Best of LessWrong: June 2017

1. Welcome to Lesswrong 2.0

Best of LessWrong: June 2017

1. Welcome to Lesswrong 2.0

Welcome to Lesswrong 2.0

Lesswrong 2.0 is a project by Oliver Habryka, Ben Pace, and Matthew Graves with the aim of revitalizing the Lesswrong discussion platform. Oliver and Ben are currently working on the project full-time and Matthew Graves is providing part-time support and oversight from MIRI.

Our main goals are to move Lesswrong to a modern codebase, add an effective moderation system, and integrate the cultural shifts that the rationality community has made over the last eight years. We also think that many of the distinct qualities of the Lesswrong community (e.g. propensity for long-form arguments, reasoned debate, and a culture of building on one another's conceptual progress) suggest a set of features unique to the new Lesswrong that will greatly benefit the community.

We are planning to be improving and maintaining the site for many years to come, but whether the site will be successful ultimately depends on whether the community will find it useful. As such, it is important to get your feedback and guidance on how the site should develop and how we should prioritize our resources. Over the coming months we want to experiment with many different content-types and page designs, while actively integrating your feedback, in an attempt to find a structure for Lesswrong that is best suited at facilitating rational discourse.

What follows is a rough summary of how we are currently thinking about the development of Lesswrong 2.0, and what we see as the major pillars of the Lesswrong 2.0 project. We would love to get your thoughts and critiques on these.

Table of Contents:

I. Modern Codebase

II. Effective Moderation

III. Discourse Norms

IV. New Features

V. Beta Feedback Period

I. Modern Codebase

The old lesswrong is one of the only successful forks of the reddit codebase (forked circa 2009). While reddit's code served as a stable platform while our community was in its initial stages, it has become hard to develop and extend because of its age, complexity and monolithic design.

Lesswrong 2.0, on the other hand, is based on modern web technologies designed to make rapid development much easier (to be precise React, GraphQL, Slate.js, Vulcan.js and Meteor). The old codebase was a pain to work with, and almost every developer who tried to contribute gave up after trying their hands at it. The new Lesswrong codebase on the other hand is built with tools that are well-documented and accessible, and is designed to have a modular architecture. You can find our Github repo here.

We hope that these architectural decisions will allow us to rapidly improve the site and turn it into what a tool for creating intellectual progress should look like in 2017.

II. Effective Moderation

Historically, LW has had only a few dedicated moderators at a time, applying crude tools, which has tended to lead to burnout and backlash. There are many many obvious things we are planning to do to improve moderation, but here are some of the top ones:

Spam defense

Any user above N karma can flag a post as spam, which renders it invisible to everyone but mods. Mods will check the queue of spam posts, deleting correct flags, and removing the power from anyone that misuses it. If it seems necessary, we will also be integrating all the cool new spam detection mechanisms that modern technology has given us in the last 8 years.

Noob defense

Historically, Lesswrong’s value has come in large part from being a place on the internet where the comments were worth reading. This was largely a result of the norms and ability of the people who were commenting on the page, with a strong culture of minimizing defensiveness, searching for the truth and acting in the spirit of double crux. To sustain that culture and level of quality, we need to set up broad incentives that are driven by the community itself.

The core strategy we are currently considering is something we’re calling the Sunshine Regiment. The Sunshine regiment is a pretty large set of trusted users who have access to reduced moderating powers, such as automatically hiding comments for other users and temporarily suspending comment threads. The goal is to give the community the tools to de-escalate conflicts and help both the users and moderators make better decisions, by giving both sides time to reflect and think and by distributing the load of draining moderation decisions.

Troll defense

The two main plans we have against trolls is to change the Karma system to something more like “Eigenkarma” and improvements to the moderator tools. In an Eigenkarma system the weights of the votes of a user depends on how many other trustworthy users have upvoted that user. For the moderator tools, one of the biggest projects is a much better data querying interface that aims to help admins notice exploitative voting behavior and other problems in the voting patterns.

III. Discourse Norms

In terms of culture, we still broadly agree with the principles that Eliezer established in the early days of Overcoming Bias and Lesswrong. The twelve virtues of rationality continue to resonate with us, and the “The Craft and the Community” sequence is still highly influential on our thinking. The team (and in particular Oliver) have taken significant inspiration from the original vision of Arbital in our ideas for Lesswrong 2.0.

That being said we also think that the culture of the rationality community has substantially changed in the last eight years, and that many of those changes were for the better. As Eliezer himself said in the opening to “Rationality: AI to Zombies”:

> It was a mistake that I didn’t write my two years of blog posts with the intention of helping people do better in their everyday lives. I wrote it with the intention of helping people solve big, difficult, important problems, and I chose impressive-sounding, abstract problems as my examples. In retrospect, this was the second-largest mistake in my approach. It ties into the first-largest mistake in my writing, which was that I didn’t realize that the big problem in learning this valuable way of thinking was figuring out how to practice it, not knowing the theory. I didn’t realize that part was the priority; and regarding this I can only say ‘Oops’ and ‘Duh.’

We broadly agree with this, and think both that the community has made important progress in that direction, and that there are still many things to improve about the current community culture. We do not aim to make the new Lesswrong the same as it was at its previous height, but instead aim to integrate many of the changes to the culture of the rationalist culture, while also re-emphasizing important old virtues that we feel have been lost in the intervening years.

We continue to think that strongly discouraging the discussion of highly political topics is the correct way to go. A large part of the value of Lesswrong comes from being a place where many people can experience something closer to rational debate for the first time in their life. Political topics are important, and not to be neglected, but they serve as a bad introduction and base on which to build a culture of rationality. We are open to creating spaces on Lesswrong where people above a certain Karma threshold can discuss political topics, but we would not want that part of the site to be visible to new users, and we would want the votes on that part of the site to be less-important for the total karma of the participating users. We want seasoned and skilled rationalists to discuss political topics, but we do not want users to seek out Lesswrong primarily as a venue to hold political debates.

As a general content guideline on the new Lesswrong: If while writing the article the author is primarily writing with the intent of rallying people to action, instead of explaining things to them, then the content is probably ill-suited for Lesswrong.

IV. New Features

You can find our short-term feature roadmap over here in this post. This is a high-level overview on our reasoning on some of the big underlying features we expect to significantly shape the nature of Lesswrong 2.0.

Content curation:

Many authors want their independence, which is one of the reasons why Scott Alexander prefers to write on SlateStarCodex instead of Lesswrong. We support that need for independence, and are hoping to serve it in two different ways:

- We are making it very easy for trusted members of the rationality community to crosspost their content to Lesswrong. We already set up an RSS-feed integration that allows admins to associate external RSS feeds with a user, so that whenever something new gets added to that RSS feed, their user account will automatically create a post with their new content on Lesswrong, and if they want us to, not only add a link but the complete text of their post (which encourages discussion on Lesswrong instead of the external blog).

- We want to give trusted authors moderation powers for the discussions on their own posts, allowing them to foster their own discussion norms, and giving them their own sphere of influence on the discussion platform. We hope this will both make the lives of our top authors better and will also create a form of competition between different cultures and moderation paradigms on Lesswrong.

Arbital-Style features and content:

Arbital did many things right, even though it never really seemed to take off. We think that allowing users to add prediction-polls is great, and that it is important to give authors the tools to create their own content that is designed to be maintained over a long period of time and with multiple authors. We also really like link previews on hover-over as well as the ability to create highly interconnected networks of concepts with overview pages.

Of the Arbital features, prediction polls are most certainly going to end up on feature list, but as of yet, it is unclear whether we want to copy any other features directly, though we expect to be inspired by many small features.

Better editor software:

The editor on Lesswrong and the EA Forum often lead to badly formatted posts. The editor didn’t deal well with copying content over from other webpages or Google docs, which often resulted in hard to read posts that could only be fixed by directly editing the HTML. We are working on an editor experience that will be flexible and powerful, while also making it hard to accidentally mess up the formatting of a post.

Sequences-like content with curated comments:

After doing a large amount of interviews with old users of Lesswrong, it became clear to us that the vast majority of top contributors on Lesswrong spent at least 3 months doing nothing else but reading the sequences and other linearly-structured content on the page, while also reading the discussion on those posts. We aim to improve that experience significantly, while also making it easier to start participating in the discussion.

Books like Rationality: AI to Zombies are valuable in that they reach an audience that was impossible to reach with the old Lesswrong, and by curating the content into an established book-like format. But we also think that something very important is lost when cutting the discussion out of the content. We aim to make Lesswrong a platform that provides sequences-like content in formats that are as easy to consume as possible, while also encouraging the user to engage with the discussion on the posts, and be exposed to critical comments, disagreements and important contradicting or supporting facts. We also hope that being exposed to the discussion will more directly teach new users how to interact with the culture of Lesswrong and to learn the art of rationality more directly by observing people struggling in conversation with difficult intellectual problems.

V. Beta Feedback Period

It’s important for us to note that we don’t think that online discussion is primarily a technical problem. Our intention in sharing our plans with you and launching a closed beta are to discover both the cultural and the and technical problems that we need to solve to build a new and better discussion platform for our community. With your feedback we're planning to rework the site, adjust our feature priorities and and make new plans for improving the culture of the new Lesswrong 2.0 community.

Far more important than implementing any particular feature is building an effective culture that has the correct social incentives. As such, our focus lies on building a community with norms and social incentives that facilitate good discourse, with a platform that does not get in the way of that. However, we do think that there are certain underlying attributes of a discussion platform that significantly shift the nature of discussions on that platform in a way that prevents or encourages good community norms to arise, i.e. Twitter's 140 character limit makes it almost impossible to have reasoned discourse. At this stage, we are still trying to figure out what the best content-types and fundamental design philosophies are that are best for giving rise and facilitating effective discussion.

That's all we have for now. Please post your ideas for features or design changes as top-level comments, and discuss your concerns and details of the suggestions in second-level comments. We will be giving significant weight to the discussion and votes in our decisions on what to work on for the coming weeks.