Best of LessWrong: May 2014

1. Truth: It's Not That Great
2. Moving on from Cognito Mentoring
3. A Dialogue On Doublethink

Best of LessWrong: May 2014

Truth: It's Not That Great

> Rationality is pretty great. Just not quite as great as everyone here seems to think it is.
>
> -Yvain, "Extreme Rationality: It's Not That Great"

> The folks most vocal about loving "truth" are usually selling something. For preachers, demagogues, and salesmen of all sorts, the wilder their story, the more they go on about how they love truth...
>
> The people who just want to know things because they need to make important decisions, in contrast, usually say little about their love of truth; they are too busy trying to figure stuff out.
>
> -Robin Hanson, "Who Loves Truth Most?"

A couple weeks ago, Brienne made a post on Facebook that included this remark: "I've also gained a lot of reverence for the truth, in virtue of the centrality of truth-seeking to the fate of the galaxy." But then she edited to add a footnote to this sentence: "That was the justification my brain originally threw at me, but it doesn't actually quite feel true. There's something more directly responsible for the motivation that I haven't yet identified."

I saw this, and commented:

> <puts rubber Robin Hanson mask on>
>
> What we have here is a case of subcultural in-group signaling masquerading as something else. In this case, proclaiming how vitally important truth-seeking is is a mark of your subculture. In reality, the truth is sometimes really important, but sometimes it isn't.
>
> </rubber Robin Hanson mask>

In spite of the distancing pseudo-HTML tags, I actually believe this. When I read some of the more extreme proclamations of the value of truth that float around the rationalist community, I suspect people are doing in-group signaling—or perhaps conflating their own idiosyncratic preferences with rationality. As a mild antidote to this, when you hear someone talking about the value of the truth, try seeing if the statement still makes sense if you replace "truth" with "information."

This standard gives many statements about the value of truth its stamp of approval. After all, *information* is pretty damn valuable. But statements like "truth seeking is central to the fate of the galaxy" look a bit suspicious. Is information-gathering central to the fate of the galaxy? You could argue that statement is *kinda* true if you squint at it right, but really it's too general. Surely it's not just any information that's central to shaping the fate of the galaxy, but information about specific subjects, and even then there are tradeoffs to make.

This is an example of why I suspect "effective altruism" may be better branding for a movement than "rationalism." The "rationalism" branding encourages the meme that truth-seeking is great we should do lots and lots of it because truth is so great. The

effective altruism movement, on the other hand, recognizes that while gathering information about the effectiveness of various interventions is important, there are tradeoffs to be made between spending time and money on gathering information vs. just doing whatever currently seems likely to have the greatest direct impact. Recognize information is valuable, but avoid analysis paralysis.

Or, consider statements like:

- Some truths don't matter much.
- People often have legitimate reasons for not wanting others to have certain truths.
- The value of truth often has to be weighed against other goals.

Do these statements sound heretical to you? But what about:

- Information can be perfectly accurate and also worthless.
- People often have legitimate reasons for not wanting other people to gain access to their private information.
- A desire for more information often has to be weighed against other goals.

I struggled to write the first set of statements, though I think they're right on reflection. Why do they sound so much worse than the second set? Because the word "truth" carries powerful emotional connotations that go beyond its literal meaning. This isn't just true for rationalists—there's a reason religions have sayings like, "God is Truth" or "I am the way, the truth, and the life." "God is Facts" or "God is Information" don't work so well.

There's something about "truth"—how it readily acts as an applause light, a sacred value which must not be traded off against anything else. As I type that, a little voice in me protests "but truth *really is* sacred"... but once we can't say there's some limit to how great truth is, hello affective death spiral.

Consider another quote, from Steven Kaas, that I see frequently referenced on LessWrong: "Promoting less than maximally accurate beliefs is an act of sabotage. Don’t do it to anyone unless you’d also slash their tires, because they’re Nazis or whatever." Interestingly, the original blog included a caveat—"we may have to count everyday social interactions as a partial exception"—which I never see quoted. That aside, the quote has always bugged me. I've never had my tires slashed, but I imagine it ruins your whole day. On the other hand, having less than maximally accurate beliefs about something *could* ruin your whole day, but it could very easily not, depending on the topic.

Furthermore, sometimes sharing certain information doesn't just have little benefit, it can have substantial costs, or at least substantial risks. It would seriously trivialize Nazi Germany's crimes to compare it to the current US government, but I don't think that means we have to promote maximally accurate beliefs about ourselves to the folks at the NSA. Or, when negotiating over the price of something, are you required to promote maximally accurate beliefs about the highest price you'd be willing to pay, even if the other party isn't willing to reciprocate and may respond by demanding that price?

Private information is usually considered private precisely because it has limited benefit to most people, but sharing it could significantly harm the person whose private information it is. A sensible ethic around information needs to be able to deal with issues like that. It needs to be able to deal with questions like: is this information

that is in the public interest to know? And is there a power imbalance involved? My rule of thumb is: secrets kept by the powerful deserve extra scrutiny, but so conversely do their attempts to gather *other people's* private information.

"Corrupted hardware"-type arguments can suggest you should doubt your own justifications for deceiving others. But parallel arguments suggest you should doubt your own justifications for feeling entitled to information others might have legitimate reasons for keeping private. Arguments like, "well truth is supremely valuable," "it's extremely important for me to have accurate beliefs," or "I'm highly rational so people should trust me" just don't cut it.

Finally, being rational in the sense of being well-calibrated doesn't necessarily require making truth-seeking a major priority. Using the evidence you have well doesn't necessarily mean gathering lots of new evidence. Often, the alternative to knowing the truth is not believing falsehood, but admitting you don't know and living with the uncertainty.

Moving on from Cognito Mentoring

Back in December 2013, Jonah Sinick and I launched Cognito Mentoring, an advising service for intellectually curious students. Our goal was to improve the quality of learning, productivity, and life choices of the student population at large, and we chose to focus on intellectually curious students because of their greater potential as well as our greater ability to relate with that population. We began by offering free personalized advising. Jonah announced the launch in a LessWrong post, hoping to attract the attention of LessWrong's intellectually curious readership.

Since then, we feel we've done a fair amount, with a lot of help from LessWrong. We've published a few dozen blog posts and have an information wiki. Slightly under a hundred people contacted us asking us for advice (many from LessWrong), and we had substantive interactions with over 50 of them. As our reviews from students and parents suggest, we've made a good impression and have had a positive impact on many of the people we've advised. We're proud of what we've accomplished and grateful for the support and constructive criticism we've received on LessWrong.

However, what we've learned in the last few months has led us to the conclusion that Cognito Mentoring is not ripe for being a full-time work opportunity for the two of us.

For the last few months, we've eschewed regular jobs and instead done contract work that provides us the flexibility to work on Cognito Mentoring, eating into our savings somewhat to cover the cost of living differences. This is a temporary arrangement and is not sustainable. We therefore intend to scale back our work on Cognito Mentoring to "maintenance mode" so that people can continue to benefit from the resources we've already collected, with minimal additional effort on our part, freeing us up to take regular jobs with more demanding time requirements.

We might revive Cognito Mentoring as a part-time or full-time endeavor in the future if there are significant changes to our beliefs about the traction, impact, and long-run financial viability of Cognito Mentoring. Part of the purpose of "maintenance mode" will be to leave open the possibility of such a revival if the idea does indeed have potential.

In this post, I discuss some of the factors that led us to change our view, the conditions under which we might revive Cognito Mentoring, and more details about how "maintenance mode" for Cognito Mentoring will look.

Reason #1: Downward update on social value

We do think that the work we've done on Cognito Mentoring so far has generated social value, and the continued presence of the website will add more value over time. However, our view has shifted in the direction of lower *marginal social value* from working on Cognito Mentoring full-time, relative to simply keeping the website live and doing occasional work to improve it. Specifically:

- It's quite possible that the lowest-hanging fruit with respect to the advisees who would be most receptive to our advice has already been plucked. We received the bulk of our advisees through LessWrong within the month after our initial posting. Other places where we've posted about our service have led to fewer advisees (more here).

- Of our website content, only a small fraction of the content gets significant traction (see our list of popular pages), so honing and promoting our best content might be a better strategy for improving social value than trying to create a comprehensive resource. This can be done while in maintenance mode, and does not require full-time effort on our part.

What might lead us to change our minds: If we continue to be contacted by large numbers of potentially high-impact people, or we get evidence that the advising we've already done has had significantly greater impact than we think it did, we'll update our social value upward.

Reason #2: Downward update on long-run financial viability

We have enough cash to go on for a few more months. But for Cognito Mentoring to be something that we work full time on, we need an eventual *steady* source of income from it. Around mid-March 2014, we came to the realization that charging advisees is not a viable revenue source, as Jonah described at the end of his post about how Cognito Mentoring can do the most good (see also this comment by Luke Muehlhauser and Jonah's response to it below the comment). At that point, we decided to focus more on our informational content and on looking for philanthropic funding.

Our effort at looking into philanthropic funding did give us a few leads, and some of them could plausibly result in us getting small grants. However, none of the leads we got pointed to potential *steady long-term income sources*. In other words, we don't think philanthropic funding is a viable long-term revenue model for Cognito Mentoring.

Our (anticipated) difficulty in getting philanthropic funding arises from two somewhat different reasons.

1. What we're doing is somewhat new and does not fit the standard mold of educational grants. Educational foundations tend to give grants for fairly specific activities, and what we're doing does not seem to fit those.
2. We haven't demonstrated significant traction or impact yet (even though we've had a reasonable amount of per capita impact, the total number of people we've influenced so far is relatively small). This circles back to Reason #1: funders' reluctance to fund us may in part stem from their belief that we won't have much social value, given our lack of traction so far. Insofar as funders' judgment carries some information value, this should also strengthen Reason #1.

What might lead us to change our minds: If we are contacted by a funder who is willing to bankroll us for over a year and also offer a convincing reason for why he/she thinks bankrolling us is a good idea (so that we're convinced that our funding can be sustained beyond a year) we'll change our minds.

Reason #3: Acquisition of knowledge and skills

One of the reasons we've been able to have an impact through Cognito Mentoring so far is that both Jonah and I have knowledge of many diverse topics related to the questions that our advisees have posed to us. But our knowledge is still woefully inadequate in a number of areas. In particular, many advisees have asked us questions in the realms of technology, entrepreneurship, and the job environment, and while we have pointed them to resources on these, firsthand experience, or close secondhand experience, would help us more effectively guide advisees. We intend to take jobs related to computer technology (in fields such as programming or data science), and these jobs might be at startups or put us in close contact with startups.

This will better position us to return to mentoring later if we choose to resume it part-time or full-time.

Knowledge and skills we acquire working in the technology sector could also help us design better interfaces or websites that can more directly address the needs of our audience. So far, we've thought of ourselves as content-oriented people, so we've used standard off-the-shelf software such as WordPress (for our main website and blog) and MediaWiki (for our information wiki). Part of the reason is that we wanted to focus on content creation rather than interface design, but part of the reason we've stuck to these is that we didn't think we could design interfaces. Once we've acquired more programming and design experience, we might be more open to the idea of designing interfaces and software that can meet particular needs of our target audience.We might design an interface that helps people study more effectively, make better life decisions, or share reviews of courses and colleges, in a manner similar to softwares or websites such as Anki or Beeminder or Goodreads. There might also be potential for a more effective online resource that teaches programming than those in existence (e.g. Codecademy). It's not clear right now whether there exists a useful opportunity of this sort that we are particularly well-suited to, but with more coding experience, we'll at least be *able* to implement an idea of this sort if we decide it has promise.

Reason #4: Letting it brew in the background can give us a better idea of the potential

If we continue to gradually add content to the wiki, and continue to get links and traffic to it from other sources, it's likely that the traffic will grow slowly and steadily. The extent of organic growth will help us figure out how much promise Cognito Mentoring has. If our wiki gets to the point of steadily receiving thousands of pageviews a day, we will reconsider reviving Cognito Mentoring as a part-time or full-time endeavor. If, on the other hand, traffic remains at approximately the current level (about a hundred pageviews a day, once we exclude spikes arising from links from LessWrong and Marginal Revolution) then the idea is probably not worth revisiting, and we'll leave it in maintenance mode.

In addition, by maintaining contact with the people we've advised, we can get more insight into the sort of impact we've had, whether it is significant over the long term, and how it can be improved. This again can tell us whether our impact is sufficiently large as to make Cognito Mentoring worth reviving.

What "maintenance mode" entails

1. **We'll continue to have contact information available, but will scale back on personalized advising**: People are welcome to contact us with questions and suggestions about content, but we will not generally offer detailed personalized responses or do research specific to individuals who contact us. We'll attempt to point people to relevant content we've already written, or to other resources we're already aware of that can address their concerns.
2. **The information wiki will remain live**, and we will continue to make occasional improvements, but we won't have a time schedule of when particular improvements have to be implemented by.
3. **Existing blog posts will remain**, but we probably won't be making many new blog posts. New blog posts will happen only if one of us has an idea that really seems worth sharing and for which the Cognito Mentoring blog is an ideal forum.

4. **We'll continue our administrative roles in the communities of existing Cognito Mentoring advisees**
5. **We'll continue periodically reviewing the progress of people we've advised so far**: This will help us get a better sense of how valuable our work has been, and can be useful should we choose to revive Cognito Mentoring.
6. **We'll continue to correspond with advisees we have so far (time permitting)**, though we'll give more priority to advisees who continue to maintain contact of their own accord and those whose activities seem to have higher impact potential.
7. **We'll try to get our best content linked from other sources, such as about.com**: Sources like about.com are targeted at the general population. We can try to get linked to from there as an additional resource for the more intellectually curious population that's outside the core focus of about.com.
8. **We'll link more extensively to other sources that people can use**: For instance, we can more emphatically point to 80,000 Hours for people who are interested in career advising in relation to effective altruist pursuits. We can point to about.com and College Confidential for more general information about mainstream institutions. We already make a number of recommendations on our website, but as we stop working actively, it becomes all the more important that people who come to us are appropriately redirected to other sources that can help them.

Conclusion and summary (TL;DR)

We (*qua* Cognito Mentoring) are grateful to LessWrong for being welcoming of our posts, offering constructive criticism, and providing us with some advisees we've enjoyed working with. We think that the work we've done has value, but don't think that there's enough marginal value from full-time work on Cognito Mentoring. We think we can do more good for ourselves and the world by switching Cognito Mentoring to maintenance mode and freeing our time currently spent on Cognito Mentoring for other pursuits. The material that we have already produced will continue to remain in the public domain and we hope that people will benefit from it. We may revisit our "maintenance mode" decision if new evidence changes our view regarding traction, impact, and long-run financial viability.

A Dialogue On Doublethink

Followup to: Against Doublethink (sequence), Dark Arts of Rationality, Your Strength as a Rationalist

---

Doublethink

> It is obvious that the same thing will not be willing to do or undergo opposites in the same part of itself, in relation to the same thing, at the same time. --Book IV of Plato's *Republic*

Can you simultaneously want sex and not want it? Can you believe in God and not believe in Him at the same time? Can you be fearless while frightened?

To be fair to Plato, this was meant not as an assertion that such contradictions are impossible, but as an argument that the soul has multiple parts. It seems we can, in fact, want something while also not wanting it. This is awfully strange, and it led Plato to conclude the soul must have multiple parts, for surely no one part could contain both sides of the contradiction.

Often, when we attempt to accept contradictory statements as correct, it causes cognitive dissonance--that nagging, itchy feeling in your brain that won't leave you alone until you admit that something is wrong. Like when you try to convince yourself that staying up just a little longer playing 2048 won't have adverse effects on the presentation you're giving tomorrow, when you know full well that's exactly what's going to happen.

But it may be that cognitive dissonance is the exception in the face of contradictions, rather than the rule. How would you know? If it doesn't cause any emotional friction, the two propositions will just sit quietly together in your brain, never mentioning that it's logically impossible for both of them to be true. When we accept a contradiction wholesale without cognitive dissonance, it's what Orwell called "doublethink".

When you're a mere mortal trying to get by in a complex universe, doublethink may be adaptive. If you want to be completely free of contradictory beliefs without spending your whole life alone in a cave, you'll likely waste a lot of your precious time working through conundrums, which will often produce even more conundrums.

Suppose I believe that my husband is faithful, and I also believe that the unfamiliar perfume on his collar indicates he's sleeping with other women without my permission. I could let that pesky little contradiction turn into an extended investigation that may ultimately ruin my marriage. Or I could get on with my day and leave my marriage intact.

It's better to just leave those kinds of thoughts alone, isn't it? It probably makes for a happier life.

Against Doublethink

Suppose you believe that driving is dangerous, and also that, while you are driving, you're completely safe. As established in Doublethink, there may be some benefits to letting that mental configuration be.

There are also some life-shattering downsides. One of the things you believe is false, you see, by the law of the excluded middle. In point of fact, it's the one that goes "I'm completely safe while driving". Believing false things has consequences.

> Be irrationally optimistic about your driving skills, and you will be happily unconcerned where others sweat and fear. You won't have to put up with the inconvenience of a seatbelt. You will be happily unconcerned for a day, a week, a year. Then CRASH, and spend the rest of your life wishing you could scratch the itch in your phantom limb. Or paralyzed from the neck down. Or dead. It's not inevitable, but it's possible; how probable is it? You can't make that tradeoff rationally unless you know your real driving skills, so you can figure out how much danger you're placing yourself in. --Eliezer Yudkowsky, Doublethink (Choosing to be Biased)

What are beliefs for? Please pause for ten seconds and come up with your own answer.

Ultimately, I think beliefs are inputs for predictions. We're basically very complicated simulators that try to guess which actions will cause desired outcomes, like survival or reproduction or chocolate. We input beliefs about how the world behaves, make inferences from them to which experiences we should anticipate given various changes we might make to the world, and output behaviors that get us what we want, provided our simulations are good enough.

My car is making a mysterious ticking sound. I have many beliefs about cars, and one of them is that if my car makes noises it shouldn't, it will probably stop working eventually, and possibly explode. I can use this input to simulate the future. Since I've observed my car making a noise it shouldn't, I predict that my car will stop working. I also believe that there is something causing the ticking. So I predict that if I intervene and stop the ticking (in non-ridiculous ways), my car will keep working. My belief has thus led to the action of researching the ticking noise, planning some simple tests, and will probably lead to cleaning the sticky lifters.

If it's true that solving the ticking noise will keep my car running, then my beliefs will cash out in correctly anticipated experiences, and my actions will cause desired outcomes. If it's false, perhaps because the ticking can be solved without addressing a larger underlying problem, then the experiences I anticipate will not occur, and my actions may lead to my car exploding.

Doublethink guarantees that you believe falsehoods. Some of the time you'll call upon the true belief ("driving is dangerous"), anticipate future experiences accurately, and get the results you want from your chosen actions ("don't drive three times the speed limit at night while it's raining"). But some of the time, if you actually believe the false thing as well, you'll call upon the opposite belief, anticipate inaccurately, and choose the last action you'll ever take.

Without any principled algorithm determining which of the contradictory propositions to use as an input for the simulation at hand, you'll fail as often as you succeed. So it makes no sense to anticipate more positive outcomes from believing contradictions.

Contradictions may keep you happy as long as you never need to use them. Should you call upon them, though, to guide your actions, the debt on false beliefs will come due. You will drive too fast at night in the rain, you will crash, you will fly out of the car with no seat belt to restrain you, you will die, and it will be your fault.

Against Against Doublethink

What if Plato was pretty much right, and we sometimes believe contradictions because we're sort of not actually one single person?

It is not literally true that Systems 1 and 2 are separate individuals the way you and I are. But the idea of Systems 1 and 2 suggests to me something quite interesting with respect to the relationship between beliefs and their role in decision making, and modeling them as separate people with very different personalities seems to work pretty darn well when I test my suspicions.

> I read Atlas Shrugged probably about a decade ago. I was impressed with its defense of capitalism, which really hammers home the reasons it’s good and important on a gut level. But I was equally turned off by its promotion of selfishness as a moral ideal. I thought that was *basically* just being a jerk. After all, if there’s one thing the world doesn’t need (I thought) it’s more selfishness.
>
> Then I talked to a friend who told me Atlas Shrugged had changed his life. That he’d been raised in a really strict family that had told him that ever enjoying himself was selfish and made him a bad person, that he had to be working at every moment to make his family and other people happy or else let them shame him to pieces. And the revelation that it was sometimes okay to consider your own happiness gave him the strength to stand up to them and turn his life around, while still keeping the basic human instinct of helping others when he wanted to and he felt they deserved it (as, indeed, do Rand characters). --Scott of Slate Star Codex in All Debates Are Bravery Debates

If you're generous to a fault, "I should be more selfish" is probably a belief that will pay off in positive outcomes should you install it for future use. If you're selfish to a fault, the same belief will be harmful. So what if you were too generous half of the time and too selfish the other half? Well, then you would want to believe "I should be more selfish" with only the generous half, while disbelieving it with the selfish half.

Systems 1 and 2 need to hear different things. System 2 might be able to understand the reality of biases and make appropriate adjustments that would work if System 1 were on board, but System 1 isn't so great at being reasonable. And it's not System 2 that's in charge of most of your actions. If you want your beliefs to positively influence your actions (which is the point of beliefs, after all), you need to tailor your beliefs to System 1's needs.

For example: The planning fallacy is nearly ubiquitous. I know this because for the past three years or so, I've gotten everywhere five to fifteen minutes early. Almost every single person I meet with arrives five to fifteen minutes late. It is very rare for someone to be on time, and only twice in three years have I encountered the (rather awkward) circumstance of meeting with someone who also arrived early.

Before three years ago, I was also usually late, and I far underestimated how long my projects would take. I knew, abstractly and intellectually, about the planning fallacy,

but that didn't stop System 1 from thinking things would go implausibly quickly. System 1's just optimistic like that. It responds to, "Dude, that is not going to work, and I have a twelve point argument supporting my position and suggesting alternative plans," with "Naaaaw, it'll be fine! We can totally make that deadline."

At some point (I don't remember when or exactly how), I gained the ability to look at the true due date, shift my System 1 beliefs to make up for the planning fallacy, and then hide my memory that I'd ever seen the original due date. I would see that my flight left at 2:30, and be surprised to discover on travel day that I was not late for my 2:00 flight, but a little early for my 2:30 one. I consistently finished projects on time, and only disasters caused me to be late for meetings. It took me about three months before I noticed the pattern and realized what must be going on.

I got a little worried I might make a mistake, such as leaving a meeting thinking the other person just wasn't going to show when the actual meeting time hadn't arrived. I did have a couple close calls along those lines. But it was easy enough to fix; in important cases, I started receiving Boomeranged notes from past-me around the time present-me expected things to start that said, "Surprise! You've still got ten minutes!"

This unquestionably improved my life. You don't realize just how inconvenient the planning fallacy is until you've left it behind. Clearly, considered in isolation, the action of believing falsely in this domain was instrumentally rational.

Doublethink, and the Dark Arts generally, applied to carefully chosen domains is a powerful tool. It's dumb to believe false things about really dangerous stuff like driving, obviously. But you don't have to doublethink indiscriminately. As long as you're careful, as long as you suspend epistemic rationality only when it's clearly beneficial to do so, employing doublethink at will is a great idea.

Instrumental rationality is what really matters. Epistemic rationality is useful, but what use is holding accurate beliefs in situations where that won't get you what you want?

Against Against Against Doublethink

There are indeed epistemically irrational actions that are instrumentally rational, and instrumental rationality is what really matters. It is pointless to believing true things if it doesn't get you what you want. This has always been very obvious to me, and it remains so.

There is a bigger picture.

Certain epistemic rationality techniques are not compatible with dark side epistemology. Most importantly, the Dark Arts do not play nicely with "notice your confusion", which is essentially your strength as a rationalist. If you use doublethink on purpose, confusion doesn't always indicate that you need to find out what false thing you believe so you can fix it. Sometimes you have to bury your confusion. There's an itsy bitsy pause where you try to predict whether it's useful to bury.

As soon as I finally decided to abandon the Dark Arts, I began to sweep out corners I'd allowed myself to neglect before. They were mainly corners I didn't know I'd neglected.

The first one I noticed was the way I responded to requests from my boyfriend. He'd mentioned before that I often seemed resentful when he made requests of me, and I'd insisted that he was wrong, that I was actually happy all the while. (Notice that in the short term, since I was probably going to do as he asked anyway, attending to the resentment would probably have made things more difficult for me.) This self-deception went on for months.

Shortly after I gave up doublethink, he made a request, and I felt a little stab of dissonance. Something I might have swept away before, because it seemed more immediately useful to bury the confusion than to notice it. But I thought (wordlessly and with my emotions), "No, look at it. This is exactly what I've decided to watch for. I have noticed confusion, and I will attend to it."

It was very upsetting at first to learn that he'd been right. I feared the implications for our relationship. But that fear didn't last, because we both knew the only problems you can solve are the ones you acknowledge, so it is a comfort to know the truth.

I was far more shaken by the realization that I really, truly was ignorant that this had been happening. Not because the consequences of this one bit of ignorance were so important, but because who knows what other epistemic curses have hidden themselves in the shadows? I realized that I had not been in control of my doublethink, that I couldn't have been.

Pinning down that one tiny little stab of dissonance took great preparation and effort, and there's no way I'd been working fast enough before. "How often," I wondered, "does this kind of thing happen?"

Very often, it turns out. I began noticing and acting on confusion several times a day, where before I'd been doing it a couple times a week. I wasn't just noticing things that I'd have ignored on purpose before; I was noticing things that would have slipped by because my reflexes slowed as I weighed the benefit of paying attention. "Ignore it" was not an available action in the face of confusion anymore, and that was a dramatic change. Because there are no disruptions, acting on confusion is becoming automatic.

I can't know for sure which bits of confusion I've noticed since the change would otherwise have slipped by unseen. But here's a plausible instance. Tonight I was having dinner with a friend I've met very recently. I was feeling s little bit tired and nervous, so I wasn't putting as much effort as usual into directing the conversation. At one point I realized we had stopped making making any progress toward my goals, since it was clear we were drifting toward small talk. In a tired and slightly nervous state, I imagine that I might have buried that bit of information and abdicated responsibility for the conversation--not by means of considering whether allowing small talk to happen was actually a good idea, but by not pouncing on the dissonance aggressively, and thereby letting it get away. Instead, I directed my attention at the feeling (without effort this time!), inquired of myself what precisely was causing it, identified the prediction that the current course of conversation was leading away from my goals, listed potential interventions, weighed their costs and benefits against my simulation of small talk, and said, "What are your terminal values?"

(I know that sounds like a lot of work, but it took at most three seconds. The hard part was building the pouncing reflex.)

When you know that some of your beliefs are false, and you know that leaving them be is instrumentally rational, you do not develop the automatic reflex of interrogating

every suspicion of confusion. You might think you can do this selectively, but if you do, I strongly suspect you're wrong in exactly the way I was.

I have long been more viscerally motivated by things that are interesting or beautiful than by things that correspond to the territory. So it's not too surprising that toward the beginning of my rationality training, I went through a long period of being so enamored with a-veridical instrumental techniques--things like willful doublethink--that I double-thought myself into believing accuracy was not so great.

But I was wrong. And that mattered. Having accurate beliefs is a ridiculously convergent incentive. Every utility function that involves interaction with the territory--interaction of just about any kind!--benefits from a sound map. Even if "beauty" is a terminal value, "being viscerally motivated to increase your ability to make predictions that lead to greater beauty" increases your odds of success.

Dark side epistemology prevents total dedication to continuous improvement in epistemic rationality. Though individual dark side actions may be instrumentally rational, the patterns of thought required to allow them are not. Though instrumental rationality is ultimately the goal, your instrumental rationality will always be limited by your epistemic rationality.

That was important enough to say again: Your instrumental rationality will always be limited by your epistemic rationality.

It only takes a fraction of a second to sweep an observation into the corner. You don't have time to decide whether looking at it might prove problematic. If you take the time to protect your compartments, false beliefs you don't endorse will slide in from everywhere through those split-second cracks in your art. You must attend to your confusion the very moment you notice it. You must be relentless an unmerciful toward your own beliefs.

Excellent epistemology is not the natural state of a human brain. Rationality is hard. Without extreme dedication and advanced training, without reliable automatic reflexes of rational thought, your belief structure will be a mess. You can't have totally automatic anti-rationalization reflexes if you use doublethink as a technique of instrumental rationality.

This has been a difficult lesson for me. I have lost some benefits I'd gained from the Dark Arts. I'm late now, sometimes. And painful truths are painful, though now they are sharp and fast instead of dull and damaging.

And it is so worth it! I have much more work to do before I can move on to the next thing. But whatever the next thing is, I'll tackle it with far more predictive power than I otherwise would have--though I doubt I'd have noticed the difference.

So when I say that I'm against against against doublethink--that dark side epistemology is bad--I mean that there is more potential on the light side, not that the dark side has no redeeming features. Its fruits hang low, and they are delicious.

But the fruits of the light side are worth the climb. You'll never even know they're there if you gorge yourself in the dark forever.