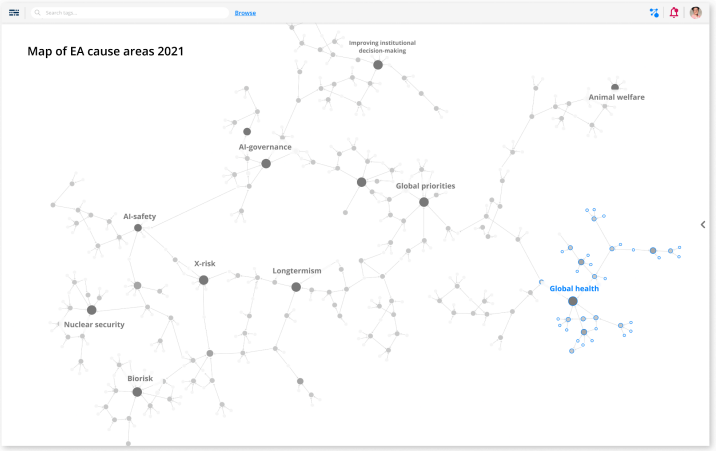
Collaborative Bayesian Networks

Last updated 04. Jan 2022

Summary

In this text, I review ideas and initiatives towards collaborative bayesian networks and present the case for my interpretation of the idea. I argue that we should build and sustain a Bayesian Knowledge Graph representing the wisdom of the EA crowd on promising cause areas. I also argue that crowdsourcing is feasible if contributors make judgments about a subject while they are reading about it anyways.



This document is very much a work in progress, and I expect the ideas to evolve and improve as new considerations demonstrate weaknesses and point in the direction of refinements. Read on with this caveat in mind, and please share any critical feedback.

Introduction

In a range of domains, such as law, geopolitical forecasting and investing, expert judgments vary a lot, even on similar and identical cases.[1] This implies that individual expert judgment is often inaccurate, and therefore unreliable.

Research in the decision sciences has uncovered several effective means by which individuals and organisations can improve the accuracy of judgments.[2] A key finding is that individuals who reason in a Bayesian manner tend to make more accurate judgments.[3] Another important finding is that the wisdom of crowds, when certain conditions are met, reliably outperform expert

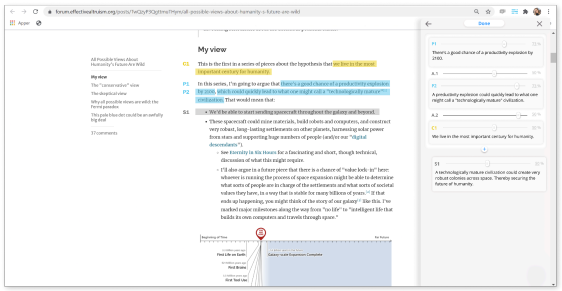
[1] See Philip Tetlock's *Expert Political Judgment* (2006) for a detailed study on the accuracy of geopolitical forecasting. For a comprehensive up-to-date overview of research in other domains, including law, medicine, recruitment, finance, and many others, see Kahneman, Sibony and Sunstein's new book *Noise* (2021, part I, chapters 1-3).

[2] Some good books on this are Daniel Kahneman's *Thinking fast and slow* (2011), Douglas Hubbard's *How to measure anything* (2014), Philip Tetlocks's *Superforecasting* (2015), Sperber and Marcier's *The Enigma of Reason* (2019) and Kahneman, Sabiny and Sunstein's *Noise* (2021, part V, chapters 18-22). See Ten Commandments for Aspiring Superforecasters - Good Judgment for key takeaways from Tetlock's research, and also Evidence on good forecasting practices from the Good Judgment Project – AI Impacts for a review of the evidence for these practices.. See How to turn down the noise that mars our decision-making for a review of the book *Noise,* and How to Measure Anything - LessWrong for a review of the book *How to Measure Anything* (Muehlhauser 2013).

[3] Here is a good and accessible intro to Bayesian thinking Bayesian Mindset (Karnofsky, 2021), which also has an audio version.

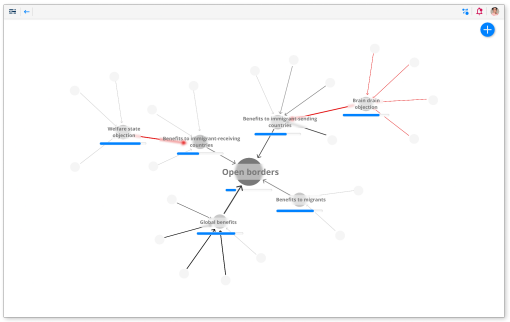
judgement in many domains.[4] In domains where it is unclear who the experts, or superforecasters are, the wisdom of crowds can be a good epistemic asset in decision making.

These results have inspired many ideas for improving decision making in EA.[5] In what follows, I review several such ideas and initiatives, and draw lessons from them.[6] I then present and defend two new ideas for crowdsourcing the wisdom of the EA crowd, and representing their knowledge in a way that is useful to decision makers in EA.



The first idea is to elicit judgments through a browser extension that lets users express beliefs about what they are reading quantitatively, while they are reading. Embedding elicitation in existing research practices like this may open estimation practises to a wider audience and is therefore a natural extension for elicitation systems targeting wider communities.

The second idea is to represent knowledge in a graph that is specifically designed to aid the sort of reasoning supported by the epistemic norms of the EA community. This is a graph in which nodes and edges have epistemic credences, and represent claims, and inferences respectively. Nodes represent priors, and new additions to the graph automatically update connected nodes through bayesian inference relations.[7]



The two ideas for elicitation and representation are separate, and can be evaluated as such. However, when combined, they have some exciting synergies. Notably, the conjunction has the potential to elicit and represent the knowledge and beliefs of the community as a whole, quantitatively. This interactive representation is continually updated as contributors research and assess ideas. Here is a prototype of how a simple graph might look like.

Table of contents

[4] See The Promise of Prediction Markets. For a compelling study of the effects of crowdsourcing in the domain of law, see (Katz et. al. 2017). See also chapter 21 of Noise (Kahneman et. al. 2021), and Corporate Prediction Markets: Evidence from Google, Ford, and Firm X * | The Review of Economic Studies | Oxford Academic (Cowgill & Zitzewitz 2015).

[5] For instance, at EAG 2021 Will MacAskill has suggested setting up a forecasting organisation. Ozzie Gooen describes several ideas in a forum post about ambitious altruistic software engineering efforts.

[6] Notably Arbital, Guesstimate, Metaculus and Roam Research, but also many others.

[7] The closest analogues I know to a Bayesian Knowledge Graph would be Guesstimate and Causal, which do the same thing, but not through a graph.

In the first chapter, I review ideas and initiatives towards improved epistemics in EA. I focus on epistemic norms, and systems for expert elicitation and knowledge representation. In these discussions I present two new ideas. A system for eliciting judgments while reading, and a system for representing beliefs graphically in a bayesian network.

In the second chapter I describe a system that combines these ideas. I then argue that this system would be a valuable addition to the EA knowledge infrastructure, with the potential to significantly promote community engagement, research, and decision-making. The first section explains how it could work through illustrative user stories, and design sketches. In a section on epistemology, I argue that the system provides valuable insights to decision makers. A collaborative bayesian network is different from other prediction platforms in that it represents inferencial relations between claims. I argue that this feature effectively mitigates the central epistemic problems of relying on aggregations of judgement. In a section on feasibility, I argue that the system incentivises contributors, and fosters community engagement by accentuating the originator of contributions, and by making the value of the contribution salient.

The third chapter is an impact assessment analysis of a project to develop a collaborative bayesian network using the importance, tractability, neglectedness (ITN) framework. The first section argues that work on new ways to represent and interact with ideas is neglected, but potentially very impactful. Moreover, in this section I also review the existing knowledge infrastructure in EA as well as some other institutions. I argue that a collaborative bayesian network is a novel contribution that would not compete with, but rather complement existing infrastructure. In the second section, I consider the tractability of a successful project to build, grow and sustain such a network. I outline a plan for building and growing it, including a listing of costs, technological risks and uncertainties. In the third section, I list potential risks and benefits, including beneficial externalities of the project itself. In the fourth and final section, I summarise everything, and add the elements of the impact assessment together.

Background

Information flows are essential to the effectiveness of groups. EA is a group in which information flows are especially important. If the people and organisations of EA are to find the most effective ways to do good, and direct resources towards actually doing it, the community as a whole needs a well-functioning knowledge infrastructure to support decision making. EA is surely one of the institutions whose decision making we should try to improve.[8]

[8] See Jess Whittlestone's introduction to the idea and rationale behind improving institutional decision making. See Stefan Torges' excellent The case for building more and better epistemic institutions in the effective altruism community - EA Forum for the case of improving decision making in EA through better knowledge infrastructure. 'Better communication channels across domains' is one of several activities recommended for improving institutional decision-making in a Founders Pledge report on the topic Longtermist institutional reform | Founders Pledge (Goth and Lerner 2021).

The knowledge infrastructure of EA

I use the word knowledge infrastructure to refer to systems and processes of knowledge production, transmission and distribution within a group. The value of knowledge in general is to support decision making, and so the function of a knowledge infrastructure is to support the decision making of the members of the group. Therefore, we should chiefly assess existing parts and potential additions to the EA knowledge infrastructure on the basis of shared norms[9] for good decision making.[10] It might be worthwhile to implement an idea to the knowledge infrastructure if it can be said to plausibly improve decision making in the group.

Philip Tetlock's research into superforecasters[11] indicates that reliable individuals and groups employ rigid epistemic practises, as opposed to intuition and heuristics. Five epistemic practises stand out.

- **Analysis**: Competent forecasters break down vague and ambiguous questions and claims into concrete parts for which sensible answers can be found.
- **Research**: A common denominator to all good superforecasters is that they spend a lot of time thinking and researching issues before they give their estimates. Doing good forecasting about an issue is a full-time job, and requires a structured and careful research method. For example, most forecasters will look to the past to identify relevant baserates first, and use the base rate as an anchor which is updated on the basis of dissimilarities with the present case.
- **Numeracy**: Most saliently perhaps, good forecasters place numerical odds on the probabilities of events occurring, or claims being true. Also, they update existing beliefs on the basis of new information in a Bayesian manner.
- **Cooperation**: When superforecasters band together to critically appraise each other's forecasts and update on the basis of discussion, they tend to do even better than when working alone.
- **System**: Supeforecasers tend to have a system for clarifying ideas, directing research, noting estimates, and updating them. Some use computer programs that they have devised on their own, and others use structured notebooks.

In the EA community, many members have taken one or more of these epistemological virtues to heart. Discussions in EA tend to employ a sophisticated vocabulary that is the result of careful analysis of central issues, and most EA's are careful to define the terms they use. Ideas expressed in EA-literature are typically precise and understandable. They also tend to be well researched. Many community members are researchers, and non-researchers who express opinions on the forum or in live discussions typically read widely on the relevant topics, and undergird their

[9] I don't explicate these norms, but rather state specific norms here and there, and leave it to the reader to assess whether my interpretation of the epistemic norms in EA is accurate. A detailed analysis would be an essay of it's own.
[10] For another approach, see Improving Institutional Decision-Making: Which Institutions? (A Framework) - EA Forum.
[11] See for instance this brief recent article from the Economist: How spooks are turning to superforecasting in the Cosmic Bazaar, or Tetlock's books *Expert Political Judgement*, and *Superforecasting*.

claims with citations to scientific studies. The credence authors assign to such claims are often expressed by the assignment of quantified estimates, saying precisely how likely the person takes it that the claim is true in numerical terms. It requires great courage to express one's ideas clearly and boldly through precise concepts and quantified estimates, as this makes it much easier for others to point out mistakes. However, EA's typically welcome corrections and feedback. It is not uncommon for EA's to publish early iterations of a writeup on their idea in order to draw feedback to inform further thinking. They see their own projects as part of a joint effort to find truth and knowledge, and the rest of the community as part of that grand epistemic enterprise.

However, even though EA's exemplify these virtues to a larger extent than most other groups, there is room for improvement, especially when the community is considered on a group level. Even though individuals tend to define terms when introducing new ideas in their own writings (especially researchers), the meanings attached to the same words can still fluctuate quite a bit in the writings of different people in the community. The word 'longtermism', for instance, can be taken to mean different things,[12] with different implications. Although EA literature draws on a rich base of research, and there are resource directories for many topics, there is still a lot of work to be done to summarize and organize relevant research in many cause areas. There are also some challenges to the assignment of numerical probability estimates in some cases. In cases of tail risks, for instance, where there are few or no baserates, it can be quite challenging to distill precise values from research with any confidence. The issue comes up, for instance, in the 80k episode with Ajeya Cotra, and is a recurrent theme in forecasting, and discussions about formal epistemology generally. There is also some reason to think that the EA-community could collaborate more effectively.

Staff at EA orgs do a lot of work to address these issues. For instance, the cause area profiles at 80k is a beacon to many when starting to research some EA related idea. The engaged participants in the wider EA information ecosystem with the forums,[13] conferences,[14] blogs,[15] podcasts,[16] and research also continuously does great work to coordinate ideas, people, and disseminate the knowledge contained in the community at large.

However, the EA community consists of many competent and knowledgeable persons, and so many EA's feel that there is still a lot of untapped knowledge contained within the minds of community members. The brunt of epistemic competence in an organization tends to reside in the heads of everyone involved.[17] Moreover, many EA's would like to contribute in some way, and

---

[12] See for instance the meanings distinguished in this paper: The case for strong longtermism - June 2021 update - EA Forum.
[13] Most saliently EA Forum and LessWrong
[14] EA Global and EAGx
[15] Eg. Bryan Tomasik's Essays on Reducing Suffering, Robin Hanson's Overcoming Bias, Gwern Branwen's Essays · Gwern.net, or Paul Christiano's The sideways view – Looking askance at reality.
[16] Eg. Home The 80,000 Hours Podcast with Rob Wiblin, Hear This Idea: About, and The FLI Podcast. See also [Crowdfunding] LessWrong podcast.
[17] In his book *Principles*, Ray Dalio describes how he built the successful investment company Bridgewater. He attributes the success of the company to the competence of the people employed, and a set of crowdsourcing methods designed to elicit radically honest judgments (using a Delphi-style method). Also, he emphasises the value of a quantitative approach involving the creation of computer models on the basis of principles for good decision making.

especially in a way such that they can use the expertise that is particular to them, and showcase their skills.

These reasons, and more besides, motivate a range of crowdsourcing projects to elicit the wisdom of the EA crowd. The overarching idea is that the most valuable asset in EA is the community of competent and highly engaged people involved, and that the most valuable projects in EA leverage this community to draw on the latent epistemic values retained by members. Additionally, the idea is that community members wish to help, and that if we can find ways for community members to contribute in a collaborative and meaningful way that matters to the decision-making of others, this might be a key driver of active engagement and growth.

Two strands of epistemic crowdsourcing projects aim to draw on the knowledge of the community to (1) organise and represent key EA ideas, and (2) elicit forecasts from community members on these ideas. In what follows, I briefly describe and discuss some approaches to these ends.

Eliciting the wisdom of crowds

As I mentioned in the introduction, scientific research indicates that aggregating the wisdom of crowds, and considering this aggregate in deliberation is a good way to improve the accuracy of judgments. So much so that Philip Tetlock, in an interview with Alexander Berger, *"... estimates that algorithmic aggregating of the predictions of a large group (e.g. 300) of typical, good-judgment forecasters can produce results nearly as accurate as a small group (e.g. 10) of superforecasters."* (2016: p3).

EA orgs fund prizes for Metaculus forecasting calls on EA-related question-clusters.[18] Moreover, Metaculus has received EA support in several rounds, and so have other prediction and aggregation platforms like Elicit, Guesstimate and Foretold.io.[19] Other prediction and aggregation ideas and platforms include S-process, Good Judgment® Open, PredictionBook, Empiricast, and prediction markets like Manifold Markets, PredictIt and Kalshi.

Michael Aird has collected quantitative estimates from EA researchers on key claims concerning existential risks in a database. In his Database of existential risk estimates - EA Forum (Aird 2019), he notes that surveys of EA researchers on controversial issues are frequently cited, and has probably played an important role in the deliberations of many EA's.[20] QURI has extended

[18] Such as Forecasting AI Progress | Metaculus and Nuclear Risk Tournament | Metaculus.
[19] See Habryka's writeup for the Long-Term Future Fund: April 2019 grant recommendations - EA Forum. EA funds has also supported other prediction platforms, including grants to Jacob Lagerros and Ozzie Gooen in the same round (see especially the writeup to Gooen's project for details). Elicit is built by Ought which is mostly funded by OpenPhil.
[20] He also explains why it might be valuable to collect estimates of this sort, and have them readily available in an easy to access database, and anticipates and rebuts a series of objections (see original post, and also this post).

this idea, and developed Metaforecast,[21] which collects quantitative judgments on issues from several forecasting platforms.[22]

There are several epistemic considerations relevant to making precise public judgments, and aggregating judgments of that sort. Here is a list of considerations I, and others,[23] think are important:

- **Accountability.** Contributors who make judgments can be held accountable for their beliefs. This is a chief part of reasoning transparency, and makes research a more high-stakes activity, which carries a suite of beneficial cognitive and epistemic benefits.
- **Training**. Contributors who make judgments get good forecasting practice.[24]
- **Accuracy**. The aggregate judgment of a group of competent individuals tends to be more accurate than individual expert judgments.[25]
- **Value of information**. In addition to accuracy, aggregation of beliefs could indicate areas of consensus and disagreement, and over time, show the progression of belief on the basis of new arguments and new findings.
    - **Value of research questions.** Knowledge concerning where there is agreement, disagreement and varying degrees of confidence can be useful for identifying which issues deserve more research, and conversely, which questions have been researched enough, because everyone already agrees on them.
    - **Value of research results**. The way we interpret research results depends on what we already believe, and this is especially true for evaluations of the value of research findings. Hindsight bias as well as bias against publishing insignificant, or "null", results systematically distorts the value of research. A mapping of belief could be used to systematically track whether new research actually changes anyone's minds, and this could then be a measure for the value of information.[26]
- **Collaboration**. Good systems for elicitation and aggregation are a great way for EAs to contribute in a way that is useful to the community, while also demonstrating their knowledge and expertise. Insights into the beliefs of community members, and also the community as a whole can be useful for finding collaborators on projects.
    - **Contribution**. Distinct way for contributors to help out, in addition to demonstrating knowledge and expertise.[27]

---

[21] See Introducing Metaforecast: A Forecast Aggregator and Search Tool - EA Forum.
[22] See also https://beliefelicitation.github.io/paper/ for more on elicitation tools.
[23] See Pathways to impact for forecasting and evaluation - EA Forum created by Nuno Sempere (2021).
[24] In the FLI podcast episode Transcript: The Art of Predicting, Anthony Aguirre, a founder of Metaculus, lists several key benefits of forecasting, including the benefit of training (Aguirre 2017: 14-16 min marks).
[25] See introduction, especially footnotes 2 and 4. See also (Aguirre 2017: 14-16 min marks). See also
[26] This is the main idea in Stefano Dellavigna, Devin Pope and Eva Vivalt's article Predict science to improve science (2019), where they explain how this could work in the domain of social science specifically. See also Vivalt's Predicting research results can mean better science and better advice (Vivalt 2019) and Eva Vivalt: Forecasting research results - EA Forum (Vivalt, linkpost for presentation from 2019) for summaries. However, the idea of measuring the value of information through a process of Bayesian updating is quite general and works in all other domains as well.
[27] See Can the EA community copy Teach for America? (Looking for Task Y). See also this May 2021 grant rationale by Max Daniel: "Contributing to a wiki is a concrete way to add value and contribute to the community that is accessible to basically all community members. My impression is that opportunities like this are in significant demand, and currently severely undersupplied by the community. If the project goes

    - **Recruitment**. Distinct way for individuals and orgs to find collaborators on projects.[28]

These considerations are, I think, quite compelling. In fact, the case is so good that several high-profile companies have experimented with prediction markets, including Google, Microsoft and others.[29] However, most companies that tried experimenting with prediction markets stopped.[30] Why? In their Prediction Markets in The Corporate Setting - EA Forum, Nuno Sempere, Misha Yagudin and Eli Lifland (2021) review attempts to use prediction markets in a corporate setting, and explore several plausible factors to an explanation why they stopped, and might not be a good idea.[31] The main ones are:

- **Feasibility**
    - The markets must have a low enough cost to create and maintain. That is, setting the system up must not be too costly, and it must not steal too much time to engage with it.[32]
- **Value to decision makers**
    - The markets must provide more value to decision-makers than the cost to create them and to subsidize predictions on them.
- **Value to contributors**
    - The markets must be attractive enough to traders to elicit accurate predictions. There are several challenges with this for standard prediction markets, and some of these problems are aggregated for EA relevant questions. A central problem here is that questions are hard to resolve.[33]
- **Other effects**
    - The markets must not have large negative side-effects, such as costs to the company's dynamics and morale.

Eli Lifland also notes several relevant considerations in his Bottlenecks to more impactful crowd forecasting - EA Forum (2021). The three bottlenecks are;

- Creating the important questions
- Incentivizing time spent on important questions
- Incentivizing forecasters to collaborate

---

well, many students, researchers, and professionals might contribute to the wiki in their spare time, and find the experience motivating and satisfying."

[28] (Aguirre 2017: 14-16 min marks). See also (Sempere 2021).

[29] See Prediction Markets in The Corporate Setting - EA Forum.

[30] Prediction Markets in The Corporate Setting - EA Forum.

[31] See also Tyler Cowen's post with 7 answers to the same question.

[32] In Oliver Habrykas writeup for a grant to Ozzie Gooen for Foretold, he similarly writes: *The biggest concerns I have with Ozzie's work, as well as the work on other prediction and aggregation platforms, is that the problem of getting people to actually use the product turns out to be very hard. Matt Fallshaw's team at Trike Apps built https://predictionbook.com/, but then found it hard to get people to actually use it. Ozzie's last project, Guesstimate, seemed quite well-executed, but similarly faltered due to low user numbers and a lack of interest from potential customers in industry. As such, I think it's important not to underestimate the difficulty of making the product good enough that people actually use it.*

[33] However, there are some ideas for how to mitigate this. See the paper Improving Judgments of Existential Risk: Better Forecasts, Questions, Explanations, Policies by Ezra Karger, Pavel D. Atanasov, Philip Tetlock :: SSRN, and also this post expressing the related idea of meta-resolution.

In each of the explanatory factors uncovered by Sempere, Yagudin and Lifland (2021), and also of the bottlenecks identified by Lifland (2021), there is much nuance and many different considerations at play. I won't go into these here, but would strongly advise the attentive reader to read them carefully, as these authors have much more experience with forecasting than I do.[34]

I now want to delve deeper into one consideration which I think is crucial in explaining why crowdsourcing isn't more common. It has to do with the context of judgement.

Contexts of judgement

The way I see it, the main reason why more people don't use prediction platforms is that plotting judgments in a platform isn't part of typical existing habits and workflows.[35] Potential users would have to develop entirely new habits and ways to approach issues. One must have very strong motivations to enter a forecasting platform, find an interesting question, and then research that question in order to plot a sensible judgement. This process can be time-consuming and laborious, and so the bar is quite high. Researchers and professional forecasters can allow themselves to research some question or set of questions for the sole purpose of finding a reasonable estimate, because that is part of their job which they are paid to do. However, most people don't have the time or capacity to do this. It is generally very hard to make people read and do research on other issues than the topics they are already interested in. The interested laymen who might use their free time to read and research issues of interest in EA would most likely want to choose what and when to read themselves. And if they at some point would be open to sharing judgments about an issue, the time for sharing would most likely be right after reading about it. Most people can't be expected to frequent prediction platforms, scouting for a question they might have a view on, and then researching it some more to build confidence.

This does not mean that we should give up the idea of crowdsourcing forecasts from the wider EA community. Most EA's read and research many of the same topics, which are typically the topics of interest to decision makers. Moreover, in some contexts, EA's are very keen on sharing their views on such topics. For instance, right after a seminar presentation many are eager to say what they think about the idea expressed. Also, people who are in the process of writing about a particular issue, tend to want to express their views on that issue in particular. Similarly, the time when I get messages from people who want to share their views on some EA related topic, is typically immediately after they have read a post or book about it. For analytical purposes, these considerations can be boiled down to the following points:

- **Timing**. It is more feasible to elicit judgments on an issue at the exact time when someone is already thinking about it.
- **Place**. It is more feasible to elicit judgments on an issue where they are already engaging with it.
- **Interest**. It is more feasible to elicit judgments on an issue from someone if they are already interested in that exact issue.

[34] Also, see Prediction-Driven Collaborative Reasoning Systems - LessWrong.
[35] See Prediction Markets in The Corporate Setting - EA Forum for a similar, and informative discussion on prediction markets in particular. The authors also emphasize the way in which existing prediction markets and forecasting platforms are hard to use. An illustrative case study for this, is the success of Robinhood, which introduces millions to day-trading through superior UX.

A similar line of reasoning is, perhaps, part of the explanation why Metaculus has now started issuing requests for *fortified essays*. The fortified essay is a text explaining a rationale that is relevant to one or more forecasting questions. Forecasting questions are embedded in the fortified essay, so that readers can make their judgments while reading, which does solve one of the problems I identified above. Moreover, the fortified essay is supposed to offer the flexibility needed to contextualize the question to be answered, and inform about the reasons that are relevant to the forecasting question.

This is a welcome addition, and goes a long way towards solving the problems of context. If fortified essays are widely read, and can be trusted to include all the main relevant reasons that bear on the important questions in need of answers in EA, this will go a long way towards solving the problem. However, this is a really tall order. Professional writers and researchers whose job it is to write texts that many people read and consider carefully have a hard time doing it, and the competition for really good writing is hard.

Considerations like these are probably what prompted the more flexible approach of making it possible for writers to embed forecasting questions into texts to be published on forums like LessWrong.[36] I think this is an important step in the direction of successfully crowdsourcing forecasting and research in EA and rationalist communities. Many authors would love to see what others think about their ideas, and so the incentives for forum writers to embed questions into their texts is good, and it doesn't require much effort from users to submit judgments on questions while reading about them.

However, many important texts aren't on the forums, and we would like to elicit judgments on these texts as well. If we could transform existing literature containing the arguments that are relevant to crucial questions in EA into something like fortified essays, this could be a more scalable and comprehensive solution to contextualizing forecasts.

It seems to me that a natural next step for advancing the cause of crowdsourcing EA community belief is to generalize the process of turning questions found in articles and other text content into actionable forecasting questions. Probably, this could best be done through web annotation software embedded as a browser extension.[37] If everyone who has the same extension installed could create, see and respond to forecasting questions in the process of reading, we might see a whole lot more forecasting responses in the EA community.[38]

---

[36] Both Metaculus and Ought has done work to this end, see Embedded Interactive Predictions on LessWrong. This was also a planned feature for Arbital.

[37] I am thinking of something along the lines of WorldBrain's Memex and Home : Hypothesis. When Grammarly, a grammar-correction device, implemented their service as a browser extension that could be activated anywhere, usage skyrocketed.

[38] Here is a comment from the EA forum expressing a similar idea.

Knowledge representation

There is great value in organising ideas systematically, and representing it succinctly in a way so that it is readily available. This is true both for individuals, and for communities, which is the reason why many organisations and communities build and sustain infrastructure for knowledge representation. I am aware of two ways to represent knowledge in the dynamic forms outlined in the previous section on elicitation. Knowledge representation through wikis, and through knowledge graphs. In this section, I note strengths and weaknesses of both approaches, and describe initiatives I am aware of.

Local Wiki and Forum

There are many reasons why a successful wiki could be valuable to EA.[39] Here are some of them (see rationale in footnotes):

- **Research**. A wiki could make it significantly easier for community members to quickly find and retrieve information that is otherwise scattered across various media.[40]
- **Onboarding**. Effective and engaging onboarding for new members.[41]
- **Contribution**. Making contributions to a wiki is a quick and easy, low-threshold way to help out, which is both good for contributors and the community as a whole.[42]
- **Paradigm**. Shared terminology and theoretical understanding of key issues makes it easier to identify key research questions, and what it would mean to answer them.[43]

Several EA wiki projects have been initiated, including priority.wiki, EAWiki, EA Concepts and the LessWrong Wiki. Unfortunately, these didn't catch on as much as one would like. The risk and downside with wikiprojects is that they can be hard to maintain, due to reasons such as these:

- **Unclear standards**. Unclear how volunteers can contribute, and what standards for articles are.[44]
- **Lack of coordination**. Resources split between multiple projects.[45]
- **Costs of contributions and lacking incentives**. A problem all crowdsourcing projects, including wikis struggle with, is how to incentivize community members to write high-quality contributions.[46]

---

[39] See for instance David Tomasik's The Value of Wikipedia Contributions in Social Sciences. Also see the rationale behind the 2020 and 2021 grants to Pablo Stafforini for a new EA wiki project, Should we use wiki to improve knowledge management within the community? - EA Forum, Announcing PriorityWiki: A Cause Prioritization Wiki - EA Forum, and Local EA Wiki Discussion. As Stefan Torges notes, the US Intelligence Community manages knowledge in a wiki. This is a good argument in itself to think that a wiki approach can be effective.

[40] See (Vollmer 2020) and (Daniel 2021).

[41] See (Daniel 2021).

[42] See (Daniel 2021).

[43] See (Daniel 2021).

[44] See (Vollmer 2019).

[45] See (Vollmer 2019). See this discussion.

[46] See (Andreev 2017), also the comments.

Despite these concerns, a new initiative to use the EA forum to build out a new wiki based on tags has garnered support from The Effective Altruism Infrastructure Fund, and the EA Forum team. An External Evaluation of the EA Wiki conducted by Nuño Sempere from Quantified Uncertainty Institute was recently published on the EA forum. Sempere does a quantitative analysis using data from google analytics to compare various metrics, such as hours spent working on the wiki, and hours spent viewing content at the wiki. However, as Sempere acknowledges, the relevant measure to evaluate impact is time-saved through the wiki, which is a metric we don't have. I would add that we should also consider the value of knowledge and understanding of reading content at the wiki which one might not easily find elsewhere, but this is a value which is even harder to measure.

The EA forum is forked from LessWrong,[47] whose original design was conceived by Eliezer Yudkowsky. Yudkowsky had an idea for another platform as well, whose vision included several ideas mentioned above, and with some affinities to the EA forum wiki system.[48] The platform was called Arbital, and was supposed to be a knowledge representation device for research and decision making, whose design was informed by the epistemic norms that are prevalent in the rationalist community around LessWrong. It follows a standard wiki-format, but also has functionality resembling that of a blog and forum, and lets users express quantitative estimates on claims, and insert links to explanations, to mention a few things. Alexei Andreev, the core contributor of that project, wrote a detailed postmortem for the Arbital project. In it, he reflects on the reasons why it failed, and engages with the LessWrong community in the comments in a very rich and informative discussion. I won't try to capture the nuance here, but a theme that Andreev and several commenters seem to find especially important has to do with motivation and incentive structures for contributing content.

Knowledge Graphs

Another way to represent knowledge is through knowledge graphs. A knowledge graph represents knowledge through a graphical representation consisting of nodes and edges. Each node represents a concept, idea or item by a name or short sentence. Edges represent relations between the content of the nodes. Popular software services like Obsidian and Roam make it easy for individuals, and teams, to take notes and add them to a knowledge graph.[49] In these systems, the nodes and edges are interactive, and can be activated to retrieve more information, or see patterns in the graph.

In EA, there are several ideas and initiatives to consolidate knowledge in knowledge graphs.[50]

---

[47] Which is forked from Reddit.
[48] The ideas underlying Arbital and other initiatives towards efficient knowledge representation systems are old, see Zettelkasten and Memex. These ideas were the basis for the first hypertext systems, including Project Xanadu and the World Wide Web.
[49] See also Ken.
[50] See Clarifying some key hypotheses in AI alignment and Modelling Transformative AI Risks (MTAIR) Project: Introduction for a Bayesian knowledge graph showing considerations in AI risks. See also Causal diagrams of the paths to existential catastrophe - EA Forum for use cases of causal graphs in existential risks. See also the AI Alignment Roam Database Project: https://forms.gle/xrvBaAJ45tZLgMCF9, Causal diagrams of the paths to existential catastrophe - EA Forum and "Epistemaps" for AI Debates? (or for other issues) - EA Forum.

In a recent forum post, Ben West proposes several EA communication project ideas. One of these ideas is to make a map of EA ideas, where, perhaps, a knowledge graph could be applicable. Videos mapping intellectual terrain are popular on youtube.[51] There is a chance that a graph could be a good way to show how EA-ideas relate to each other as well. Knowledge graphs are elegant, and if they are neatly structured, can give a sense of intellectual aesthetic satisfaction.

The Long-Term Future Fund funded Roam due to the potential general epistemic gains to be had from advancing their platform.[52] Although the format of a knowledge graph is different from text-based wikis, a knowledge graph can have the same functional properties as a wiki. In fact, if we chose to develop a knowledge graph representing key ideas in EA, the graph wouldn't have to compete with a wiki at all. A knowledge graph could just be hooked onto a wiki, such that nodes in the graph link to concepts and articles in the wiki. A knowledge graph does not replace the text-articles in a wiki, but is rather an alternative to the interface for searching, navigating and finding content on the wiki. In this respect, knowledge graphs contribute two key benefits:

- Navigability.
- Standardisation.

Navigability

File and folder and tag-based systems require a basic familiarity with relevant keywords to navigate between entries. In a knowledge graph, the relationship between ideas are represented by edges (lines), and ideas that are relevant to each other are closer in proximity. As such, it is possible to explore an intellectual terrain without knowing what one is looking for.

In simple graphs, the content of each node is a sentence or two. In interactive knowledge graphs, nodes can be activated to elicit one, or several additional levels of information. One could for instance envisage a structure in which the first level of a node contains a name (1-3 words),[53] the second level is a sentence or two, the third level is a full paragraph, and a fourth level is a full wiki-article or link to a full wiki-article.

Interactive knowledge graphs, where it is possible to zoom in and out, and choose an appropriate level for the granularity of information about particular nodes, are especially easy to navigate, since it doesn't require a lot of reading to explore.

Standardisation

The greatest strength of knowledge graphs has to do with standardisation of content. Due to the streamlined structure of text content for the nodes in a graph, it is easier to collaborate with others in creating a knowledge graph than writing longer text entries. The same applies to the edges, as smart and clear conventions for the semantics of edges can give clear guidelines for additions to the graph.

---

[51] See for instance The Map of Mathematics.
[52] See Long-Term Future Fund: August 2019 grant.
[53] In venture capital, it is common to present business ideas in slide-decks where visualization, keywords and compact sentences replace long-form text content. Knowledge graph applies similar ideas but in a more general form.

However, when graphs expand, and especially when several people are involved in building the same graph, the graph quickly becomes disorganized if there aren't clear standards. In most knowledge graphs, there are no rules as to what the contents of nodes should be, or what the edges in the graph represent.

Another key strength of knowledge graphs is their flexibility. It is easy to delete or modify the edges between nodes in a graph, without affecting the content of the node. In text-based entries, as new information comes in, it is often necessary to modify the text-content, which is a more cumbersome task.

Some of the greatest challenges to wikis had to do with unclear **standards**, **lack of coordination**, and **costs of contributions and lacking incentives**. Knowledge graphs cannot solve the problem of incentives alone, but at the higher levels of the nodes in a graph, it is easier to enforce clear standards, and since these are more compact than wiki entries, the cost of a valuable contribution is even lower. Also, a knowledge graph mapping EA ideas does not compete, but rather complements existing wikis, and would benefit from drawing on the content of existing initiatives.

A Bayesian Knowledge Graph

A solution to the aforementioned problem of standardisation in knowledge graphs is to introduce limiting rules as to what nodes and edges may represent. Similar to the way a map should accentuate details of particular importance to the purpose of the map, a graph ought to accentuate details that are relevant to the purpose of the graph. Typical knowledge graphs show how ideas, actors and theories are related to each other. Such graphs are useful for note-taking, and for association. However, they are not particularly useful for reasoning. If we would like a knowledge graph to aid reasoning, it ought to represent the elements of reasoning, namely **claims and inferences**.

We could add nuance to the kind of reasoning we would like to see. In the EA community, there is wide support for a broad set of epistemic norms.[54] A distinctive set of these norms are related to what we may call Bayesian reasoning. **Bayesians think in terms of graded beliefs**, and like to quantify probability estimates concerning the likely truth of claims. In other words, a Bayesian thinker does not simply believe or disbelieve some claim. Rather, they believe it to a greater or lesser extent. A system designed to aid Bayesian reasoning, then, should let users express beliefs in degrees, as probabilistic estimates. A precise way to do this is to use a numerical point scale, going from 0-1. However, to make Bayesian reasoning more accessible, it might be smart to use an ordinal scale, with the option to use a numerical point scale, or slider, instead.

In addition to graded beliefs, **an important epistemic property of judgments is the confidence in which some person expresses it.** I might be inclined to say that there is a 90% chance that the democrats will win the next US presidential election, however, this might be based on almost nothing, and so I might be very unsure whether the judgment is true.

[54] See the five points above. Also, see Reasoning Transparency, and A conversation with Philip Tetlock, March 8, 2016 Participants Summary Calibration software Bayesian question clusters, or the writing guidelines to the EA forum: How to use the Forum - EA Forum.

Confidence is an important epistemic property, and a good system for representing beliefs should express confidence in some scale. Here too, a numerical point scale would be most precise, but an ordinal scale is probably sufficient for most uses. **Another way to express confidence is to allow judgments to be expressed as distributions**. Sometimes we would like to say that we are 90% confident that there is a 70-97% chance that the next US presidential election will be won by the democrats, for instance.

Reason – Inference, explanation and argument

A central part of Bayesian reasoning is to update beliefs on the basis of new information, and so we would like some mechanism for this as well. That mechanism is Bayesian inference. When quantitative forecasts come as isolated responses to questions, they are like black boxes. One simply has to trust the forecaster. We should of course try to take a step back sometimes and consider the judgments of others in our deliberations, but we should not let majorities or expert opinions we don't understand override our inside view on all occasions. Moreover, in many of the fields that are of interest to EA, it is unclear who the experts are, and what constitutes relevant expertise. For many empirical questions where some forecaster may establish a good track-record, it is less problematic to simply trust the expert. It would be a whole lot easier to trust others' judgment if we could understand and assess their reasons for judging differently than ourselves.

Whenever I encounter someone who holds a different belief than me on some issue, I am typically interested in knowing what **reasons** they have for holding this contrary belief. Do they know something that I don't? Am I misunderstanding some important argument? The mere fact that they disagree is not very interesting, the interesting part is their reasons for thinking differently.

Chains of reasoning are especially useful in a Bayesian system. For longer lines of reasoning, seeing how the probabilities assigned to earlier premises in the chain impact the outcomes in the end is quite valuable, as most people (even experts) tend to fail when intuiting this calculation.

The point of the fortified essay is to give the rational context of a judgment, that is, the grounds, research or supporting arguments that bear on the likely truth of the question to be answered. The fact that the relevant forecasts were made on the basis of a fortified essay, or was embedded in a LessWrong post, gives the decision maker some insight into the grounds for the forecasts. However, the context in itself does not reveal which arguments were consequential in the judgments of the various forecasters, which, in fact, is a key component of making these judgments perspicuous.[55] Moreover, the context doesn't in itself guarantee that the full rationale for the forecast is perspicuous to the decision maker, since most forecasters probably draw on other arguments besides the ones mentioned in the essay or post. What is needed, is a flexible system for elicitation that lets users link an argument from one text, to the conclusion of another.

The considerations concerning the epistemic value of reasons seem decisive to me, and stake a direction for further improvements to the EA knowledge infrastructure. We should not only

[55] See Reasoning Transparency | Open Philanthropy (Muehlhauser 2017: section 3.2).

make it possible for community members to register their thoughts concerning crucial questions, but also to model and assess the arguments that bear on these questions. If we built out a software infrastructure for estimating both the likely truth of key claims, and the arguments that are relevant to that estimation, the output would be a whole lot more valuable to researchers and decision makers than individual forecasts alone.

The practice of analyzing the reasoning in a text is called argument-analysis. Argument analysis is the practice of identification, interpretation, clarification and evaluation of reasoning. This practice is the key component of most university courses on critical thinking/informal logic,[56] and the practices advocated in such courses closely aligns with norms of reasoning transparency adhered to by some EA's and EA orgs.[57] Argument analysis can be more or less detailed, complex, and cumbersome, which entail costs to the person conducting the analysis. However, the practice is quite valuable for the person conducting the analysis, as it makes them more aware of the grounds for their ideas, and also makes it much easier for others to understand how they think.

Such features could be part of a knowledge graph if users could express beliefs about claims through judgments, and update these beliefs by adding new claims to the graph with inferential relations to the claim with the belief to be updated. A graph of this kind, that is specifically suited to support Bayesian reasoning by giving quantitative weights expressing belief in claims, and weights expressing the conditional probabilities between claims, is called a Bayesian Network or a Bayesian Knowledge Graph.[58]

Other ideas and initiatives

The idea of a Bayesian Knowledge Graph is not original to me. In fact, there exists commercial software for making such graphs.[59] In 2019, Anthony Aguirre, cofounder of Metaculus, also received a LTFF grant to develop something like a Bayesian Knowledge Graph to be part of Metaculus. Here is Habryka's grant writeup (2019), and here is a quote from the application:

"*The second expansion would link questions into a network. Users would express links between questions, from very simple ("notify me regarding question Y when P(X) changes substantially) to more complex ("Y happens only if X happens, but not conversely", etc.) Information can also be gleaned from what users actually do. The strength and character of these relations can then generate different graphical models that can be explored interactively, with the ultimate goal of a crowd-sourced quantitative graphical model that could structure event relations and propagate new information through the network.*"

However, these plans seem not to have materialized, yet.[60]

---

[56] Here is a document I have written on argument analysis with some notes on argument analysis from contemporary textbooks on critical thinking.
[57] See for instance Reasoning Transparency | Open Philanthropy.
[58] See Beard et. al.'s An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards - ScienceDirect (2020:. p7) for more on this. A thread in the comments to Michael Aird's post about the existential risk database also emphasise this point (Kristoffersson 2019).
[59] See Products - Analytica and Norsys Software Corp. Also, Kauffman Fellows, a VC investment training program, also apply similar methods in training for reasoning procedures to investment decisions.
[60] I have received some indication that Metaculus might move on this after all (Jan 2022).

The same idea was also a key motivation for the founders of Roam, which they express in their white paper from the winter of 2017/2018. LTFF has also supported Roam in multiple rounds.[61] However, at the end of the paper, the authors has put in a recent note, saying that:

"*We wrote this White Paper in Winter of 2017/2018, and while it still reflects much of our vision for the tool, some of the features -- particularly related to [[Bayesian Reasoning]], [[Argument Analysis]], and [[Prediction Markets]] we found to be in-fact much lower priority for the researchers and decision makers whose thinking we aim to assist -- and risked making the tool too complex for the more significant use cases.*"

This is quite understandable. Roam is a for-profit organization aiming to make a self-sustaining commercial note-taking product for a wide target audience. For this, the sort of Bayesian Knowledge Graph envisioned here is not viable. For the Bayesian Knowledge Graph to be successful, it would be necessary to put severe restrictions on the semantics of the graph, which would make it unusable for simple note-taking.

David Manheim and a group of researchers was supported on a project to research and represent knowledge on AI risks using Bayesian Graphing software (Analytica).[62] They have now mapped out AI risk debates, and going forward, plan to elicit expert judgments and map them onto the Bayesian Knowledge Graph. The graph is not publicly available yet, so it is still a bit early to make calls on its utility, but the project itself seems quite promising.

Beneficial properties

A Bayesian Knowledge Graph can be designed to have some very powerful epistemic properties. The most important ones have to do with updating and personalized epistemic nudges.

Updating

If someone adds a node with an inference relation to another node in the network, the credence of the other node is automatically updated in light of the new node. If the updated node supports yet another node, this is updated as well, and so on. It is generally very hard for most people to update their beliefs in light of new evidence, especially for long chains of reasoning. A Bayesian Knowledge Graph can help with this, both for individuals, and for representing the collective knowledge of an organisation where a researcher in one field may submit a finding with implications for a fellow researcher working in another field. As such, a Bayesian Knowledge Graph may help solve some of the biggest problems of specialization and knowledge silos.

Personalised Nudges

Another exciting potential of Bayesian Knowledge Graphs has to do with the potential uses of data. The data structure can be used to automatically generate a personalised feed or newsletter nudging users to improve epistemic practices. Here are some examples of the type of nudges I have in mind (it would be easy to generate more):

---

[61] See August 2019: Long-Term Future Fund Grants and Recommendations and Long-Term Future Fund: November 2019 short grant writeups - EA Forum.
[62] See Habryka's writeup May 2021: Long-Term Future Fund Grants.

- Find points on which other users disagree with you, and link to arguments they find persuasive for disbelieving something you believe, which you haven't considered yet.
- Major updates (10% or more) of people you follow or agree with a lot.
- New arguments that bear on issues you have a view on, which others find convincing.
- Controversial implications of your views that you haven't considered yet.
- Issues you have a view on, which many others have commented on, and updated on lately.

Collaborative Bayesian network

In the previous chapter I described and discussed crowdsourcing projects aiming to improve the knowledge infrastructure in EA. I discussed ideas and initiatives for eliciting quantitative judgments and ideas and initiatives towards organising and representing this knowledge.

From this discussion, I identified the following benefits of wikis:

- **Research**. Easier to quickly find and retrieve information.
    - **Arguments**. See overview of the most important arguments bearing on a claim.
    - **Navigability**. Navigate intellectual terrain without understanding keywords.
    - **Standardisation**. Simplified formats for representing ideas.
- **Onboarding**. Effective and engaging onboarding for new community members.
- **Contribution**. Quick and easy way to make a contribution.
- **Paradigm**. Shared terminology and understanding.

I then argued that knowledge graphs built on top of a wiki could enhance navigability of content, and that specific features of knowledge graphs to do with standardisation of format could mitigate some of the problems with wiki projects. Moreover, while retaining the benefits of wikis, a collective knowledge graph can also be customized to bolster reasoning and decision making in accordance with the epistemic norms adhered to in EA. The way to do this is to elicit and aggregate quantitative judgments on relevant issues. I identified several benefits to crowdsourcing and aggregating quantitative judgments, including the following:

- **Accountability.** Contributors can be held accountable for their beliefs.
- **Training**. Contributors who make judgments get good forecasting practice.
- **Accuracy**. The aggregate of multiple judgments is more reliable than single judgments.
- **Value of information**. Patterns of belief indicate the value of questions and results.
    - **Value of research questions.** Confidence, disagreement and consensus in beliefs is relevant to the value of research questions.
    - **Value of research results**. Tracking of belief revision can be used to measure the value of research results.
- **Collaboration**. Elicitation and aggregation can connect people to projects.
    - **Contribution**. Contributors may demonstrate knowledge and expertise.[63]

[63] "Contributing to a wiki is a concrete way to add value and contribute to the community that is accessible to basically all community members. My impression is that opportunities like this are in significant demand, and currently severely undersupplied by the community. If the project goes well, many students,

- **Recruitment**. Projects can find collaborators to projects.[64]

In the discussion on elicitation, I argued that a browser extension for web-based text annotation would be a good solution to what I perceive to be the greatest problems of prediction platforms. If users could register their thoughts on what they are reading, while they are reading, more people would perhaps submit judgments to a common platform.

Moreover, I also argued that judgments on the likely truth of claims should be supplanted by an overview of the most important reasons that bear on these claims. All scientific articles have an explanatory or argumentative structure, and so do most of the literature that is relevant to EA cause areas. It is just as important to map the grounds for judgments, as the judgments themselves.

A graph structure incorporating both claims (beliefs) and arguments (inferences), is a graph in which every node is a claim that can be forecasted, and every edge (line) is an inferential relation showing what ideas depend on each other, one could use the graph to follow, or express lines of reasoning. Putting these semantic constraints on the graph would make the whole structure less informative than text-articles, however, nodes could contain links to relevant text content to ameliorate this downside.

A graph of this sort can be called a collaborative bayesian network. It is a representation of knowledge that aims to fulfill, or at least partially fulfill, many of the same functions of a wiki and a prediction platform, simultaneously.[65] In what follows, I'll refer to this knowledge infrastructure system for eliciting and representing knowledge in EA as the *Solon system,* or simply *Solon*.[66]

In addition to the beneficial characteristics of wikis and prediction platforms outlined above, a browser extension for web-based text annotation hooked onto a Bayesian knowledge graph, contributes new exciting epistemic benefits of its own:

- **Updating**. Bayesian updating, reverberating throughout the network.
- **Personalised Nudges**. Personalised nudges for improved epistemics based on data.

In what follows, I want to present my vision for Solon through user-stories and sketches. The idea here is to try to paint a realistic story about how the Solon system could work, and draw evocative images to make the benefits listed above more concrete. I then want to argue that the expected value of a project to develop Solon is high, despite significant epistemic and practical concerns associated with scepticism, cascades and the feasibility of attracting people to the platform, and sustaining it. I end this chapter with a cost-effectiveness analysis.

---

researchers, and professionals might contribute to the wiki in their spare time, and find the experience motivating and satisfying." — May 2021 grant rationale by Max Daniel

[64] (Aguirre 2017: 14-16 min marks).

[65] The idea for this project was sparked by some closing thoughts of Michael Aird, in his: Crucial questions for longtermists - EA Forum.

[66] After Solon the wise, who laid the grounds for the first Greek democracy. A highly successful crowdsourcing initiative for a relatively small, but highly efficient community, that went on to conquer most of the then known world.

Use-cases and user stories

In this section, I will try to illustrate paradigmatic use-cases for the Solon system through user stories. That is, I will tell a story of how using Solon was helpful for different people, in different ways. The stories are also intended to support the general claim that a system like this would be useful to its users and the community, and as such, will replace what would otherwise be an in-depth theoretical discussion of the benefits of the system. Although I think there are more good use-cases, I'll focus on two stories that express the most important benefits and use-cases.

Reader and decision maker

David, a university student of philosophy, is active at LessWrong and the Effective Altruism forum. At the behest of a recent forum-post, he installed the Solon browser extension for analysis and estimation. One afternoon, he reads a fresh argumentative text on the EA forum about the expected impact of open borders, inspired by Bryan Caplan's work. After he has read through it, he activates Solon. David is the first to consider the text with the extension, so he is given the option to conduct a logical analysis of the text. He takes it upon himself to identify the main conclusion of the text, and the arguments given in its favor. It takes about 40 minutes to analyze the 6 page text in terms of a series of claims, and inferential relations between those claims using Solon's built-in functions. After this, he registers his best guesstimates concerning the likely truth of the claims, and the inferential strength of the arguments.

On the other side of the globe, Melissa, an EA-aligned social science researcher finds the same post about open borders. She finds the topic interesting, and decides to read the post. Melissa has also installed and activated Solon. When she opens the post, she is notified that someone has given an analysis of it, and is given the option of seeing the analysis while reading. She chooses to do this, which makes David's analysis visible to her (but not his estimates). Annotated parts of text are marked, and cards for claims and inferences are displayed to the right of the main text, in a similar fashion to comments in google docs.

Melissa draws on her knowledge about the topic, and quickly notes her credences while reading. She also quickly evaluates the arguments, registers a counterargument to one of the inferences, and leaves a comment explaining why she estimates a key claim to be most likely false. All in all, the time she spent reading and assessing was about two times longer than it would have taken to merely read it. After she has read through the post, and assessed all claims and arguments from David's analysis, Solon offers the option to see what others think.

Of the 200 people who read the post, 7 people, in addition to David and Melissa, assessed the reasoning with Solon. Melissa was surprised to see that the others mostly agreed with her takes. However, one claim in particular had drawn controversy. The claim was that world GDP would double if all countries adopted an open border policy. On this claim, there were three comments, each representing a reason or objection to the relevant claim. Melissa suspected that the other estimators perhaps hadn't considered a relevant recent study, so she also commented, and referenced the study. After this, she activates a function in Solon to see the analysis as a knowledge graph.

A new tab opens, where the cards from the analysis are now represented as nodes and edges in a graph. In addition to the claims and inferences in the original post, there are also additional nodes and edges going to and from the ones from the post about anti-aging. Melissa identifies the controversial claim about GDP doubling, and has a quick look at the main nodes supporting and undermining it. One of the arguments that another commenter had inserted had been judged to be quite strong by other estimators. The argument was supported by several other reasons, some of which were supported by still more reasons. Melissa was unsure, and curious about the claim, and so she followed the chain of reasoning in the graph. Most of the reasons and nodes were self-explanatory or reasons that she was familiar with already. She simply skipped past these. Others were new and intriguing. For some of these, she activated the nodes to see if others had written more about them. For some of them, she also followed the url-link in the node to have a look at the text document which the node was originally taken from. After some light research on the double GDP claim, she felt a whole lot more confident that this might actually be true. In the graph, she put weights on the various arguments bearing on it, which then automatically updated her credence for that claim, and automatically updated the credence for the main conclusion as well, that open borders is a good idea.

Melissa now accesses a function to represent her personal knowledge graph. Most nodes and edges in the graph now disappear, and the weights only represent her own beliefs. She spends a couple of minutes contemplating the updates to prior beliefs automatically engendered by the new additions to her network, which is now highlighted in her graph. She is quite confident that her views are quite reasonable, and that the network is coherent, which is a great source of satisfaction for her. Although Melissa is happy in her job right now, she has for some time been thinking that it would also be exciting to work on policy questions related to immigration. She has chosen to share the graph, so that anyone can see her perspectives on the research literature she has read and assessed with Solon. The fact that her thoughts about many of these complicated thorny issues are shared publicly, gives her extra motivation to read, and critically engage with the research literature on immigration and open borders.

The person who had extended David's original analysis of the text on open borders, to include references and arguments concerning the claim that open borders globally would lead to a one time doubling of GDP globally, was Robert. When Melissa updated her beliefs on the basis of Robert's additional research, a bar in Roberts Solon dashboard updated to reflect the epistemic impact of his research contribution. Next to this impact barometer was a scale representing his forecasting score, which automatically updates when forecasted claims resolve.[67] Robert opens a button to see additional stats, where alignment with other EA's are shown, as well as confidence levels in various topics where he has been active.

Robert is a funding officer at an EA aligned fund, and is currently in the process of evaluating a grant application. The application is for an initiative to do with open borders. Robert is quite knowledgeable about the topic, but there are some important parts of the overarching argument that he is a bit unsure about. Moreover, he doesn't know any specialists on this topic in particular. Robert enters Solon's graph view, and plots the reasoning in the grant application,

[67] Most claims that are relevant to EA don't resolve. The score is only calculated for the ones that do. The forecasting-scoring system might just be imported from Metaculus, or it could be a Brier score.

which largely overlaps with other claims associated with the text on open borders that David analyzed.

At this point, about 13 people[68] have assessed some of the crucial claims and arguments that are relevant to the open borders claim. There is wide consensus amongst the other estimators on some of the claims Robert is unsure about, and fierce disagreement on a few others. Through a filtering mechanism, Robert chooses to differentially weight the estimates of the estimators on the basis of their backgrounds. Melissa, being a social science researcher with a graph network which demonstrates a fair deal of knowledge about the topic, is judged to be twice as reliable about this particular topic, than David, in Robert's particular choice of weights. On this differentially weighted aggregate of estimates, a few claims in the network stand out as in need of further research, and clarification. On one of these claims, Melissa and another researcher with a strong relevant academic background has made contradictory judgments. Robert reaches out to both of them, references the particular claim that he is unsure about, and asks if they can elaborate on their forecast, and if they have the time to chat about the grant. After this, Robert feels more confident about the grant. His judgment now accords with the beliefs of the community as whole, with a few exceptions where he is quite confident that he can explain how their reasoning is mistaken, with support from experts in the field.

Seminar

The past year, Toby attended a variety of international security webinars/workshops on topics like great power competition (US-China competition), nuclear strategy, grand strategy, etc. and often ended the webinars thinking "What did they even say? Where did they disagree? What did I even learn or change my mind about?"

The next event he attended was a conference organized by an EA organization on the topic of fish welfare. Christian, who was the central organiser, had decided to try Solon at the webinar, and a workshop tied to the seminar.

Jacob, a presenter, was asked by Christian to analyse the research paper on which his presentation was based with Solon. Since Jacob already knew his own paper very well, it was quite easy for him to analyse his paper in the Solon framework.[69]

When introducing Jacob for his talk, Christian shared a link to the Solon graph representing the main points of the talk. As Jacob went through his presentation slides, Toby sometimes switched over to the graph, to note questions. He also used a function to insert an objection to a central point in the talk anonymously. He also upvoted a few other questions and objections, and used

[68] 10-12 individual predictions is enough to reduce noise, and give impressive results, see Daniel Katz' comment to this article: Crowdsourcing "can accurately predict court decisions 80% of time" says study.
[69] If speakers or note-takers were to try to map some of the general points of the discussion it would be easier to see what were the key claims/ideas (vs. what they did not say or intend to say) and where the speakers agreed, disagreed, were uncertain, etc. Of course, it might also reveal that actually the speakers didn't have many substantive points, although one would hope that using such mapping would encourage participants to clearly delineate their points (or just not pretend like they are making distinct points when in fact they are just repackaging/presenting the same idea with different language to make themselves stand out).

the graph to keep track of earlier points of the talk that were relevant to the current slide. For several claims and arguments, he also gave a numerical credence.

After the talk, Christian opened the floor for questions. Two questions had garnered a host of upvotes during the talk, so Christian started by asking the audience whether the originators of any of these questions wanted to ask it themselves. One member of the audience answered, but Christian had to ask the other question himself. As these questions were answered, this clarified things quite a bit. Toby saw a way to disambiguate a central point in the talk, and so he went to the graph to suggest a change to a node and one of the arguments. He also raised his hand in the webinar to make sure his way of framing things was correct.

Since Jacob had given quantitative estimates expressing his confidence in the central claims and arguments, a large part of the discussion revolved around how likely it was that central claims were true. Toby, and several others in the audience had also expressed their credences. This made it easy for everyone to shy away from issues in the talk about which everyone already agreed. These numerical credences, and especially the points on which there was much spread in the credences given, was useful to Jacob in deciding the direction of future research.

The epistemology of aggregating judgement

In what follows, I will discuss several considerations concerning the idea that a collection of quantitative estimates drawn from the effective altruist community promotes epistemic value to the community. Most of these considerations are drawn from this thread, and Michael Aird's Potential downsides of using explicit probabilities - EA Forum (2020), The importance and challenges of estimating existential risk - Michael Aird - transcript (2020) and Ways of describing the "trustworthiness" of probabilities, (2021). The discussion has also benefited from Denis Drescher's question How might better collective decision-making backfire? - EA Forum, and the responses to it.

In the discussion to follow, I'll treat these considerations as objections to Solon. My strategy of defence will be to deflect objections by showing that they overgeneralize, or to draw lessons from the objection that might then inform the further development and design of the Solon system.[70] A recurring theme is that elicitation of reasons in addition to forecasts is the solution to the epistemic problems associated with merely surveying quantitative estimates on difficult questions.

[70] The idea here is that we use the objections as something like design specifications that inform the further development of the design of the Solon system. It is good practice to do this before actually creating a system at all. However, this general strategy is even more effective once a system is up and running. There is much to be learned from creating a machine to emulate good ideas about decision making, and thereafter to modify the machine or decision algorithms on the basis of the results until we learn how best to construct it to give effective guidance. In his book *Principles,* Ray Dalio repeatedly emphasise the value of creating a machine or explicit model of one's reasoning in order to learn where one is wrong, and to learn from past mistakes. The Solon system is an attempt to do this for EA thinking.

Disagreement and scepticism

First off, one might be concerned about disagreement. There is usually a very large amount of disagreement between people.[71] The fact that there is a large amount of disagreement is considered by some to be a convincing reason to think that there is no objective answer, or at least no objective answer available, concerning the issue in question.[72] Disagreement could indicate that we are clueless, and in that case, estimations won't help.

Even if one is inclined to think that disagreement is itself a good reason to think that we are clueless, it will then be quite valuable to find out where there is disagreement. This form of knowledge helps us more clearly identify what people disagree about and to what extent, and could make that disagreement more salient. That could help us avoid assuming other people all think like us or like that one person who happens to have written about [topic]. And it can also help us see that there are some things on which people do tend to agree (e.g. AI is a bigger deal than gamma ray bursts).[73]

The strongest and most decisive reply denies that disagreement gives us any strong reason to think that there is no answer available to us. Many people may be wrong, and in most cases we can give compelling explanations why they are wrong. To feel the force of the argument from disagreement, we would need to see some additional evidence suggesting that those who disagree cannot, even in principle, come to agree. The typical strategy for the sceptic is to show that we cannot acquire knowledge about the issue in question. In the case of longtermist questions, two seemingly forceful reasons[74] to think this are the facts that we lack baserates for long-term predictions, and the fact that long-term predictions have a terrible track record.

Lacking baserates

The first step to a good prediction is to consider relevant baserates. However, for many longtermist questions, there seem to be no relevant baserates on which to base our prediction. One might think that this means that we cannot find the answer to the question, since we don't have a good place to start.

Even if there are no relevant baserates for a question, this does not imply that we cannot give sensible answers to it. The simple reason is that there are many other methods to answer questions. Another reply is that there most likely are relevant, indirect baserates that we can use to evaluate longtermist questions. This is not the place to get into detailed methodological discussions on all the longtermist questions. I'll be happy to leave this by saying that it is at least not obvious that a lack of baserates for longtermist questions is a sufficient reason to think that we cannot give sensible answers to them.

---

[71] See Daniel: 2020

[72] Such arguments are sometimes used in philosophical debates, although few academic philosophers find them convincing. See for instance Loeb's "Moral realism and the argument from disagreement" (1996). See also Kane B's excellent video introduction to the topic: The Epistemology of Disagreement 1.

[73] Thanks to Michael Aird for pointing me in the direction of this reply.

[74] I have these from Aird's The importance and challenges of estimating existential risk - Michael Aird - transcript.

Are long-term predictions unreliable?

In Philip Tetlock's studies on forecasting, geopolitical forecasts of the type used in the Good Judgment tournaments have a terrible track record when ranging further than 5 years into the future. This terrible track record seems to suggest that we cannot know how the future looks more than 5 years into the future. In an interview with Philip Tetlock, Alexander Berger recalls:

> *Professor Tetlock's research shows that it is difficult for experts to make good predictions about outcomes five or more years into the future. Predictions beyond the lifespan of the person making them may be entertaining but should not be treated as credible.There's little or no evidence that people are capable of reliable non-trivial multi-decade predictions, and there is a strong track record of such predictions failing.* - (Berger 2014: p3).

However, the geopolitical questions in the Good Judgment tournaments are much more detailed and nuanced than the types of forecasts about the future that are relevant to the crucial questions in EA. I think it would be fair to say that they are harder than some of the questions we want answers to, since we can allow for long time-frames, and general overarching events. Moreover, as Tetlock also admits, we do make sensible predictions about the stock-market, and some types of questions are similarly predictable (Berger 2014: p3).

Many authors has written on this topic,[75] and I think it's fair to say that the issue is not settled. A question here is whether quantitative judgments are better or worse than

....

Wide / narrow sampling (bias vs lack of expertise)

Solon is likely to produce narrow sampling, since it primarily samples opinions on particular topics from people who are disposed to read about those topics on their own. These will typically be people who care about and think that these topics and issues are important in some way. Max Daniel gives an illustrative example: "For instance, the participants at the 2008 x-risk conference might be especially inclined to think x-risk is likely."[76]

Conversely, one might worry that the sample of responses is too wide.[77] The concern here is that the people sampled don't have reasoned, stable views, and that their estimation therefore don't really provide any epistemic value to the aggregate. Here is Daniel again:

Based on my own experience of filling in such surveys and anecdotal feedback, I'm not sure how much to trust the answers if at all. I think many ***people simply don't have stable views*** *on the quantitative values one wants to ask about, and essentially 'make up' an answer that may be mostly determined by psychological substitution.*

And he adds nuance in another comment in the same thread:

---

[75] See How Feasible Is Long-range Forecasting? | Open Philanthropy and Long-range forecasting - EA Forum.
[76] Daniel: 2020
[77] See The Paradox of Expert Opinion for a good discussion of this point.

I think surveying more people on what their x-risk credences are will have ~zero or even negative epistemic value for the purpose of improving our x-risk estimates. Instead, we'd need to identify specific research questions, have people spend a long time doing the required research, and then ask those specific people. (So e.g. I think Ord's estimate have positive epistemic value, and they also would if he stated them in a survey - the point is that this is because he has spent a lot of time deriving these specific estimates. But if you survey people, even longtermist researchers, most of them won't have done such research, and even if they have lots of thoughts on relevant questions if you ask them to give a number they haven't previously derived with great care they'll essentially 'make it up'.)

So, is Solon likely to produce a sample that is too narrow, or too wide? What can we do to deflect or ameliorate the worries of wide/narrow samples?

On the issue of whether Solon will give us a narrow or wide sampling, I actually think we'll get a fairly balanced sample. The reason is that Solon elicits estimates from users about the content that they are reading anyways. This means that the sample will consist of people who are interested in the topic, and who are reading about it. This might be experts, but might also be interested, reasonable researchers who take an interest in the topic, without them staking their career on x-risk being very important, for instance. It seems to me that this is exactly the sample we want.

What if we found the above not to be true? What if Solon tended to be either narrow in a bias-inducing way, or wide in a problematic way? If this turns out to be the case, we could use resources in Solon to mitigate it. A perk of the Solon system is that it collects information about the users who give estimates. If user accounts on Solon hooked onto EA-Hub for instance, we would have additional information regarding the research track-record of respondents, which would tell us whether the sample was narrow or wide in relevant senses. With insights into the background of the estimators and the rationale behind a forecast, we have tools to evaluate the aggregate. We could use these insights to differentially weight estimates from people with different backgrounds (and forecasting track-records). Metaculus already does this form of differential weighting, and also accounts for common forms of bias in their aggregation schemes. The Metaculus aggregate tends to be more accurate than most expert forecasters in the long run, and the Solon aggregate will be informed by even more relevant information.

Moreover, Solon elicits reasons as well as forecasts. Analysis of some aggregate of forecasts in Solon would give us the possibility to see if the people giving similar estimates are relying on the same set of reasons, or whether they account for different reasons in their deliberations. If two estimates or sets of estimates rely on different sets of reasons, but support the same conclusion, the estimates or sets of estimates constitute different strands of evidence pointing in the same direction, which then is a stronger form of evidence.[78] If, for some topic, we think common sense is misleading, we may disregard any estimator who hasn't taken all key arguments into account, or who doesn't have a background in the relevant topic.

Aristotle rightly said that inquiry begins with the opinions of the many and the wise. In other words, when researching some issue, we should look at what the majority of people think about

[78] This idea is similar to the concept of 'double counting'.

it, and what experts think about it. However, as Aristotle also says, we should not be content to simply review the opinions of experts and majorities, we should also think about the issue for ourselves. The real value in Solon is not the forecasts themselves, but the overview of arguments that bear on them, which puts decision makers in a better position to make good judgments.[79]

Social dynamics

There are several ways in which social dynamics can lead to bad decisions. In this section, I'll discuss anchoring, informational cascades, polarization and signaling.

Polarization

Groups tend to polarize when talking together. This is certainly a concern in EA, and could potentially increase due to mechanisms in Solon. I don't really have a strong response to this, other than claiming that the harms of intellectual polarization are, in this case, outweighed by the goods of stronger community and exchanges of ideas.

Signaling

For some topics, people might give false estimates to signal popular beliefs. For instance on controversial topics like race and IQ. To mitigate the potentially deleterious effects of signaling, it might be a good idea to hide people's opinions on a select class of controversial topics. However, I am not sure about this at all, and would appreciate input on this.

Anchoring

Anchoring is the phenomenon in which a judgment is based on a particular reference point, or anchor. If Solon presented reasoners with aggregates of EA judgments, there is some reason to think that this would constitute an anchor for the judgments of others, preventing them from approaching the issue in an independent manner.

The easiest answer here is to design Solon in a way in which users would have to give their own independent impressions before accessing others' thinking. One could then let them give their all things considered estimate after seeing what the others believe, but the fact that this second judgment comes after seeing the aggregate is tracked. There is some evidence that this Delphi-type technique, in which each person gives her own estimate, and then can see what other people think, and revise in light of their estimates as well, works well, at least in small groups FLI Podcast: On Superforecasting with Robert de Neufville: (2020: 1:17-8,).

If there is no database giving the community access to the wisdom of the EA crowd, people coming to EA questions would likely anchor to individuals instead, or worse, a survey of EA research crowd wisdom from a conference in 2008. There are plenty of anchors around. I should like to see an argument as to why this particular possible anchor is more likely to bias EA's more than the others. In fact, I should think that an additional anchor-candidate within EA could constitute an alternative point of view which could serve to express the thoughts of a silent

[79] Anthony Aquirre endorses a similar point in a FHI podcast, where he says: "The important thing in forecasting is the process leading to the predictions. We need software for helping people follow good processes, and to distinguish between forecasts based on good process, and no process" The Art Of Predicting With Anthony Aguirre And Andrew Critch.

majority, which then could be a check to balance the influence of thought leaders in the community.

Information Cascades

From lessWrong: "An information cascade occurs when people signal that they have information about something, but actually based their judgment on other people's signals, resulting in a self-reinforcing community opinion that does not necessarily reflect reality."[80]

There is certainly a chance that access to knowledge about what others think will lead to further information cascades in the EA community. However, the specific design of Solon does go some way towards mitigating this effect, again, because it accounts for reasons. However, I am tempted to offer a stronger response. Recall the lesson from Aristotle I mentioned above: Inquiry begins with the thoughts and research of others. I am partial to a view according to which it is simply a mistake to think that we come to form well-reasoned judgments about issues without engaging with the thinking of others about it. That is, I don't think we ever form valuable impressions that are fully independent. The right approach then, is to anchor to the best thinking on a subject, and then to consider the reasons that bear on it independently, which might lead one to think that others are mistaken in some way, and that a better view is possible. It is, of course, a mistake to only defer to others instead of thinking for oneself, which is what happens in information cascades. But the solution is not to limit access to the thinking of others, but rather to promote epistemic norms of independent thinking and careful and reasonable engagement with the research and justification that bear on the relevant issues.

Vagueness and ambiguity

A great problem of science has to do with clarity. In the context of forecasting, Max Daniel is right to assert that:

It's very hard to figure out what exactly to ask about. E.g. how to operationalize different types of AI risk? Even once you've settled on some operationalization, people will interpret it differently. It's very hard to avoid this.[81]

This is the problem of vagueness and ambiguity. It is, as Daniel says, a very hard problem. Modern analytic philosophers spend entire careers in attempts to give clear definitions of central concepts in new or emerging fields. Survey makers, with limited time and resources, surveying opinion on emerging topics without a paradigm of shared concepts and theories, don't stand a chance. However, if we take these issues seriously, we can design Solon to mitigate the problems of vagueness and ambiguity.

With Solon, claims to be forecasted are drawn from text documents. In these documents, claims are contextualized, and this context provides the grounds for disambiguation and clarification. In a software-system like Solon, it is possible to implement systems for clarification, either through

[80] Information Cascades, Deference and social epistemology - EA Forum,
[81] Daniel: 2020

prompts, check-lists, and other guidelines, or through language technologies for disambiguation.[82]

A crucial difference between Solon and surveys, is that one cannot change the phrasing of question claims in surveys, whereas this is possible in Solon. Careful research papers are often careful to delineate clear meanings to the central claims in an argument, but sometimes others see room for improving the clarity and relevance of an important claim. Solon also crowdsources this analytical work, making it possible for observers and forecasters to suggest more precise expressions of central claims. This, of course, invites additional problems, since any prior forecast assumes the prior expression. If the new expression is adopted, the forecasts on the prior expression will be lost, but the clarity will in many cases be worth it. Perhaps users who had given a prior forecast could be notified that a critical mass of other users had adopted a new expression for the claim, and be elicited to make a new forecast for the new expression. This solution allows room for collective, incremental improvement, although in some cases we might see a mess of different versions of the same claim. However, even though this would be a problem, it is arguably not different from the state we are in already. Making the problem explicit gives us better grounds to address it.

Feasibility

I have argued that the Solon system has the potential to contribute great epistemic value to the EA community, if we can manage to build it and foster a community of active users around it. These are big if's. In this section I discuss the feasibility of building and sustaining Solon, with a special emphasis on community-building aspects as this is the most challenging part of the project.

I start off by considering barriers and incentives for contributors to invest time to help building an EA knowledge graph, and expressing their opinions on EA ideas. I actually think AI can be a huge help to mitigate a key barrier. In a section on data and AI, I note a few thoughts about the structure of data that is created in Solon, and how methods from machine learning can be used to turn this data into attractive features of the Solon System. After this, I touch on a few considerations to do with getting the system off the ground, and present ideas for doing so effectively.

Fostering a community

In a comment, Edo Erad, who has been involved in several discussions about the idea of a local EA wiki, recounts several points about running a wiki that he gathered from Chris Watkins, who has experience doing just that.[83]

1. To kick things off, the wiki should have at least one person that can commit to managing the project, someone who is technically skilled (perhaps paid), and about 5 people who can be counted on as core contributors.

[82] In my notes on argument analysis, I have written a chapter on clarification which outlines best practices for disambiguation and clarification of words and sentences for evaluation. Some of this material can inform resources of the sort outlined here.
[83] See How to make a successful new wiki - Appropedia (Chris Watkins, Teratornis: 2009).

2. It needs to be acknowledged in the community as a respected place to point to. Should be coordinated in advance with people who are likely to use it as a reference and to people who are interested in contributing to it.
3. Before starting a wiki, make sure that there would actually be enough users.
4. Work out the license in advance. Especially if importing materials from other sources (say previous wikis, or the forum).
5. It's best to set a culture where people can just write things in without much formatting. Other people can take care of readability later. Also, promote editing whenever the user feels like something needs to be added/adjusted/deleted - it's better to move fast and break things as it is easy to fix.

Jimmy Wales, founder of Wikipedia, also liste five elements he considers essential to successful wiki-projects.

1. Mechanisms for effective collaboration (wiki software provides revision control, ability to revert unconstructive edits; but there must be a dedicated community of users to continuously monitor a wiki).
2. Online identities (pseudo-identities are fine as long as they are stable).
3. Shared vision. The community of participants must share the same vision of what they are trying to build.
4. Flat hierarchies, with the fewest possible barriers to participation.
5. Speed. People must be able to see results from their work with the least administrative delay.

The fuel for all crowdsourcing initiatives is the motivation of contributors, and this is the single most important bottleneck for a successful platform for knowledge infrastructure. Important drivers of motivation are intrinsic, monetary and social factors, and key barriers are shame, time and effort. In what follows I expand a bit on each of these drivers and barriers, and note a few concrete ways to enhance, boost, ameliorate, or mitigate them.

Motivation

We may group motivations in three broad categories: social, intrinsic and extrinsic motivations. These distinctions are pragmatic, and are not intended to be sharp. I'll quickly go through these, and discuss how to think about them at the end.

Intrinsic

Some, but certainly not all, can be intrinsically motivated to contribute knowledge to systems like Solon. A non-exhaustive list of intrinsic motivations include learning, altruism and mastery.

Learning and mastery

Even non-competitive practices, like calligraphy, can be motivating to practitioners due to an urge to master the craft. This is a motivating factor of many researchers as well, and might be a motivation for contributors to Solon.

There is some reason to think that careful analysis of the texts one is reading, and staking one's beliefs in public estimates will be an effective way to learn the material one is reading about.

Although it might be naive to think that most people read or do research to learn (rather than to signal intelligence for having read), I think it is likely that at least some people would find the additional learning effects of using Solon motivating. Some people might contribute to Solon because they think they improve their own judgment by doing so. If we think that this factor is significant, we might want to introduce measures to enhance the learning benefits of interacting and contributing to Solon.

Measures

To accentuate the learning benefits of using Solon, we could introduce features to notify, and sement best practices of logic and forecasting in contributors.[84]

- Checklists for good judgment is one such measure that can significantly enhance judgment. An optional feature could provide checklists for biases, evaluating arguments, and forecasting, to improve judgment.[85]
- Repeat calibration tests could also help contributors learn more from using Solon.[86]
- With Solon, it is possible to introduce features to identify unusual, or controversial beliefs. That is, beliefs in which one's point of view differs from most other people whom one usually agrees with. Such features could make it easy to go over one's web of beliefs to track anomalies, and update others when there are arguments or other information they are missing.
- Follow-functions could make it easier to see how one's thinking is in line with, or contrary to, researchers whose opinions one respects.
- Epistemic nudges. The data structure in Solon can be used to automatically generate a personalised feed or newsletter nudging users to improve epistemic practices. Here are some examples of the type of nudges I have in mind (it would be easy to generate more):
    - Find points on which other users disagree with you, and link to arguments they find persuasive for disbelieving something you believe, which you haven't considered yet.
    - Major updates (10% or more) of people you follow.
    - New arguments that bear on issues you have a view on, which others find convincing.
    - Controversial implications of your views that you haven't considered yet.
    - Issues you have a view on, which many others have commented on, and updated on lately.

Altruism

Many effective altruists are intrinsically motivated by altruistic motives. If community members see contributions to the graph as an effective way to do good, they just might contribute. Making the epistemic effects of contributions to Solon perspicuous, then, could motivate altruistically motivated EA's. In addition to this direct effect on the beliefs on others, the system as a whole may be perceived to have beneficial effects on the decision making of the institution. If contributors feel like this is the case, and that their contributions matter to the decision making

[84] The improvement of community epistemics is, in fact, a focus of CEA's at the moment, as is exemplified in this role: Expression of Interest: Epistemics Project Manager.
[85] See *Noise* chapters 19 and 22 to see a case for checklists.
[86] See New web app for calibration training | Open Philanthropy.

of the organisation as a whole, which again could act as a model for the decision making of other institutions, this might be very motivating.

Measure
Introduce features to track changes in community belief. One could have a function where users could leave a reason why they chose their selected estimate, or why they significantly updated a belief. In cases of hingy beliefs, if several people update just a bit, this can have vast cascading effects propagating throughout the network, which might be very consequential from an organisational perspective.

Social

In an enlightening LW comment to the Arbital Postmortem, Jan Kulveit shares his experience from being a community builder of the Czech Wikipedia Chapter. In the comment, he emphasises the importance of **community design**. The following paragraph is worth quoting in full:

> *Of course, with most wikipedists, somewhere in the background is an altruistic motivation to help with the aggregation of human knowledge and create something like Encyclopedia Galactica. But on a day-to-day basis, what helps keeping people motivated is working with other dedicated people, receiving feedback, being able to see others interacting with your edits and improving further, and even some forms of conflicts. Also valuable editing leads to increasing your weight in the community, you can gain various social goods, responsibility, various functions, and of course power.* (Kulveit 2017).

For most people, in most contexts, social motivations are stronger than any other forms of motivation.[87] Two important forms of social motivation are recognition and competition.

Competition

In the last 5-10 minutes of an interview with the superforecaster Balkan Devlen on the podcast Global Guessing, Devlen emphasises the way superforecasters typically are motivated by competitive factors. Similarly to the way athletes and gamers invest their time to become the best in their sport or game of choice, researchers and forecasters are motivated to best understand and predict how the world works. Objective scoring systems are typical ways to make a practice competitive. Other ideas include scoreboards, and competitions. For instance, one could envisage a prize for the conference presentation that most updated the beliefs of other participants to the conference.

Measure

[87] In an interesting comment, EdoArad summarizes a study concerning the motivations for contributing user-generated content: "In summary, the empirical results paint a somewhat different picture of sustained contribution than originally hypothesized. Specifically, sustained contributors appear to be motivated by a perception that the project needs their contributions (H1); by abilities other than domain expertise (H2); by personal rather than social motives (H3 & 4); and by intrinsic enjoyment of the process of contributing (H7) rather than extrinsic factors such as learning (H6). In contrast, meta-contributors seem to be motivated by social factors (H3 & 4), as well as by intrinsic enjoyment (H7)". See also this comment.

- A measure to incentivize contributors through competition could be to introduce objective scoring systems like a Brier Score, a scoring system similar to the one used at Metaculus,[88] and/or a score for consistency.
- Another measure could be to introduce scoreboards and organise competitions around epistemics within the EA community.

Recognition

Perhaps the strongest motivator for contributions to collaborative projects is recognition.[89] Social platforms like Reddit, Facebook, Strava, Instagram, Youtube or LessWrong[90] and the EA Forum, have **karma systems** which give users the option to recognize each other's contributions. These systems effectively curate content, and motivate users to contribute in a way that benefits the network as a whole. As David_Althaus and kokotajlod notes in Incentivizing forecasting via social media - EA Forum:

Many people seem to care strongly about how many views and likes their content gets. To increase their follower count, many people spend great effort on improving their thumbnails, video-editing, and so on. There seem to be hundreds of videos on how to "game" the Youtube algorithm, many of them with more than a million views. Spending a few hours on learning how to make forecasts doesn't seem inherently more difficult or less enjoyable.

Measure

- A measure to incentivize contributors through recognition is to introduce karma systems for types of contributions where objective measures are inapplicable.
- One might also import techniques from social media, like 'follow functions', to allow contributors to follow each other's work, and give each other well-deserved recognition for contributions.

External

Although social and intrinsic factors are sufficient to sustain many social systems, they often require the additional motivating factors of external rewards to get up and running. Once up and running, external factors also motivate contributors to make the system run effectively.

Monetary

An easy and effective way to motivate contributors is to pay them. Monetary compensation is a good way to create content of the exact sort one would like to see. However, monetary compensation is costly, and should be replaced by other forms of motivation when possible. Also, there is some risk that monetary compensation distorts the forms of judgment one would like to elicit for a Bayesian knowledge graph. For these reasons, monetary compensation is best suited for creating an initial critical mass of content to make the graph useful to decision makers, and perhaps for editing and moderating roles. One could also motivate users to contribute high-quality content, and do well in competitive rankings that are also useful to the community through prizes.

---

[88] Metaculus FAQ.
[89] See this comment for a good explication of this point.
[90] See Habryka's post on LessWrong 2.0 (for many nuanced reflections on how karma systems can be used to promote a desirable culture.

Measures

- Project based employment for contributing content for initial critical mass.
- Later on, stable employment for moderators.
- Prizes for desirable engagement with the system.

Career

Many EA's contemplate careers in effective organizations, including EA orgs. This group might be motivated by the idea of showcasing their knowledge and competencies. Through Solon, users may contribute to the epistemic progress of the EA community as a whole, and simultaneously represent the knowledge and beliefs of contributors. Since Solon can be used to assess what one is reading, usage over time tracks and displays the knowledge of contributors, and also the nuance of their thoughts concerning what they are reading. One could conceivably hook personal knowledge graphs onto the EA Hub account of contributors, which could then show a map of the beliefs of these individuals. One could also envisage programs for issuing certificates to users who have read and engaged with key EA literature.[91] It is easy to merely claim to have read and interacted with a research literature, and it is hard to assess how thoroughly someone has done so. However, with Solon, individuals can quite easily demonstrate their knowledge about a subject matter through their personal graph (evaluators can check and see if the estimates of the contributor are sensible).

Measures

- Represent assessments and contributions of individuals in a personal knowledge graph.
- Showcase the epistemic impacts of contributions.
- Issue certificates for having completed courses or having read and interacted with research literature.

Barriers

The main barriers to making contributions are shame, time and effort.

Shame

Another key barrier is shame. If there is a high bar for the quality of contributions to the system, some people might choose not to contribute because they don't think that their contributions can live up to the standards of the community. This is a real problem, and the Solon system in many ways accentuates it, and makes it easier for community members to judge each other on the basis of conflicting beliefs.

Mitigation

A way to partly mitigate shame is to allow users to contribute without their contributions being visible to everyone. One could make it so that users could contribute in a way that is only visible to moderators, and perhaps a few select friends, as well as contributing to the aggregate for the relevant claims and inferences.

[91] See a related discussion of this idea here: Certificates of impact - EA Forum.

Time and effort

In the Solon system, contributors analyse and/or assess the reasoning in a text while reading it. Analysing and assessing are distinct activities. Analysing a text involves the identification of key claims and arguments. Assessing the reasoning in a text involves giving quantitative estimates concerning the likely truth of key claims, and the strength of the arguments. For some texts, it is harder to sensibly assess the reasoning in the text. In other texts, it is harder to even follow the reasoning, but quite easy to assess it once one has found it. Luckily, most EA's have a very clear writing style in which conclusions typically appear at the beginning of a text, and in which arguments are typically inserted as easy to identify bullet points. However, some texts are of course harder to analyse and assess. Also, many EA's have some experience or at least some understanding of the practice of forecasting difficult questions, and so have a much easier time estimating probabilities than people in general.

Crafting forecasting questions is hard work,[92] and answering them in a nuanced manner is time-consuming as well. It should come as no surprise then, that analyzing texts, and rewriting annotated sentences found in texts as forecasting questions can be hard work. Moreover, giving reasonable estimations concerning the strength of arguments, and weighing them up against each other, is a whole lot more complex than merely reading about them.

Mitigation

- Promote clear writing so that the key ideas in texts are easy to understand and evaluate.
- Establish a culture for giving '*prima facie*' estimations (first impressions) of key claims and arguments. As Toby Ord puts it, estimates should indicate the right order of magnitude, not necessarily be an exact estimation of the precise probability of the correctness of a claim, or the rational strength of an argument.
- Include a high-quality onboarding process to reduce complexity and effort for users.
- Automate the process of analysis through language technologies.

Mess

Crowdsourcing projects quickly become messy. If a lot of people contribute to building a knowledge graph, it is likely to become messy. A common way to ameliorate this problem is by moderation. However, moderation can be expensive, especially if the culture of contributors isn't very good. There are technical ways to ameliorate the problems of mess as well, such as versioning control, similar to git. However, this too requires moderation.

Measures

- Moderation
- Versioning control through moderation

[92] At Metaculus, staff is responsible for crafting most of the questions.

Impact assessment

In this chapter, I attempt an impact assessment analysis of Solon using the importance, tractability, neglectedness framework. The first section argues that work on new ways to represent and interact with ideas is a neglected, but potentially very impactful. Moreover, in this section I also review the existing knowledge infrastructure in EA as well as some other institutions. I argue that Solon is a novel contribution that would not compete with, but rather complement existing infrastructure. In the second section, I consider the tractability of a successful project to build, grow and sustain Solon. I outline a plan for building and growing Solon, including a listing of costs, technological risks and uncertainties. In the third section, I list potential risks and benefits of the Solon system, including beneficial externalities of the project to build Solon. In the fourth and final section, I summarize everything, and add the elements of the impact assessment together.

Neglectedness

F

Tractability

F

Estimating while reading - A small experimental study

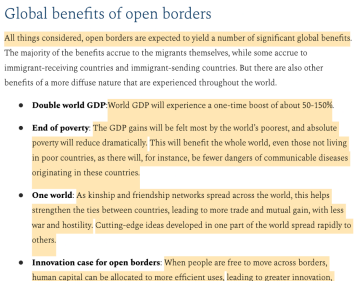
In the fall of 2021, I conducted a small experimental study to find out whether it would be time-consuming or cumbersome to estimate while reading (see section on elicitation above).

Experimental design

The structure of the experiment was as follows. I invited a bunch of friends and acquaintances to participate in a three-step experiment, expected to take a total of 2 hours. 14 people agreed to participate in the experiment, and 12 completed all 3 steps. Several participants were associated with EA Norway, some were friends from philosophy studies, and others were friends and acquaintances with no, or next to no familiarity with logic or forecasting. Participants assented to their data being processed and used for the purposes of this study in an anonymised fashion.

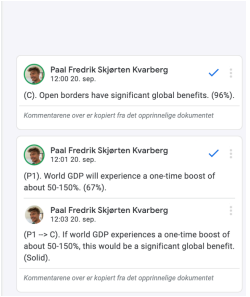
The tasks for the first two steps of the experiment involved analysis and estimation of Bryan Kaplan's blog posts on open borders. I invited participants to a google document with instructions and two tasks. Participants were divided in two groups, group 1 and group 2, and were given different texts to engage with.

The instructions briefly explained the concept of quantitative estimation, and inferential strength, and how to express credences of confidence and strength. Confidence was to be expressed numerically on a 0-1 range, and the strength of arguments was to be expressed in five natural language terms in an ordinal scale: very weak, weak, good, solid and waterproof. I demonstrated what I had in mind with an example.


Global benefits of open borders

All things considered, open borders are expected to yield a number of significant global benefits. The majority of the benefits accrue to the migrants themselves, while some accrue to immigrant-receiving countries and immigrant-sending countries. But there are also other benefits of a more diffuse nature that are experienced throughout the world.

- **Double world GDP**: World GDP will experience a one-time boost of about 50-150%.
- **End of poverty**: The GDP gains will be felt most by the world's poorest, and absolute poverty will reduce dramatically. This will benefit the whole world, even those not living in poor countries, as there will, for instance, be fewer dangers of communicable diseases originating in these countries.
- **One world**: As kinship and friendship networks spread across the world, this helps strengthen the ties between countries, leading to more trade and mutual gain, with less war and hostility. Cutting-edge ideas developed in one part of the world spread rapidly to others.
- **Innovation case for open borders**: When people are free to move across borders, human capital can be allocated to more efficient uses, leading to greater innovation,



The first task was the same for both groups. It was to read a short text outlining a very general case for open borders, and to assign credences to claims and inferences. The text included a main claim, and four overarching lines of argument and very compactly formulated summaries of these lines of argument organised in a bullet list. In google docs, I had already marked the conclusion, as well as the main arguments. The participant was then to judge the likely truth of the claims, and the strength of the arguments. The second task was similar, but in this case, the two groups received different texts to analyse, and had to identify the claims and arguments themselves. After having completed the two tasks, participants answered 9 questions about their experience doing the 2 tasks.

In the second part of the experiment, participants read and assessed 3 short texts similar to the ones from the first step of the experiment. One of the texts was taken from the analysis that participants had done themselves in step 1. Group 1 assessed texts analysed by group 2, and vice versa. After participants had assessed all three texts, they answered 7 questions.

Step 3 of the experiment was a semi-structured interview, a brief demonstration, and a looser conversation on the experiences of the participant. Questions focused on whether participants could see themselves estimating while reading, and what would have to be true for them to engage in the kinds of activities of the experiment. Questions were also specifically crafted to elicit the right interpretation of the answers to questions, and of the results. I also showed a prototype of how a knowledge graph could look like which displayed the estimations of the other participants. We then talked about the graph, and I asked if the participants would use a graph for research if there was a detailed graph for an issue they cared about.

Results

I haven't conducted a thorough analysis of the data, although I might if I think it'd be worth it. Here are some patterns in responses that are relevant to the goals of the study.

- Argument analysis
    - Most participants thought it was easy to analyse the texts in the experiment, and experienced this as quite straightforward.
    - In my assessment, the analyses of the participants were OK. There were minor mistakes here and there, but they mostly did a good job. Estimators who assessed texts with mistakes often pointed this out (sometimes they disagreed with me, but more often with the other participants).
- Estimation

    - Almost everyone agreed that it was hard work to assign numerical credences to claims and arguments. On average, participants gave estimation a ¾ on a 1-5 difficulty scale. There was some variation, with some participants using less than 10 minutes on step two, while others used more than 30 minutes. Also, some participants reported that it was frustrating to assign credences, while others thought it was a fun exercise.
    - Most participants gave fairly sensible estimations, and agreed on many claims. However, some particular claims were quite controversial.
    - Participants who were familiar with forecasting had an easier time assigning numerical credences to claims, though some were confused by the idea of assigning weights to inferences.
- Motivation
    - Most participants said that they probably wouldn't analyse and assess texts regularly.
    - Most participants said that they were open to submitting judgments on questions they knew something about, and cared about.
    - Most participants said that they might analyse a text they had themselves written if there was a chance that others might give feedback in the form of quantitative credences on it.
- Graph
    - When I showed participants the graph, almost everyone thought it was neat. Almost all participants said that they would use a graph like that for research if it was available, and had good content for a topic they cared about.

Limitations and confounds

There are several limitations to the study:

- Sample. There were only 12 participants, and the sample were friends and acquaintances of mine.
- The texts had a very clear argumentative structure, which meant that it was easier to analyse these texts than it is to analyse many other texts.
    - However, in the EA community, norms of reasoning transparency, including specific norms for expressing reasons in terms of bullet points or paragraphs are prevalent. This means that communication in EA often has a clear argumentative structure, which to some extent ameliorates this limitation.
- Participants were instructed to read the same set of texts on immigration and open borders. Many participants responded that they appreciated the texts, and thought it was interesting to read about.
    - However, the experimental setup involved a deadline (not strictly enforced, but still) for reading and assessing the content. And participants could not choose what to read and evaluate themselves. This meant that participants did not have the freedom of choosing time and topic for analysis and assessment, which they would in a real use-case of the Solon system.
- Participants had to read two heavy paragraphs of instructions on a form of assessment that most participants had never seen before, and also only got a single example to consider. In a real use-case, participants would have seen several examples of the

analyses and assessments of others before trying it out for themselves, perhaps making it easier to get into.
- Participants did not have the motivation of external recognition for their work in the experiment, since the setup promised that the results would be anonymized. This might have removed parts of the excitement of estimation from the participants, influencing their experience.

Conclusion

The clearest result was that participants who were familiar with, and interested in, forecasting and/or logic were positive, whereas participants who weren't thought the exercise was cumbersome. There are several limitations to the study, so I wouldn't update heavily on the basis of it. However, the study does constitute something like a proof of concept. Estimating while reading is not as pleasurable as reading without estimating. However, the kinds of people who like to read carefully and think through issues in more detail might be open to estimating while reading for literature they find engaging and know something ab
out.

Synthesis and tentative conclusion

f

Getting off the ground

Any software project has to have a plan to gain traction, and get off the ground. This is especially true for systems with network properties, where the value of the system is dependent on the number of participants and quality of contributions. To get the system started, one must be able to **build** it, or at least to build a minimum viable version of it. It is then necessary to get to a **critical mass** of participants and content so that the system is valuable to the users. After this, it is necessary to find a good way to introduce the system to the community. To find the right **beachheads**.[93]

Building Solon - Technological risk

Solon is a digital system with several advanced features which must be developed by a team of product designers and software engineers in close collaboration with relevant stakeholders in EA. In software development projects like this, the greatest risks of failure are market risks: that no-one wants to use the product, but there is also a technological risk involved. That the team is unable to develop the required functionality for the product in an efficient and satisfactory manner. The feasibility of development has a lot to do with the team, which I won't go into here, but also to do with product features. A relevant question to ask in this regard, then, is: are there any reasons to think that Solon will be especially hard to develop?

Solon has two main components: knowledge graph and browser extension for text annotation. A superficial reason to believe that the development of both of these components is tractable, at least in principle, is the existence of products with analogous functionality. However, it might be that the development of these products was quite hard, and expensive. It is famously hard to

[93] Some ideas in this section are inspired by the Arbital post-mortem discussion.

estimate the time and cost of software development projects, even for experts, and I am quite unsure about this myself. I currently tend to think that the required functionality is tractable for a reasonable investment, but I would appreciate any feedback on this.

Critical mass of content - Scraping and collaborating

In the Solon system, different types of users can draw value from it in different ways. Decision makers draw value from the content. In order for Solon to be valuable to decision makers, they must be confident that they might find content in Solon that is relevant and valuable to their deliberative processes. Content contributors, on the other hand, mostly derive value from signaling their expertise to decision makers. In order for Solon to be valuable to them, they must be confident that decision makers use Solon, and see their contributions. In other words, Solon is a two-sided marketplace of ideas. To get it off the ground, then, the first step is to bring a critical mass of relevant content onto the platform, so that it could start working as a resource to decision makers.[94]

The best way to do this, I think, is to start a project in which a paid project leader, and a group of volunteers, collect existing data from the EA information ecosystem, and manually plot estimates and link the central ideas of EA together into a Bayesian Knowledge Graph. This would involve Michael Aird's database for existential risk estimates, forecasts from Metaculus on EA-related questions, and also estimates from other prediction platforms. It would also involve a mapping of key arguments to support or undermine these forecasts. It might be good to try to map out some key documents, or books. Posts like Are we living at the most influential time in history? - EA Forum and Growth and the case against randomista development - EA Forum are good examples. Books could be *Doing Good Better*, *Superintelligence,* or *The Precipice*, which could serve to bind ideas in EA together. The point of this initial project would be to create the basic infrastructure of the EA knowledge graph, which would both serve an epistemic function, and which it might be interesting for EA's to consider, and contribute estimations to. This curated beginning to the graph would also constitute an example of how we would like the graph to look like, and how to build on it.

In order for Solon to be a valuable addition to the EA knowledge infrastructure, it is of paramount importance that the system can be integrated into, and work well in conjunction with, instead of in opposition to, related platforms. Questions issued at other prediction platforms, should, to the largest possible extent, be imported to Solon with links back to the original platform.[95] Moreover, the reverse should also be true. Other prediction platforms should be able to draw on Solon data. Solon should be seen as a distinct way to consider information. Another perspective one might take, when assessing some issue. Not as a contender to replace existing forums or platforms, which fulfill distinct roles.[96]

[94] See Arbital postmortem for a nuanced discussion of this.
[95] Such as, for instance, the predictions to the recent Nuclear Risk Horizons Project on Metaculus.
[96] As I mentioned earlier, I think the target user groups of prediction platforms is different from the group targeted by the Solon system. Existing prediction platforms target hard core power-users, whereas Solon aims to harness the beliefs and considerations of a wider base.

Beachheads

When core software infrastructure and content is up and running, it is time to introduce Solon to the community. There should be clear paths of entry that are likely to engage members. Here are some tentative ideas for how to do this.

Forum

The first thing to do once the platform is up and running is to announce on LessWrong and the EA Forum. However, it might be a good idea to try to get some publicity in other media as well, both within and outside EA. If there is at least one good example of a use-case where traversing the graph can be useful would go a long way towards telling a good story, and getting good publicity for Solon.

Volunteers

Another idea is to explicitly ask for help within the EA community. If we can demonstrate the epistemic value of the graph to decision makers, whether it is just a way to navigate topics to find relevant sources of research, or whether the overview of relevant arguments, and assessments of key claims and arguments are helpful in decision making, volunteers might put in some hours of contributions if the value is clear to them. Volunteer contributions can be a good way to kickstart Solon.[97]

Organized EA activities

If the idea is so good that volunteers might want to help, organizers of EA activities might be susceptible as well. One could ask organizers, volunteers or staff, to introduce the Solon graph, and use it for reasoning tasks in seminars, reading groups, or teaching, for instance through the EA virtual program.[98]

Mapping institute

At EAG 2021, Will MacAskill suggested we might set up a forecasting institute to "host, maintain and develop currently existing forecasting platforms since most existing platforms are relatively small and rely on the work of volunteers". If a high-quality graph seems useful, then it might be worth it to set up a research institute to make quality assessments, and map out intellectual territory of interest.

Importance

F

Other institutions

Improving Institutional Decision-Making: Which Institutions? (A Framework) - EA Forum.

[97] For instance through Impact CoLabs.
[98] For instance through Virtual Programs.

There is some reason to think that there are a few use-cases for the Solon system in other domains. A conglomerate of investment companies and foundations recently organized a hackathon with a $300k prize to create software along the lines of a Bayesian Knowledge Graph.[99] Also, Kauffman Fellowship, an investor-training program, applies graphs to complex decision making.[100] Moreover, The Decision Lab and TGG Group, associated with, among others, Philip Tetlock and Daniel Kahneman respectively, do some consulting.

Few companies take in the learnings of the frontiers of decision science the same way the EA community does, but as these examples show, there might be a market beyond EA for something like Solon. If this is the case, the project might, in time, pay for itself. If successful, then, the system might generate some revenue to pay salaries for a small team of moderators.

Improving democratic deliberation processes

An important factor to healthy democracies are good democratic debates in which many participate. See Towards a longtermist framework for evaluating democracy-related interventions - EA Forum for a thorough analysis of this from a longtermist perspective.

The Solon system could perhaps work as a deliberate mini-public for democratic deliberation. See Deliberation May Improve Decision-Making - EA Forum by Neil Dullaghan (2019) for a good introduction to the idea of deliberative mini-publics, and an evaluation and assessment along longtermist lines, including critique. along the lines suggested here. See criticisms of mini-publics here.

Tech for epistemics has been used in fora of this sort. See Polis - Opinion | A Strong Democracy Is a Digital Democracy and Kialo.

The European Union
Jun 16, 2021 Artificial intelligence, big data and democracy ID.

See also
https://www.founderspledge.com/stories/longtermist-institutional-reform
Medium Investigation: Democratic and Institutional Reform [EXTERNAL]
Longtermist institutional reform [EXTERNAL]

Data and AI

The Solon system creates many forms of useful data.[101] In this subsection I'll explore a few concrete use cases

[99] https://project4634147.tilda.ws/.
[100] Applying Decision Analysis to Venture Investing.
[101] The data from Solon could potentially be useful to other EA-aligned groups as well. One such group, is Ought. See their Automating reasoning about the future at Ought for a description of some of their plans, and insights into why Solon data could be useful to their pursuits.

Argument mining

Researchers in the interdisciplinary field of argument-mining apply machine learning (ML) to the concepts of informal logic. This rapidly growing subfield of natural language processing (NLP) can be divided into three subtasks: detection, segmentation and relation prediction, with each field requiring different parts of the NLP tool set.[102] The field is relevant to various applications in many research areas, with perhaps the most prestigious one being the holy grail of AI Research: an artificial general intelligence that can reason and think at a level comparable to, or better than humans (AGI). Various approaches have been applied to solve these tasks, including systems incorporating all the subtasks,[103] as well as several subtask specific applications.[104]

In recent years, some of the scientific innovations from this field have been applied in concrete use cases to improve public debate.[105] Use cases like these indicate great potential for improving public debate through argument technology, and the potential is far from exhausted. However, to unleash the true potential of such technologies, they must become more accurate.

In most areas of machine learning (ML), the bottleneck for increased accuracy is data. In argument mining, data scarcity is especially severe, due to the complexity and labour intensity of the annotation task. In addition, the available benchmarks have been dominated by narrow domain-specific tasks, i.e. a different set of labels for each new data set. This, combined with deep models that are hard to interpret and that suffers from the inability to handle long dependency concepts, makes argument mining a very challenging task. However, the advent of the transformer model,[106] and subsequent contextual word-embedding methodologies allows the development of models that are sensitive enough to catch the subtle nuances of informal logic with much smaller datasets than was previously possible. Moreover, theoretical work towards the development of an argument-interchange format,[107] and construction of a database for argument-mining corpora with interchangeable annotation schemes makes cross-label datasets available for the same training tasks.[108]

---

[102] For a review of the current state of argument-mining research, see Cabrio and Villata (2018) "Five Years of Argument Mining: a Data-driven Analysis", Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Survey track. Pages 5427-5433 and Lawrence, John and Reed, Chris (2019) "Argument Mining: A Survey", Computational Linguistics Volume 45, Number 4.

[103] Eger, S., Daxenberger, J., & Gurevych, I. (2017). Neural end-to-end learning for computational argumentation mining

[104] 18. Ruiz-Dolz, R., Heras, S., Alemany, J., & García-Fornes, A. (2020). Transformer-Based Models for Automatic Identification of Argument Relations: A Cross-Domain Evaluation..

[105] See Reason-Checking Fake News | November 2020 | Communications of the ACM, Prta: A System, and The Evidence Toolkit. See also: Argument technology for debating with humans. We may think of these initiatives as early machine protoforms of argument checking. See Stefan Schubert and ClearerThinking's Fact-Checking 2.0 | Philosophy, Logic and Scientific Method for more on this.

[106] A. Vaswani *et al.*, "Attention Is All You Need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008.

[107] See (Lawrence et al. 2012): Lawrence, John; Bex, Floris; Reed, Chris; Snaith, Mark (2012). Verheij, Bart; Szeider, Stefan; Woltran, Stefan (eds.). *AIFdb: infrastructure for the argument web*. IOS Press. pp. 515–516..

[108] See http://www.aifdb.org/search.

The process of analysing texts in Solon involves marking a part of text, a claim, and linking it to another claim, thereby forming an inferential relationship between the two claims (an argument). This process is analogous to the process of creating the type of linguistic training data that is needed to apply machine learning to detection and identification tasks in argument mining. We can design Solon to collect and store user generated data in a format that is readily available for machine learning. With techniques from argument mining, and data from Solon, we can then train models that may detect the claims and arguments in text automatically. Existing models are already useful, but with more data, they will become continuously more accurate, up until the point at which they can outperform and replace human analysis of these texts.

At this point, we have an artificially intelligent system that understands the linguistic patterns characteristic of informal logical structures, i.e. human reasoning. Such a system has several useful applications, including:

- **Analysing texts.** Detecting claims and arguments, so that contributors don't have to.
- **Summarising texts**. Representing the main line of reasoning in a Knowledge Graph.
- **Writing enhancement**. Giving writers an overview of the reasoning in their own texts.
- **Proto-argument checking**. Detects arguments, and highlights fallacies or possible fallacies.

Learning EA values and beliefs // Explainable decision support

The Solon graph represents the beliefs of the EA community, and consists of a range of values on a bunch of issues that are related to each other. Data of this kind can be used to develop artificially intelligent systems for decision support, which reflects the values and beliefs of the community.

AI driven decision-support systems have been derided for being black-boxes, and for relying on biassed or otherwise untrustworthy forms of reasoning. The clearest example of this is the judicial decision-support system Compas, which, due to the nature of data on which it was trained, predicted twice as many false positives for recidivism in black offenders than for white offenders.

Besides judicial applications, artificially intelligent decision-support systems can be used to amplify decision making in health, finance, science and politics, to pick some. However, in all of these important areas, it is necessary to trust the systems that are used. Analogously to the way we hold each other accountable, we need a way to hold AI systems accountable.

In the *Critique of Pure Reason,* Immanuel Kant explained the core of agency as being a *transcendental unity of apperception*. In plain english, the idea is that the defining feature of agency has to do with thinking before acting and weighing relevant considerations before believing. In this middle step between input and output, humans may stand back and draw on the full range of their beliefs and worldview in order to assess and interpret the input, in order to reach sensible conclusions. This step, which is lacking in the immediate and instinctual decision-processes of animals, and the simple rule based processes of most machines, is perhaps the very essence of agency as we know it. To develop a responsible machine agent, it is necessary to develop a machine that may stand back and reason critically about how it should interpret it's data. We are just now beginning to develop machines with such capabilities. However, the AI

systems that most closely resemble agents in this sense are black boxes, whose rules for converting inputs to outputs are opaque to us. To have an agent in which we can trust, it is necessary that it can explain the line of reasoning behind it's advice. Through the Solon system, we collect data of a kind that can be used to develop machine agents of this kind.

The Solon dataset consists of claims and inferences with numerical values indicating credences. If the dataset is large enough, and covers main topics in EA, for instance, one could query EA related issues and get answers that reflect the values in the graph. We could, for instance, ask the system: "is inheritance tax a good idea?", and Solon could say something like: "yes, if taken to mean X, **I think it is**. Would you like to know why I think this?", if the user then said yes, Hylas could say: "The best arguments for believing this are X and Y, and the most consequential consideration is Z. I think Z can be rebutted because that argument relies on K being true, and K relies on P and Q, which are speculative at best. Moreover, X and Y have solid foundations, and are widely accepted by domain experts. Would you like to delve deeper into any of these considerations?" From here the user could choose how to proceed with the issue. The important takeaway is that Solon would have a point of view (reflecting the graph of course), and would be able to defend that view, and be held accountable for the implications of those views.[109] Hylas would in this sense have **agency** in the sense of being aware of **reasons**, and it would be possible to learn how it thinks, and thereby to build **trust and understanding**.[110]

Values are rarely explicitly spelled out in discussions, and are unlikely to form separate nodes in the Solon network. However, even though values are rarely spelled out explicitly in argumentative texts, they are tacitly expressed through inferencial relationships. The estimated strength of arguments according to the form: 'Measure X is expected to result in K QALY's, therefore we should do X instead of Y.', expresses a consequentialist leaning. The values assigned to ethical inferences is an expression of values, and can be learned by a machine. This is a way to train a system for decision-support human values.

Parameter optimization through reinforcement learning

A problem with teaching human values to machines is that there isn't a consensus on what the human values are. Although we might just want to train machines to follow whatever values humans tend to follow, it might be interesting to explore whether algorithmic methods can be used to promote epistemic values of the graph as a whole.

An idea would be to maximise key epistemic principles on which there is wide consensus.

- A key property of some claims is that they are either true or false.
- Coherence. An epistemic virtue of belief networks is coherence, which should be maximised if possible.
- Occam's Razor, i.e. the theoretical virtue of simplicity.

[109] It would also be possible to see how the experts and laymen who have registered their beliefs about these issues think.

[110] By keeping a score of how often a user agrees with its evaluation, Hylas would be able to constantly update both the way advice is given, and the knowledge graph itself. For the user, this would feel like coming to terms with the system, and it takes in one's perspective, moderating its own view on the background of discussion, just like a sensible discussion partner.

- Theoretical conservatism.

In Solon, numerical values represent the aggregated beliefs of the EA community. Using reinforcement learning techniques, we may test a wide variety of distributions of credences on claims and inferences, and see how best to organize the values of the graph to optimize a set of epistemic parameters, like the ones mentioned above, for instance. It would work like this. All credences in the graph are replaced by random numbers, where all binary propositions are given exact positive or negative values. The resulting random distribution is then automatically evaluated through a scoring system, involving a chosen set of epistemic parameters (for ones above for instance). If the score is higher than any earlier attempts, the new graph is saved as the best found to date. This process is then repeated until the machine finds an optimal distribution that harmonizes with our chosen epistemic principles. There is some reason to think that this process could have epistemic value, but I am very unsure about this.

Conclusion and summary of considerations

In this chapter I have presented the case for Solon. In the first section, I considered several objections to the idea that Solon would improve decision making in EA, and argued that Solon would contribute positive epistemic value even though some of the concerns could not be entirely mitigated. Next, I considered whether it would be feasible to build and sustain Solon, and argued that it would, if we took extra care to design the system in a way that catered to the motivations of contributors. I then wrote about several other considerations, including long-term effects of the system being adopted and used by other institutions, and in the public sphere, as well as potentials for being used to train AI systems human values.

In this section, I try to summarize the most important considerations that bear on the evaluation of Solon. I try to break overarching considerations or strands of argument into concrete benefits, downsides or risks.

System Benefits

- **Improve decision making**. First and foremost, Solon is meant to improve decision making in EA through a Bayesian knowledge graph that is specifically designed to aid reasoning in accordance with the epistemic norms of the EA community. These are the mechanism by which I think Solon could improve decision making:
    - **Overview**. Easy to navigate map of key claims and arguments in EA
        - **Research**. More effective research for all members.
            - **Arguments**. See overview of the most important arguments bearing on a claim.
            - **Navigability**. Navigate intellectual terrain without understanding keywords.
            - **Standardisation**. Simplified formats for representing ideas.
        - **Onboarding**. Effective and engaging onboarding for new members.
        - **Paradigm**. Shared terminology and understanding.

    - **Aggregating judgments**. Eliciting and aggregating judgments on key claims and arguments has at least two beneficial implications.
        - **Accountability.** Contributors can be held accountable for their beliefs.
        - **Training**. Contributors who make judgments get good forecasting practice.
        - **Accuracy**. The aggregate of multiple judgments is more reliable than single judgments.
        - **Value of information**. Patterns of belief indicate the value of questions and results.
            - **Value of research questions.** Confidence, disagreement and consensus in beliefs is relevant to the value of research questions.
            - **Value of research results**. Tracking of belief revision can be used to measure the value of research results.
    - **Updating**. Bayesian updating, reverberating throughout the network.
    - **Personalised Nudges**. Personalised nudges for improved epistemics.
- **Representing community knowledge and belief**. Secondly, the system is a platform where EA's may contribute in a way that is useful to the community, while also demonstrating their knowledge and expertise to the community, and others. In other words, Solon may serve to reveal facts about the members of the EA community, as well as properties of the group as a whole. This has two types of benefit:
    - **Contribution**. Distinct way to demonstrate knowledge and expertise for contributors.
    - **Recruitment**. Distinct way for EA individuals and orgs to find collaborators on projects
- **AI**. Thirdly, Solon is designed to elicit data of a sort that has applications in AI, including:
    - Argument mining.
    - Explainable decision support systems
- **Other institutions**. Furthly, Solon could improve decision making in other institutions, including:
    - Other philanthropic organizations
    - Government agencies
    - Private companies
    - The public sphere

System Risks

There are also risks associated with a system like Solon.

- Due to information cascades and related worries, Solon serves to increase groupthink and other negative epistemic biases in the EA community.
- The openness implied by the system as a whole leads to stigma and shame and other negative emotions for members of the community.[111]
- Outsiders might be put off by a clear overview of the beliefs of the members of the EA community. The typical EA contributor might have somewhat absurd views, which, if brought into the light, would be off-putting to others.

[111] See Prediction Markets in The Corporate Setting - EA Forum for a more nuanced discussion on this.

- If researchers from EA orgs clearly represented the view of their organization in Solon, differences would be more apparent, which could lead to bickering, more criticism and less cooperation.
- Programming mistakes in Solon might lead to wrong calculations in the Bayesian network, and could potentially lead decision makers to make the wrong decision.

Project Costs

However, building Solon, and getting it off the ground, is likely to be quite costly. The entire project, including opportunity costs for unpaid volunteers might cost between $0.5-2M, or more.

Project Risks

Moreover, there are several risks associated with a project towards building and sustaining something like Solon, including, but not limited to the following.

- The project team might fail to collaborate with other actors in this space, resulting in competition that is deleterious to the field as a whole.
- The idea is much more costly to develop and maintain than anticipated, so much so that the implementation is netto negative.
- Execution might otherwise fail, and this might discourage other attempts to proceed with other valuable projects in this space.
- The base rate for projects like this is quite low, as there are many who have tried something along the lines suggested in this document, without succeeding. However, there are some projects that have succeeded as well, including GJP.

Project benefits in case of failure

However, even if such a project fails, there might still be some value to the project.

- Informational value
    - As the analysis in the first chapter shows, there are several related initiatives in this space that point in the general direction of something like Solon. Even if the project fails, it will at least have explored the two ideas underlying Solon, and might give valuable insights into the problems with projects of this kind.
- Experience to the project team.
    - Whoever ends up working on the project is likely to be value aligned, and so is likely to work on other EA-value aligned projects in the future. Even if the project fails, the team would have gained valuable experience which makes it more likely to succeed in another project in the future.

Concluding remarks

Building and sustaining Solon, or something like it, is an enormous challenge. Many people have tried to do something like it, without reaching the full potential of the vision underlying projects like it. In a 2019 Long Term Future Fund grant writeup, Habryka notes

> *I do think that the road to building knowledge aggregation platforms will include many failed projects and many experiments that never get traction; as such, I do think that one should not*

over-update on the lack of users for some of the existing platforms. As a positive counterexample, the Good Judgment Project seems to have a consistently high number of people making predictions.

The road to