# SquirrelInHell

2017-03-28

## Make Your Observations Pay Rent

### 第一

Elon Musk said during the panel at the Asilomar conference ("Beneficial AI 2017 Conference"):

> *[...] Everyone is already superhuman. And a cyborg. The limitation is one of bandwidth. We're bandwidth constrained, particularly on output. Our input is much better, but our output is extremely slow. If you want to be generous, you could say maybe it's a few hundred bits per second, or a kilobyte, something like that, output. The way we output is, we have our little meat-sticks, that we move very slowly and push buttons, or tap a little screen. And that's just extremely slow. Compare it to a computer which can communicate at the terabit level. Very big orders of magnitude differences. Our input is much better because of vision, but even that could be enhanced significantly. I think the two things that are needed for a future that we would look at and conclude is good, most likely, is we, we have to solve that bandwidth constraint. With a direct neural interface. I think a high bandwidth interface is the cortex. [...]*

I realized that when I had originally heard the statement quoted above, I had failed *horribly* to make even the most obvious predictions based on it. Oh, and one of my friends posted on Facebook to the effect of "funny, it looks like Elon Mush has different models from everyone else in the room!". So this part was addidtionally rubbed in my face in the most obvious way possible. And I still failed, utterly and completely.

So this post is a post-mortem of my failure. But first, let's look at where all the pieces came from.

### 第二

E. Yudkowsky points out in the Sequences that to have accurate beliefs, we need to make them "pay rent" in terms of anticipated experience. The lesson is clear enough - if a belief makes me anticipate something, and then this thing happens, then hooray - I had an accurate belief. If it doesn't happen, then oops - at least I have learned something. And if I don't anticipate anything in particular based on my belief, well then, why waste bits of storage space in my memory on such a belief?

So now we have a tool to get rid of all those pesky wrong beliefs. All it takes it to realize that a belief predicts inaccurately, or doesn't predict anything at all. Then we can delete the bad unwanted belief, and we are done! Back in the happy epistemic paradise.

Now internalizing this lesson is valuable, and a big step towards epistemic rationality. But also, reality is messy, and there are places where we don't want to be *too* quick about putting strain on our fresh and barely-started-to-form beliefs. Especially when we are looking at a new topic, brainstorming etc., it makes sense to just generate a lot of rough ideas, and sort through them later. Trying to judge all of them online, before they make their first baby steps out in the world, would just be carnage. It leads to another failure mode, in which we are too scared to think so we just don't.

And those fresh tentative ideas are not very *dangerous* to our epistemics - when exposed to more learning, the inaccurate ones just fade away, leaving space for other

Blog Archive

January 2018 (3)
December 2017 (3)
November 2017 (1)
October 2017 (2)
September 2017 (1)
August 2017 (2)
July 2017 (1)
May 2017 (2)
April 2017 (1)
March 2017 (2)
January 2017 (2)
November 2016 (1)
October 2016 (1)
September 2016 (2)
August 2016 (1)
April 2016 (1)
March 2016 (1)

More by SquirrelInHell

- AI Safety Comics
- Android Apps
- Be Well Tuned
- Rationality Updates

beliefs to flourish. They don't really put up a fight like all those entrenched crony beliefs.

There's something even more comforting about just *observing* phenomena. There is a way in which we can just look at something and try to take it in, which mostly involves focusing attention on it and letting the brain do its thing. And in many cases it's enough - if we convince our brains that something is important and worth absorbing, the brains will probably be happy to just do that. At the end of this process we have a bunch of *observations*, which are indeed a very empirical kind of belief (of course they are still "beliefs" in the broader Bayesian sense). Observations feel very *safe* for my epistemic instincts. Beliefs in the form of "I observed X in situation Y" have a pretty good track record of not being biased. So there's no pressure to put them under intense scrutiny.

## 第三

Now we're getting to the vegetable of the matter. When I watched the video quoted above, it seems to me that I executed something that I would like to call a "cow stare". Like a cow watching a fashion show, or something else that requires human-level understanding. Despite all its unyielding intensity the "cow stare" does not carry enough intelligence to do anything useful.

And the way I suspect it happened is that I fell back on the "just observe" pattern, without realizing that the situation calls for much more. In many cases, the "just observe" mode of interacting is actually enough, and my brain just automatically chooses to run the predictive processes in the background. If I stare at a smudge on a wall, my brain just goes "oh, there's a smudge. I wonder how it happened", and from there it somehow moves on to ponder stuff like "will I see the same kind of smudge around air vents in other rooms?", and "does the smudge-generating process approach an equilibrium?" etc.

So the problem is not that observing something carefully, and leaving it at that, is always stupid and doesn't work. The problem is that it works, until the problem is too hard. It's more like *failing to match the level of thought to the level of problem*. And I think that yes, with enough staring, my brain does eventually come up with useful thoughts even about high-order, complicated phenomena. But in those cases, it's just *way too slow*. If the balance in unfavourable, i.e. the staring is not very intense and the stared-at-object is very high-order, we are indeed approaching a "cow stare" level of usefulness.

And one thing that seems to help, is to make beliefs pay rent. Only, like, *actually getting it* this time. On a micro-scale it feels like every time I watch something, my thoughts pull strongly to run forward in time, and also I *refuse to give up when it's hard*. In case of Musk, I might wonder "what will Elon Musk do next?", and immediately think "how can I ever know *that*!" with all the false modesty of refusing to think. So instead, I can just ask the obvious questions, such as "well, what if he does exactly what he said? how likely is that?".

*Note: All of the above was prompted when I found out today that Elon Musk is starting a new company called Neuralink. I did not yet have a confirmation of the company's mission, but the name is suggestive enough.*

**No comments:**

**Post a Comment**

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)