The LessWrong Review

The LessWrong 2018 Review

LessWrong is currently doing a major review of 2018 — looking back at old posts and considering which of them have stood the tests of time. It has three phases:

- Nomination *(ends Dec 1st at 11:59pm PST)*
- Review *(ends Dec 31st)*
- Voting on the best posts *(ends January 7th)*

Authors will have a chance to edit posts in response to feedback, and then the moderation team will compile the best posts into a physical book and LessWrong sequence, with $2000 in prizes given out to the top 3-5 posts and up to $2000 given out to people who write the best reviews.

Helpful Links:

- *Top 2018 posts sorted by karma*
- *2018 posts aggregated by month*
- *You can see nominated posts here*
- *Voting Results*

This is the first week of the LessWrong 2018 Review – an experiment in improving the LessWrong Community's longterm feedback and reward cycle.

This post begins by exploring the motivations for this project (first at a high level of abstraction, then getting into some more concrete goals), before diving into the details of the process.

Improving the Idea Pipeline

In his LW 2.0 Strategic Overview, habryka noted:

> We need to build on each other’s intellectual contributions, archive important content, and avoid primarily being news-driven.
>
> We need to improve the signal-to-noise ratio for the average reader, and only broadcast the most important writing
>
> [...]
>
> Modern science is plagued by severe problems, but of humanity’s institutions it has perhaps the strongest record of being able to build successfully on its previous ideas.
>
> The physics community has this system where the new ideas get put into journals, and then eventually if they’re important, and true, they get turned into textbooks, which are then read by the upcoming generation of physicists, who then write new papers based on the findings in the textbooks. All good scientific fields have good textbooks, and your undergrad years are largely spent reading them.

Over the past couple years, much of my focus has been on the **early-stages** of LessWrong's idea pipeline – creating affordance for off-the-cuff conversation, brainstorming, and exploration of paradigms that are still under development (with features like shortform and moderation tools).

But, the beginning of the idea-pipeline is, well, not the end.

I've written a couple times about what the later stages of the idea-pipeline might look like. My best guess is still something like this:

> I want LessWrong to encourage extremely high quality intellectual labor. I think the best way to go about this is through escalating positive rewards, rather than strong initial filters.
>
> Right now our highest reward is getting into the curated section, which... just isn't actually that high a bar. We only curate posts if we think they are making a good point. But if we set the curated bar at "extremely well written and extremely epistemically rigorous and extremely useful", we would basically never be able to curate anything.
>
> My current guess is that there should be a "higher than curated" level, and that the general expectation should be that posts should only be put in that section after getting reviewed, scrutinized, and most likely rewritten at least once.

I still have a lot of uncertainty about the right way to go about a review process, and various members of the LW team have somewhat different takes on it.

I've heard lots of complaints about mainstream science peer review: that reviewing is often a thankless task; the quality of review varies dramatically, and is often entangled with weird political games.

Meanwhile: LessWrong posts cover a variety of topics – some empirical, some philosophical. In many cases it's hard to directly evaluate their truth or usefulness. LessWrong team members had differing opinions on what sort of evaluation is most useful or practical.

I'm not sure if the best process is more open/public (harnessing the wisdom of crowds) or private (relying on the judgment of a small number of thinkers). The current approach involves a mix of both.

What I'm most confident in is that the review should focus on older posts.

New posts often feel exciting, but a year later, looking back, you can ask if it *actually* has become a helpful intellectual tool. (I'm also excited for the idea that, in future years, the process could also include reconsidering previously-reviewed posts, if there's been something like a "replication crisis" in the intervening time)

Regardless, I consider the LessWrong Review process to be an experiment, which will likely evolve in the coming years.

Goals

Before delving into the process, I wanted to go over the high level goals for the project:

1. Improve our longterm incentives, feedback, and rewards for authors

2. Create a highly curated "Best of 2018" sequence / physical book

3. Create common knowledge about the LW community's collective epistemic state regarding controversial posts

Longterm incentives, feedback and rewards

Right now, authors on LessWrong are rewarded essentially by comments, voting, and other people citing their work. This is fine, as things go, but has a few issues:

- Some kinds of posts are quite valuable, but don't get many comments (and these disproportionately tend to be posts that are more proactively rigorous, because there's less to critique, or critiquing requires more effort, or building off the ideas requires more domain expertise)
- By contrast, comments and voting both nudge people towards posts that are clickbaity and controversial.
- Once posts have slipped off the frontpage, they often fade from consciousness. I'm excited for a LessWrong that rewards Long Content, that stand the tests of time, as is updated as new information comes to light. (In some cases this may involve editing the original post. But if you prefer old posts to serve as a time-capsule of your post beliefs, adding a link to a newer post would also work)
- Many good posts begin with an "epistemic status: thinking out loud", because, at the time, they were just thinking out loud. Nonetheless, they turn out to be quite good. Early-stage brainstorming is good, but if 2 years later the early-stage-brainstorming has become the best reference on a subject, authors should be encouraged to change that epistemic status and clean up the post for the benefit of future readers.

The aim of the Review is to address those concerns by:

- Promoting old, vetted content directly on the site.
- Awarding prizes not only to authors, but to reviewers. It seems important to directly reward high-effort reviews that thoughtfully explore both how the post could be improved, and how it fits into the broader intellectual ecosystem. (At the same time, *not* having this be the final stage in the process, since building an intellectual edifice requires four layers of ongoing conversation)
- Compiling the results into a physical book. I find there's something... literally *weighty* about having your work in printed form. And because it's much harder to edit books than blogposts, the printing gives authors an extra incentive to clean up their past work or improve the pedagogy.

A highly curated "Best of 2018" sequence / book

Many users don't participate in the day-to-day discussion on LessWrong, but want to easily find the best content.

To those users, a "Best Of" sequence that includes not only posts that seemed exciting at the time, but distilled reviews and followup, seems like a good value proposition. And meanwhile, helps move the site away from being time-sensitive-newsfeed.

Common knowledge about the LW community's collective epistemic state regarding controversial posts

Some posts are highly upvoted because everyone agrees they're true and important. Other posts are upvoted because they're more like exciting hypotheses. There's a lot of disagreement about which claims are actually true, but that disagreement is crudely measured in comments from a vocal minority.

The end of the review process includes a straightforward vote on which posts seem (in retrospect), useful, and which seem "epistemically sound". This is not the *end* of the conversation about which posts are making true claims that carve reality at it's joints, but my hope is for it to ground that discussion in a clearer group-epistemic state.

Review Process

Nomination Phase

1 week (Nov 20th – Dec 1st)

- Users with 1000+ karma can nominate posts from 2018, describing how they found the post useful over the longterm.
- The nomination button is in the post dropdown-menu (available at the top of posts, or to the right of their post-item)
- For convenience, you can review posts via:
  - a list of all 2018 posts, sorted by karma
  - if you want a more in-depth overview, 2018 posts clustered by month

Review Phase

4 weeks (Dec 1st – Dec 31st)

- Authors of nominated posts can opt-out of the review process if they want.
  - *They also can opt-in, while noting that they probably won't have time to update their posts in response to critique. (This may reduce the chances of their posts being featured as prominently in the Best of 2018 book)*
- Posts with sufficient* nominations are announced as contenders.
  - *We're aiming to have 50-100 contenders, and the nomination threshold will be set to whatever gets closest to that range*
- For a month, people are encouraged to look at them thoughtfully, writing comments (or posts) that discuss:
  - How has this post been useful?
  - How does it connect to the broader intellectual landscape?
  - Is this post epistemically sound?
  - How could it be improved?
  - What further work would you like to see people do with the content of this post?
- A good frame of reference for the reviews are shorter versions of LessWrong or SlatestarCodex book reviews (which do a combination of epistemic spot checks, summarizing, and contextualizing)

- Authors are encouraged to engage with reviews:
    - Noting where they disagree
    - Discussing what sort of followup work they'd be interested in seeing from others
    - Ideally, updating the post in response to critique they agree with

Voting Phase

1 Week (Jan 1st – Jan 7th)

Posts that got at least one review proceed to the voting phase. The details of this are still being fleshed out, but the current plan is:

- Users with 1000+ karma rate each post on a 1-10 scale, with 6+ meaning "*I'd be happy to see this included in the 'best of 2018*'" roundup, and 10 means "*this is the best I can imagine*"
- Users are encouraged to (optionally) share the reasons for each rating, and/or share thoughts on their overall judgment process.

Books and Rewards

Public Writeup / Aggregation

Soon afterwards (hopefully within a week), the votes will all be publicly available. A few different aggregate statistics will be available, including the raw average, and potentially some attempt at a "karma-weighted average."

Best of 2018 Book / Sequence

Sometime later, the LessWrong moderation team will put together a physical book, (and online sequence), of the best posts and most valuable reviews.

This will involve a lot of editor discretion – the team will essentially take the public review process and use it as input for the construction of a book and sequence.

I have a lot of uncertainty about the shape of the book. I'm guessing it'd include anywhere from 10-50 posts, along with particularly good reviews of those posts, and some additional commentary from the LW team.

Note: This may involve some custom editing to handle things like hyperlinks, which may work differently in printed media than online blogposts. This will involve some back-and-forth with the authors.

Prizes

- Everyone whose work is featured in the book will receive a copy of it.
- There will be $2000 in prizes divided among the authors of the top 3-5 posts (judged by the moderation team)
- There will be **up to** $2000 in prizes for the best 0-10 reviews that get included in the book. (The distribution of this will depend a bit on what reviews we get and how good they are)
- *(note: LessWrong team members may be participating as reviewers and potentially authors, but will not be eligible for any awards)*

The Review Phase

LessWrong is currently doing a major review of 2018 — looking back at old posts and considering which of them have stood the test of time. Info about what features we added to the site for writing reviews is in December's monthly updates post.

There are three phases:

- ***Nomination** (completed)*
- ***Review** (ends Dec 31st [EDIT: Jan 13th])*
- ***Voting** on the best posts (ends January 7th [EDIT: Jan 13th])*

We're now in the Review Phase, and there are 75 posts that got two or more nominations. The full list is here. Now is the time to dig into those posts, and for each one ask questions like "What did it add to the conversation?", "Was it epistemically sound?" and "How do I know these things?".

The LessWrong team will award $2000 in prizes to the reviews that are most helpful to them for deciding what goes into the Best of 2018 book.

If you're a nominated author and for whatever reason don't want one or more of your posts to be considered for the Best of 2018 book, contact any member of the team - e.g. drop me an email at benitopace@gmail.com.

Creating Inputs For LW Users' Thinking

The goal for the next month is for us to try to figure out which posts we think were the best in 2018.

Not which posts were talked about a lot when they were published, or which posts were highly upvoted at the time, but which posts, with the benefit of hindsight, you're most grateful for being published, and are well suited to be part of the foundation of future conversations.

This is in part an effort to reward the best writing, and in part an effort to solve the bandwidth problem (there were more than 2000 posts written in 2018) so that we can build common knowledge of the best ideas that came out of 2018.

With that aim, when *I'm* reviewing a post, the main question I'm asking myself is

> What information can I give to other users to help them think clearly and accurately about whether a given post should be added to our annual journal?

A large part of the review phase is about producing inputs for our collective thinking. With that in mind, I've gathered some examples of things you can write that are help others understand posts and their impacts.

1) Personal Experience Reports

There were a lot of examples of this in the nomination phase, which I found really useful, and would find useful to read more of. Here are some examples:

Raemon:

> This post... may have actually had the single-largest effect size on "amount of time I spent thinking thoughts descending from it."

Joh N. Swentworth

> This post (and the rest of the sequence) was the first time I had ever read something about AI alignment and thought that it was actually asking the right questions. It is not about a sub-problem, it is not about marginal improvements. Its goal is a gears-level understanding of agents, and it directly explains why that's hard. It's a list of everything which needs to be figured out in order to remove all the black boxes and Cartesian boundaries, and understand agents as well as we understand refrigerators.

Swimmer963:

> Used as a research source for my EA/rationality novel project, found this interesting and useful.

David Manheim:

> Until seeing this post, I did not have a clear way of talking about common knowledge. Despite understanding the concept fairly well, this post made the points more clearly than I had seen them made before, and provided a useful reference when talking to others about the issue.

Eli Tyre:

> One of my favorite posts, that encouraged me to rethink and redesign my honesty policy.

ryan_b:

> I have definitely linked this more than any other post.

More detail is also really great. I'd definitely encourage the above users to be more thorough about how the ideas in the post impacted them. Here's a nomination that had a bunch more detail about how the ideas have affected them.

jacobjacob:

> In my own life, these insights have led me to do/considering doing things like:
>
> • not sharing private information even with my closest friends -- in order for them to know in future that I'm the kind of agent who can keep important information (notice that there is the counterincentive that, in the moment, sharing secrets makes you feel like you have a stronger bond with someone -- even though in the long-run it is evidence to them that you are less trustworthy)
>
> • building robustness between past and future selves (e.g. if I was excited about and had planned for having a rest day, but then started that day by work and being really excited by work, choosing to stop work and decide to rest such that

different parts of me learn that I can make and keep inter-temporal deals (*even if* work seems higher ev in the moment))

- being more angry with friends (on the margin) -- to demonstrate that I have values and principles and will defend those in a predictable way, making it easier to coordinate with and trust me in future (and making it easier for me to trust others, knowing I'm capable of acting robustly to defend my values)

- thinking about, in various domains, "What would be my limit here? What could this person do such that I would stop trusting them? What could this organisation do such that I would think their work is net negative?" and then looking back at those principles to see how things turned out

- not sharing passwords with close friends, even for one-off things -- not because I expect them to release or lose it, but simply because it would be a security flaw that makes them more vulnerable to anyone wanting to get to me. It's a very unlikely scenario, but I'm choosing to adopt a robust policy across cases, and it seems like useful practice

A special case here is data from the author themselves, e.g. “Yeah, this has been central to my thinking” or “I didn’t really think about it again” or “I actually changed my mind and think this is useful but wrong”. I would generally be excited for users to review their own posts now that they've had ~1.5 years of hindsight, and I plan to do that for all the posts I've written that were nominated.

If a post had a big or otherwise interesting impact on you, consider writing that up.

2) Big Picture Analysis (e.g. Book Reviews)

There are lots of great book reviews on the web that really help the reader understand the context of the book, and explain what it says and adds to the conversation.

Some good examples on LessWrong are be the reviews of Pearl's Book of Why, The Elephant in the Brain, The Secret of Our Success, Consciousness Explained, Design Principles of Biological Circuits, The Case Against Education (part 2, part 3), and The Structure of Scientific Revolutions.

Many of these reviews do a great job of things like

- Talking about how the post fits into the broader conversation on that topic
- Trying to pass the ITT of the author by explaining how they see the world
- Looking at that same topic through their own worldview
- Pointing out places they see things differently and offering alternative hypotheses.

A review of some LessWrong posts would be that time Scott reviewed Inadequate Equilibria. Oh, and don’t forget that time Scott reviewed Inadequate Equilibria.

Many of the posts we’re reviewing are shorter than most of the reviews I linked to, so it doesn’t apply literally, but much of the spirit of these reviews is great. Also check out others short book reviews and consider writing something in that style (e.g. SSC, Thing of Things).

Consider picking a book review style you like and applying it to one of the nominated posts.

3) Testing Subclaims (e.g. Epistemic Spot Checks)

Elizabeth Van Nostrand has written several posts in this style.

- Epistemic Spot Check: The Role of Deliberate Practice in the Acquisition of Expert Performance
- Epistemic Spot Check: Full Catastrophe Living (Jon Kabat-Zinn)
- Epistemic Spot Check: The Dorito Effect (Mark Schatzker)

For another example, in Scott's review of Secular Cycles, one way he tried to think about the ideas in the book was to gather a bunch of alternative data sets on which to test some of the author’s claims.

These things aren't meant to be full reviews of the entire book or paper, or advice on overall how to judge it. They take narrower questions that are definitively answerable, like is a random sample of testable claims literally true, and answers them as fully as possible.

If there is an important subclaim of a post you think you can check out, consider trying to verify/falsify the claim and writing up your results and partial results.

Go forth and think out loud!

Quadratic voting for the 2018 Review

This is an old post, but still contains useful information on quadratic voting for the LessWrong Review.

LessWrong is currently reviewing the posts from 2018, and I'm trying to figure out how voting should happen. The new hotness that all your friends are talking about is quadratic voting, and after thinking about it for a few hours, it seems like a pretty good solution to me.

I'm writing this post primarily for people who know more about this stuff to show me where the plan will fail terribly for LW, to suggest UI improvements, or to suggest an alternative plan. If nothing serious is raised that changes my mind in the next 7 days, we'll build a straightforward UI and do it.

I've not read anything about it, so briefly, what is quadratic voting?

I'm just picking it up, so I'll write a short explanation it as I understand it, and will update this if it's confused/mistaken.

As I understand it, the key insight behind quadratic voting is that everyone can cast multiple votes, but that *the marginal cost of votes is increasing*, rather than staying constant.

With other voting mechanisms where cost stays constant, people are always incentivised to vote repeatedly for their favourite option. This is similar to how, under most scoring rules, people bet *all* their chips on the outcome they think is most likely, rather than spread their money directly proportional to their actual probability mass. You have to think carefully to design a proper scoring rule to incentivise users to write down their full epistemic state.

With quadratic voting, instead of spending all your votes on your favourite option, the more you spend on an option the more it costs to spend more on that option, and other options start looking more worthwhile. Your first vote costs 1, your second vote costs 2, your nth votes costs n. And you have a *limited* amount of cost you can pay, so you start having to make new comparisons about where the marginal cost should be spent.

Concretely, there are two things for a person voting in this system to keep a track of:

- *Votes* is the total votes you make.
- *Cost* is the total cost you pay for your votes

If you votes are

- Post A: *5 votes*
- Post B: *8 votes*
- Post C: *1 vote*

Then your numbers are:

- *Votes* is 5 + 8 + 1 = 14
- *Cost* is (1 + 2 + 3 + 4 + 5) + (1 + 2 + 3 + 4 + 5 + 6 + 7 + 8) + (1) = 54

This is a system that not only invites voters to give a rank ordering, but to give your price for marginal votes on different posts. This information makes it much more easy to combine everyone's preferences (though how much you care about everyone's utilities is a free variable in the system. Democracies weight everyone equally, and we can consider alternatives like karma-weighting on LW).

It's called quadratic voting because the sum of all numbers up to n is like half of n^2. I think (but don't know) that it doesn't really matter how much each vote costs, as long as it's increasing, because then it causes you to calculate your price for marginal votes on different options. "Price Voting" might have been a good name for it.

Other explanations I've seen frame it that the marginal vote gets counted as less, rather than it costing more, but I'm pretty sure this has an identical effect.

Vitalik Buterin has written more about it here.

How would this work in the LessWrong 2018 Review?

So, we're trying to figure out what were the best posts from 2018, in an effort to build common knowledge of what the progress we made was, and also reward thinkers for having good ideas. We'll turn the output into a sequence and, with a bit of editing, I'll also make it into a physical book.

(Note: Content will only be published with author's explicit consent, obviously.)

To do this vote, I'd suggest the following setup:

- All users over a certain karma threshold (probably 1000 karma, which is ~500 users) are given an input page where they can vote on all posts that were nominated that year.
- Their total Cost users can spend is set to 500.
- We will keep this open for users for 2 weeks, during which time they can cast their votes. You can save your votes and also come back to edit them any time in the two weeks.
- While the votes are being cast, there will be a 'snapshot' published 1 week in, showing what the final result would be if the vote were taken that day. This will help users understand what output the votes are connected to, and help to figure out what posts you want to write reviews for (i.e. writing things to help other users understand why a post is being undervalued/overvalued according to you).
- At the end, I'll publish a few versions of aggregating the votes, such as karma-weighted, not-karma-weighted, and maybe some other ways of combining rank orderings. We'll do something like select the top N posts where the word-count sums to ~100k words (aka a 350 page book) to be in the final sequence. I expect this will be around 25-30 posts.

Some objections

What if I have reason to think a post is terrible?

Negative voting is also allowed, where a negative votes costs the same as a positive vote i.e. 4 votes costs the same as -4 votes. I think this probably deals with such cases fine.

What about AI alignment writing, or other writing that I feel I cannot evaluate?

So let me first deal with the most obvious case here, which is AI alignment writing. I think these posts are important research from the perspective of the long-term future of civilization, but also they aren't of direct interest to most users of LessWrong, and more importantly many users can't personally tell which posts are good or bad.

For one, I think that alignment ideas are very important and I plan to personally work on it in a focused way, and so I'm not too worried about it not all getting recognised in this review process. I'd ideally like to make books of the Embedded Agency, Iterated Amplification, and Value Learning sequences, for example, so I'm not too bothered if the best ideas from each of those aren't in the *LessWrong 2018* book.

For two, I think that I think there is a deep connection between AI alignment and rationality, and that much key stuff for thinking about both (bayes, information theory, embedded agency, etc) has been very useful for me in thinking about both AI alignment and personal decision making. I think that some of the best alignment content consists of deep insights that many rationalists will find useful, so I do think some of it will pass this bar.

For three, I still think I trust users to have a good sense of what content is valuable. I think I have a sense of what alignment content is useful in part because I trust other users in a bunch of ways (Paul Christiano, Wei Dai, Abram Demski, etc). There's no rule saying you can't update on the ideas and judgement of people you trust.

Overall I trust users and their judgments quite a bit, and if users feel they can't tell if a post is good, then I think I want to trust that judgment.

Is this too much cognitive overhead for LW users?

I am open to ideas for more efficient voting systems.

A key thing about spreading it over two weeks, means that users don't have to do it in a single sitting. Personally, I feel like I can cast my quadratic votes on 75 posts in about 20 mins, and then will want to come back to it a few days later to see if it still feels right. However Ray found it took him more like 2 hours of heavy work, and feels like the system will be too complicated for most people to get a handle on.

I think the minimum effort is fairly low, so I expect most users to have a fine time, but I'm happy to receive public and private feedback about this. In general I likely want to make a brief survey for afterwards, about how the whole process went.

Voting Phase of 2018 LW Review

For the past 1.5 months on LessWrong, we've been doing a major review of 2018 — looking back at old posts and asking which of them have stood the test of time.

The LessWrong 2018 Review has three major goals.

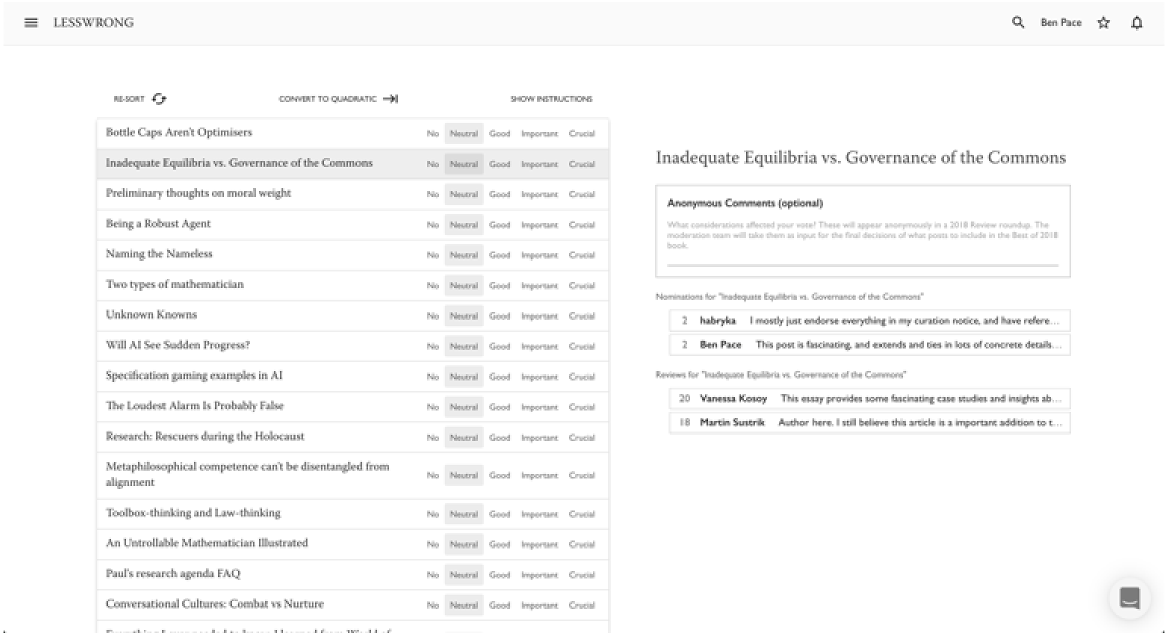
- First, it is an experiment in improving the LessWrong community's longterm feedback and reward cycle.
- Second, it is an attempt to build common knowledge about the best ideas we've discovered on LessWrong.
- Third, after the vote, the LessWrong Team will compile the top posts into a physical book.

We spent about 2 weeks nominating posts, and 75 posts received the the 2 required nominations to pass through that round. (See all nominations on the nominations page.) Then we spent a month reviewing them, and users wrote 72 reviews, a number of them by the post-authors themselves. (See all reviews on the reviews page.)

And finally, as the conclusion of all of this work, we are now voting on all the nominated posts. Voting is open for 12 days, and will close on Sunday, January 19th. (We'll turn it off on Monday during the day, ensuring all timezones get it throughout Sunday.)

The vote has a simple first section, and a detailed-yet-optional second section based on quadratic voting. If you are one of the 430 users with 1000+ karma, you are eligible to vote, then now is the time for you to participate in the vote by following this link.

For all users and lurkers, regardless of karma, the next 12 days are your last opportunity to write reviews for any nominated posts in 2018, which I expect will have a significant impact on how people vote. As you can see, all reviews are highlighted when a user is voting on a post. (To review a post, go to the post and click "Write A Review" at the top of the post.)



Bottle Caps Aren't Optimisers	No	Neutral	Good	Important	Crucial
Inadequate Equilibria vs. Governance of the Commons	No	Neutral	Good	Important	Crucial
Preliminary thoughts on moral weight	No	Neutral	Good	Important	Crucial
Being a Robust Agent	No	Neutral	Good	Important	Crucial
Naming the Nameless	No	Neutral	Good	Important	Crucial
Two types of mathematician	No	Neutral	Good	Important	Crucial
Unknown Knowns	No	Neutral	Good	Important	Crucial
Will AI See Sudden Progress?	No	Neutral	Good	Important	Crucial
Specification gaming examples in AI	No	Neutral	Good	Important	Crucial
The Loudest Alarm Is Probably False	No	Neutral	Good	Important	Crucial
Research: Rescuers during the Holocaust	No	Neutral	Good	Important	Crucial
Metaphilosophical competence can't be disentangled from alignment	No	Neutral	Good	Important	Crucial
Toolbox-thinking and Law-thinking	No	Neutral	Good	Important	Crucial
An Untrollable Mathematician Illustrated	No	Neutral	Good	Important	Crucial
Paul's research agenda FAQ	No	Neutral	Good	Important	Crucial
Conversational Cultures: Combat vs Nurture	No	Neutral	Good	Important	Crucial

Inadequate Equilibria vs. Governance of the Commons

Anonymous Comments (optional)

What considerations affected your vote? These will appear anonymously in a 2018 Review roundup. The moderation team will take them as input for the final decisions of what posts to include in the Best of 2018 book.

Nominations for "Inadequate Equilibria vs. Governance of the Commons"

2	**habryka**	I mostly just endorse everything in my curation notice, and have refere…
2	**Ben Pace**	This post is fascinating, and extends and ties in lots of concrete details…

Reviews for "Inadequate Equilibria vs. Governance of the Commons"

20	**Vanessa Kosoy**	This essay provides some fascinating case studies and insights ab…
18	**Martin Sustrik**	Author here. I still believe this article is a important addition to t…

This is the end of this post. If you'd like to read more detailed instructions about how to vote, the rest of the text below contains instructions for how to use the voting system.

How To Vote

Sorting Posts Into Buckets

The first part of voting is sorting the nominated posts into buckets.

The five buckets are: No, Neutral, Good, Important, Crucial. Sort the posts however you think is best.

The key part is the *relative weighting* of different posts. For example, it won't make a difference to your final vote if you put every post in 'crucial' or every post in 'good'.

Fine-Tuning Your Votes

The system we're using is quadratic voting (as I discussed a few weeks ago).

Once you're happy with your buckets, click 'Convert to Quadratic'. At this point the system converts your buckets roughly into their quadratic equivalents.

The system will only assign integer numbers of votes, which means that it will likely only allocate around 80-90% of the total votes available to you. If you vote on a smaller number of posts (<10), the automatic system may not use your entire quadratic voting budget.

If you're happy with how things look, you can just leave at this point, and your votes will be saved (you can come back any time before the vote closes to update them). But if you want to allocate 100% of your available votes, you'll likely need to do fine-tuning.

There are two key parts of quadratic voting you need to know:

- First, you have a *limited budget* of votes.
- Second, votes on a post have *increasing* marginal cost.

This means that your first vote costs 1 point, your second vote on that post costs 2 points, your third costs 3 points. Your *nth* vote costs *n* points.

You have 500 points to spend. You can see how many points you've spent at the top of the posts.

The system will automatically weight the buckets differently. For example, I just did this, and I got the following weightings:

- Good: 2 votes.
- Important: 4 votes.
- Crucial: 9 votes.
- Neutral: 0 votes.
- No: -4 votes.

(Note that negative votes cost the same as positive votes. The first negative vote costs 1 point, the second negative vote costs 2 points, etc.)

You'll see your score at the top of the page. (When I arrived on the fine-tuning page, the system had spent about 416 points, which meant I had a significant number of votes left to buy.)

Once you're happy with the balance, just close the page; your votes will be saved.

You can return to this page anytime until voting is over, to reconfigure your weights.

Leaving Comments (Anonymous)

There's a field to leave anonymous thoughts on a post. All comments written here will be put into a public Google doc, and be linked to from the post that announces the result of the vote. If you want to share your thoughts, however briefly, this is a space to do that.

I will likely be making a book of 2018 posts, and if I do I will use the votes as my guide to what to include, so I'll definitely be interested in reading through people's anonymous thoughts and feelings about the 2018 LW posts.

Extra Notes

Additional info on voting

If you'd like to go back to the buckets stage, hit “Return to basic voting”. If you do this, all of your fine-tuning will be thrown out the window, and the system will re-calculate your weights entirely based on the new buckets you assign.

I find it really valuable to be able to see the posts in the order I've currently ranked them, so there's a button at the top to re-order the posts, which I expect to be clicked dozens of times by each user.

If you click on a post, all nominations and reviews for that post will appear in a box on the side of the page. You may want to read these to make a more informed decision when voting.

The results

The voting will have many outputs. Once we've had the time to analyse the data, we'll include a bunch of data/graphs, all anonymised, such as:

- For each winning post and each bucket, how many times people picked that bucket
- The individual votes each post got
- The results if the karma cutoff was 10,000 karma rather than 1,000
- The output of people's buckets compared with the output of people's quadratic fine-tuning
- The mean and standard deviation of the votes

Vote Here (if you have more than 1000 karma)

Please Critique Things for the Review!

I've spent a lot of time defending LW authors' right to have the conversation they want to have, whether that be early stage brainstorming, developing a high context idea, or just randomly wanting to focus on some particular thing.

LessWrong is not only a place for finished, flawless works. Good intellectual output requires both Babble and Prune, and in my experience the best thinkers often require idiosyncratic environments in order to produce and refine important insights. LessWrong is a **full-stack intellectual pipeline.**

But the 2018 Review is supposed to be *late stage* in that pipeline. We're pruning, not babbling here, and criticism is quite welcome. We're deliberately offering just as much potential prize money ($2000) to reviewers as to the top-rated authors.

Nominated authors had the opportunity to opt out of the review process, and none of them did. Getting nominated is meant to feel something like "getting invited to the grown-ups table", where your ideas are subjected to serious evaluation, and that scrutiny is seen as a sign of respect.

In my current expectations, the Review is one of the primary ways that LessWrong ensures high epistemic standards. But how well that plan works is proportional to how much effort critics put into it.

The Review and Voting Phases will continue for another until January 19th. During that time, review-comments will appear on the voting page, so anyone considering how to vote on a given post will have the opportunity to see critiques. The reviews will appear abridged initially, so I'd aim for the first couple sentences to communicate your overall takeaway.

The Review norms aren't "literally anything goes" – ridicule, name-calling etc still aren't appropriate. I'd describe the intended norms for reviews as "professional". But, posts nominated for the Review should treated as something like "the usual frontpage norms, but with a heavier emphasis on serious evaluation."

I'm still not sure precisely what the rules/guidelines should be about what is acceptable for the final Best of 2018 Book. In some cases, a post might make some important points, but also make some unjustified claims. (I personally think Local Validity as Key to Sanity and Civilization falls in this category). My current best guess is that it'd be fine if such posts end up in the book, but I'd want to make sure to also include reviews that highlighted any questionable statements.

Happy Critiquing!

2018 Review: Voting Results!

The votes are in!

59 of the 430 eligible voters participated, evaluating 75 posts. Meanwhile, 39 users submitted a total of 120 reviews, with most posts getting at least one review.

Thanks a ton to everyone who put in time to think about the posts - nominators, reviewers and voters alike. Several reviews substantially changed my mind about many topics and ideas, and I was quite grateful for the authors participating in the process. I'll mention Zack_M_Davis, Vanessa Kosoy, and Daniel Filan as great people who wrote the most upvoted reviews.

In the coming months, the LessWrong team will write further analyses of the vote data, and use the information to form a sequence and a book of the best writing on LessWrong from 2018.

Below are the results of the vote, followed by a discussion of how reliable the result is and plans for the future.

Top 15 posts

1. Embedded Agents by Abram Demski and Scott Garrabrant
2. The Rocket Alignment Problem by Eliezer Yudkowsky
3. Local Validity as a Key to Sanity and Civilization by Eliezer Yudkowsky
4. Arguments about fast takeoff by Paul Christiano
5. The Costly Coordination Mechanism of Common Knowledge by Ben Pace
6. Toward a New Technical Explanation of Technical Explanation by Abram Demski
7. Anti-social Punishment by Martin Sustrik
8. The Tails Coming Apart As Metaphor For Life by Scott Alexander
9. Babble by alkjash
10. The Loudest Alarm Is Probably False by orthonormal
11. The Intelligent Social Web by Valentine
12. Prediction Markets: When Do They Work? by Zvi
13. Coherence arguments do not imply goal-directed behavior by Rohin Shah
14. Is Science Slowing Down? by Scott Alexander
15. A voting theory primer for rationalists by Jameson Quinn and Robustness to Scale by Scott Garrabrant

Top 15 posts not about AI

1. Local Validity as a Key to Sanity and Civilization by Eliezer Yudkowsky
2. The Costly Coordination Mechanism of Common Knowledge by Ben Pace
3. Anti-social Punishment by Martin Sustrik
4. The Tails Coming Apart As Metaphor For Life by Scott Alexander
5. Babble by alkjash
6. The Loudest Alarm Is Probably False by Orthonormal
7. The Intelligent Social Web by Valentine
8. Prediction Markets: When Do They Work? by Zvi
9. Is Science Slowing Down? by Scott Alexander
10. A voting theory primer for rationalists by Jameson Quinn
11. Toolbox-thinking and Law-thinking by Eliezer Yudkowsky
12. A Sketch of Good Communication by Ben Pace
13. A LessWrong Crypto Autopsy by Scott Alexander
14. Unrolling social metacognition: Three levels of meta are not enough. by Academian
15. Varieties Of Argumentative Experience by Scott Alexander

Top 10 posts about AI

(The vote included 20 posts about AI.)

1. Embedded Agents by Abram Demski and Scott Garrabrant
2. The Rocket Alignment Problem by Eliezer Yudkowsky
3. Arguments about fast takeoff by Paul Christiano
4. Toward a New Technical Explanation of Technical Explanation by Abram Demski
5. Coherence arguments do not imply goal-directed behavior by Rohin Shah
6. Robustness to Scale by Scott Garrabrant
7. Paul's research agenda FAQ by zhukeepa
8. An Untrollable Mathematician Illustrated by Abram Demski
9. Specification gaming examples in AI by Vika
10. 2018 AI Alignment Literature Review and Charity Comparison by Larks

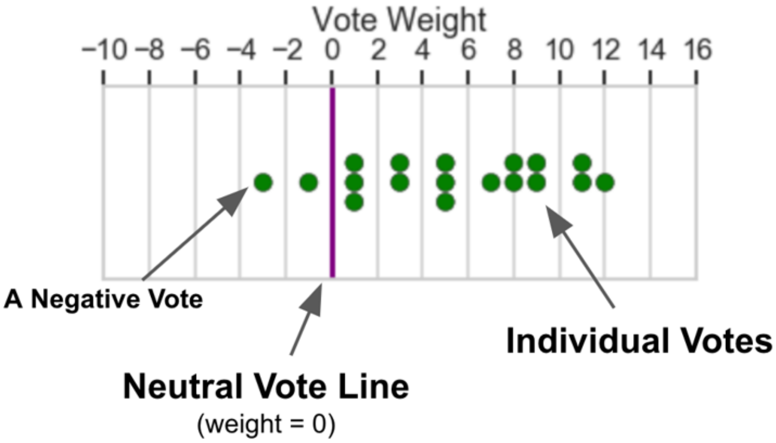
The Complete Results

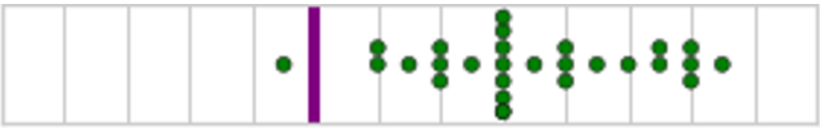
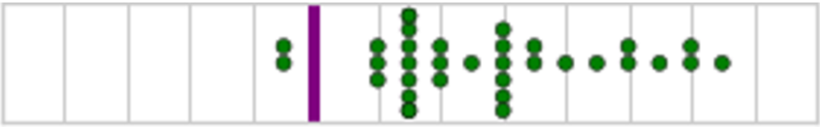
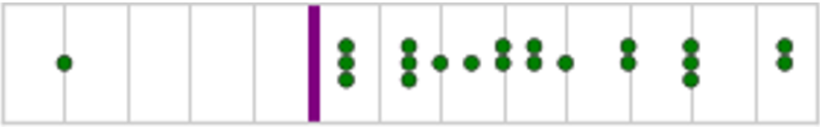
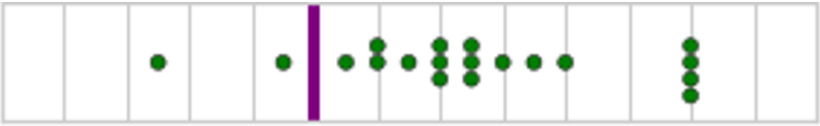
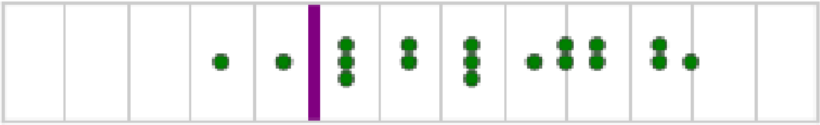
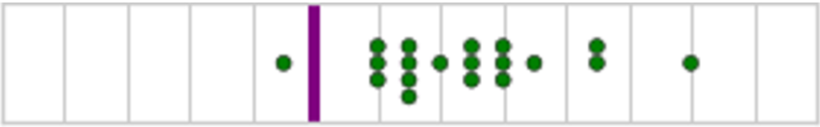
Click Here If You Would Like A More Comprehensive Vote Data Spreadsheet

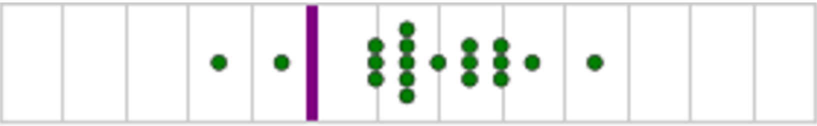
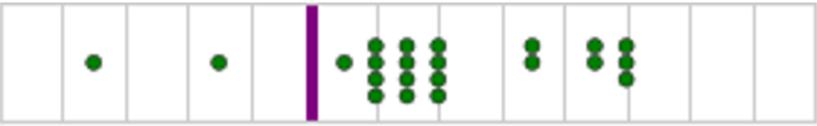
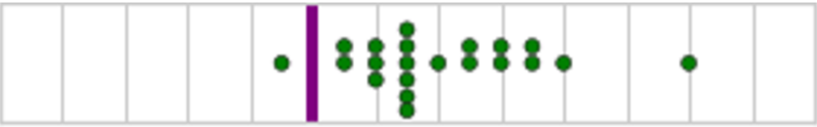
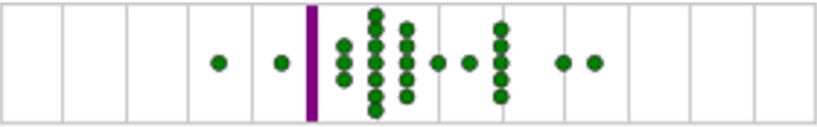
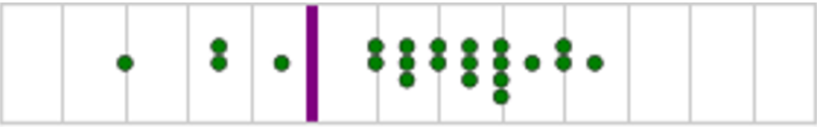
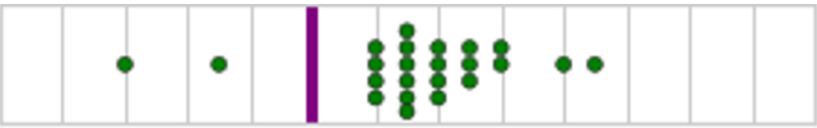
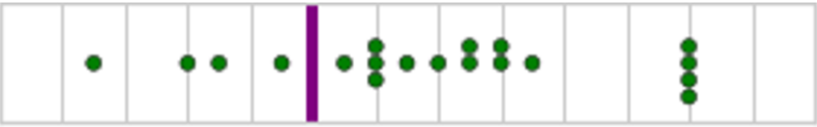
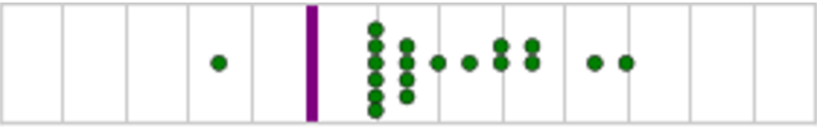
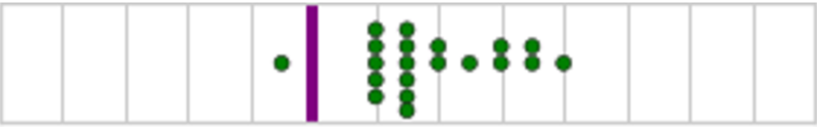
To help users see the spread of the vote data, we've included *swarmplot* visualizations.

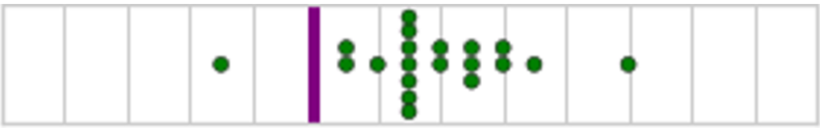
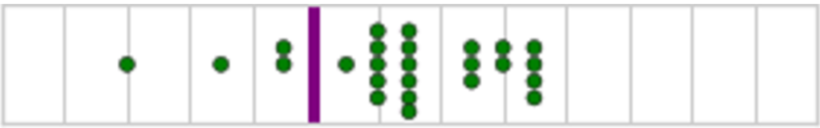
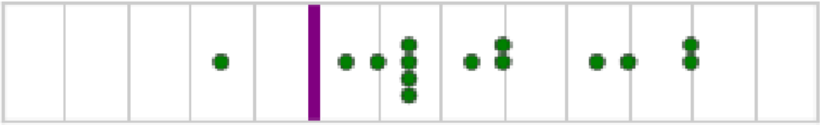
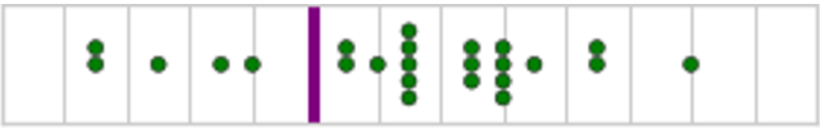
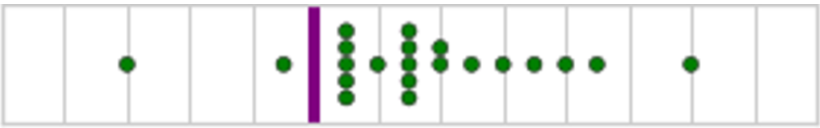
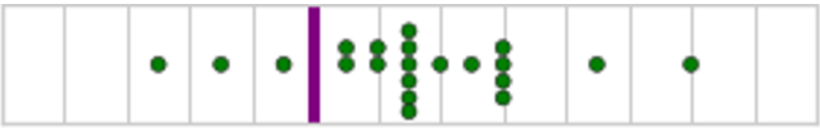
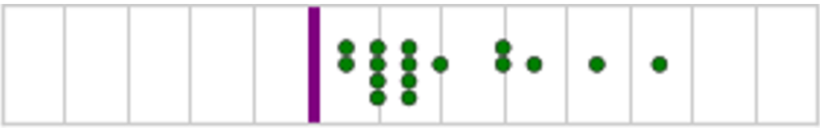
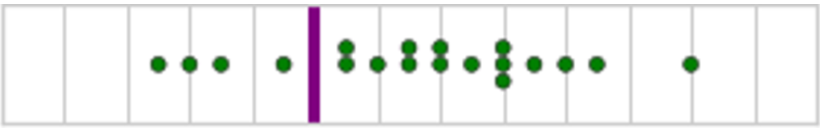
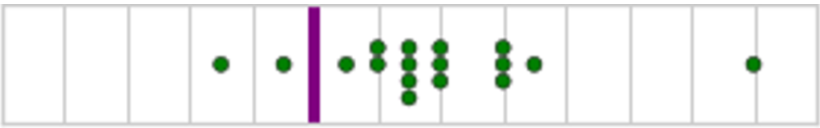
- For space reasons, only votes with weights between -10 and 16 are plotted. This covers 99.4% of votes.
- Gridlines are spaced 2 points apart.

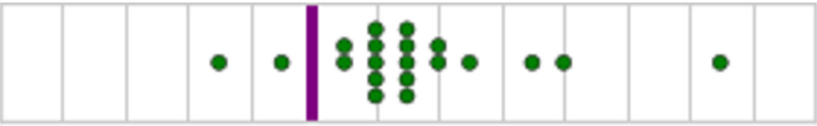
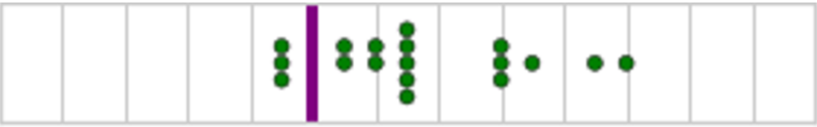
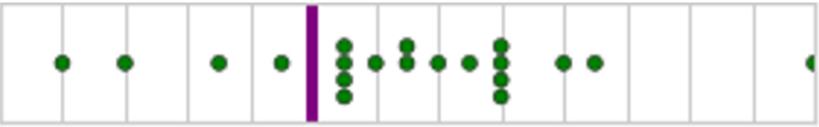
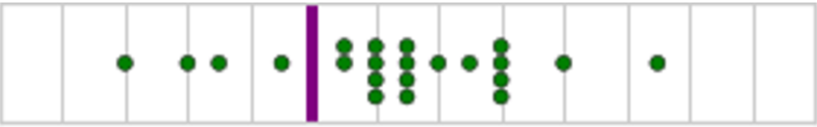
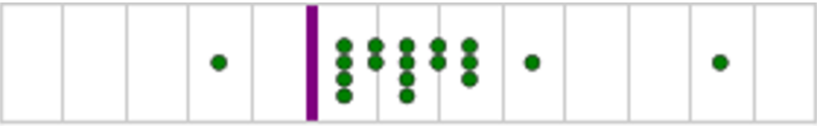
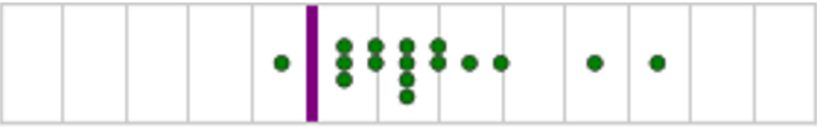
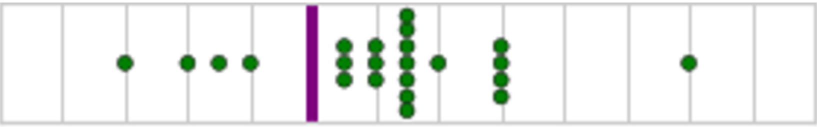
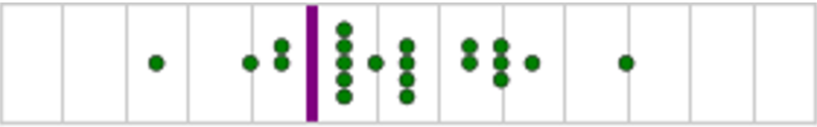
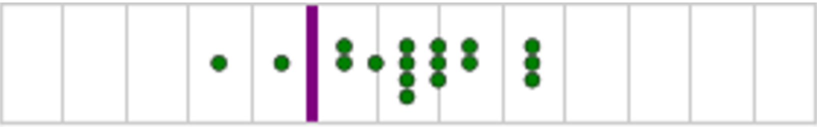
- Concrete illustration: The plot immediately below has 18 votes ranging in strength from -3 to 12.

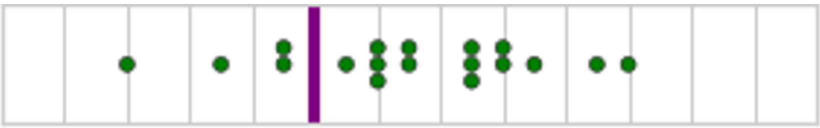
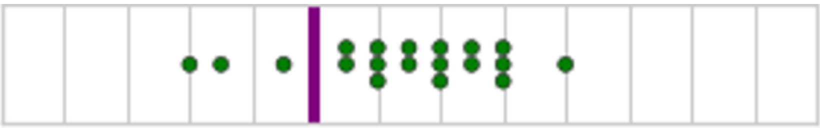
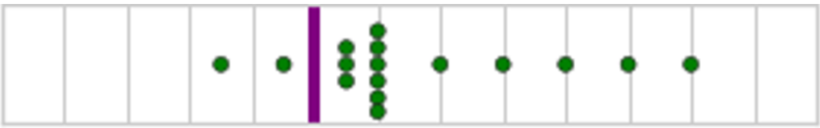
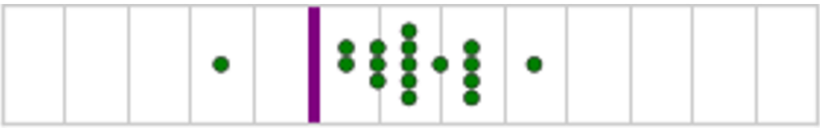
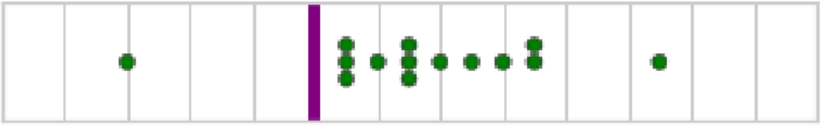
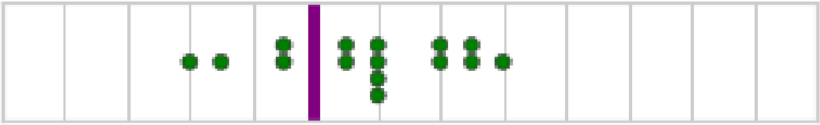
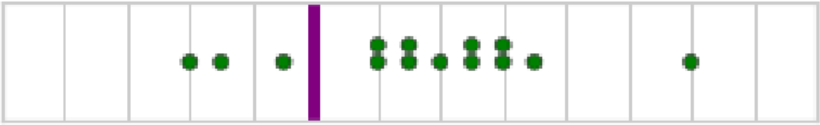
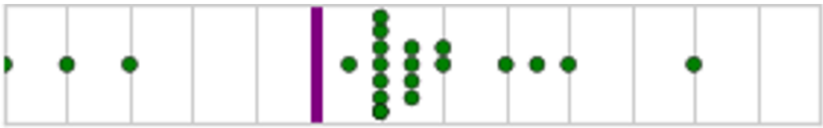
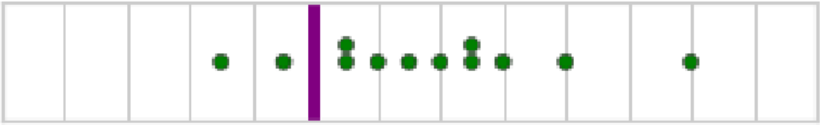


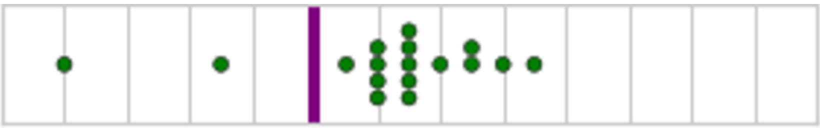
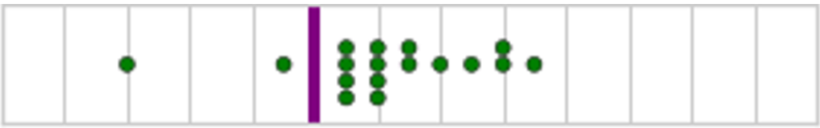
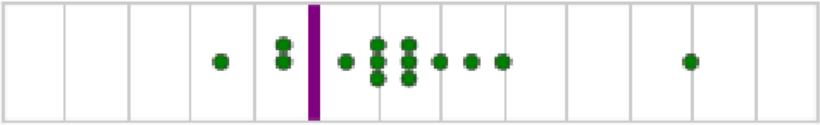
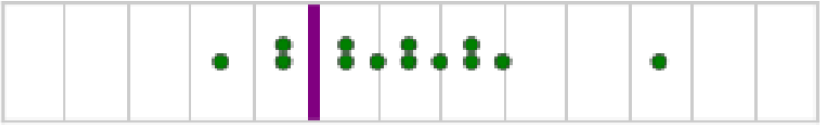
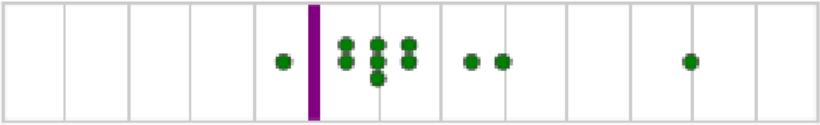
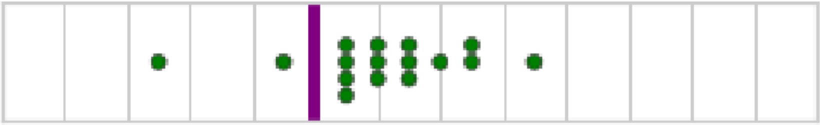
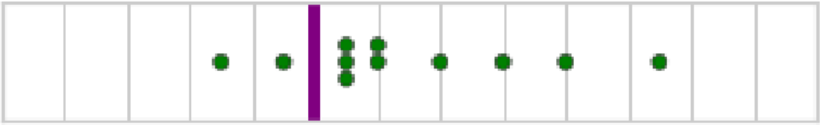
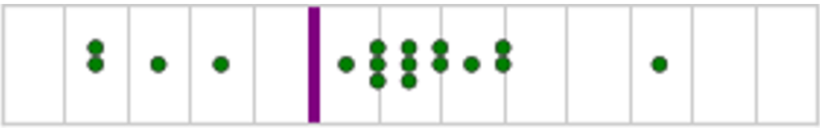
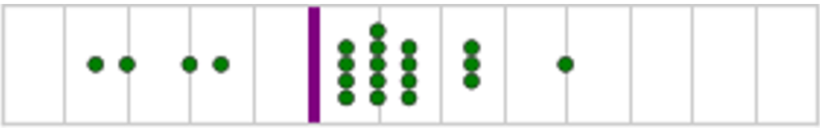
#	Post Title	Total	Vote Spread
1	Embedded Agents	209	 (One outlier vote of +17 is not shown)
2	The Rocket Alignment Problem	183	
3	Local Validity as a Key to Sanity and Civilization	133	
4	Arguments about fast takeoff	98	
5	The Costly Coordination Mechanism of Common Knowledge	95	
6	Toward a New Technical Explanation of Technical Explanation	91	

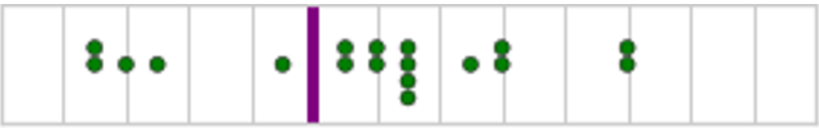
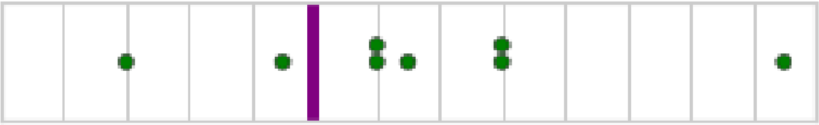
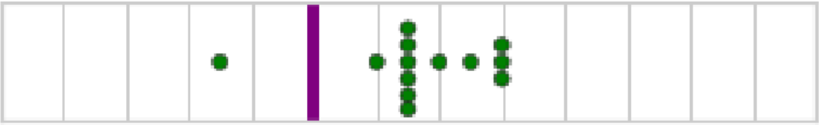
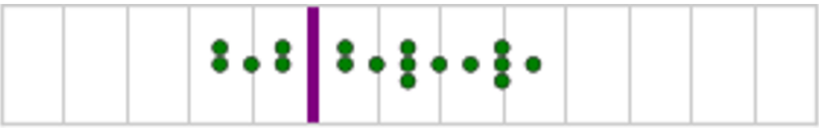
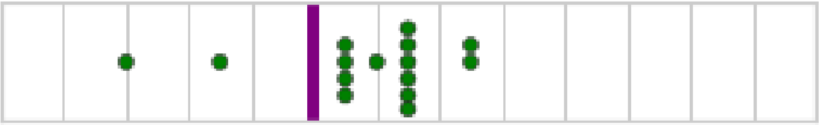
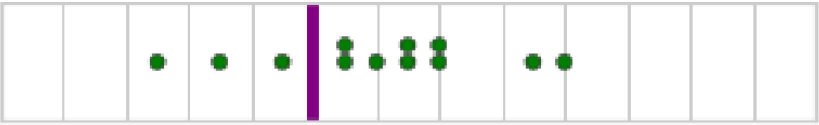
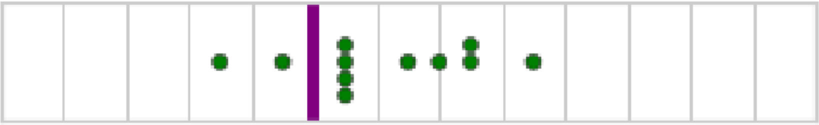
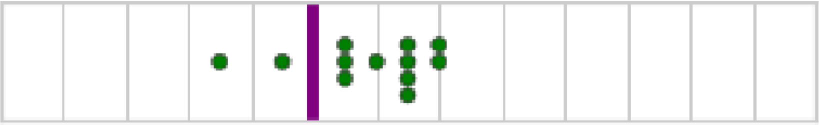
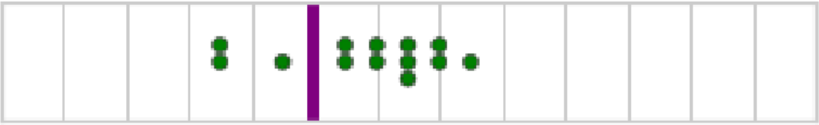
#	Post Title	Total	Vote Spread
7	Anti-social Punishment	90	 (One outlier vote of +20 is not shown)
8	The Tails Coming Apart As Metaphor For Life	89	
9	Babble	85	
10	The Loudest Alarm Is Probably False	84	
11	The Intelligent Social Web	79	
12	Prediction Markets: When Do They Work?	77	
13	Coherence arguments do not imply goal-directed behavior	76	
14	Is Science Slowing Down?	75	
15	Robustness to Scale	74	

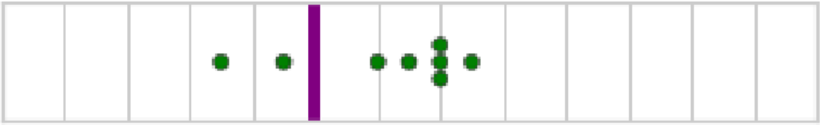
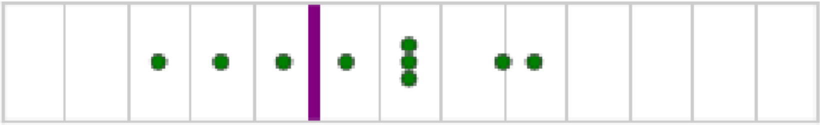
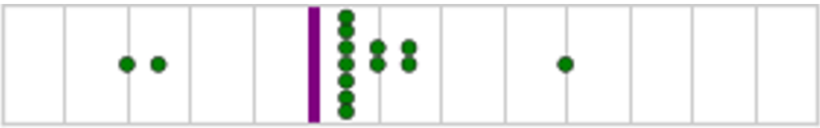
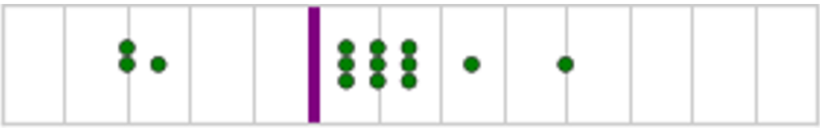
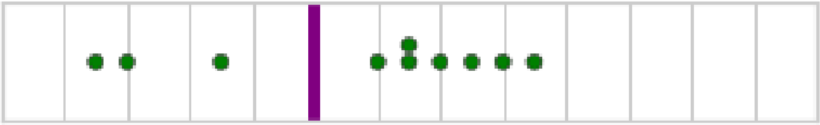
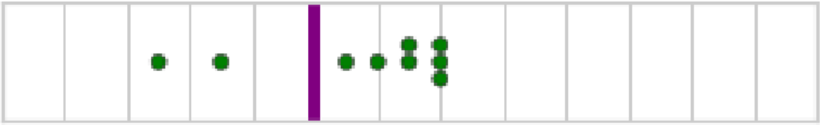
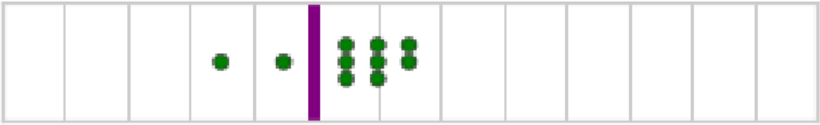
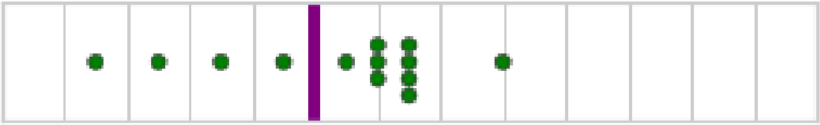
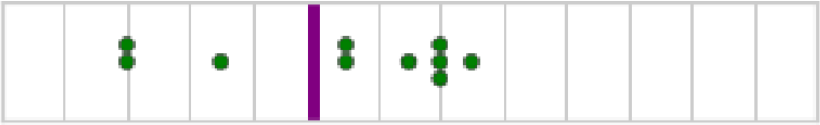
#	Post Title	Total	Vote Spread
15	A voting theory primer for rationalists	74	
17	Toolbox-thinking and Law-thinking	73	
18	A Sketch of Good Communication	72	
19	A LessWrong Crypto Autopsy	71	
20	Paul's research agenda FAQ	70	
21	Unrolling social metacognition: Three levels of meta are not enough.	69	
22	An Untrollable Mathematician Illustrated	65	
23	Specification gaming examples in AI	64	
23	Will AI See Sudden Progress?	64	

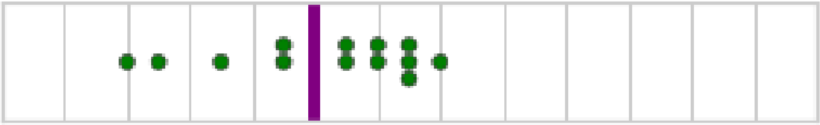
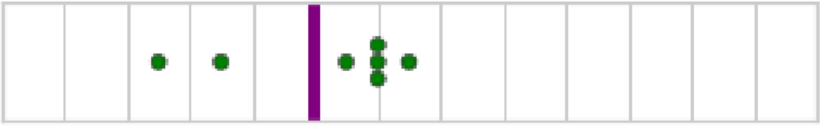
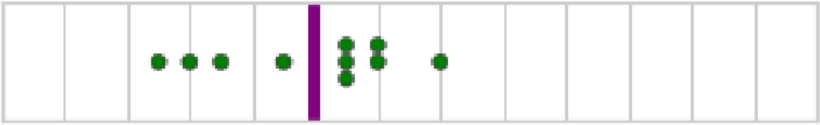
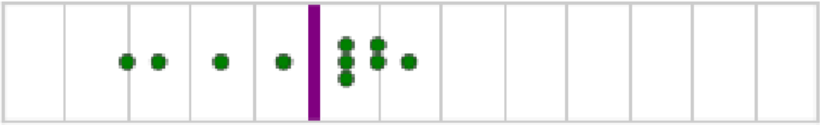
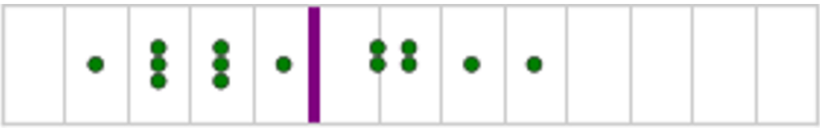
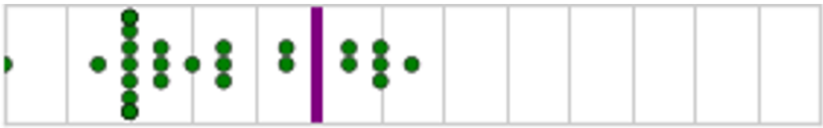
#	Post Title	Total	Vote Spread
23	Varieties Of Argumentative Experience	64	
26	Meta-Honesty: Firming Up Honesty Around Its Edge-Cases	62	
27	My attempt to explain Looking, insight meditation, and enlightenment in non-mysterious terms	60	
27	Naming the Nameless	60	
27	Inadequate Equilibria vs. Governance of the Commons	60	
30	2018 AI Alignment Literature Review and Charity Comparison	57	
31	Noticing the Taste of Lotus	55	
31	On Doing the Improbable	55	
31	The Pavlov Strategy	55	

#	Post Title	Total	Vote Spread
31	Being a Robust, Coherent Agent (V2)	55	
35	Spaghetti Towers	54	
36	Beyond Astronomical Waste	51	
36	Research: Rescuers during the Holocaust	51	
38	Open question: are minimal circuits daemon-free?	48	
38	Decoupling vs Contextualising Norms	48	 (One outlier vote of +23)
40	On the Loss and Preservation of Knowledge	47	
41	Is Clickbait Destroying Our General Intelligence?	46	
42	What makes people intellectually active?	43	

#	Post Title	Total	Vote Spread
43	Why everything might have taken so long	40	
44	Challenges to Christiano’s capability amplification proposal	39	
45	Public Positions and Private Guts	38	
46	Clarifying "AI Alignment"	36	
46	Expressive Vocabulary	36	
48	Bottle Caps Aren't Optimisers	34	
49	Argue Politics* With Your Best Friends	32	
50	Player vs. Character: A Two-Level Model of Ethics	30	
51	Conversational Cultures: Combat vs Nurture (V2)	29	

#	Post Title	Total	Vote Spread
51	Act of Charity	29	
53	Optimization Amplifies	27	
53	Circling	27	 (One outlier vote of -17)
55	Realism about rationality	25	 (Two outliers of -30 and +18)
55	Caring less	25	
57	Lessons from the Cold War on Information Hazards: Why Internal Communication is Critical	24	
57	The Bat and Ball Problem Revisited	24	
59	Argument, intuition, and recursion	21	
59	Unknown Knowns	21	

#	Post Title	Total	Vote Spread
61	**Competitive Markets as Distributed Backprop**	18	
62	**Towards a New Impact Measure**	14	
62	**Explicit and Implicit Communication**	14	
62	**On the Chatham House Rule**	14	
62	**Historical mathematicians exhibit a birth order effect too**	14	
66	**Everything I ever needed to know, I learned from World of Warcraft: Goodhart’s law**	13	
67	**The funnel of human experience**	11	
68	**Understanding is translation**	9	
69	**Preliminary thoughts on moral weight**	7	

#	Post Title	Total	Vote Spread
70	Metaphilosophical competence can't be disentangled from alignment	3	
71	Two types of mathematician	2	
72	How did academia ensure papers were correct in the early 20th Century?	-2	
73	Birth order effect found in Nobel Laureates in Physics	-5	
74	Give praise	-10	
75	Affordance Widths	-142	 (One outlier of -29)

How reliable is the output of this vote?

For most posts, between 10-20 people voted on them (median of 17). A change by 10-15 in a post's score is enough to move a post up or down around 10 positions within the rankings. This is equal to a few moderate strength votes from two or three people, or an exceedingly strong vote from a single strongly-feeling voter. This means that the system is somewhat noisy, though it seems to me very unlikely that posts at the very top could end up placed much differently.

The vote was also affected by two technical mistakes the team made:

1. *The post-order was not randomized.* For the first half of the voting period, the posts on the voting page appeared in order of number of nominations (least to most) instead of appearing randomly, thereby giving more visual attention to the first ~15 or so posts (these were posts with 2 nominations). Ruby looked into it and says that 15-30% more people cast votes on these earlier-appearing posts compared to those appearing elsewhere in the list. *Thanks to gjm for identifying this issue*.
2. *Users were given some free negative votes.* When calculating the cost of users' votes, we used a simple equation, but missed that it produced an off-by-one error for negative numbers. Essentially, users got a free 1-negative-vote-weight on all the posts to which they had voted on negatively. To correct for this, for those who had exceeded their budget - 18 users in total - we reduced the strength of their negative votes by a single unit, and for those who had not spent all their points their votes were unaffected. This didn't affect the rank-ordering very much, a few posts changed by 1 position, and a smaller number changed by 2-3 positions.

The effect size of these errors is not certain since it's hard to know how people would have voted counterfactually. My sense is that the effect is pretty small, and that the majority of noise in the system comes from elsewhere.

Finally, we discarded exactly one ballot, which spent 10,000 points on voting instead of the allotted 500. Had a user gone over by a small amount e.g. 1-50 points, we had planned to scale their votes down to fit the budget. However when someone's allocation was so extreme, we were honestly unsure what adjustment to their votes they would have wanted, as if their points had been normalised down to 500, the majority of their votes would have been adjusted to zero. (This decision was made without knowing the user who cast the ballot or which posts were affected.)

Overall, I think the vote is a good indicator to about 10 places within the rankings, but, for example, I wouldn't agonise over whether a post is at position #42 vs #43.

The Future

This has been the first LessWrong Annual Review. This project was started with the vision of creating a piece of infrastructure that would:

1. Create common knowledge about how the LessWrong community feels about various posts and topics and the progress we've made.
2. Improve our longterm incentives, feedback, and rewards for authors.
3. Help create a highly curated "Best of 2018" Sequence and Book.

The vote reveals much disagreement between LessWrongers. Every post has at least five positive votes and every post had at least one negative vote – except for An Untrollable Mathematician Illustrated by Abram Demski, which was evidently just too likeable – and many people had strongly different feelings about many posts. Many of these seem more interesting to me than the specific ranking of the given post.

In total, users wrote 207 nominations and 120 reviews, and many authors updated their posts with new thinking, or clearer explanations, showing that both readers and authors reflected a lot (and I think changed their mind a lot) during the review period. I think all of this is great, and like the idea of us having a Schelling time in the year for this sort of thinking.

Speaking for myself, this has been a fascinating and successful experiment - I've learned a lot. My thanks to Ray for pushing me and the rest of the team to actually do it this year, in a move-fast-and-break-things kind of way. The team will be conducting a *Review of the Review* where we take stock of what happened, discuss the value and costs of the Review process, and think about how to make the review process more effective and efficient in future years.

In the coming months, the LessWrong team will write further analyses of the vote data, award prizes to authors and reviewers, and use the vote to help design a sequence and a book of the best writing on LW from 2018.

I think it's awesome that we can do things like this, and I was honestly surprised by the level of community participation. Thanks to everyone who helped out in the LessWrong 2018 Review - everyone who nominated, reviewed, voted and wrote the posts.

Reviewing the Review

We just spent almost two months reviewing the best posts of 2018. It was a lot of development work, and many LW users put in a lot of work to review and vote on things.

We've begun work on the actual printed book, which'll be distributed at various conferences and events as well as shipped to the featured authors. I expect the finished product to influence the overall effect of the Review. But meanwhile, having completed the "review" part, I think there's enough information to start asking:

Was it worth it? Should we do it again? How should we do it differently?

Was it worth it? Should we do it again?

My short answer is "yes and yes." But I have some caveats and concerns.

My own goals for the Review were:

1. Actually identify the best posts
2. Improve longterm incentive structures and feedback systems
    1. Give users a reason to improve old content
3. Check our collective epistemic state on controversial posts
4. Figure out *how* to evaluate blogposts
5. Create a shared sense of ownership over LW's intellectual pipeline
6. Evaluate LessWrong as a site and project

Some of those (1, 4, 6) I feel able to evaluate independently, others depend on how other people felt about the process, and how much value they got for the effort they put in. It also depends on what the counterfactual actions the Review is being compared to.

But overall, I personally found the process very rewarding. Nominating, reviewing and voting gave me a clearer sense of how various ideas fit together, and what LessWrong had accomplished in 2018. This involved a fair effort on my part (several hours of re-reading, thinking, comparing), but it felt both enjoyable and worthwhile.

Identifying the best posts

I think the review did a decent job at this. The obvious comparison is "what were the top karma posts of 2018?". Could we have saved ourselves a ton of work by checking that? This is somewhat confounded by the fact that we changed our karma system partway through 2018 (reducing the power of the average upvote, after initially increasing it at the beginning of LW 2.0). At some point we plan to re-run the voting history using the current upvote strengths, which will give us clearer information there.

Meanwhile, comparing the top karma posts of 2018 to the top-voted posts in the review, there are some obvious differences. (Most obviously, "Arbital Postmortem" didn't end up featured in the review, while being the top-scoring post)

I think the simple act of filtering on *“Which posts had at least some people who had made use of them, and were willing to endorse them?”* was a key factor.

I think there was a lot of room to improve here. I felt that the voting process seemed, at least in some cases, more like “how prestigious should this post be?”, rather than giving me a clear sense of “how useful was this post?”.

Next year, I’d like to experiment with other voting processes that help disentangle “should this post be in a public-facing Best of LW book?” from “Was this post valuable?”, and “Did it reflect good intellectual practices?”.

Improving Longterm Incentives and Feedback

This is hardest to evaluate right now. But it’s also where I’m expecting most of the Review process’s value to lie. A decade from now, my bet is that LessWrong will have demonstrably better epistemics and value if we keep doing some form of a longterm, retrospective review process.

But for this year, we saw at least some outcomes in this space.

First, as a nominated author, I found it pretty rewarding to see some posts of mine getting discussed, and that some of them had had a longterm impact on people. I’m guessing that generalizes.

Second, at least 4 people I know of deliberately updated their post for the review. Two of those people were on the LW team, and one of them was me, so, well, I’m not going to count that too strongly. But meanwhile lots of authors gave self-reviews that reflected their updated thoughts.

Third, we saw a number of reviews that gave critical feedback, often exploring not just the ideas in the post but how the post should fit conceptually into an overall worldview.

Not all of those reviews were clearly valuable. But I think the clearest sign of counterfactually valid and valuable reviews were:

- Abram’s review of Rationality Realism (where a lot of latent disagreement came to light, followed by in depth discussion of that disagreement)
- A review by Bucky which looked into a cited paper in Zvi’s Unknown Unknowns. This was a particular type of intellectual labor that I was hoping to come out of the review process, which I expected to not happen much by default.

Checking out epistemic state on controversial posts

I think this had more room for improvement. The aforementioned Rationality Realism discussion was great, and I think there was at least some progress in, say, Vaniver’s post on Circling that acted as a sort-of-review for Unreal’s 2018 post.

I don’t have a strong sense that any debates were settled. But, well, I do think it takes an absurdly long time to resolve disagreements even when people are trying hard in good faith.

I think we did at least get a clearer sense of how controversial each post was, from the voting process, and that seems like a good starting place.

Figure Out *How* To Evaluate Blogposts

There were a diverse array of blogposts. Some of them benefited from conceptual, philosophical debate. Some of them benefited from statistical review of scientific papers.

A few of them had implied empirical claims, which would be pretty hard to check. I still hope that someone investigates more thoroughly El Tyrei's question about "Has there been a memetic collapse?", which looks into some of Eliezer's assumptions in Local Validity and Is Clickbait Destroying Our General Intelligence?. But, to be fair, it's a lot of work, it's confusing how to even go about it, and right now I don't think we've really offered good enough rewards for answering it thoroughly.

Overall, we got a fair number of people who worked on reviews, but a small number of people did most of the work. A couple people noted that reviewing felt like "work", and I think the strength of internet forums is making intellectual work feel like play.

I am uncertain how to take all of this into account for next year.

Shared sense of ownership over LW's intellectual pipeline

I don't have a clear sense of how this worked out. I know that participating in the review process increased my own sense of partial-ownership over the intellectual process, and I have some sense that the other people participated most heavily in the process felt something of that. But I'm not sure how it worked overall.

This goal was less *necessary* than some of the other goals but still seems useful for the longterm health of the site.

Evaluate LessWrong as a site

While engaging in the review process, I skimmed posts from 2017, and 2019 as well as digging significantly into the nominated posts from 2018. This gave me some sense of LW's overall output trajectory. This *doesn't* necessarily give us clear common knowledge of the community's *collective* epistemic state, but I found it at least useful for myself to form an opinion on "how is LW doing?"

One thing I found was that in 2017, there were relatively few unique authors that I was excited – many of the most exciting things were posts from Eliezer's Inadequate Equilibria sequence, which already had been made into a book, and then maybe… 5 other authors among posters I was particularly excited about?

In 2018, there was a much more diverse array of popular authors. Eliezer is present as one of the top contributors, but there were around 40 authors featured in the review, and even if you're just focusing on the top half that's a healthier base of authorship.

I think we have aways to go – 2018 was when LW 2.0 officially launched and it was still hitting its stride. My rough sense (partially informed by the metrics we've started tracking) is that 2019 was a slight improvement, and that things have started picking up in particular towards the end of 2019.

In the 2018 Review, my overall sense is that there were many "quite solid" posts on the subject of general rationality, and coordination. Meanwhile, I think a lot of very *concrete* progress was made on Alignment. It looks (from the outside) like the field went from a position of not really having any common language to communicate, to establishing several major paradigms of thought with clear introduction sequences.

Some people have expressed some sense that… Alignment posts don't quite feel like they count. Partly because the Alignment Forum is [sort of] a separate website, and partly because they're just a lot less accessible. I think it's admittedly frustrating to have a lot of high-context, technical content that's hard for the average user to participate in.

But, it still seems quite important, and I think it is sufficient to justify LessWrong 2.0's existence. Much of it is tightly interwoven with the study of rationality and agency. And much of it is precisely the sort of hard, confusing problem that LessWrong was founded to help solve.

I do hope for more accessible "rationality" oriented content to build momentum on LessWrong. I think some progress was made on that in 2019, and, well, we'll see next year hopefully how that looked in retrospective.

Problems with Execution

Too much stuff

There were 75 posts that made it into the review process. This seemed roughly the right amount of contenders, but more than people could easily think about at once. One way or another, we need to do a better job of directing people's attention.

Options I can see include:

- Have a higher nomination threshold, somehow aiming for closer to 50 posts. (for reference, this year, 23 posts had 3+ nominations rather than 2. I'm assuming more posts would have gotten 3 nominations if we had stated that explicitly as a requirement.
    - *I don't think it's reasonable to cull the initial pool to less than 50, especially if post volume grows each year.*
- Somehow culling the nomination pool some other way, partway through the process.
    - *I don't currently have good ideas on how to do this*
- Direct user's attention to posts they'd previously engaged with (i.e. views, upvotes, comments)
    - *I'm pretty confident we'll do this next year, but it doesn't seem sufficient*
- Direct people's attention to a smaller group of posts at a time. Maybe every few days, direct people's attention to a different set of posts
    - *This seems potentially promising, but 75 is still just a lot of posts to get through. It'd be 18 posts per week if the review period was a month, 9 if it were 2 months.*
- Radically restructure the thing somehow (perhaps doing rolling reviews every few months rather than a single all-encompassing one?)
    - *Rolling reviews feel… less exciting somehow. But I could imagine this turning out to be right approach.*
- Randomly assigning each user a smaller number of posts to focus on. (Perhaps each user gets 5 or 10 posts they're supposed to evaluate, and if they've evaluated those ones, they a) get to have their votes counted in the larger tally, b) they are then welcome to review other posts that they're excited about)

A lot of the options feel like good ideas, but insufficient. But maybe if you combine them together you get something workable.

Voting vs Reviewing

This year, we initially planned to separate out the reviewing stage and the voting stage. Ben and I ended up deciding to have the Voting Phase overlap the end of the Review Phase. I can't remember all the reasons for this but they included "we still wanted more reviews overall, and we expected people who show up to do voting to end up doing some reviews along the way", and "some people might want to update their votes in response to views, and vice versa."

I think it's plausible that next year it might be better to just have the Vote and Review phases completely overlapping. (In particular if we end up doing something like "assign each user 10 random posts to evaluate." I imagine it being a fairly hard task to "do a serious review of 10 posts", but to be fairly achievable to "think about 10 posts enough to cast some votes on them, and if you end up writing reviews in the meanwhile that'd be great."

Voting vs Survey

As I mentioned earlier: A worry I have with our voting system this year is that it felt more to me like "ranking posts by prestige" than "ranking them by truth/usefulness." It so happens that "prestige in LW community" does prioritize truth/usefulness and I think the outcome mostly tracked that, but I think we can do better.

I'm worried because:

- I'd expected, by default, for these things to be jumbled together.
- Whatever you thought of Affordance Widths, it seems unlikely for it's "-140" score to be based on the merits of the post, rather than people not wanting the author represented in a Best of LW book. (This isn't obviously wrong: reputational effects matter and it's fine to give people an outlet for that, but I think it's better to ask those questions separately from questions of truth, and usefulness. It currently seems to me that if an unpopular author wrote an obviously important piece, the review wouldn't be able to determine that)

I also noticed conflict between *"which posts really make sense to showcase to the rest of the world?"* and *"which posts do we want to reward for important internal updates, that were either very technical, or part of high context ongoing conversations."*

So I'd like to frame it more as a survey next year, with questions like:

- Have you thought about the ideas in this post in the past year?
- Do the ideas in this post seem important?
- Does this post demonstrate good epistemics?
- Should this post appear in a public-facing Best of LW book?
- Should this post appear in an inward-facing, high-context LW Journal?

(With options for "yes/no/neutral", perhaps with Strong No and Strong Yes)

The main reason *not* to do this is that in many cases the answers may be similar, enough that they feel annoyingly redundant rather than helpfully nuanced. But I currently lean towards "try it at least once." I'm hoping this would prompt users to give honest answers rather than trying to strategically vote.

If we ended up combining the Review and Voting Phases, this might come along with text-boxes for Reviews. Possibly just one catch-all textbox. Or, possibly broken up into multiple freeform questions, such as:

- How has this post been helpful to you?
- What advice do you have to the author to improve this post?
- What thoughts do you have for other voters to help them evaluate this post?
- What further work or questions would you like to see done in this post?

Nomination Phase was a bit confusing

I originally hoped the nominations phase would include detailed information on why the posts were good. Instead most nominations ended up fairly brief, and people gave more nuanced positive thoughts in the review phase.

I think in many cases it was useful for nominations to include reasons why, to give other nominators some context for why they might consider seconding the nomination. But in some cases that just felt superfluous.

At the very least, next year I'd include a more prominent "endorse nomination" button, that makes it lower effort for subsequent nominators to increase the nomination count. It's possible that including reasons for nominations isn't necessary, and we can handle that as part of the Review step.

Alignment Review?

It did seem like the Alignment Forum would have benefited from having a somewhat different review process. My current guess is that next year, there would be a LessWrong Review, and an Alignment Review, and some of the content is overlapping but they're optimized separately.

It's possible the Alignment Review might include things *not* published on the Alignment Forum (and, in fact, it'd be a fine outcome if it were concluded that the most important Alignment progress happened elsewhere). In the months leading up to the review, AF users might be encouraged to make link posts for 2019 Alignment Content that they think was particularly important.

Further Feedback?

Did you have your own take on the review process? Were there particular problems with execution? Having seen how the review process fit together, do you have overall concerns about the approach or the ontology?

I'm interested in feedback on whatever levels stand out to you.