he dosen't yet know about chromosomal crossover, That if there is no wizard gene, but rather a muggle gene, and the muggle gene sometimes gets hit by cosmic rays and ceases to function thereby producing a non-muggle allele, then some of the muggle vs. wizard alleles in the wizard population that got there from muggleborns will be repairable via chromosomal crossover, thus sometimes causing two wizards to give birth to a squib. Furthermore this will happen more frequently in wizards who have recent muggleborn ancestry. I wonder if Lucius told Draco that when Draco told him about Harry's theory of genetics. Anyway, this concludes my strictly personal speculations. It's not in the text, so it's not real unless it's in the text somewhere. 'Opinion of God', Not 'Word of God'. But this concludes my personal speculations on the origin of magic, and the nature of the "wizard gene". [A]

# Can we talk about mental illness?

For a site extremely focused on fixing bad thinking patterns, I've noticed a bizarre lack of discussion here. Considering the high correlation between intelligence and mental illness, you'd think it would be a bigger topic.

I personally suffer from Generalized Anxiety Disorder and a very tame panic disorder. Most of this is focused on financial and academic things, but I will also get panicky about social interaction, responsibilities, and things that happened in the past that seriously shouldn't bother me. I have an almost amusing response to anxiety that is basically my brain panicking and telling me to go hide under my desk.

I know lukeprog and Alicorn managed to fight off a good deal of their issues in this area and wrote up how, but I don't think enough has been done. They mostly dealt with depression. What about rational schizophrenics and phobics and bipolar people? It's difficult to find anxiety advice that goes beyond "do yoga while watching the sunrise!" Pop psych isn't very helpful. I think LessWrong could be. What's mental illness but a wrongness in the head?

Mental illness seems to be worse to intelligent people than your typical biases, honestly. Hiding under my desk is even less useful than, say, appealing to authority during an argument. At least the latter has the potential to be useful. I know it's limiting me, and starting cycles of avoidance, and so much more. And my mental illness isn't even that bad! Trying to be rational and successful when schizophrenic sounds like a Sisyphusian nightmare.

I'm not fighting my difficulties nearly well enough to feel qualified to author my own posts. Hearing from people who are managing is more likely to help. If nothing else, maybe a Rational Support Group would be a lot of fun.

# Twenty basic rules for intelligent money management

**1. Start investing early in life.**

The power of compound interest means you will have much more money at retirement if you start investing early in your career. For example, imagine that at age eighteen you invest $1,000 and earn an 8% return per year. At age seventy you will have $54,706. In contrast, if you make the same investment at age fifty you will have a paltry $4,661 when you turn seventy.

Many people who haven't saved for retirement panic upon reaching middle age. So if you are young don't think that saving today will help you only when you retire, but know that such savings will give you greater peace of mind when you turn forty.

When evaluating potential marriage partners give bonus points to those who have a history of saving. Do this not because you want to marry into wealth, but because you should want to marry someone who has discipline, intelligence and foresight.

**2. Maintain a diversified portfolio.**

By purchasing many different types of investments you reduce your financial risk. Even a single seemingly stable stock can easily fall by 70% in a single year. In contrast, a broad investment portfolio is extremely unlikely to decline in value by such a gigantic amount unless something truly horrible happens to the entire world's economy. As the saying goes, "don't put all of your eggs in one basket."

**3. Consider buying an index fund.**

Index funds provide cheap and easy ways to acquire a diversified stock portfolio. An index fund is a mutual fund that invests in every stock in its index. So, for example, an S&P 500 index fund will purchase all 500 stocks in the S&P 500 index, which consists of the 500 largest publicly traded stocks in the United States.

**4. Don't forget about foreign stocks.**

To achieve optimal portfolio diversification you need to invest in foreign stocks. The bigger a nation's economy, the more money you should put in its stock market. You can buy index funds that invest in foreign stocks. By investing in diverse foreign securities from many nations you will probably reduce the chance that your portfolio will suffer a sudden huge decline.

**5. Don't try to out-guess the market.**

Ordinary investors, and indeed even most professional investors, are horrible at figuring out which individual stocks will outperform the entire market. For reasons I won't go into, economists have overwhelming evidence that without inside information not available to the general public an investor can't accurately predict which stocks will do well.

If you try to out-guess the market you might get lucky and earn a superior return. But on average you will do worse compared to someone who holds a diversified portfolio. Furthermore, by placing a large bet on a few stocks you will necessarily violate many of the other investing rules listed here and so will most likely

pay a financial penalty. True, it can be fun to take a chance and gamble on one stock. But you probably shouldn't allow the excitement of gambling to infect your investment decisions. If you want to gamble bet a few dollars on blackjack, but stick to a sound, diversified, investment strategy.

You may know people who brag about how they made a killing in the stock market by calculating which stocks would do well. Keep in mind, however, that they might be telling you only about their profitable stock choices and not their losses. Furthermore, even if someone has on average beaten the market, he probably just got lucky. After all, although people do win lotteries, such winners don't really have any special abilities at guessing which lotto numbers will come up.

A few professional investors, such as multi-billion-dollar hedge funds, might well have means of earning superior returns by investing in just a few financial securities. But if they do possess financial superpowers they won't share them with you for less than a lot of money. Also, such investors probably earn their above-average returns by investing in exotic financial instruments, such as derivative securities, that you don't have access to.

**6. Don't take on "stupid" risks.**

You may have heard that financial markets compensate investors for taking on risks. This is true, but it doesn't apply to *stupid* risks.

Imagine you work for a construction company. You learn that the owner pays higher wages for employees who do work on the top of tall, unfinished buildings because such work is extremely risky. Construction companies have to pay more to workers who undertake the most perilous tasks or else no laborer would be willing to do such dangerous work.

This week, say, you want to make a lot of money. You understand that the construction company pays the highest wages to workers who do the most dangerous jobs. So you intend to work on the top of the tallest skyscraper *while drunk*! You figure that since this is extremely risky you should get a large bonus. But of course management pays only for risks it needs someone to undertake, and they obviously don't need anyone to labor while under the influence.

Stocks on average pay higher returns than government bonds because otherwise everyone would buy the bonds and no one would take the risk of owning stocks. Since markets need people to buy stocks, they must compensate those who do by giving them, on average, higher returns. But the market doesn't need anyone to hold an undiversified portfolio. If you do, you are taking on risks that benefit no one and so you won't get paid for it. Holding an undiversified portfolio and expecting to get a high average return because of all the risk you are taking on is analogous to working on a skyscraper while drunk and expecting your employer to pay you a premium because you are increasing the riskiness of your job.

**7. Understand the dangers of actively managed mutual funds.**

**7(A) On average, actively managed funds do worse than passively managed funds do.**

Index funds are *passively* managed because the funds don't try to guess which stocks will do well. In contrast, *actively* managed mutual funds do try to identify stocks that will outperform the market. Most actively managed mutual funds, however, do much worse than broad-based index funds such as S&P 500 index funds.

**7(B). Actively managed mutual funds have high fees.**

Mutual fund fees have a tremendous impact on long-term investment performance. Imagine that the market

goes up by 8% a year. One mutual fund charges fees of .2% a year; another charges fees of 1.5% per year. Pretend that you invest $1,000 in both funds. After thirty years you will have $9,518 in the first fund but only $6,614 in the second.

Mutual funds often charge high fees to pay expensive MBAs to pick stocks. But MBAs are not on average, any good at out-guessing the stock market. So when you buy a high fee mutual fund you are wasting money on MBAs.

Index funds often have the lowest fees because these funds don't try to out-guess the market and so don't need to hire expensive (and useless) stock guessing MBAs. Still, some index funds do charge high fees and so should be avoided at all costs.

**7(C). Survivorship bias artificially inflates the mutual fund industry's past performance.**

Let's say I start 100 mutual funds. Each fund will randomly select a few stocks to invest in. Almost certainly at least one of my funds will get lucky and earn a high return. After a few years I will identify the one that did the best and market this fund to consumers. I will quietly close down the other 99 funds. When attracting customers for my one surviving fund I will claim that its fantastic past performance proves I'm an investment genius. Of course, since all my stock picks were random I have not demonstrated any investment skill. If you evaluate only the mutual fund that survives it will indeed appear that I'm an investment wizard. But such "survivorship bias" corrupts the evaluation.

Mutual funds that do very poorly shut down. Consequently, if we just take the average past returns of mutual funds that exist today we would get an estimate of performance that overstates the overall investment skills of the mutual fund industry.

**8. Don't engage in much stock trading.**

When you trade a stock you pay a fee. And as a previous investment tip explains, in the long run fees decimate investment performance. Furthermore, when you trade stocks that are not in a tax-preferred plan (such as a 401(k) plan) you often pay extra taxes.

**9. Invest in tax-preferred vehicles such as 401K plans.**

The U.S. government gives tremendous tax benefits to those who invest in certain tax-advantaged vehicles such as 401(k), 403(b), or IRA plans. (Restated: The U.S. government imposes a "stupidity tax" on investors who don't take advantage of tax preferred plans.) These plans have yearly contribution limits, so a wise investment strategy is to put as much as the government allows into the plans you are eligible to contribute to.

**10. Avoid credit card debt.**

High interest rate credit card debt is financial cancer. Each month you should pay off your full credit card balance to avoid such financial sickness. If you can't, however, call your credit card company to negotiate a better rate.

Credit card companies love customers who (a) have lots of debt, but (b) make only their minimum payment each month. If you are such a customer then call your credit card provider and tell it that because of the high interest rates it charges you want to transfer your balance to a card from another company. Chances are your credit card provider will offer you a lower rate to keep you as a customer. The credit card industry is highly competitive. Use this to your advantage if you can't pay off your total balance each month.

**11. Always take full advantage of matching contribution pension plans.**

Some employers will match a worker's contribution to his retirement account. These matching plans always have some upper limit after which the employer will no longer match contributions. For example, an employer might deposit fifty cents into your 401(k) account for every dollar you put in as long as you have put in less than $6,000. After you have put in $6,000 your employer won't match any additional money you put into your retirement account.

You should always take full advantage of matching pension plans because they offer the best rate of return of any investment. For example, with the 50% plan described above you get an immediate and risk free 50% return on your investment. Putting less than $6,000 in this hypothetical plan is the equivalent to telling your employer that it should keep some of the money it was prepared to give you.

Many employers don't offer matching contribution pension plans.

**12. Be cautious about investing in your employer's stock.**

Companies love for employees to buy lots of their stock because such stock-owning laborers care more about the company's profitability. But it's financially perilous to buy your firm's stock because if the firm goes under you lose not only your job but also part of your savings.

Some companies, however, offer significant financial enticements to employees who do buy their stock. If these enticements are large enough you should seriously consider giving in. But understand that by buying your company's stock you *are* taking on significant risk.

**13. Remember that your home is a very risky asset.**

It's temping to think that your home is a much more solid investment than your stocks because you can actually touch your home. But as with individual stocks, the value of a single home can fall rapidly . This is especially true if you have a mortgage.

Imagine, for example, that you buy a $400,000 home and pay for it with $40,000 in cash and a $360,000 mortgage. So you have a $360,000 debt on a home worth $400,000, meaning that you have $40,000 in home equity. Now assume that the value of your home falls by 10% and becomes worth $360,000. Since you still owe $360,000 on the home, your equity in it has fallen to zero! A 10% fall in the value of your home has obliterated your home equity.

A home is inherently risky because it's an undiversified asset. Mortgage debt magnifies this risk.

This doesn't mean you shouldn't buy a home. The favorable tax treatment of mortgage interest makes it worthwhile for most adult Americans to be home owners.

**14. Learn how your financial advisor gets paid.**

People respond to incentives. If, for example, your stock broker receives a fee every time you make a stock trade then he may well advise you to make more trades then you should. If your real estate agent gets paid the same whether you buy after looking at five or twenty-five houses then she has an incentive to get you to make a quick home buying decision.

**15. Buy life insurance if your family relies on your income or time.**

If your family relies on your income you owe it to them to buy life insurance. No one likes to think that he could die but, alas, all men are mortal. If you rely on your spouse's income make sure that he/she has life insurance. A husband should get life insurance before his wife becomes pregnant in case the worst happens. Homemakers who don't earn any income but have dependent children should still buy life insurance because if they die their spouses may have to hire someone to do many of the tasks that the homemaker had previously done.

**16. You and your spouse should have disability insurance.**

You might well impose a greater financial burden on your family if you become seriously disabled than if you die. If you die you stop bringing in income, but you also (after your funeral) stop consuming. If, however, you can't work because of a disability, you not only won't be earning money but you will also require a significant amount of financial support from your family. To (partially) protect your family from this burden you should purchase disability insurance. Most people get such insurance through their employer, so speak to your company's human resources department about getting disability insurance.

**17. As you approach retirement consider putting much of your new savings into safe government bonds.**

Sudden falls in the stock market have a greater impact on those close to retirement than on younger investors who can ride out the inevitable ups and downs of the market. So as you approach retirement you should consider putting much of your new savings into safe government bonds.

**18. Women usually live longer than men and so need to save more for retirement.**

Since women live about six years longer than men do, they will on average need more money to finance a comfortable retirement.

**19. Keep in mind that you might live a lot longer than your grandparents will/did.**

Over the next forty years scientists might develop many successful anti-aging treatments. Because of the possibility of such technologies, a forty-year-old alive today has a non-trivial chance of still being alive one hundred years from now. So when deciding how much to save for retirement take into account that you might spend a heck of a lot more time in retirement than any of your grandparents will/did.

**20. Determine how much people in your job make.**

Your employer knows how much people in your position are paid. If you don't you're at a disadvantage in salary negotiations. Many people are uncomfortable discussing salaries with co-workers. Overcome such discomfort to find out if you deserve a raise.

Originally published on Google's Knol

# Half-assing it with everything you've got

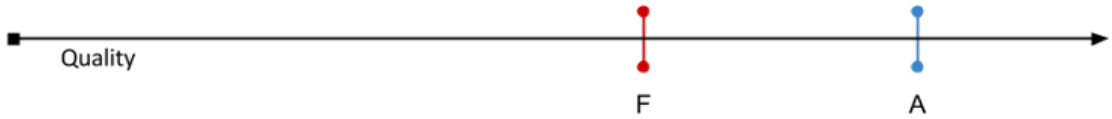
This is a linkpost for http://mindingourway.com/half-assing-it-with-everything-youve-got/

I hang out around a lot of effective altruists. Many of them are motivated primarily by something like guilt (for having great resource and opportunity while others suffer) or shame (for not helping enough). Hell, many of my non-EA friends are primarily motivated by guilt or shame.

I'm not going to criticize guilt/shame motivation: I have this policy where, when somebody puts large amounts of effort or money towards making the world a better place, I try really hard not to condemn their motives. Guilt and shame may be fine tools for jarring people out of complacence. However, I worry that guilt and shame are unhealthy long-term motivators. In many of my friends, guilt and shame tend to induce akrasia, reduce productivity, and drain motivation. So over the next few weeks, I'll be writing a series of posts about removing guilt/shame motivation and replacing it with something stronger.

# 1

Say you're a college student, and you have a paper due. The quality of the paper will depend upon the amount of effort you put in. We'll say that you know the project pretty well: you can get an A with only moderate effort, and with significant effort you could produce something much better than the usual A-grade paper.




The education environment implicitly attempts to convince students that their preferences point ever rightward along this line. Parents and teachers say things like "you should put in your best effort," and they heap shame upon people who don't strive to push ever rightward along the quality line.

People generally react to this coercion in one of two ways. The first group (the "slackers") rejects the implication that quality=preferences. These are the people who don't care about the class, who complain constantly about the useless pointless work they have to do, who half-ass the assignment and turn in something that either barely passes or fails entirely. Slackers tend to resent the authority forcing them to write the paper.

The second group (the "tryers") are the ones who accept the premise that quality=preferences, and strive ever rightwards on the quality line. Tryers include people of all ability levels: some struggle as hard as they can just to get a C, others flaunt their ability to produce masterpieces. Some try to curry favor with the teacher, others are perfectionists who simply can't allow themselves to turn in anything less than their best effort. Some of them are scrupulous people, who feel guilty even after getting an A, because they know they could have done better, and think they should have. Some are humble, some are show-offs, but all of them are pushing rightward.

Society has spent a *lot* of time conditioning us to think of the tryers as better than the slackers. Being a tryer is a virtue. Slackers are missing the point of education; why are they even there? The tryers are going to go places, the slackers will never amount to anything.

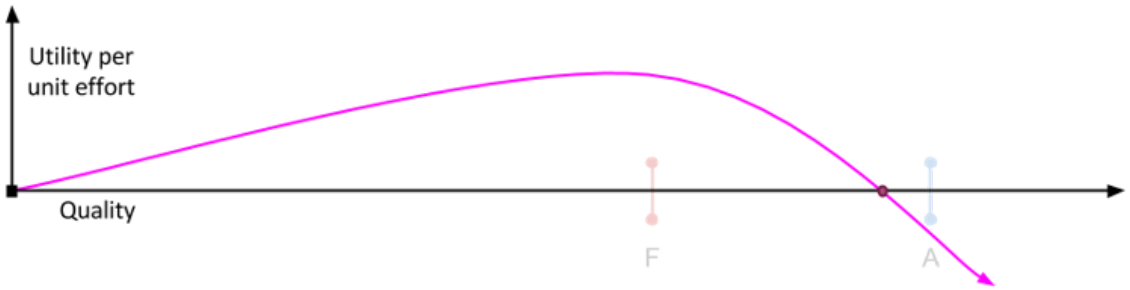
But in fact, both groups are doing it wrong.

If you want to be highly effective, *remember what you're fighting for.*

And, spoiler alert, you aren't fighting for "write a high-quality paper." That would be a pretty silly thing to fight for.

What is your goal in taking this class? Perhaps you're doing it thanks to a combination of social pressure (your parents said to), social inertia (everybody else goes to college), and a vague belief that this is the path towards a good job and a comfortable life. Or perhaps you're there because you want good grades so you can acquire lots of money and power which you will use to fight dragons. Or perhaps you're there out of a genuine thirst for knowledge. But no matter why you're there, your reason for being there will pick out a single target point on the quality line. Your goal, then, is to hit that quality target — no higher, no lower.

Your preferences are not "move rightward on the quality line." Your preferences are to *hit the quality target with minimum effort.*




If you're trying to pass the class, then pass it with minimum effort. Anything else is wasted motion.

If you're trying to ace the class, then ace it with minimum effort. Anything else is wasted motion.

If you're trying to learn the material to the fullest, then mine the assignment for all its knowledge, and don't fret about your grade. Anything else is wasted motion.

If you're trying to do achieve some combination of good grades (for signalling purposes), respect (for social reasons), and knowledge (for various effects), then pinpoint the minimum quality target that gets a good grade, impresses the teacher, and allows you to learn the material, and hit that as efficiently as you can. Anything more is wasted motion.

Your quality target may be significantly left of F — if, say, you've already passed the class, and this assignment doesn't matter. Your quality target may be significantly to the right of A — if, say, you're there to learn the material, and grade inflation means that it's much easier to produce an A-grade paper than it is to complete the assignment in the maximally informative way. But no matter what, your goals will induce a quality target.

Both the slackers *and* the tryers are pursuing lost purposes. The slackers scoff at the tryers, who treat an artificial quality line like it's their actual preferences and waste effort over-achieving. The tryers scoff at the slackers, who are taking classes but refusing to learn. And both sides are right! Because both sides are wasting motion.

The slackers fail to deploy their full strength because they realize that the quality line is not their preference curve. The tryers deploy their full strength at the wrong target, in attempts to go as far right as possible, wasting energy on a fight that is not theirs. So take the third path: *remember what you're fighting for.* Always deploy your full strength, in order to hit your quality target as fast as possible.

Half-ass everything, with everything you've got.

(My teachers used to say that I could do great things if only I applied myself. I used to tell them that if they wanted me to apply more effort, they would need to invent higher letter grades.)

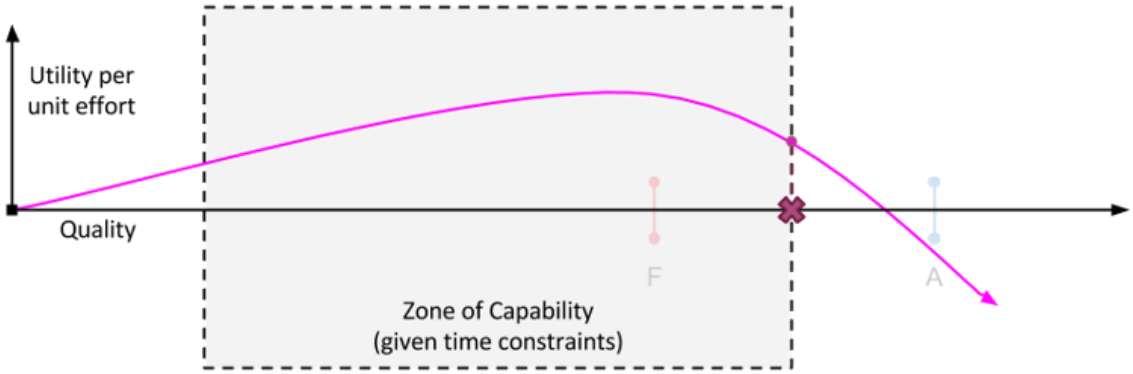
# 2

A common objection arises here:

> *Some things are too important to "half-ass." Some things are simply worth fighting for with your full strength. It's one thing to half-ass a homework assignment, and another thing entirely to half-ass saving a life. Sometimes you want to push as far right as you can on the quality curve.*

This is both true and false, because it is mixed up. Given any project, *always* aim no higher than the quality target, and always strive for minimum expenditure of effort. It doesn't matter whether you're writing a term paper, pulling a person out of a burning house, or creating a galaxy-spanning human civilization — the goal is always to achieve some quality target with minimum effort. Negentropy is scarce.

That said, the quality target can be *really really high.* In fact, the quality target is sometimes unatainably high. Often, we simply aren't capable of hitting our quality targets, and in those cases, we *do* want to push as far right along the preference curve as we can.




This can occur naturally whenever you work on something difficult relative to your skill level, or in competitive situations, or if you're signalling your ability to work hard. But don't get confused. Even if you write for the love of writing, you eventually have to stop editing and call it finished. Even if you're getting somebody out of a burning building, you eventually stop putting effort towards ensuring that they survive in favor of putting that effort towards saving other dying people instead. Even if you're building an intergalactic civilization, you need to trade off energy spent building the civilization against energy spent living in it.

There are goals for which you cannot achieve your quality targets, and in those cases, you will push ever rightwards. But too many people automatically assume that, when an authority figure describes a quality line, they're "supposed to" push as far right as possible. They think they "should" care about quality. This is silly: real world problems are not about producing the highest-quality products. In all walks of life, the goal is to hit a quality target with minimum effort.

This is of course only a fuzzy and inaccurate description of reality. The relative costs of time, effort, energy, attention, and quality are generally in flux, and change with both information

and circumstance. The essential point is to be able to differentiate between the implicit quality line highlighted circumstance, and your actual preference curve.

---

Let me be clear about what I'm *not* saying. If you're taking a college course, I'm *not* telling you that you should be scraping by by only the barest of margins. If you're saving a life, I'm *not* telling you to prefer speed over caution. In general, I'm decidedly **not** saying that you must always identify the worst outcome that you'd grudgingly accept as your target.

What I am saying is, don't conflate the quality line with the preference curve. Don't get confused when the teacher labels one quality-point "pass" and another "fail," for these are just labels, and your deeper goals are likely only tangentially related to those labels. Remember what you're trying to achieve, identify your quality target, and aim for that: no higher, no lower.

(Also, remember that the planning fallacy exists! If you shoot for a D, you might get an F. Humans tend to be overconfident. When you pick your targets, be cautious, and leave yourself comfortable margins.)

# 3

The common slacker objection goes:

> *But what if "get the minimum passing grade as quick as possible" is also boring? What if this task, too, is meaningless?*

Then get out of college!

I personally find that shooting for the minimum acceptable quality is usually *fun*. Doing the homework assignment is boring, but finding a way to get the homework assignment up to an acceptable level *with as little total effort as possible* is an interesting optimization problem that actually engages my wits, an optimization problem which both my inner perfectionist and my inner rebel can get behind.

But sometimes, after remembering what you're fighting for, the whole project will still seem worthless. Sometimes, the goal of getting the minimum passing grade with minimum effort will still stink of somebody else trying to pass off their arbitrary metric as your true preferences. In that case, consider dropping the class.

More generally, if there's no variation on "achieve such-and-such a goal with minimum effort" that seems worth doing, then you may need to abandon that goal entirely.

---

By contrast, the common tryer objection runs as follows:

> *But I'm a perfectionist! I physically can't stop caring about a low-quality product. I'm compelled to do my best.*

Great! Harness the perfectionist within you, and point it towards the goal of hitting your target with minimum effort.

Instead of being a perfectionist about the paper, be a perfectionist about *writing* the paper. Be a perfectionist about identifying good strategies, about abandoning sunk costs, about killing your darlings, about noticing when you're done. Be a perfectionist about wasting no attention. Be a perfectionist about learning from your mistakes. Perfectionism can be a powerful tool, but there's no need to point it at overachieving on metrics you don't care about.

# 4

Attempting to hit a quality target with the least possible effort is, in a sense, a much more difficult task than pushing as far right on the quality-line as possible. One always could push further right on the quality line with more time: when one is trying to write a great paper, they always *could* correct their flaws with more time and energy. But when one is trying to produce a paper with minimal wasted motion, mistakes are irrevocable. Time cannot be un-wasted.

In this sense, switching from being a tryer to a whole-assed-half-asser may lead to more guilt and shame than usual, if you start feeling guilty about every wasted motion.

However, I see far too many people feeling guilt and shame about not having pushed far enough along the quality line. They feel guilty about not putting effort towards their job (which they hate); they feel guilty about not being a good enough friend (when they are nearly at breakdown themselves); they feel guilty about not fulfilling their parent's expectations (which are ridiculous and uninformed). In order to replace guilt and shame with intrinsic motivation, it is first necessary to break the slacker/tryer dichotomy. If you've got to feel guilty, please feel guilty about missing your own targets, rather than feeling guilty about not adopting some arbitrary quality line as your true preferences. The former type of guilt is the one that I have a shot at addressing.

(Scrupulous people: in the interim, please don't feel guilty about wasting motion! Treat it like an important part of the human action process rather than something to be ashamed of. Future posts will expand on this idea.)

# 5

Most people seem to have two modes of working on problems: the slacker path-of-least-resistance "coasting" mode, and the tryer make-a-masterpiece "overachiever" mode. When faced with a problem, most people either put in the minimum effort necessary to scrape by without pissing off the relevant authorities, or else they pour their heart and soul into the task.

Almost everybody spends some time in both modes. Some people overachieve in history class and coast in grammar class. Some people overachieve at work and coast in their relationship. In fact, most heartwarming bad-students-can-be-good-people-too stories are about how students who are slacking in most domains are secretly trying really hard when it comes to dance/sports/music/number theory.

This, of course, is another piece of tryer propaganda: "Don't worry," the movies say, "these slackers aren't bad people, because they're secretly tryers in other domains!" As if you're only a good person if you can adopt *some* arbitrary quality line as your true preferences.

Most people are trapped in the slacker/tryer dichotomy. They either do as little as they can get away with or as much as they can manage. They're either aiming for barely acceptable or they're aiming to be the best. Very few people seem able to pick a target in the middle and then pursue it with *everything they've got.* Very few people seem capable of deploying their *full strength* to hit "mediocre" as efficiently as possible.

Reject the dichotomy. Keep your eye on the preference curve. And remember that the preference curve says this, and only this:

*Succeed, with no wasted motion.*

The slacker in you rebels against pointless tasks, and the tryer in you wants perfection. So satisfy both: aim for the minimum necessary target, and move there as efficiently as possible.

And if ever you forget what it means to "succeed" in one context or another, take a moment to pause and remember what you're fighting for.

# Minds: An Introduction

You're a mind, and that puts you in a pretty strange predicament.

Very few things get to be minds. You're that odd bit of stuff in the universe that can form predictions and make plans, weigh and revise beliefs, suffer, dream, notice ladybugs, or feel a sudden craving for mango. You can even form, *inside your mind*, a picture of your whole mind. You can reason about your own reasoning process, and work to bring its operations more in line with your goals.

You're a mind, implemented on a human brain. And it turns out that a human brain, for all its marvelous flexibility, is a lawful thing, a thing of pattern and routine. Your mind can follow a routine for a lifetime, without ever once noticing that it is doing so. And these routines can have great consequences. When a mental pattern serves you well, we call that "rationality."

You exist as you are, hard-wired to exhibit certain species of rationality and certain species of irrationality, because of your ancestry. You, and all life on Earth, are descended from ancient self-replicating molecules. This replication process was initially clumsy and haphazard, and soon yielded replicable *differences* between the replicators. "Evolution" is our name for the change in these differences over time.

Since some of these reproducible differences impact reproducibility—a phenomenon called "selection"—evolution has resulted in organisms suited to reproduction in environments like the ones their ancestors had. Everything about you is built on the echoes of your ancestors' struggles and victories.

And so here you are: a mind, carved from weaker minds, seeking to understand your own inner workings, that they can be improved upon—improved upon relative to *your* goals, and not those of your designer, evolution. What useful policies and insights can we take away from knowing that this is our basic situation?

## Ghosts and Machines

Our brains, in their small-scale structure and dynamics, look like many other mechanical systems. Yet we rarely think of our minds in the same terms we think of objects in our environments or organs in our bodies. Our basic mental categories—belief, decision, word, idea, feeling, and so on—bear little resemblance to our physical categories.

Past philosophers have taken this observation and run with it, arguing that minds and brains are fundamentally distinct and separate phenomena. This is the view the philosopher Gilbert Ryle called "the dogma of the Ghost in the Machine."[1] But modern scientists and philosophers who have rejected dualism haven't necessarily replaced it with a better predictive model of how the mind works. *Practically* speaking, our purposes and desires still function like free-floating ghosts, like a magisterium cut off from the rest of our scientific knowledge. We can talk about "rationality" and "bias" and "how to change our minds," but if those ideas are still imprecise and unconstrained by any overarching theory, our scientific-sounding language won't

protect us from making the same kinds of mistakes as those whose theoretical posits include spirits and essences.

Interestingly, the mystery and mystification surrounding minds doesn't just obscure our view of *humans*. It also accrues to systems that seem mind-like or purposeful in evolutionary biology and artificial intelligence (AI). Perhaps, if we cannot readily glean what we are from looking at ourselves, we can learn more by using obviously *in*human processes as a mirror.

There are many ghosts to learn from here—ghosts past, and present, and yet to come. And these illusions are real cognitive events, real phenomena that we can study and explain. If there *appears* to be a ghost in the machine, that appearance is itself the hidden work of a machine.

The first sequence of *The Machine in the Ghost*, "The Simple Math of Evolution," aims to communicate the dissonance and divergence between our hereditary history, our present-day biology, and our ultimate aspirations. This will require digging deeper than is common in introductions to evolution for non-biologists, which often restrict their attention to surface-level features of natural selection.

The third sequence, "A Human's Guide to Words," discusses the basic relationship between cognition and concept formation. This is followed by a longer essay introducing Bayesian inference.

Bridging the gap between these topics, "Fragile Purposes" abstracts from human cognition and evolution to the idea of minds and goal-directed systems at their most general. These essays serve the secondary purpose of explaining the author's general approach to philosophy and the science of rationality, which is strongly informed by his work in AI.

## Rebuilding Intelligence

Yudkowsky is a decision theorist and mathematician who works on foundational issues in Artificial General Intelligence (AGI), the theoretical study of domain-general problem-solving systems. Yudkowsky's work in AI has been a major driving force behind his exploration of the psychology of human rationality, as he noted in his very first blog post on *Overcoming Bias*, The Martial Art of Rationality:

> Such understanding as I have of rationality, I acquired in the course of wrestling with the challenge of Artificial General Intelligence (an endeavor which, to actually succeed, would require sufficient mastery of rationality to build a complete working rationalist out of toothpicks and rubber bands). In most ways the AI problem is enormously more demanding than the personal art of rationality, but in some ways it is actually easier. In the martial art of mind, we need to acquire the real-time procedural skill of pulling the right levers at the right time on a large, pre-existing thinking machine whose innards are not end-user-modifiable. Some of the machinery is optimized for evolutionary selection pressures that run directly counter to our declared goals in using it. Deliberately we decide that we want to seek only the truth; but our brains have hardwired support for rationalizing falsehoods. [...]

> Trying to synthesize a personal art of rationality, using the science of rationality, may prove awkward: One imagines trying to invent a martial art using an abstract theory of physics, game theory, and human anatomy. But humans are not reflectively blind; we do have a native instinct for introspection. The inner eye is not sightless; but it sees blurrily, with systematic distortions. We need, then, to apply the science to our intuitions, to use the abstract knowledge to correct our mental movements and augment our metacognitive skills. We are not writing a computer program to make a string puppet execute martial arts forms; it is our own mental limbs that we must move. Therefore we must connect theory to practice. We must come to see what the science means, for ourselves, for our daily inner life.

From Yudkowsky's perspective, I gather, talking about human rationality without saying anything interesting about AI is about as difficult as talking about AI without saying anything interesting about rationality.

In the long run, Yudkowsky predicts that AI will come to surpass humans in an "intelligence explosion," a scenario in which self-modifying AI improves its own ability to productively redesign itself, kicking off a rapid succession of further self-improvements. The term "technological singularity" is sometimes used in place of "intelligence explosion;" until January 2013, MIRI was named "the Singularity Institute for Artificial Intelligence" and hosted an annual Singularity Summit. Since then, Yudkowsky has come to favor I.J. Good's older term, "intelligence explosion," to help distinguish his views from other futurist predictions, such as Ray Kurzweil's exponential technological progress thesis.[2]

Technologies like smarter-than-human AI seem likely to result in large societal upheavals, for the better or for the worse. Yudkowsky coined the term "Friendly AI theory" to refer to research into techniques for aligning an AGI's preferences with the preferences of humans. At this point, very little is known about when generally intelligent software might be invented, or what safety approaches would work well in such cases. Present-day autonomous AI can already be quite challenging to verify and validate with much confidence, and many current techniques are not likely to generalize to more intelligent and adaptive systems. "Friendly AI" is therefore closer to a menagerie of basic mathematical and philosophical questions than to a well-specified set of programming objectives.

As of 2015, Yudkowsky's views on the future of AI continue to be debated by technology forecasters and AI researchers in industry and academia, who have yet to converge on a consensus position. Nick Bostrom's book *Superintelligence* provides a big-picture summary of the many moral and strategic questions raised by smarter-than-human AI.[3]

For a general introduction to the field of AI, the most widely used textbook is Russell and Norvig's *Artificial Intelligence: A Modern Approach*.[4] In a chapter discussing the moral and philosophical questions raised by AI, Russell and Norvig note the technical difficulty of specifying good behavior in strongly adaptive AI:

> [Yudkowsky] asserts that friendliness (a desire not to harm humans) should be designed in from the start, but that the designers should recognize both that their own designs may be flawed, and that the robot will learn and evolve over time. Thus the challenge is one of mechanism design—to define a mechanism for evolving AI systems under a system of checks and balances, and to give the systems utility functions that will remain friendly in the face of such changes. We

> can't just give a program a static utility function, because circumstances, and our desired responses to circumstances, change over time.

Disturbed by the possibility that future progress in AI, nanotechnology, biotechnology, and other fields could endanger human civilization, Bostrom and Ćirković compiled the first academic anthology on the topic, *Global Catastrophic Risks*.[5] The most extreme of these are the *existential risks*, risks that could result in the permanent stagnation or extinction of humanity.[6]

People (experts included) tend to be *extraordinarily bad* at forecasting major future events (new technologies included). Part of Yudkowsky's goal in discussing rationality is to figure out which biases are interfering with our ability to predict and prepare for big upheavals well in advance. Yudkowsky's contributions to the *Global Catastrophic Risks* volume, "Cognitive biases potentially affecting judgement of global risks" and "Artificial intelligence as a positive and negative factor in global risk," tie together his research in cognitive science and AI. Yudkowsky and Bostrom summarize near-term concerns along with long-term ones in a chapter of the *Cambridge Handbook of Artificial Intelligence*, "The ethics of artificial intelligence."[7]

Though this is a book about *human* rationality, the topic of AI has relevance as a source of simple illustrations of aspects of human cognition. Long- term technology forecasting is also one of the more important applications of Bayesian rationality, which can model correct reasoning even in domains where the data is scarce or equivocal.

Knowing the design can tell you much about the designer; and knowing the designer can tell you much about the design.

We'll begin, then, by inquiring into what our own designer can teach us about ourselves.

---

**1. Gilbert Ryle, *The Concept of Mind* (University of Chicago Press, 1949).**

**2. Irving John Good, "Speculations Concerning the First Ultraintelligent Machine," in Advances in Computers, ed. Franz L. Alt and Morris Rubinoff, vol. 6 (New York: Academic Press, 1965), 31–88, doi:10.1016/S0065- 2458(08)60418- 0.**

**3. Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014).**

**4. Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. (Upper Saddle River, NJ: Prentice-Hall, 2010).**

**5. Bostrom and Ćirković, *Global Catastrophic Risks*.**

**6. An example of a possible existential risk is the "grey goo" scenario, in which molecular robots designed to efficiently self-replicate do their job too well, rapidly outcompeting living organisms as they consume the Earth's available matter.**

**7. Nick Bostrom and Eliezer Yudkowsky, "The Ethics of Artificial Intelligence," in *The Cambridge Handbook of Artificial Intelligence*, ed. Keith Frankish and William Ramsey (New York: Cambridge University Press, 2014).**

# Rationality: An Introduction

In the autumn of 1951, a football game between Dartmouth and Princeton turned unusually rough. A pair of psychologists, Dartmouth's Albert Hastorf and Princeton's Hadley Cantril, decided to ask students from both schools which team had initiated the rough play. Nearly everyone agreed that Princeton hadn't started it; but 86% of Princeton students believed that Dartmouth had started it, whereas only 36% of Dartmouth students blamed Dartmouth. (Most Dartmouth students believed "both started it.")

When shown a film of the game later and asked to count the infractions they saw, Dartmouth students claimed to see a mean of 4.3 infractions by the Dartmouth team (and identified half as "mild"), whereas Princeton students claimed to see a mean of 9.8 Dartmouth infractions (and identified a third as "mild").[1]

When something we value is threatened—our world-view, our in-group, our social standing, or something else we care about—our thoughts and perceptions rally to their defense.[2,3] Some psychologists go so far as to hypothesize that the human ability to come up with explicit justifications for our conclusions evolved *specifically* to help us win arguments.[4]

One of the basic insights of 20th-century psychology is that human behavior is often driven by sophisticated unconscious processes, and the stories we tell ourselves about our motives and reasons are much more biased and confabulated than we realize. We often fail, in fact, to realize that we're doing any story-telling. When we seem to "directly perceive" things about ourselves in introspection, it often turns out to rest on tenuous implicit causal models.[5,6] When we try to argue for our beliefs, we can come up with shaky reasoning bearing no relation to how we first arrived at the belief.[7] Rather than trusting explanations in proportion to their predictive power, we tend to trust *stories* in proportion to their psychological appeal.

How can we do better? How can we arrive at a realistic view of the world, when we're so prone to rationalization? How can we come to a realistic view of our mental lives, when our thoughts *about* thinking are also suspect?

What's the *least* shaky place we could put our weight down?

## The Mathematics of Rationality

At the turn of the 20th century, coming up with simple (e.g., set-theoretic) axioms for arithmetic gave mathematicians a clearer standard by which to judge the correctness of their conclusions. If a human or calculator outputs "2 + 2 = 4," we can now do more than just say "that seems intuitively right." We can explain *why* it's right, and we can prove that its rightness is tied in systematic ways to the rightness of the rest of arithmetic.

But mathematics lets us model the behaviors of physical systems that are a lot more interesting than a pocket calculator. We can also formalize *rational belief in general*, using probability theory to pick out features held in common by all successful forms of

inference. We can even formalize *rational behavior in general* by drawing upon decision theory.

Probability theory defines how we would ideally reason in the face of uncertainty, if we had the requisite time, computing power, and mental control. Given some background knowledge (*priors*) and a new piece of evidence, probability theory uniquely and precisely defines the best set of new beliefs (*posterior*) I could adopt. Likewise, decision theory defines what action I should take based on my beliefs. For any consistent set of beliefs and preferences I could have, there is a decision-theoretic answer to how I should then act in order to satisfy my preferences.

Suppose you find out that one of your six classmates has a crush on you—perhaps you get a letter from a secret admirer, and you're sure it's from one of those six—but you have no idea which of the six it is. Bob happens to be one of those six classmates. If you have no special reason to think Bob's any likelier (or any less likely) than the other five candidates, then what are the odds that Bob is the one with the crush?

Answer: The odds are 1:5. There are six possibilities, so a wild guess would result in you getting it right once for every five times you got it wrong, on average.

We can't say, "Well, I have no idea who has a crush on me; maybe it's Bob, or maybe it's not. So I'll just say the odds are fifty-fifty." Even if we would rather say "I don't know" or "Maybe" and stop there, the right answer is still 1:5. This follows from the assumption that there are six possibilities and you have no reason to favor one of them over any of the others.[8]

Suppose that you've *also* noticed you get winked at by people ten times as often when they have a crush on you. If Bob then winks at you, that's a new piece of evidence. In that case, it would be a mistake to stay skeptical about whether Bob is your secret admirer; the 10:1 odds in favor of "a random person who winks at me has a crush on me" outweigh the 1:5 odds against "Bob has a crush on me."

It would *also* be a mistake to say, "That evidence is so strong, it's a sure bet that he's the one who has the crush on me! I'll just assume from now on that Bob is into me." Overconfidence is just as bad as underconfidence.

In fact, there's only one viable answer to this question too. To change our mind from the 1:5 prior odds in response to the evidence's 10:1 likelihood ratio, we multiply the left sides together and the right sides together, getting 10:5 posterior odds, or 2:1 odds in favor of "Bob has a crush on me." Given our assumptions and the available evidence, guessing that Bob has a crush on you will turn out to be correct 2 times for every 1 time it turns out to be wrong. Equivalently: the probability that he's attracted to you is 2/3. Any other confidence level would be inconsistent.

It turns out that given very modest constraints, the question "What should I believe?" has an objectively right answer. It has a right answer when you're wracked with uncertainty, not just when you have a conclusive proof. There is always a correct amount of confidence to have in a statement, even when it looks more like a "personal belief" instead of an expert-verified "fact."

Yet we often talk as though the existence of uncertainty and disagreement makes beliefs a mere matter of taste. We say "that's just my opinion" or "you're entitled to your opinion," as though the assertions of science and math existed on a different and

higher plane than beliefs that are merely "private" or "subjective." To which economist Robin Hanson has responded:[9]

> You are never entitled to your opinion. Ever! You are not even entitled to "I don't know." You are entitled to your desires, and sometimes to your choices. You might own a choice, and if you can choose your preferences, you may have the right to do so. But your beliefs are not about you; beliefs are about the world. Your beliefs should be your best available estimate of the way things are; anything else is a lie. [ . . . ]
>
> It is true that some topics give experts stronger mechanisms for resolving disputes. On other topics our biases and the complexity of the world make it harder to draw strong conclusions. [ . . . ]
>
> But never forget that on any question about the way things are (or should be), and in any information situation, there *is* always a best estimate. You are only entitled to your best honest effort to find that best estimate; anything else is a lie.

Our culture hasn't internalized the lessons of probability theory—that the correct answer to questions like "How sure can I be that Bob has a crush on me?" is just as logically constrained as the correct answer to a question on an algebra quiz or in a geology textbook.

Our brains are kludges slapped together by natural selection. Humans aren't perfect reasoners or perfect decision-makers, any more than we're perfect calculators. Even at our best, we don't compute the *exact* right answer to "what should I think?" and "what should I do?"[10]

And yet, knowing we can't become *fully* consistent, we can certainly still get better. Knowing that there's an ideal standard we can compare ourselves to—what researchers call Bayesian rationality—can guide us as we improve our thoughts and actions. Though we'll never be perfect Bayesians, the mathematics of rationality can help us understand *why* a certain answer is correct, and help us spot exactly where we messed up.

Imagine trying to learn math through rote memorization alone. You might be told that "10 + 3 = 13," "31 + 108 = 139," and so on, but it won't do you a lot of good unless you understand the pattern behind the squiggles. It can be a lot harder to seek out methods for improving your rationality when you don't have a general framework for judging a method's success. The purpose of this book is to help people build for themselves such frameworks.

## Rationality Applied

The tightly linked essays in *How to Actually Change Your Mind* were originally written by Eliezer Yudkowsky for the blog *Overcoming Bias*. Published in the late 2000s, these posts helped inspire the growth of a vibrant community interested in rationality and self-improvement.

*Map and Territory* was the first such collection. *How to Actually Change Your Mind* is the second. The full six-book set, titled *Rationality: From AI to Zombies*, can be found on *Less Wrong* at **http://lesswrong.com/rationality**.

One of the rationality community's most popular writers, Scott Alexander, has previously observed:[11]

> [O]bviously it's useful to have as much evidence as possible, in the same way it's useful to have as much money as possible. But equally obviously it's useful to be able to use a limited amount of evidence wisely, in the same way it's useful to be able to use a limited amount of money wisely.

Rationality techniques help us get more mileage out of the evidence we have, in cases where the evidence is inconclusive or our biases are distorting how we interpret the evidence.

This applies to our personal lives, as in the tale of Bob. It applies to disagreements between political factions and sports fans. And it applies to philosophical puzzles and debates about the future trajectory of technology and society. Recognizing that the same mathematical rules apply to each of these domains (and that in many cases the same cognitive biases crop up), *How to Actually Change Your Mind* freely moves between a wide range of topics.

The first sequence of essays in this book, **Overly Convenient Excuses**, focuses on probabilistically "easy" questions—ones where the odds are extreme, and systematic errors *seem* like they should be particularly easy to spot.…

From there, we move into murkier waters with **Politics and Rationality**. Politics—or rather, mainstream national politics of the sort debated by TV pundits—is famous for its angry, unproductive discussions. On the face of it, there's something surprising about that. Why do we take political disagreements so personally, even though the machinery and effects of national politics are often so distant from us in space or in time? For that matter, why do we not become *more* careful and rigorous with the evidence when we're dealing with issues we deem important?

The Dartmouth-Princeton game hints at an answer. Much of our reasoning process is really rationalization—story-telling that makes our current beliefs feel more coherent and justified, without necessarily improving their accuracy. **Against Rationalization** speaks to this problem, followed by **Seeing with Fresh Eyes**, on the challenge of recognizing evidence that doesn't fit our expectations and assumptions.

In practice, leveling up in rationality often means encountering interesting and powerful new ideas and colliding more with the in-person rationality community. **Death Spirals** discusses some important hazards that can afflict groups united around common interests and amazing shiny ideas, which rationalists will need to overcome if they're to translate their high-minded ideas into real-world effectiveness. *How to Actually Change Your Mind* then concludes with a sequence on **Letting Go**.

Our natural state *isn't* to change our minds like a Bayesian would. Getting the Dartmouth and Princeton students to notice what they're actually seeing won't be as easy as reciting the axioms of probability theory to them. As philanthropic research analyst Luke Muehlhauser writes in "The Power of Agency":[12]

> You are not a Bayesian homunculus whose reasoning is "corrupted" by cognitive biases.
>
> You just *are* cognitive biases.

Confirmation bias, status quo bias, correspondence bias, and the like are not tacked on to our reasoning; they are its very substance.

That doesn't mean that debiasing is impossible. We aren't perfect calculators underneath all our arithmetic errors, either. Many of our mathematical limitations result from very deep facts about how the human brain works. Yet we can train our mathematical abilities; we can learn when to trust and distrust our mathematical intuitions; we can shape our environments to make things easier on us. And if we're wrong today, we can be less so tomorrow.

---

[1] Albert Hastorf and Hadley Cantril, "They Saw a Game: A Case Study," *Journal of Abnormal and Social Psychology* 49 (1954): 129–134, http://www2.psych.ubc.ca/~schaller/Psyc590Readings/Hastorf1954.pdf.

[2] Emily Pronin, "How We See Ourselves and How We See Others," *Science* 320 (2008): 1177–1180.

[3] Robert P. Vallone, Lee Ross, and Mark R. Lepper, "The Hostile Media Phenomenon: Biased Perception and Perceptions of Media Bias in Coverage of the Beirut Massacre," *Journal of Personality and Social Psychology* 49 (1985): 577–585, http://ssc.wisc.edu/~jpiliavi/965/hwang.pdf.

[4] Hugo Mercier and Dan Sperber, "Why Do Humans Reason? Arguments for an Argumentative Theory," *Behavioral and Brain Sciences* 34 (2011): 57–74, http://hal.archives-ouvertes.fr/file/index/docid/904097/filename/MercierSperberWhydohumansreason.pdf.

[5] Richard E. Nisbett and Timothy D. Wilson, "Telling More than We Can Know: Verbal Reports on Mental Processes," *Psychological Review* 84 (1977): 231–259, http://people.virginia.edu/~tdw/nisbett&wilson.pdf.

[6] Eric Schwitzgebel, *Perplexities of Consciousness* (MIT Press, 2011).

[7] Jonathan Haidt, "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment," *Psychological Review* 108, no. 4 (2001): 814–834, doi:10.1037/0033-295X.108.4.814.

[8] We're also assuming, unrealistically, that you can really be certain the admirer is one of those six people, and that you aren't neglecting other possibilities. (What if more than one of your classmates has a crush on you?)

[9] Robin Hanson, "You Are Never Entitled to Your Opinion," *Overcoming Bias* (Blog), 2006, http://www.overcomingbias.com/2006/12/you_are_never_e.html.

[10] We lack the computational resources (and evolution lacked the engineering expertise and foresight) to iron out all our bugs. Indeed, even a maximally efficient reasoner in the real world would still need to rely on heuristics and approximations. The best possible computationally tractable algorithms for changing beliefs would still fall short of probability theory's consistency.

[11] Scott Alexander, "Why I Am Not Rene Descartes," *Slate Star Codex* (Blog), 2014, http://slatestarcodex.com/2014/11/27/why-i-am-not-rene-descartes/.

[12] Luke Muehlhauser, “The Power of Agency,” *Less Wrong* (Blog), 2011, http://lesswrong.com/lw/5i8/the_power_of_agency/.

# Beginnings: An Introduction

This, the final book of *Rationality: From AI to Zombies*, is less a conclusion than a call to action. In keeping with *Becoming Stronger*'s function as a jumping-off point for further investigation, I'll conclude by citing resources the reader can use to move beyond these sequences and seek out a fuller understanding of Bayesianism.

This text's definition of normative rationality in terms of Bayesian probability theory and decision theory is standard in cognitive science. For an introduction to the heuristics and biases approach, see Baron's *Thinking and Deciding*.[1] For a general introduction to the field, see the *Oxford Handbook of Thinking and Reasoning*.[2]

The arguments made in these pages about the *philosophy* of rationality are more controversial. Yudkowsky argues, for example, that a rational agent should one-box in Newcomb's Problem—a minority position among working decision theorists.[3] (See Holt for a nontechnical description of Newcomb's Problem.[4]) Gary Drescher's *Good and Real* independently comes to many of the same conclusions as Yudkowsky on philosophy of science and decision theory.[5] As such, it serves as an excellent book-length treatment of the core philosophical content of *Rationality: From AI to Zombies*.

Talbott distinguishes several views in Bayesian epistemology, including E. T. Jaynes's position that not all possible priors are equally reasonable.[6,7] Like Jaynes, Yudkowsky is interested in supplementing the Bayesian optimality criterion for belief revision with an optimality criterion for priors. This aligns Yudkowsky with researchers who hope to better understand general-purpose AI via an improved theory of ideal reasoning, such as Marcus Hutter.[8] For a broader discussion of philosophical efforts to naturalize theories of knowledge, see Feldman.[9]

"Bayesianism" is often contrasted with "frequentism." Some frequentists criticize Bayesians for treating probabilities as subjective states of belief, rather than as objective frequencies of events. Kruschke and Yudkowsky have replied that frequentism is even more "subjective" than Bayesianism, because frequentism's probability assignments depend on the intentions of the experimenter.[10]

Importantly, this philosophical disagreement shouldn't be conflated with the distinction between Bayesian and frequentist data analysis methods, which can both be useful when employed correctly. Bayesian statistical tools have become cheaper to use since the 1980s, and their informativeness, intuitiveness, and generality have come to be more widely appreciated, resulting in "Bayesian revolutions" in many sciences. However, traditional frequentist methods re- main more popular, and in some contexts they are still clearly superior to Bayesian approaches. Kruschke's *Doing Bayesian Data Analysis* is a fun and accessible introduction to the topic.[11]

In light of evidence that training in statistics—and some other fields, such as psychology—improves reasoning skills outside the classroom, statistical literacy is directly relevant to the project of overcoming bias. (Classes in formal logic and informal fallacies have not proven similarly useful.)[12,13]

## An Art in its Infancy

We conclude with three sequences on individual and collective self- improvement. “Yudkowsky’s Coming of Age” provides a last in-depth illustration of the dynamics of irrational belief, this time spotlighting the author’s own intellectual history. “Challenging the Difficult” asks what it takes to solve a truly difficult problem—including demands that go beyond epistemic rationality. Finally, “The Craft and the Community” discusses rationality groups and group rationality, raising the questions:

- Can rationality be learned and taught?
- If so, how much improvement is possible?
- How can we be confident we’re seeing a real effect in a rationality intervention, and picking out the right cause?
- What community norms would make this process of bettering ourselves easier?
- Can we effectively collaborate on large-scale problems without sacrificing our freedom of thought and conduct?

Above all: What’s missing? What should be in the next generation of rationality primers—the ones that replace this text, improve on its style, test its prescriptions, supplement its content, and branch out in altogether new directions?

Though Yudkowsky was moved to write these essays by his own philosophical mistakes and professional difficulties in AI theory, the resultant material has proven useful to a much wider audience. The original blog posts inspired the growth of *Less Wrong*, a community of intellectuals and life hackers with shared interests in cognitive science, computer science, and philosophy. Yudkowsky and other writers on *Less Wrong* have helped seed the effective altruism movement, a vibrant and audacious effort to identify the most high-impact humanitarian charities and causes. These writings also sparked the establishment of the Center for Applied Rationality, a nonprofit organization that attempts to translate results from the science of rationality into useable techniques for self-improvement.

I don’t know what’s next—what other unconventional projects or ideas might draw inspiration from these pages. We certainly face no shortage of global challenges, and the art of applied rationality is a new and half-formed thing. There are not many rationalists, and there are many things left undone.

But wherever you’re headed next, reader—may you serve your purpose well.

---

**1. Jonathan Baron, *Thinking and Deciding* (Cambridge University Press, 2007).**

**2. Keith J. Holyoak and Robert G. Morrison, *The Oxford Handbook of Thinking and Reasoning* (Oxford University Press, 2013).**

**3. Bourget and Chalmers, “What Do Philosophers Believe?”**

**4. Holt, “Thinking Inside the Boxes.”**

**5. Gary L. Drescher, *Good and Real: Demystifying Paradoxes from Physics to Ethics* (Cambridge, MA: MIT Press, 2006).**

**6. William Talbott, “Bayesian Epistemology,” in *The Stanford Encyclopedia of Philosophy*, Fall 2013, ed. Edward N. Zalta.**

**7. Jaynes, *Probability Theory*.**

**8. Marcus Hutter, *Universal Artificial Intelligence: Sequential Decisions Based On Algorithmic Probability* (Berlin: Springer, 2005), doi:10.1007/b138233.**

**9. Richard Feldman, "Naturalized Epistemology," in *The Stanford Encyclopedia of Philosophy*, Summer 2012, ed. Edward N. Zalta.**

**10. John K. Kruschke, "What to Believe: Bayesian Methods for Data Analysis," *Trends in Cognitive Sciences* 14, no. 7 (2010): 293–300.**

**11. John K. Kruschke, *Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan* (Academic Press, 2014).**

**12. Geoffrey T. Fong, David H. Krantz, and Richard E. Nisbett, "The Effects of Statistical Training on Thinking about Everyday Problems," *Cognitive Psychology* 18, no. 3 (1986): 253–292, doi:10.1016/0010- 0285(86)90001- 0.**

**13. Paul J. H. Schoemaker, "The Role of Statistical Knowledge in Gambling Decisions: Moment vs. Risk Dimension Approaches," *Organizational Behavior and Human Performance* 24, no. 1 (1979): 1–17.**

# The World: An Introduction

Previous essays have discussed human reasoning, language, goals, and social dynamics. Mathematics, physics, and biology were cited to explain patterns in human behavior, but little has been said about humanity's place in nature, or about the natural world in its own right.

Just as it was useful to contrast humans as *goal-oriented systems* with inhuman processes in evolutionary biology and artificial intelligence, it will be useful in the coming sequences of essays to contrast humans as *physical systems* with inhuman processes that *aren't* mind-like.

We humans are, after all, built out of inhuman parts. The world of atoms looks nothing like the world as we ordinarily think of it, and certainly looks nothing like the world's conscious denizens as we ordinarily think of them. As Giulio Giorello put the point in an interview with Daniel Dennett: "Yes, we have a soul. But it's made of lots of tiny robots."[1]

*Mere Reality* collects seven sequences of essays on this topic. The first three introduce the question of how the human world relates to the world revealed by physics: "Lawful Truth" (on the basic links between physics and human cognition), "Reductionism 101" (on the project of scientifically explaining phenomena), and "Joy in the Merely Real" (on the emotional, personal significance of the scientific world-view). This is followed by two sequences that go into more depth on specific academic debates: "Physicalism 201" (on the hard problem of consciousness) and "Quantum Physics and Many Worlds" (on the measurement problem in physics). Finally, the sequence "Science and Rationality" and the essay A Technical Explanation of Technical Explanation tie these ideas together and relate them to scientific practice.

The discussions of consciousness and quantum physics illustrate the relevance of reductionism to present-day controversies in science and philosophy. For those interested in a bit of extra context, I'll say a few more words about those two topics here. For those eager to skip ahead: skip ahead!

# Minds in the World

**Can we ever know what it's like to be a bat?**

We can certainly develop better cognitive models for predicting bat behav-ior, or more fine-grained models of bat neurology—but it isn't obvious that this would tell us what echolocation subjectively feels like, or what flying feels like, *from the bat's point of view*.

Indeed, it seems as though we could never even be certain that there *is* anything it's like to be a bat. Why couldn't an unconscious automaton replicate all the overt behaviors of a conscious agent to arbitrary precision? (Philosophers call such automata "zombies," though they have little in common with the zombies of folklore—who are *quite visibly* different from conscious agents!)

A race of alien psychologists would run into the same problem in trying to model *human* consciousness. They might arrive at a perfect predictive model of what we say and do when we see a red rose, but that wouldn't mean that the aliens fully understand what redness feels like "from the inside."

Running with examples like these, philosophers like Thomas Nagel and David Chalmers have argued that third-person cognitive and neural models can never fully capture first-person consciousness.[2,3] No matter how much we know about a physical system, it is always logically possible, on this view, that the system has no first-person experiences. Traditional dualism, with its immaterial souls freely floating around violating physical laws, may be false; but Chalmers insists on a weaker thesis, that consciousness is a "further fact" not fully explainable by the physical facts.

A number of philosophers and scientists have found this line of reasoning persuasive.[4] If we feel this argument's intuitive force, should we grant its conclusion and ditch physicalism?

We certainly shouldn't reject it just because it *sounds strange* or feels vaguely unscientific. But how does the argument stand up to a *technical* understanding of how explanation and belief work? Are there any hints we can take from the history of science, or from our understanding of the physical mechanisms underlying evidence? "Physicalism 201" will return to this question.

# Worlds in the World

Quantum mechanics is our best mathematical model of the universe to date, powerfully confirmed by a century of tests. The theory posits a complex- numbered "probability amplitude," so called because a specific operation (squaring the number's absolute value—the Born rule) lets us probabilistically predict phenomena at small scales and extreme energy levels. This amplitude changes deterministically in accord with the Schrödinger equation. In the process, it often enters odd states called "superpositions."

Yet when we perform experiments, the superpositions seem to vanish without a trace. When we aren't looking, the Schrödinger equation appears to capture everything there is to know about the dynamics of physical systems. When we *are* looking, though, this clean determinism is replaced by Born's probabilistic rule. It's as though the ordinary laws of physics are suddenly suspended whenever we make "observations." As John Stewart Bell put the point:

> It would seem that the theory is exclusively concerned about "results of measurements" and has nothing to say about anything else. What exactly qualifies some physical systems to play the role of the "measurer"? Was the wavefunction of the world waiting to jump for thousands of millions of years until a single-celled living creature appeared? Or did it have to wait a little longer, for some better qualified system . . . with a PhD?

Everyone agrees that this strange mix of Schrödinger and Born's rules has proved empirically adequate. However, the question of exactly *when* Born's rule enters the mix, and what it all *means*, has produced a chaos of different views on the nature of quantum mechanics.

Early on, the Copenhagen school—Niels Bohr and other originators of quantum theory—splintered into several standard ways of talking about the experimental results and the odd formalism used to predict them. Some, taking the theory's focus on "measurements" and "observations" quite literally, proposed that consciousness plays a fundamental role in physical law, intervening to cause complex amplitudes to "collapse" into observables. Others, led by Werner Heisenberg, advocated a non-realistic view according to which physics is about our states of knowledge rather than about any objective reality. Yet another Copenhagen tradition, summed up in the slogan "shut up and calculate," warned against metaphysical speculation of all kinds.

Yudkowsky uses this scientific controversy as a proving ground for some central ideas from previous sequences: map-territory distinctions, mysterious answers, Bayesianism, and Occam's Razor. Since he is not a physicist—and neither am I—I'll provide some outside sources here for readers who want to vet his arguments or learn more about his physics examples.

Tegmark's *Our Mathematical Universe* discusses a number of relevant ideas in philosophy and physics.[5] Among Tegmark's more novel ideas is his argument that all consistent mathematical structures exist, including worlds with physical laws and boundary conditions entirely unlike our own. He distinguishes these Tegmark worlds from multiverses in more scientifically mainstream hypotheses—e.g., worlds in stochastic eternal inflationary models of the Big Bang and in Hugh Everett's many-worlds interpretation of quantum physics.

Yudkowsky discusses many-worlds interpretations at greater length, as a response to the Copenhagen interpretations of quantum mechanics. Many-worlds has become very popular in recent decades among physicists, especially cosmologists. However, a number of physicists continue to reject it or maintain agnosticism. For a (mostly) philosophically mainstream introduction to this debate, see Albert's *Quantum Mechanics and Experience*.[6] See also the *Stanford Encyclopedia of Philosophy*'s introduction to "Measurement in Quantum Theory,"[7] and their introduction to several of the views associated with "many worlds" in "Everett's Relative-State Formulation"[8] and "Many-Worlds Interpretation."[9]

On the less theoretical side, Epstein's *Thinking Physics* is a great text fortraining physical intuitions.[10] It's worth keeping in mind that just as one can understand most of cognitive science without understanding the nature of subjective awareness, one can understand most of physics without having a settled view of the ultimate nature (and size!) of the physical world.

---

1. Daniel C. Dennett, Freedom Evolves (Viking Books, 2003).
2. David J. Chalmers, The Conscious Mind: In Search of a Fundamental Theory (New York: OxfordUniversity Press, 1996).
3. Thomas Nagel, "What Is It Like to Be a Bat?," Philosophical Review 83, no. 4 (1974): 435–450, http://www.jstor.org/stable/2183914.
4. In a survey of Anglophone professional philosophers, 56.5% endorsed physicalism, 27.1% endorsed anti-physicalism, and 16.4% endorsed other views (e.g., "I don't know").[11] Most philosophers reject the metaphysical possibility of Chalmers's "zombies," but there is no consensus about why, exactly, Chalmers's zombie argument fails. Kirk summarizes contemporary positions on phenomenal consciousness, giving arguments that resemble Yudkowsky's against the possibility of knowing or referring to irreducible qualia.[12]
5. Max Tegmark, Our Mathematical Universe: My Quest for the Ultimate Nature of

Reality (Random House LLC, 2014).
6. David Z. Albert, Quantum Mechanics and Experience (Harvard University Press, 1994).
7. Henry Krips, “Measurement in Quantum Theory,” in The Stanford Encyclopedia of Philosophy, Fall 2013, ed. Edward N. Zalta.
8. Jeffrey Barrett, Everett’s Relative-State Formulation of Quantum Mechanics, ed. Edward N. Zalta, http://plato.stanford.edu/archives/fall2008/entries/qm- everett/.
9. Lev Vaidman, “Many-Worlds Interpretation of Quantum Mechanics,” in The Stanford Encyclopediaof Philosophy, Fall 2008, ed. Edward N. Zalta. 838
10. Lewis Carroll Epstein, Thinking Physics: Understandable Practical Reality, 3rd Edition (Insight Press, 2009).
11. David Bourget and David J. Chalmers, “What Do Philosophers Believe?,” Philosophical Studies (2013): 1–36.
12. Robert Kirk, Mind and Body (McGill-Queen’s University Press, 2003).

# Ends: An Introduction

Value theory is the study of what people care about. It's the study of our goals, our tastes, our pleasures and pains, our fears and our ambitions.

That includes conventional morality. Value theory subsumes things we *wish* we cared about, or would care about if we were wiser and better people—not just things we already do care about.

Value theory also subsumes mundane, everyday values: art, food, sex, friendship, and everything else that gives life its affective valence. Going to the movies with your friend Sam can be something you value even if it's not a *moral* value.

We find it useful to reflect upon and debate our values because how we act is not always how we wish we'd act. Our preferences can conflict with each other. We can desire to have a different set of desires. We can lack the will, the attention, or the insight needed to act the way we'd like to.

Humans do care about their actions' consequences, but not consistently enough to formally qualify as agents with utility functions. That humans don't act the way they wish they would is what we mean when we say "humans aren't instrumentally rational."

## Theory and Practice

Adding to the difficulty, there exists a gulf between how we *think* we wish we'd act, and how we *actually* wish we'd act.

Philosophers disagree wildly about what we want—as do psychologists, and as do politicians—and about what we ought to want. They disagree even about *what it means* to "ought" to want something. The history of moral theory, and the history of human efforts at coordination, is piled high with the corpses of failed Guiding Principles to True Ultimate No-Really-This-Time-I-Mean-It Normativity.

If you're trying to come up with a *reliable* and *pragmatically useful* specification of your goals—not just for winning philosophy debates, but (say) for designing safe autonomous adaptive AI, or for building functional institutions and organizations, or for making it easier to decide which charity to donate to, or for figuring out what virtues you should be cultivating—humanity's track record with value theory does not bode well for you.

*Mere Goodness* collects three sequences of blog posts on human value: "Fake Preferences" (on failed attempts at theories of value), "Value Theory" (on obstacles to developing a new theory, and some intuitively desirable fea- tures of such a theory), and "Quantified Humanism" (on the tricky question of how we should *apply* such theories to our ordinary moral intuitions and decision-making).

The last of these topics is the most important. The cash value of a normative theory is how well it translates into normative practice. Acquiring a deeper and fuller understanding of your values should make you better at actually fulfilling them. At a

bare minimum, your theory shouldn't *get in the way* of your practice. What good would it be, then, to know what's good?

Reconciling this art of applied ethics (and applied aesthetics, and applied economics, and applied psychology) with our best available data and theories often comes down to the question of when we should trust our snap judgments, and when we should ditch them.

In many cases, our explicit models of what we care about are so flimsy or impractical that we're better off trusting our vague initial impulses. In many other cases, we *can* do better with a more informed and systematic approach. There is no catch-all answer. We will just have to scrutinize examples and try to notice the different warning signs for "sophisticated theories tend to fail here" and "naive feelings tend to fail here."

## Journey and Destination

A recurring theme in the pages to come will be the question: *Where shall we go? What outcomes are actually valuable?*

To address this question, Yudkowsky coined the term "fun theory." Fun theory is the attempt to figure out what our ideal vision of the future would look like—not just the system of government or moral code we'd ideally live under, but the kinds of adventures we'd ideally go on, the kinds of music we'd ideally compose, and everything else we ultimately want out of life.

Stretched into the future, questions of fun theory intersect with questions of *transhumanism*, the view that we can radically improve the human condition if we make enough scientific and social progress.[1] Transhumanism occasions a number of debates in moral philosophy, such as whether the best long-term outcomes for sentient life would be based on *hedonism* (the pursuit of pleasure) or on more complex notions of *eudaimonia* (general well-being). Other futurist ideas discussed at various points in *Rationality: From AI to Zombies* include *cryonics* (storing your body in a frozen state after death, in case future medical technology finds a way to revive you), *mind uploading* (implementing human minds in synthetic hardware), and large-scale space colonization.

Perhaps surprisingly, fun theory is one of the more neglected applications of value theory. Utopia-planning has become rather passe—partly because it smacks of naiveté, and partly because we're empirically *terrible* at translating utopias into realities. Even the word *utopia* reflects this cynicism; it is derived from the Greek for "non-place."

Yet if we give up on the quest for a true, feasible utopia (or *eutopia*, "good place"), it's not obvious that the cumulative effect of our short-term pursuit of goals will be a future we find valuable over the long term. Value is not an inevitable feature of the world. Creating it takes work. Preserving it takes work.

This invites a second question: *How shall we get there? What is the relationship between good ends and good means?*

When we play a game, we want to enjoy the process. We don't generally want to just skip ahead to being declared the winner. Sometimes, the journey matters more than

the destination. Sometimes, the journey is *all* that matters.

Yet there are other cases where the reverse is true. Sometimes the end-state is just too important for "the journey" to factor into our decisions. If you're trying to save a family member's life, it's not necessarily a *bad* thing to get some enjoyment out of the process; but if you can increase your odds of success in a big way by picking a less enjoyable strategy . . .

In many cases, our values are concentrated in the outcomes of our actions, and in our future. We care about the way the world will end up looking— especially those parts of the world that can love and hurt and want.

How do detached, abstract theories stack up against vivid, affect-laden feelings in those cases? More generally: What is the moral relationship between actions and consequences?

Those are hard questions, but perhaps we can at least make progress on determining what we *mean* by them. What are we building into our concept of what's "valuable" at the very start of our inquiry?

---

**1. One example of a transhumanist argument is: "We could feasibly abolish aging and disease within a few decades or centuries. This would effectively end death by natural causes, putting us in the same position as organisms with negligible senescence—lobsters, Aldabra giant tortoises, etc. Therefore we should invest in disease prevention and anti-aging technologies." This idea qualifies as transhumanist because eliminating the leading causes of injury and death would drastically change human life.**

**Bostrom and Savulescu survey arguments for and against radical human enhancement, e.g., Sandel's objection that tampering with our biology too much would make life feel like less of a "gift."[2,3] Bostrom's "History of Transhumanist Thought" provides context for the debate.[4]**

**2. Nick Bostrom, "A History of Transhumanist Thought," *Journal of Evolution and Technology* 14, no. 1 (2005): 1–25, http://www.nickbostrom.com/papers/history.pdf.**

**3. Michael Sandel, "What's Wrong With Enhancement," Background material for the President's Council on Bioethics. (2002).**

**4. Nick Bostrom and Julian Savulescu, "Human Enhancement Ethics: The State of the Debate," in *Human Enhancement*, ed. Nick Bostrom and Julian Savulescu (2009).**