Best of LessWrong: November 2016

1. On the importance of Less Wrong, or another single conversational locus
2. A Return to Discussion
3. Epistemic Effort

Best of LessWrong: November 2016

1. On the importance of Less Wrong, or another single conversational locus
2. A Return to Discussion
3. Epistemic Effort

On the importance of Less Wrong, or another single conversational locus

Epistemic status: My actual best bet. But I used to think differently; and I don't know how to fully explicate the updating I did (I'm not sure what fully formed argument I could give my past self, that would cause her to update), so you should probably be somewhat suspicious of this until explicated. And/or you should help me explicate it.

It seems to me that:

1. The world is locked right now in a deadly puzzle, and needs something like a miracle of good thought if it is to have the survival odds one might wish the world to have.

2. Despite all priors and appearances, our little community (the "aspiring rationality" community; the "effective altruist" project; efforts to create an existential win; etc.) has a shot at seriously helping with this puzzle. This sounds like hubris, but it is at this point at least partially a matter of track record.[1]

3. To aid in solving this puzzle, we must probably find a way to think together, accumulatively. We need to think about technical problems in AI safety, but also about the full surrounding context -- everything to do with understanding what the heck kind of a place the world is, such that that kind of place may contain cheat codes and trap doors toward achieving an existential win. We probably also need to think about "ways of thinking" -- both the individual thinking skills, and the community conversational norms, that can cause our puzzle-solving to work better. [2]

4. One feature that is pretty helpful here, is if we somehow maintain a single "conversation", rather than a bunch of people separately having thoughts and sometimes taking inspiration from one another. By "a conversation", I mean a space where people can e.g. reply to one another; rely on shared jargon/shorthand/concepts; build on arguments that have been established in common as probably-valid; point out apparent errors and then have that pointing-out be actually taken into account or else replied-to).

5. One feature that really helps things be "a conversation" in this way, is if there is a single Schelling set of posts/etc. that people (in the relevant community/conversation) are supposed to read, and can be assumed to have read. Less Wrong used to be a such place; right now there is no such place; it seems to me highly desirable to form a new such place if we can.

6. We have lately ceased to have a "single conversation" in this way. Good content is still being produced across these communities, but there is no single locus of conversation, such that if you're in a gathering of e.g. five aspiring rationalists, you can take for granted that of course everyone has read posts such-and-such. There is no one place you can post to, where, if enough people upvote your writing, people will reliably read and respond (rather than ignore), and where others will call them out if they later post reasoning that ignores your evidence. Without such a locus, it is hard for conversation to build in the correct way.

(And hard for it to turn into arguments and replies, rather than a series of non sequiturs.)

It seems to me, moreover, that Less Wrong used to be such a locus, and that it is worth seeing whether Less Wrong or some similar such place[3] may be a viable locus again. I will try to post and comment here more often, at least for a while, while we see if we can get this going. Sarah Constantin, Ben Hoffman, Valentine Smith, and various others have recently mentioned planning to do the same.

I suspect that most of the value generation from having a single shared conversational locus is not captured by the individual generating the value (I suspect there is much distributed value from having "a conversation" with better structural integrity / more coherence, but that the value created thereby is pretty distributed). Insofar as there are "externalized benefits" to be had by blogging/commenting/reading from a common platform, it may make sense to regard oneself as exercising civic virtue by doing so, and to deliberately do so as one of the uses of one's "make the world better" effort. (At least if we can build up toward in fact having a single locus.)

If you believe this is so, I invite you to join with us. (And if you believe it isn't so, I invite you to explain why, and to thereby help explicate a shared body of arguments as to how to actually think usefully in common!)

---

[1] By track record, I have in mind most obviously that AI risk is now relatively credible and mainstream, and that this seems to have been due largely to (the direct + indirect effects of) Eliezer, Nick Bostrom, and others who were poking around the general aspiring rationality and effective altruist space in 2008 or so, with significant help from the extended communities that eventually grew up around this space. More controversially, it seems to me that this set of people has probably (though not indubitably) helped with locating specific angles of traction around these problems that are worth pursuing; with locating other angles on existential risk; and with locating techniques for forecasting/prediction (e.g., there seems to be similarity between the techniques already being practiced in this community, and those Philip Tetlock documented as working).

[2] Again, it may seem somewhat hubristic to claim that that a relatively small community can usefully add to the world's analysis across a broad array of topics (such as the summed topics that bear on "How do we create an existential win?"). But it is generally smallish groups (rather than widely dispersed millions of people) that can actually bring analysis together; history has often involved relatively small intellectual circles that make concerted progress; and even if things *are* already known that bear on how to create an existential win, one must probably still combine and synthesize that understanding into a smallish set of people that can apply the understanding to AI (or what have you).

It seems worth a serious try to see if we can become (or continue to be) such an intellectually generative circle; and it seems worth asking what institutions (such as a shared blogging platform) may increase our success odds.

[3] I am curious whether Arbital may become useful in this way; making conversation and debate work well seems to be near their central mission. The Effective Altruism

Forum is another plausible candidate, but I find myself substantially more excited about Less Wrong in this regard; it seems to me one must be free to speak about a broad array of topics to succeed, and this feels easier to do here. The presence and easy linkability of Eliezer's Less Wrong Sequences also seems like an advantage of LW.

Thanks to Michael Arc (formerly Michael Vassar) and Davis Kingsley for pushing this/related points in conversation.

A Return to Discussion

Epistemic Status: Casual

It's taken me a long time to fully acknowledge this, but people who "come from the internet" are no longer a minority subculture. Senators tweet and suburban moms post Minion memes. Which means that talking about trends in how people socialize on the internet is not a frivolous subject; it's relevant to *how people interact, period*.

There seems to have been an overall drift towards social networks over blogs and forums in general, and in particular things like:

- the drift of commentary from personal blogs to "media" aggregators like *The Atlantic*, *Vox*, and *Breitbart*
- the migration of fandom from LiveJournal to Tumblr
- Facebook and Twitter as the places where links and discussions go

At the moment I'm not empirically tracking any trends like this, and I'm not confident in what exactly the major trends are — maybe in future I'll start looking into this more seriously. Right now, I have a sense of things from impression and hearsay.

But one thing I *have* noticed personally is that people have gotten *intimidated* by more formal and public kinds of online conversation. I know quite a few people who *used* to keep a "real blog" and have become afraid to touch it, preferring instead to chat on social media. It's a weird kind of perfectionism — nobody ever imagined that blogs were meant to be masterpieces. But I do see people fleeing towards more ephemeral, more stream-of-consciousness types of communication, or communication that involves no words at all (reblogging, image-sharing, etc.) There seems to be a fear of becoming too visible as a distinctive writing voice.

For one rather public and hilarious example, witness Scott Alexander's flight from LessWrong to LiveJournal to a personal blog to Twitter and Tumblr, in hopes that somewhere he can find a place isolated enough that nobody will notice his insight and humor. (It hasn't been working.)

What might be going on here?

Of course, there are pragmatic concerns about reputation and preserving anonymity. People don't want their writing to be found by judgmental bosses or family members. But that's always been true — and, at any rate, social networking sites are often *less* anonymous than forums and blogs.

It might be that people have become more afraid of trolls, or that trolling has gotten worse. Fear of being targeted by harassment or threats might make people less open and expressive. I've certainly heard many writers say that they've shut down a lot of their internet presence out of exhaustion or literal fear. And I've heard serious enough horror stories that I respect and sympathize with people who are on their guard.

But I don't think that really explains why one would drift towards more ephemeral media. Why short-form instead of long-form? Why streaming feeds instead of searchable archives? Trolls are not known for their patience and rigor. Single tweets can attract storms of trolls. So troll-avoidance is not enough of an explanation, I think.

It's almost as though the issue were *accountability.*

A blog is almost a perfect medium for personal accountability. It belongs to *you*, not your employer, and not the hivemind. The archives are easily searchable. The posts are permanently viewable. Everything embarrassing you've ever written is there. If there's a comment section, people are free to come along and poke holes in your posts. This leaves people vulnerable in a certain way. Not just to trolls, but to *critics*.

You can preempt embarrassment by declaring that you're doing something shitty anyhow. That puts you in a position of safety. I think that a lot of online mannerisms, like using all-lowercase punctuation, or using really self-deprecating language, or deeply nested meta-levels of meme irony, are ways of saying "I'm cool because I'm not putting myself out there where I can be judged. Only pompous idiots are so naive as to think their opinions are *actually valuable*."

Here's another angle on the same issue. If you earnestly, explicitly say what you think, in essay form, and if your writing attracts attention at all, you'll attract swarms of earnest, bright-but-not-brilliant, mostly young white male, commenters, who want to share their opinions, because (perhaps naively) they think their contributions will be welcomed. It's basically just "oh, are we playing a game? I wanna play too!" If you *don't* want to play with them — maybe because you're talking about a personal or highly technical topic and don't value their input, maybe because your intention was just to talk to your friends and not the general public, whatever — you'll find this style of interaction aversive. You'll read it as sealioning. Or mansplaining. Or"well, actually"-ing.

I think what's going on with these kinds of terms is something like:

> Author: "Hi! I just said a thing!"
>
> Commenter: "Ooh cool, we're playing the Discussion game! Can I join? Here's my comment!" (Or, sometimes, "Ooh cool, we're playing the Verbal Battle game! I wanna play! Here's my retort!")
>
> Author: "Ew, no, I don't want to play with you."

There's a bit of a race/gender/age/educational slant to the people playing the "commenter" role, probably because our society rewards some people more than others for playing the discussion game. Privileged people are more likely to assume that they're automatically welcome wherever they show up, which is why others tend to get annoyed at them.

Privileged people, in other words, are more likely to think they live in a *high-trust society*, where they can show up to strangers and be greeted as a potential new friend, where open discussion is an important priority, where they can trust and be trusted, since everybody is playing the "let's discuss interesting things!" game.

The unfortunate reality is that most of the world doesn't look like that high-trust society.

On the other hand, I think the ideal of open discussion, and to some extent the *past reality* of internet discussion, is a lot more like a high-trust society where everyone is playing the "discuss interesting things" game, than it is like the present reality of social media.

A lot of the value generated on the 90's and early 2000's internet was built on *people who were interested in things, sharing information about those things with like-minded individuals*. Think of the websites that were just catalogues of information about someone's obsessions. (I remember my family happily gathering round the PC when I was a kid, to listen to all the national anthems of the world, which some helpful net denizen had collated all in one place.) There is an enormous shared commons that is produced when people are playing the "share info about interesting stuff" game. Wikipedia. StackExchange. It couldn't have been motivated by pure public-spiritedness — otherwise people wouldn't have produced so much free work. There are lower motivations: the desire to show off how clever you are, the desire to be a know-it-all, the desire to correct other people. And there are higher motivations — obsession, fascination, the delight of infodumping. This isn't some higher plane of civic virtue; it's just ordinary nerd behavior.

But in ordinary nerd behavior, there are some unusual strengths. Nerds are playing the "let's have discussions!" game, which means that they're unembarrassed about sharing their take on things, *and* unembarrassed about holding other people accountable for mistakes, *and* unembarrassed about being held accountable for mistakes. It's a sort of happy place between perfectionism and laxity. Nobody is supposed to get everything right on the first try; but you're supposed to respond intelligently to criticism. Things *will get poked at*, inevitably. Poking is *friendly* behavior. (Which doesn't mean it's not also aggressive behavior. Play and aggression are always intermixed. But it doesn't have to be understood as *scary, hostile, enemy*.)

Nerd-format discussions are definitely not costless. You get discussions of advanced/technical topics being mobbed by clueless opinionated newbies, or discussions of deeply personal issues being hassled by clueless opinionated randos. You get endless debate over irrelevant minutiae. There are reasons why so many people flee this kind of environment.

But I would say that these disadvantages are necessary evils that, while they might be possible to mitigate somewhat, go along with having a genuinely *public* discourse and *public* accountability.

We talk a lot about social media killing privacy, but there's also a way in which it kills *publicness*, by allowing people to curate their spaces by personal friend groups, and retreat from open discussions. In a public square, any rando can ask an aristocrat to explain himself. If people hide from public squares, they can't be exposed to Socrates' questions.

I suspect that, especially for people who are even minor VIPs (my level of online fame, while modest, is enough to create some of this effect), it's tempting to become *less available* to the "public", less willing to engage with strangers, even those who seem friendly and interesting. I think it's worth fighting this temptation. You don't get the gains of open discussion if you close yourself off. You may not capture all the gains yourself, but that's how the tragedy of the commons works; a bunch of people have to cooperate and trust if they're going to build good stuff together. And what that means, concretely, on the margin, is taking more time to explain yourself and engage intellectually with people who, from your perspective, look dumb, clueless, crankish, or uncool.

Some of the people I admire most, including theoretical computer scientist Scott Aaronson, are notable for taking the time to carefully debunk crackpots (and offer them the benefit of the doubt in case they are in fact correct.) Is this activity a great ROI for a brilliant scientist, from a narrowly selfish perspective? No. But it's praiseworthy, because it contributes to a truly open discussion. If scientists take the time to investigate weird claims from randos, they're doing the work of proving that science is a universal and systematic way of thinking, not just an elite club of insiders. In the long run, it's very important that somebody be doing that groundwork.

Talking about interesting things, with friendly strangers, in a spirit of welcoming open discussion and accountability rather than fleeing from it, seems really underappreciated today, and I think it's time to make an explicit push towards building places online that have that quality.

In that spirit, I'd like to recommend LessWrong to my readers. For those not familiar with it, it's a discussion forum devoted to things like cognitive science, AI, and related topics, and, back in its heyday a few years ago, it was *suffused* with the nerdy-discussion-nature. It had all the enthusiasm of late-night dorm-room philosophy discussions — except that some of the people you'd be having the discussions with were among the most creative people of our generation. These days, posting and commenting is a lot sparser, and the energy is gone, but I and some other old-timers are trying to rekindle it. I'm crossposting all my blog posts there from now on, and I encourage everyone to check out and join the discussions there.

(Cross-posted from my blog, https://srconstantin.wordpress.com/)

Epistemic Effort

Epistemic Effort: Thought seriously for 5 minutes about it. Thought a bit about how to test it empirically. Spelled out my model a little bit. I'm >80% confident this is worth trying and seeing what happens. Spent 45 min writing post.

I've been pleased to see "Epistemic Status" hit a critical mass of adoption - I think it's a good habit for us to have. In addition to letting you know how seriously to take an individual post, it sends a signal about what sort of discussion you want to have, and helps remind other people to think about their own thinking.

I have a suggestion for an evolution of it - "Epistemic Effort" instead of status. Instead of "how confident you are", it's more of a measure of "what steps did you actually take to make sure this was accurate?" with some examples including:

- Thought about it musingly
- Made a 5 minute timer and thought seriously about possible flaws or refinements
- Had a conversation with other people you epistemically respect and who helped refine it
- Thought about how to do an empirical test
- Thought about how to build a model that would let you make predictions about the thing
- Did some kind of empirical test
- Did a review of relevant literature
- Ran an Randomized Control Trial

[Edit: the intention with these examples is for it to start with things that are fairly easy to do to get people in the habit of thinking about how to think better, but to have it quickly escalate to "empirical tests, hard to fake evidence and exposure to falsifiability"]

A few reasons I think this (most of these reasons are "things that seem likely to me" but which I haven't made any formal effort to test - they come from some background in game design and reading some books on habit formation, most of which weren't very well cited)

- People are more likely to put effort into being rational if there's a relatively straightforward, understandable path to do so
- People are more likely to put effort into being rational if they see other people doing it
- People are more likely to put effort into being rational if they are rewarded (socially or otherwise) for doing so.
- It's not obvious that people will get _especially_ socially rewarded for doing something like "Epistemic Effort" (or "Epistemic Status") but there are mild social rewards just for doing something you see other people doing, and a mild personal reward simply for doing something you believe to be virtuous (I wanted to say "dopamine" reward but then realized I honestly don't know if that's the mechanism, but "small internal brain happy feeling")
- Less Wrong etc is a more valuable project if more people involved are putting more effort into thinking and communicating "rationally" (i.e. making an effort to

make sure their beliefs align with the truth, and making sure to communicate so other people's beliefs align with the truth)
- People range in their ability / time to put a lot of epistemic effort into things, but if there are easily achievable, well established "low end" efforts that are easy to remember and do, this reduces the barrier for newcomers to start building good habits. Having a nice range of recommended actions can provide a pseudo-gamified structure where there's always another slightly harder step you available to you.
- In the process of writing this very post, I actually went from planning a quick, 2 paragraph post to the current version, when I realized I should really eat my own dogfood and make a minimal effort to increase my epistemic effort here. I didn't have that much time so I did a couple simpler techniques. But even that I think provided a lot of value.

Results of thinking about it for 5 minutes.

- It occurred to me that explicitly demonstrating the results of putting epistemic effort into something might be motivational both for me and for anyone else thinking about doing this, hence this entire section. (This is sort of stream of conscious-y because I didn't want to force myself to do so much that I ended up going 'ugh I don't have time for this right now I'll do it later.')
- One failure mode is that people end up putting minimal, token effort into things (i.e. randomly tried something on a couple doubleblinded people and call it a Randomized Control Trial).
- Another is that people might end up defaulting to whatever the "common" sample efforts are, instead of thinking more creatively about how to refine their ideas. I think the benefit of providing a clear path to people who weren't thinking about this at all outweights people who might end up being less agenty about their epistemology, but it seems like something to be aware of.
- I don't think it's worth the effort to run a "serious" empirical test of this, but I do think it'd be worth the effort, if a number of people started doing this on their posts, to run a followup informal survey asking "did you do this? Did it work out for you? Do you have feedback."
- A neat nice-to-have, if people actually started adopting this and it proved useful, might be for it to automatically appear at the top of new posts, along with a link to a wiki entry that explained what the deal was.

Next actions, if you found this post persuasive:

Next time you're writing any kind of post intended to communicate an idea (whether on Less Wrong, Tumblr or Facebook), try adding "Epistemic Effort: " to the beginning of it. If it was intended to be a quick, lightweight post, just write it in its quick, lightweight form.

After the quick, lightweight post is complete, think about whether it'd be worth doing something as simple as "set a 5 minute timer and think about how to refine/refute the idea". If not, just write "thought about it musingly" after Epistemic Status. If so, start thinking about it more seriously and see where it leads.

While thinking about it for 5 minutes, some questions worth asking yourself:

- If this were wrong, how would I know?
- What actually led me to believe this was a good idea? Can I spell that out? In how much detail?

- Where might I check to see if this idea has already been tried/discussed?
- What pieces of the idea might you peel away or refine to make the idea stronger? Are there individual premises you might be wrong about? Do they invalidate the idea? Does removing them lead to a different idea?