Against Rationalization II

Why a New Rationalization Sequence?

This is the first in a five-post mini-sequence about rationalization, which I intend to post one-per-day. And you may ask, why should we have such a sequence?

What is Rationalization and Why is it Bad?

For those of you just tuning in, rationalization is when you take a conclusion you want to reach and try to come up with a argument that concludes it. The argument looks very similar to one in which you started from data, evaluated as well as you could, and reached this conclusion naturally. Almost always similar enough to fool the casual observer, and often similar enough to *fool yourself.*

If you're deliberately rationalizing for an outside audience, that's out-of-scope for this sequence. All the usual ethics and game theory apply.

But if you're involuntarily rationalizing and fooling yourself, then you've failed at epistemics. And your arts have turned against you. Know a lot about scientific failures? Now you can find them in *all* the studies you didn't like!

Didn't Eliezer Already Do This?

Eliezer wrote the against rationalization sequence back in 2007/8. If you haven't read it, you probably should. It does a good job of describing what rationalization is, how it can happen, and how bad it can be. It does *not* provide a lot of tools for you to use in protecting yourself from rationalization. That's what I'll be focusing on here.

And, besides, if we don't revisit a topic this important every decade or so with new developments, then what is this community for?

Is There Hope?

Periodically, I hear someone give up on logical argument completely. "You can find an argument for anything," they say, "Forget logic. Trust [your gut / tradition / me] instead." Which brushes over the question of whether the proposed alternative is any better. There is no royal road to knowledge.

Still, the question needs answering. If rationalization looks just like logic, can we ever escape Cartesian Doubt?

The Psychiatrist Paradox

A common delusion among grandiose schizophrenics in institutions is that they are themselves psychiatrists. Consider a particularly underfunded mental hospital, in which the majority of people who "know" themselves to be psychiatrists are wrong. No examination of the evidence will convince them otherwise. No matter how overwhelming, some reason to disbelieve will be found.

Given this, should any amount of evidence suffice to convince you that you are such a psychiatrist?

I am not aware of any resolution to this paradox.

The Dreaming Paradox

But the Psychiatrist Paradox is based on an *absolute* fixed belief and *total* rationalization as seen in theoretically ideal schizophrenics. (How closely do real-world schizophrenics approximate this ideal? That question is beyond the scope of this document.) Let's consider people a little more reality-affiliated: the dreaming.

Given that any evidence of awakeness is a thing that can be dreamed, should you ever be more than 90% confident you're awake? (Assuming 16 hours awake and 2 dreaming in a typical 24 hour period.)

(Boring answer: forget confidence, always *act on the assumption* that you're awake because it's erring on the side of safety. We'll come back to this thought.)

(Also boring: most lucid dreaming enthusiasts report they do find evidence of wakefulness or dreaminess which dreams never forge. Assume you haven't found any for yourself.)

Here's my test: I ask my computer to prime factor a large number (around ten digits) and check it by hand. I can dream many things, but I'm not going to dream that my computer doesn't have the `factor` program, nor will I forget how to multiply. And I *can't* dream that it factored correctly, because I can't factor numbers that big.

You can't outsmart an absolute tendency to rationalize, but you can outsmart a *finite* one. Which, I suspect, is what we mostly have.

A Disclaimer Regarding Authorship

Before I start on the meat of the sequence (in the next post) I should make clear that not all these ideas are mine. Unfortunately, I've lost track of which ones are and which aren't, and of who proposed the ones which aren't. And the ones that aren't original to me have still gone through me enough to not be entirely as their original authors portrayed them.

If I tried to untangle this mess and credit properly, I'd never get this written. So onward. If you wish to fix some bit of crediting, leave a comment and I'll try to do something sensible.

Beyond Rationalization

Much of what appears here also applies to ordinary mistakes of logic. I'll try to tag such as they go.

The simplest ideal of thinking deals extensively with uncertainty of external facts, but trusts its own reasoning implicitly. Directly imitating this, when your own reasoning is not 100% trustworthy, is a bad plan. Hopefully this sequence will provide some alternatives.

Next: Red Flags for Rationalization

Red Flags for Rationalization

Previously: Why a New Rationalization Sequence?

What are Red Flags?

A red flag is a warning sign that you may have rationalized. Something that is practical to observe and more likely in the rationalization case than the non-rationalization.

Some are things which are likely to cause rationalization. Others are likely to be caused by it. One on this list is even based on common cause. (I don't have any based on selection on a common effect, but in theory there could be.)

How to Use Red Flags

Seeing a red flag doesn't necessarily mean that you have rationalized, but it's evidence. Likewise, just because you've rationalized doesn't mean your conclusion is wrong, only that it's not as supported as you thought.

So when one of these flags raises, don't give up on ever discovering truth; don't stop-halt-catch-fire; *definitely* don't invert your conclusion.

Just slow down. Take the hypothesis that you're rationalizing seriously and look for ways to test it. The rest of this sequence will offer tools for the purpose, but just paying attention is half the battle.

A lot of these things can be present to a greater or lesser degree, so you'll want to set thresholds. I'd guess an optimal setting has about 1/3 of triggers be true. High enough that you keep doing your checks seriously, but low because the payoff matrix is quite asymmetrical.

Basically use these as trigger-action planning. Trigger: anything on this list. Action: spend five seconds doing your agenty best to worry about rationalization.

Conflict of Interest

This is a classic reason to distrust *someone else*'s reasoning. If they have something to gain from you believing a conclusion separate from that conclusion being true, you have reason to be suspicious. But what does it mean for you to gain from you believing something apart from it being true?

Not Such Great Liars

Probably the simplest reason is that you need to deceive someone else. If you're not a practiced liar, the easiest way to do this is to deceive yourself.

Simple example: you're running late and need to give an estimate of when you'll arrive. If you say "ten minutes late" and arrive twenty minutes late, it looks like you hit another ten minutes' worth of bad luck, whereas saying "twenty minutes" looks like your fault. You're not good at straight-up lying, but if you *can convince yourself* you'll only be ten minutes late, all is well.

Unendorsed Values

Values aren't simple, and you aren't always in agreement with yourself. Let's illustrate this with examples:

Perhaps you believe that health and long life are more important that fleeting pleasures like ice cream, but there's a part of you that has a short time preference and knows ice cream is delicious. That part would love to convince the rest of you of a theory of nutrition that holds ice cream as healthy.

Perhaps you believe that you should follow scientific results wherever the evidence leads you, but it seems to be leading someplace that a professor at Duke predicted a few months ago, and there's a part of you that hates Duke. If that part can convince the rest of you that the data is wrong, you won't have to admit that somebody at Duke was right.

Wishful Thinking

A classic cause of rationalization. Expecting good things feels better than expecting bad things, so you'll want to believe it will all come out all right.

Catastrophizing Thinking

The opposite of wishful thinking. I'm not sure what the psychological root is, but it seems common in our community.

Conflict of Ego

The conclusion is: therefore I am a good person. The virtues I am strong at are the most important, and those I am weak at are the least. The work I do is vital to upholding civilization. The actions I took were justified. See Foster & Misra (2013) on Cognitive Dissonance and Affect.

Variant: therefore we are good people. Where "we" can be any group membership the thinker feels strongly about. Note that the individual need not have been involved in the virtue, work or action to feel pressure to rationalize it.

This is particularly insidious when "we" is defined partly by a large set of beliefs, such as the Social Justice Community or the Libertarian Party. Then it is tempting to rationalize that *every position* "*we*" *have ever taken was correct*.

In my experience, the communal variant is more common than the individual one, but that may be an artifact of my social circles.

Reluctance to Test

If you have an opportunity to gain more evidence on the question and feel reluctant, this is a bad sign. This one is illustrated by Harry and Draco discussing Hermione in HPMOR .

Suspicious Timing

Did you stop looking for alternatives as soon as you found this one?

Similarly, did you spend a lot longer looking for evidence on one side than the other?

Failure to Update

This was basically covered in Update Yourself Incrementally and One Argument Against An Army. The pattern of failing to update because the weight of evidence points the other way is a recognizable one.

The Feeling of Doing It

For some people, rationalization has a distinct subjective experience that you can train yourself to recognize. Eliezer writes about it in Singlethink and later refers to it as "don't even start to rationalize".

If anyone has experience trying to develop this skill, please leave a comment.

Agreeing with Idiots

True, reversed stupidity is not intelligence. Nevertheless, if you find yourself arriving at the same conclusion as a large group of idiots, this is a suspicious observation that calls for an explanation. Possibilities include:

- It's a coincidence: they got lucky. This can happen, but the more complex the conclusion, the less likely.
- They're not all that idiotic. People with terrible overall epistemics can still have solid understanding within their comfort zones.
- It's not really the same conclusion; it just sounds it when both are summarized poorly.
- You and they rationalized the conclusion following the same interest.

Naturally, it is this last possibility that concerns us. The less likely the first three, the more worrying the last one.

Disagreeing with Experts

If someone who is clearly established as an expert in the field (possibly by having notable achievements in it) disagrees with you, this is a bad sign. It's more a warning sign of bad logic in general than of rationalization in particular, but rationalization is a common cause of bad logic, and many of the same checks apply.

Next: Avoiding Rationalization

Avoiding Rationalization

Previously: Red Flags for Rationalization

It is often said that the best way to resist temptation is to avoid it. This includes the temptation to rationalize. You can avoid rationalization by denying it the circumstances in which it would take place.

When you avoid rationalization, you don't need to worry about infinite regress of metacognition, nor about overcompensation. As a handy bonus, you can demonstrate to others that you weren't rationalizing, which otherwise involves a lot of introspection and trust.

Here are three ways to do it:

Double Blinding

Identify the thing that would control your rationalization and arrange not to know it.

The trope namer is experimental science. You might be tempted to give better care to the experimental than the control group, but not if you don't know which is which. In many cases, you can maintain the blinding in statistical analysis as well, comparing "group A" to "group B" and only learning which is which after doing the math.

Similarly, if you are evaluating people (e.g. for a job) and are worried about subconscious sexism (or overcompensation for it), write a script or ask a friend to strip all gender indicators from the application.

Unfortunately, this technique requires you to *anticipate* the rationalization risk. Once you notice you might be rationalizing, it's usually too late to double-blind.

End-to-End Testing

A logical argument is a series of mental steps that build on each other. If the argument is direct, every step must be correct. As such, it is somewhat similar to a computer program. You wouldn't write a nontrivial computer program, look it over, say "that looks right" and push it to production. You would test it. With as close as possible to a full, end-to-end test before trusting it.

What does testing look like for an argument? Take a claim near the end of the argument which can be observed. Go and observe it. Since you have a concrete prediction, it should be a lot easier to make the observation.

Once you've got that, you don't need the long chain of argument that got you there. So it doesn't matter if you were rationalizing along the way. The bit at the end of the argument which you haven't chopped off still needs scrutinizing, but it's shorter, so you can give each bit more attention.

Suppose you're organizing some sort of event, and you want to not bother planning cleanup because you'll just tell all the crowd at the end "ok, everybody clean up now". You expect this to work out because of the known agentiness and competence of the crowd, overlap with HOPE attendees, estimates of the difficulty of cleanup tasks.... There's a lot of thinking that could have gone wrong, and a lot of communal pride pressuring you to rationalize. Instead of examining each question, try asking these people to clean up in some low-stakes situation.

(Do be careful to actually observe. I've seen people use their arguments to make predictions, then update on those predictions as if they were observations.)

This approach is also highly effective against non-rationalization errors in analysis as well. It's also especially good at spotting problems involving unknown unknowns.

The Side of Safety

Sometimes, you don't need to know.

Suppose you're about to drive a car, and are estimating whether you gain nontrivial benefit from a seatbelt. You conclude you will not, but note this could be ego-driven rationalization causing you to overestimate your driving talent. You *could* try to re-evaluate more carefully, or you could observe that the *costs* of wearing a seatbelt are trivial.

When you're uncertain about the quality of your reasoning, it makes sense to have a probability density function of posteriors for a yes-no question. But when the payoff table is lopsideded enough, you might find the vast bulk of the PDF is on one side of the decision threshold. And then you can just decide.

Next: Testing for Rationalization

Testing for Rationalization

Previously: Avoiding Rationalization

So you've seen reason to suspect you might be rationalizing, and you can't avoid the situation, what now?

Here are some tests you can apply to see whether you were rationalizing.

Reverse the Consequences

Let's explain this one via example:

Some Abstinence Educators like to use the "scotch tape" model of human sexuality. In it, sex causes people to attach to each other emotionally, but decreasingly with successive partners, just like tape is sticky, but less sticky when reused. Therefore, they say, you should avoid premarital sex because it will make you less attached to your eventual spouse.

Do you think this is a reasonable summary of human sexuality? Are people basically scotch tape?

Suppose the postscript had been: therefore you should have lots of premarital sex, so that you're not irrationally attached to someone. That way, when you believe you are in love and ready to commit, you really are.

Does this change your views on the scotch tape model? For many people, it does.

If so, then your views on the model are not driven by your analysis of its own merits, but by either your desire to have premarital sex, or your reluctance to admit Abstinence Educators could ever be right about anything.

(Or, possibly, your emotional revulsion at premarital sex or your affiliation to Abstinence Educators. The point of this section is unaffected.)

The point here is to break the argument into the piece to be evaluated, and the consequence of that piece which logically shouldn't effect the first part's validity but somehow does.

If the consequences seem hard to spin backwards, put on your Complete Monster Hat for a quick role-play. Suppose you think third-world charity is breaking the economies it goes to, and therefore you should keep your money for yourself, but this could be a rationalization from an unendorsed value (greed). Imagine yourself as a Dick Dastardly, a mustache-twirling villain who's trying to maximize suffering. Does Mr. Dastardly give generously to charity? Probably not.

I don't want to get into an analysis of economic aid here. If contemplating Mr Dastardly gives you a more complex result like "I should stop treating all third-world economic aid as equivalent" and not a simple "I should give", then the intuition pump is *working as intended*. Because it's helping you build a more accurate world-model.

Conservation of Expected Evidence

The examples in the original Conservation of Expected Evidence post cover this pretty well.

To put it in imperative terms, imagine you'd observed the opposite. If you wouldn't update the opposite way, something has gone wrong. If you wouldn't update as much, this must be balanced by having been surprised when you learned this.

Note that "opposite" can be a little subtle. There can be u-shaped response curves, where "too little" and "too much" are both signs of badness, and only "just right" updates you in favor of something. But unless such a curve is well-known beforehand, the resulting model takes a complexity penalty.

Ask a Friend

A classic way of dealing with flaws in your own thinking is to get someone else's thinking.

Ideally someone with uncorrelated biases. This is easiest to find when you have a personal connection weighing on your reasoning, and it's easy to find someone without one. (In extreme cases, this can be recusal: make the unconnected person do all the work.)

Be careful when asking that you don't distort the evidence as you present it. Verbosity is your friend here.

You may even find that your friend doesn't need to say anything. When you reach the weak-point in your argument, you'll feel it.

One's Never Alone with a Rubber Duck

If your friend doesn't need to say anything, maybe they don't need to be there. Programmers refer to this as "rubber duck debugging".

This has the advantage that your friend can be whomever you want. You can't actually run your ideas past Richard Feynman for a double-checking, for several reasons, but you can certainly run them past imaginary Richard Feynman. The ideal person for this is someone whose clear thinking you respect, and whose respect you want (as this will throw your social instincts into the fray at finding flaws).

Be sure, if attempting this that you explain your *entire* argument. In imagination, it's possible to fast-forward through the less interesting parts, but those could easily be where the flaw is.

Next: Using Expert Disagreement

Using Expert Disagreement

Previously: Testing for Rationalization

One of the red flags was "disagreeing with experts". While all the preceding tools apply here, there's a suite of special options for examining this particular scenario.

The "World is Mad" Dialectic

Back in 2015, Ozymandias wrote:

> I think a formative moment for any rationalist– our “Uncle Ben shot by the mugger” moment, if you will– is the moment you go “holy shit, everyone in the world is fucking insane.”
>
> First, you can say “holy shit, everyone in the world is fucking insane. Therefore, if I adopt the radical new policy of not being fucking insane, I can pick up these giant piles of utility everyone is leaving on the ground, and then I win.”
>
> Second, you can say “holy shit, everyone in the world is fucking insane. However, none of them seem to realize that they’re insane. By extension, I am probably insane. I should take careful steps to minimize the damage I do.”
>
> I want to emphasize that these are not mutually exclusive. In fact, they’re a dialectic (…okay, look, this hammer I found is really neat and I want to find some place to use it).

(I would define a "dialectic" as two superficially opposite statements, both of which are true, in such a way that resolving the apparent paradox forces you to build a usefully richer world-model. I have not run this definition past Ozy, much less Hegel.)

To which Eliezer replied in 2017:

> But, speaking first to the basic dichotomy that’s being proposed, the whole point of becoming sane is that your beliefs *shouldn’t* reflect what sort of person you are. To the extent you’re succeeding, at least, your beliefs should just reflect how the world is.

Good News, Everyone!

I did an empirical test and found no sign of a general factor of trusting your own reasoning.

But I did find lots of disagreement. What are people deciding based on?

Types of Disagreement

Explicit

The expert actually said the opposite of your conclusion.

This is the simplest form of disagreement. The expert could still be wrong, or lying, or communicating unclearly and not actually saying what you think they said. But at least there's no complications from the form of disagreement itself.

The "communicating unclearly" hypothesis should always be high in your probability space. Communication is hard. (It could be spun as "you misunderstood", but this makes the theory less pleasant without altering its substance, so generally don't bother.)

The lying theory is tricky, as it can explain anything. A good lying-based theory should explain why the expert told that particular lie out of millions of possibilities. Without that, the theory contains a contrived co-incidence, and should be penalized accordingly.

With explicit disagreement, you may be able to find the expert's reasoning. If so, try to find the crux and recurse. If you can't follow the reasoning, this is a sign that the expert *either* understands something important that you don't, *or* is spouting complete bullshit. Look for signs that the expert has substance, such as real-world accomplishments or specific endorsements from notably perceptive and honest people.

Superlative or Silence

If an expert oncologist says "glycolisis inhibitors are the best available treatment for this sort of cancer" -- and you think it might be better to use oil-drop immersion to sequence the tumor cells, find the root mutation, and CRISPR an aggressive protease into that spot -- then technically you are disagreeing. This is similar to if she is asked for something better than glycolisis inhibitors and says nothing.

But it's not really an active disagreement. Most likely she's never *considered* the sequence-and-CRISPR possibility.

(Alternative theory: she doesn't consider a therapy to be "available" if it isn't well-tested and endorsed by appropriate parties, even if it is something an agenty patient with moderate wealth could probably obtain. Communication is hard!)

Nobody considers every possibility. What people who try usually do in practice is depend on a large community to do the search. This means that if an individual disagrees with you via superlative or silence, your real disagreement is with the community they depend on.

Implied by Action

An expert takes actions which suggest a disagreement.

A venture capitalist declines to invest in a cancer-curing startup. Does this mean he thinks the cure doesn't work?

Maybe there's other factors in the decision. Maybe he thinks the founder has no head for business, and the company is fatally underestimating regulatory hurtles, but he would invest in the same technology in better hands. Unless you have a very close seat of observation, this looks the same to you.

Or maybe his values are not what you think. Maybe he's very focused on short-term returns, and a company which is five years away from revenue is of no interest to him.

Types of Experts

Individuals

By far the most straightforward.

Do consider how much of an expert the person is on this exact question. Many experts have broad knowledge, which comes at an opportunity cost of depth. Do you know more about medicine than a randomly chosen doctor? Almost certainly not. (Unless you are a doctor yourself, of course.) Do you know more about one particular disease that said doctor doesn't specialize in? Perhaps one that you have? Entirely possible. More about the exact variation which you have? (Cases of a disease are not identical.) Quite likely.

Also consider how much effort the expert put into forming their opinion. Do they have skin in the game? Is it their job to worry about questions like this? (And if so, how dutiful are they?) If this is their way of brushing you off? (If so, respect that decision and go ponder elsewhere.) The further down the list, the more likely they could be wrong because they don't care enough to be right.

Institutions or Communities

Eliezer just wrote an entire book on the subject. You'd think I wouldn't have much to add.

In fact I do. In addition to everything he wrote, consider the *size* of the community. This is especially relevant to superlative or silence disagreements, where the key question is "Did anybody evaluate this?"

The ratio of particle physicists to types of fundamental particles is about a thousand to one, and the number of ways those particles can interact is pretty limited. The odds that you've thought of a good question about particle physics which no established physicist already considered are pretty low. (Unless you have studied deeply in the field, of course.)

The ratio of entomologists to species of insect is closer to one to a thousand. Asking a good question about insects which no entomologist has seriously considered is pretty easy.

(The ratio of intestinal microbiologists to species in the intestinal microbiome is unknown -- we *don't have a good estimate* of the latter.)

I suspect this is an important piece of why Eliezer was able to outdo the psychiatric community regarding Seasonal Affective Disorder. True, there are a lot of psychiatric researchers, but there are a lot of conditions for them to study, and SAD isn't a high priority. Once you zoom in to those working on treatmentment-resistant-SAD, the numbers aren't that high, and things fall through the gaps.

An order-of-magnitude fermi estimate can be useful here.

Traditions or Biology

Traditions and biology have something in common: they are formed by evolution.

Evolution does not share your values. You can often tell what its values are, but it's a constant effort to remember this.

It also has a tendency to take undocumented dependencies on circumstances. It can be *really hard* to figure out which aspects of the environment cause a solution evolution picked to be the ideal one.

Markets

Zvi has written about this at length.

One thing I'd add is low-friction. This is especially important for refining odds far from 50%. If someone's offering a bet on "pi will equal 3 tomorrow" and I can buy "no" for $0.96 and get $1.00 when it resolves, but the betting site will take 5% of my winnings as a fee, I'm not going to bet. So the odds on the proposition will stay at the absurdly high 4%.

Tying up money that has been bet counts as a type of friction. If $0.96 will get me $1.00 in a year with no fee, but index funds are getting 5%/year and I can't do both, I have very little reason to take the bet. Labor can also be a type of friction, especially if bet sizes are capped. If I could gain $100 but it would require five hours of wrestling with terrible UIs or APIs (or studying cryptocurrencies), there's little incentive to bet.

Zvi describes how friction like this can drive people away from a market and make it too small to be useful. And that may well be the primary effect. But it can also cause asymmetric distortions, driving all probabilities toward the middle.

Against Rationalization II: Sequence Recap

Previously: Eliezer's "Against Rationalization" sequence

I've run out of things to say about rationalization for the moment. Hopefully there'll be an Against Rationalization III a few years from now, but ideally some third author will write it.

For now, a quick recap to double as a table of contents:

- Why a New Rationalization Sequence? -- What is rationalization and why is it worth fighting?
- Red Flags for Rationalization -- When you see these signs, you should worry. And do something. Even though sometimes you'll conclude everything is fine.
- Avoiding Rationalization -- If a chain of logic is endangered by rationalization, the best thing to do is not need that chain of logic.
- Testing for Rationalization -- If you can't avoid needing the logic, try using these tools to check if it's dangerously tainted by rationalization.
- Using Expert Disagreement -- In the particular case that you're worried because you disagree with the experts, try these tools