Governance and Epistemics resources

Governance
Given our current best understanding of the universe, and our preferences, what collective actions should we take next?
What policies should we establish and enforce?
What projects should we fund?

Epistemics and Truth-seeking
How should we go about understanding the truth about the universe?
How should we resolve disagreements about what is true in a way which leads to improved accuracy?
Preference Understanding
Given that the universe contains us, what are our values/preferences?
Given that it seems that we do not always fully understand our own preferences or the preferences of others, how can we figure out our own and others preferences more efficiently and accurately?

Governance

A post with motivations for trying to improve governance
https://www.lesswrong.com/posts/SCs4KpcShb23hcTni/ideal-governance-for-companies-countries-and-more

An interview talking about the broken state of modern bureaucracy
Dominic Cummings - How Dysfunctional Govt Killed 1000s in COVID

Less well known stuff

Putting less known stuff first, as less well known stuff is more likely to add to your knowledge if you've already been tracking some of this.

Minimum Partial Consensus Voting

Fair group decisions via non-deterministic proportional consensus
Jobst Heitzig, Forest W. Simmons and Sara M. Constantino Springer Nature 2021
paper: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3751225

code: https://github.com/pik-gane/vodle

pros:

- pretty great at finding a good, fair decision that will satisfy the majority of voters the majority of the time
- can automatically find better solutions than would be immediately obvious

cons:

- complicated, so a harder system for voters to trust than a simpler system like approval voting (also harder to explain, to operate as the electoral body, etc.)
- involves some randomness in outcome. I think this is a good thing, but some people find it upsetting to have a voting method which is nondeterministic given a set of input votes. Indeed, you need a trustworthy overseeable auditable source of randomness pre-agreed upon. Something like the 7th digit precision of celsius temperature reading at a specific weather station, that everyone will be able to check for themselves and which is protected from tampering.

YouCongress

https://www.lesswrong.com/posts/4KjiZeAWc7Yv9oyCb/tackling-moloch-how-youcongress-offers-a-novel-coordination
https://youcongress.com/
Transferable, aggregatable representation authority with a mix of human and AI representatives

polis-alternative-that-uses-preference-ranking

argh, I forgot the name of the site and now can't find it again
pros:

- polis alternative, but takes into account relative rankings of policies, thus gets at a finding preferences of people in a better way that Polis. Polis asks whether you agree with a given statement. Pair-comparison presents you with two options and tells you to select the preferred one (or declare them equal, or invalid).
- like Polis, it aims to gather a large number of user submitted proposals, which are intended to change in distribution to include more compromises as the voting progresses and commonalities/blockers become revealed.
- can be fruitfully combined with Polis. Indeed, a joint system would probably be ideal
- gives you a sort of ELO ranking of proposals

cons:

- if you weren't able to come up with a combined system, I'd maybe just use Polis
- there is not an easy way to trade-off some aspects of things like the exclusiveness of compared proposals. Exclusiveness of pair: to what degree can two proposals not both be done?

- also doesn't necessarily include the important aspect of costs (financial and otherwise, like externalities such as pollution) of various proposals. If you don't include that aspect manually, then you find that expensive proposals with more upside might do better. You could try to control for this afterwards with some sort of cost ranking, but that doesn't do a good job of taking people's opinions about externalities into account.
- what about as-of-yet uncertain or completely unknown externalities and costs? budget overrun probability? success probability?
- unlike quadratic voting, doesn't have a mechanism for minorities with strong narrow opinions to stand up to majorities with broader weaker opinions.

Champagne voting

How I built the SoME3 voting system with graph theory
GitHub - fcrozatier/champagne
"Champagne is a voting platform for massive competitions. It implements a peer ranking algorithm where people vote by comparing pairs of entries to let the best ones bubble up the surface."
pros:

- I think this is potentially a good system to be combined with others. For example, what if this was used for a pre-primary system, to nominate the primary candidates for a political party? It would allow the party to have thousands of candidates in the initial pool which would then get narrowed down to 10 or so for the primary. Then another method would select the 1 to win the primary, maybe MPC voting or approval voting.
- another good use could be for deciding between a large number of drafts of proposals for a given issue. if thousands of people each review just two or three drafts, or even portions of drafts, you could then do math™ on their preferences to complete the graph

cons:
not everyone gets to vote on everything, so it's not ideal as a full election system on its own

TrueSkill ranking system
https://trueskill.org/

WeBuildAI

WeBuildAI: Participatory Framework for Algorithmic Governance: Proceedings of the ACM on Human-Computer Interaction: Vol 3, No CSCW
community algorithm building, to create a 'digital representative'

Personal AI Advocates

AI agents which are able to represent your values and needs in a discussion about legislation, and make voting recommendations to you.
Personal Value Model Representatives

well known stuff

Polis

https://github.com/compdemocracy/polis
https://www.openrightsgroup.org/publications/democratic-innovations-polis-and-the-political-process/

pros:
- good for seeking common ground between polarized groups, thus good for making peace and finding sensible compromises
- seeks to gather proposals (opinions) from people as the voting progresses, encouraging users to come up with new viewpoints that might be liked or disliked by everyone. Views liked by everyone can rise up as important compromises. Inspires creativity in this way.

cons:
- bad for highlighting diversity of opinions and seeking pure truth (maximal correctness) rather than common opinions
- not a decision making method in itself, but a sort of precursor. A way of brainstorming potential compromise solutions which could then be voted on as policies.

approval voting

pros:
- simple to understand how to vote (easy to explain, thus helps with voter trust, and usability in elections where a subset of the voters have low education)
- simple to score (helps with voter trust in the system b/c unlikely to be mis-scored by sneaky shenanigans)
- deterministic given the fixed set of placed votes (as compared to systems designed with a random element like mvc)
- simple to operate as an electoral body (e.g. makes this a good choice for informal votes in small groups)

cons
- there are some strictly superior methods like MPC
- only slightly better than ranked choice, so there is an ongoing debate about which of these two methods to use

- doesn't capture the complete picture of voter preferences such as their ranking over options (e.g. ranked choice), or quadratic value preferences over options (e.g. quadratic voting)

quadratic voting

https://www.radicalxchange.org/concepts/plural-voting/
pros:

- limited in use, but good for certain choices, especially around funding when you already have a fixed set of items which are the potentially fundable proposals, and a fixed budget available to be allocated.
- potentially a good way to allocate community resources, with the price for a given resource going up exponentially as you take more of it? Or for producing externalities? There are various issues with these ideas still to be worked out.

cons:

- some fundamental limitations such as needing the set of things and the allocatable budget to be pre-decided by some other method.

ranked-choice voting

pros:

- better than first-past-the-post

cons:

- I prefer approval voting in most cases for a variety of reasons, a big one being its simplicity for understanding how to do it and how the results are calculated. I think the weirdness in scoring ranked-choice voting can lead to counter-intuitive results and thus impairs people's ability to trust it as an electoral method.
- If you want a complicated and strictly superior-to-approval-voting method, then min consensus voting is my go-to rather than ranked-choice.

futarchy

pros

- good for epistemic health, rewarding being correct over group-think or loud-malcontent-proclamations-drowning-out-sensible-opinions
- similar competence boost as meritocracy, but inherently harder for corruption to capture

cons

- complicated to implement well on a broad scale (small experiments could be done)
- untested in practice
- potential for weird edge cases and manipulations which may need specific implementation rules derived from real world testing to overcome. Not ready for large scale, high-stakes deployment like Approval voting is.

https://cip.org/whitepaper

Meritocracy

some system for selecting the most competent people to govern specific areas in which they can be shown to have expertise
pros

- higher level of competence, if the selection process works

cons

- hard to avoid corruption capture of the test mechanism

Sortition

taking a representative random sample of people from the population being governed, and pay them to make some of the decisions.
reliance on people of average competence and education, rather than the best

- pros
    - harder for corruption to capture
    - better representation of the population's values
- cons
    - lack of expertise
    - regression to the mean of competence and education
    - may do a poor job of selecting legislation which will actually have the effects they desire on realizing their values in the world

Other and notes

combo of meritocracy and sortition
could you take a random sample of grad students working in the general area of the decision being made? then you get a higher mean level of education and competence

https://citizenos.com/

https://deliberation.stanford.edu/

https://newpublic.org/directory
https://airtable.com/appgA6QrWMpXmDp9X/shr7JFPTJt1C9j0qA/tblwlRhU0Icd8p5iJ

https://www.radicalxchange.org/

Delphi Method

Loomio
https://github.com/loomio/loomio

Consul
https://consulproject.org/
https://github.com/consuldemocracy/consuldemocracy

Decidim
https://decidim.org/
https://github.com/decidim/decidim
https://xabier.barandiaran.net/research/technopolitical-autonomy/

Using LLMs to extract pseudo-polls of opinion from social media
https://arxiv.org/abs/2309.06029

a nice community budget deciding platform.
https://ethelo.com/how-it-works/

https://cip.org/whitepaper

stratified random selection
https://selection.newdemocracy.com.au/

https://wwviews.org/the-world-wide-views-method/

https://www.peopledemocracy.com/

https://www.lesswrong.com/posts/MDQnDGEKQuCAggRLe/histocracy-open-effective-group-decision-making-with

---

Governance involves certain key factors

- taxation and resource allocation (collecting taxes, spending on public projects)
- rules and enforcement (a monopoly on force, or non-physical coercion, like seizing bank accounts)
- collective bargaining and compromise (what rules to make? how to structure enforcement? what taxes to collect? what public works to fund?)

Variations on the federated archipelago concept
Imagine you have an overarching government which manages a minimal set of rules (anti-violence, property rights) but is minimalistic (libertarian).
Then imagine you have some 'partial governments' within it. These organizations wouldn't have the right to use force, but could exert other sorts of coercion. For example, they could have membership rules that involved using a community bank and agreeing that the organization had the right to seize your funds in case of certain rule violations.

There could be public works which allowed for both members and non-members to use them, by charging a premium pay-per-use fee to the non-members. This could be made fairly seamless with use of an id card, and associated biometrics like hand/finger scans and retinal scans, and memberships registered by this id. You don't have to be 'on file' with the org if you aren't a member, you just declare you aren't a member and pay the per-use fee.

There are already things like this with things like season-passes for public transit. I'm just imagining taking this further. A collective org could have lots of shared goods, like forms of transport from trains to rental cars to electric scooters, healthcare, education, pensions.

Some things are more 'real estate'-bound or 'land-bound' than 'person-bound'. For instance, fire stations and fire insurance, ambulance and emergency care, other emergency responses like earthquake, urgent physical-force policing. Or, on the more routine side, plumbing and energy infrastructure.

What about vehicle insurance? You'd need the provider to have access to information about the individual driver, and the context of the drive. You could send a notice to the user's phone about how much insurance and rental fees would be for various available cars. If a young driver wants to rent a fancy fast car, that'll cost them a lot more per mile than renting a frumpy unstylish car with limited acceleration and top-speed-governor installed.

Vaniver13y250

I like the idea of measurement. The problem, though, is that you get what you measure, not what you wanted to measure.

Suppose Charlie is risk-averse, and only approves projects with a 95% chance of meeting expectations. David is risk-neutral, and will approve projects that have a positive EV that are significantly higher than other available projects. Oftentimes, they're speculative and will only exceed expectations about 10% of the time, since they only have about a 10% chance of succeeding.

Charlie will get about nine times as many votes as David, eventually. If David votes against Charlie's projects as too bland and too low EV, this will go *even worse* for David, as eventually only Charlie's projects will be approved and David will be recorded as pessimistic on all of them.

Decision-making is not a logistic regression problem, and so I am pessimistic about logistic regression approaches applied to it. I agree that measuring decision-making ability is a very important task, but approaches like Market-Based Management seem far more promising.

[-]HonoreDB13y20

If the organization is risk-averse, it doesn't want risk-neutral voters to gain influence. If it's risk-neutral, then it should incorporate opportunity costs when judging projects in hindsight. Furthermore, if in hindsight a rejected project still appears to have had a high positive EV, the org should register the rejection of the project as a mistake.

Reply
[-][anonymous]13y
90

Suppose the organisation is risk-neutral, and Charlie abstains from the sub-95% chance projects rather than rejecting them (in a large organisation that makes many decisions you can't expect everyone to vote on everything). He also rejects the sub-5% projects.

By selectively only telling you what you already knew, Charlie builds up a reputation of being a good predictor, as opposed to David, who is far more often wrong but who is giving actual useful input.

Reply
[-]Vaniver13y
20

> Furthermore, if in hindsight a rejected project still appears to have had a high positive EV, the org should register the rejection of the project as a mistake.

This misses the heart of that criticism: mistakes have different magnitudes.

badger13y110

Your tease excited me since I recently started grappling with this issue. Unfortunately, I'm underwhelmed. If the group deals only with binary decisions, participants have a single underlying competency, participation can be suitably restricted, participants don't have strong biases, decisions can be reliably assessed right or wrong, etc, then you have an elegant solution.

> There are some clear advantages to histocracy over futarchy: most relevantly, I believe histocracy will work well on a small scale, while prediction markets require a large crowd. Given enough time and participation, histocracy will inevitably beat a market. There's less moral hazard, and less vulnerability to manipulation.

These claims seem completely unfounded. Prediction markets don't require a crowd. If implemented through a market-maker, you can get by with a single participant. PMs have issues, especially when used to make decisions, but this proposal is rife with manipulation opportunities -- accumulating competency and "spending" it to sway decisions, manipulating if a decision is counted as a success or judged at all, altering the order of decisions to accumulate competency or harm that of others, collusion to build competency of at least one individual (worthwhile since the weights are convex). The worse manipulation of prediction markets I'm aware of that wouldn't also apply to this is for traders to mislead others for later profit, which wouldn't affect the final probabilities used for decisions.

Besides PMs, there is work being done in Bayesian truth serum, peer prediction, and other collective revelation mechanisms that don't require verification of results for scoring, but still result in truthful answers.

DanielLC13y50

I think you could do better if you look at correlations.

For example, if Alice and Bob always give the same vote, then their combined votes should be exactly the same as if the vote from one of them if only one was on the board.

Another possible improvement: Instead of voting yea or nea, you give a probability. If you give a higher probability, your vote counts more, and it helps you more if you're right, but it hurts you more if you're wrong.

Nathan: I think this Value Kaleidoscope could be quite useful in producing a 'Big 5' style test, but for values rather than personality.

Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties

Authors: Taylor Sorensen, Liwei Jiang, Jena Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, Yejin Choi

Abstract: Human values are crucial to human decision-making. Value pluralism is the view that multiple correct values may be held in tension with one another (e.g., when considering lying to a friend to protect their feelings, how does one balance honesty with friendship?). As statistical learners, AI systems fit to averages by default, washing out these potentially irreducible value conflicts. To improve AI systems to better reflect value pluralism, the first-order challenge is to explore the extent to which AI systems can model pluralistic human values, rights, and duties as well as their interaction. We introduce ValuePrism, a large-scale dataset of 218k values, rights, and duties connected to 31k human-written situations.

Epistemics and Truth-Seeking

not much here yet. I've seen some interesting things, but not recorded them

prediction markets (see also futarchy)

self-prediction-tracking
calibration training

examining and describing the trustworthiness of scientific findings
better meta reviews
more data sharing

https://m.youtube.com/watch?v=OeQR0pyp7HM

value discovery

Exploring my personal values
https://docs.google.com/document/d/1OUjSI9_gJYQzw6PiMnUFi44O6V-avu7SYn2_CIM65Mg/edit?usp=drivesdk

Values of humanity
In order to be able to govern in accordance with the reflectively endorsed values of the constituency, these values must be accurately discovered. Democratic governments do some of this via things like polling. I think it could be done much better. This is a science which can be improved.
I am interested in trying to build models both of the space of values of all humans, and the subset of those values expressed within a particular human. I think there could be value in a governance system which had a debate-agent for each constituent, trained on that person's values. These agents could debate in a giant congress to find the best available compromise.
https://www.lesswrong.com/posts/QuuuxHwvHiEwcjaDd/constituency-sized-ai-congress
https://www.alignmentforum.org/posts/BDTZBPunnvffCfKff/uncovering-latent-human-wellbeing-in-llm-embeddings

What Would Jiminy Cricket Do? Towards Agents That Behave Morally
https://arxiv.org/abs/2110.13136

Think again - Adam Grant
https://stitcher.simplecastaudio.com/aa9f2648-25e9-472a-af42-4e5017da38cf/episodes/4c107e4c-89c2-4c97-afd0-afeed10db009/audio/128/default.mp3?aid=rss_feed&awCollectionId=aa9f26

48-25e9-472a-af42-4e5017da38cf&awEpisodeId=4c107e4c-89c2-4c97-afd0-afeed10db009&feed=N5eKDxJI

Appendix

Links and info

https://www.lesswrong.com/posts/cJv8rBSshrR82NRET/a-case-for-superhuman-governance-using-ai post by Ozzie Gooen

Some comments from Holden's LessWrong governance post

comment on the ideal governance post by aviv
"the work of Claudia Chwalisz and her team at the OECD, which I've found immensely valuable—not only their excellent reports, summaries, etc. but an Airtable documenting hundreds of real-world governance experiments "

lukeprog

Some other literature OTOH:

- Collective Reflective Equilibrium in Practice
- Not "ideal," but exploring what's possible: *Legal Systems Very Different from Ours*
- There's a pretty large literature on various forms of "deliberative democracy," e.g. see here and here
- I would guess there's been interesting discussions of ideal governance in the context of DAOs

topherhunt

the minor but growing trend towards self-management organizational structures, teal organizations, Holacracy, or Sociocracy.

I have some experience with Holacracy, and while I would never call it a cure-all, I feel strongly about the relevance of its driving principles to the question of what an ideal governance system would look like -- eg. a structure of nested units/teams with high levels of local autonomy, a

unique method of making governance decisions on how to change said structure, mechanisms that privilege "moving forward" over "inaction due to conflictual gridlock", fluid process for defining and appointing power-holding "roles" to individuals, etc.

Jason Brennan

I'd like to try enlightened preference voting in Denmark or New Hampshire.

How it works:
1. Everyone votes for their preferred thing (whatever is being voted on).
2. Everyone somehow registers their demographic data.
3. Everyone takes a 30-question quiz on basic political information.

With 1-3, we then estimate what a demographically identical public would have voted for if it had gotten a perfect score on the quiz. We do that instead of what the majority/plurality actually voted for.

There are lots of details here I'm not getting into, but that's what I'd want to try. No one's done it to actually decide policy, but researchers have been doing this in labs for a long time with good results.

Nathan HB

Epistocracy as described by Jason Brennan is similar to things I've been thinking about. Also sortition, which is related but different. Here are some of my thoughts:

If we have a system of 'extrapolating volition' of segments of the population by taking representative samples of people and paying them generously to learn and think and debate on a key issue for a couple of weeks, then give their best full answer as to what should be done... And then also quiz them on their factual understanding of relevant systems, and register their predictions about outcomes... I think this is in some sense a more 'fair' look at what that segment of the population would think if given plenty of time to think.

Then from there we can make a model which lets us average these opinions and extrapolate (compensate for biases in the predictions, etc.), but in order to check that we've 'extrapolated well' we should check our extrapolations with different members of that population segment. This process can be repeated several times until we get good agreement on the extrapolation. I think extrapolating is a good idea, but I worry it would be vulnerable to abuse if you didn't then have the step of the people who are being extrapolated for approving of the extrapolation.

Also, I have some thoughts along the same lines of Jason Brennan's criticism of democracy as being 'a system of pushing around the minority.' Growing up in the liberal Quaker tradition, I've spent a lot of time in groups of people trying to do decision making via consensus. The quaker

consensus process allows for some small portion of the group to 'stand aside' to allow a decision to go through that they disagree with, but this is uncommon. When it does happen, it's usually less that 5% of the group. It occurs more often in groups of over >200 people than in the more typical consensus-seeking groups of 20-100 people.

I feel like the insights from this process for me are that it doesn't scale well, and takes a lot of effort per-decision even when it works, but does have some really nice properties of finding better agreements. Often, the time taken to hear everyone's point of view and share evidence thoroughly means that better solutions are found than anyone even came to the meeting with in the first place. So, this consensus process could inform our governance process indirectly, through finding related techniques which work with larger groups (for instance the Polis Process for crowd-sourcing consensus mentioned by Audrey Tang on the 80k hrs podcast) or by selecting small representative-population-sample committees of 50-100 people and having them go through a consensus process to draft a statement of group opinion on a topic. I think this is valuable in addition to the 'individual opinion sample' mentioned above, because my experience is that people who talk face-to-face with each other on a contentious issue in the context of consensus process tend to come up with better (more win-win) and more empathetic compromises through understanding each others' points of view and deeply felt emotions. It could also be a useful process for a small decision-making group like a board of trustees.

https://adelaybeingreborn.wordpress.com/ophelimo/

The Structure of Voter Preferences
Benjamin Radcliff
https://doi.org/10.2307/2131996
https://www.jstor.org/stable/2131996

Abstract

Economic models of voting typically assume the transitivity of individual-level preferences. Other conditions, such as single-peakedness or dichotomization, are also sometimes postulated. Despite the ubiquity of these assumptions, there is a paucity of empirical tests using real-world elections with mass electorates. Using CPS data, I address these issues in the context of U.S. presidential elections from 1972 to 1984. It is maintained that (a) the traditional assumption of transitivity is empirically plausible, even with a large number of alternatives, and while preferences are (b) predominantly not dichotomous, they do (c) tend to be single-peaked along a traditional ideological dimension.

ALIGNING AI WITH SHARED HUMAN VALUES

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, Jacob Steinhardt

ABSTRACT

We show how to assess a language model's knowledge of basic concepts of morality. We introduce the ETHICS dataset, a new benchmark that spans concepts in justice, well-being, duties, virtues, and commonsense morality. Models predict widespread moral judgments about diverse text scenarios. This requires connecting physical and social world knowledge to value judgements, a capability that may enable us to steer chatbot outputs or eventually regularize open-ended reinforcement learning agents. With the ETHICS dataset, we find that current language models have a promising but incomplete ability to predict basic human ethical judgements. Our work shows that progress can be made on machine ethics today, and it provides a steppingstone toward AI that is aligned with human values.