Context: Concepts & artifacts for AI-augmented democratic innovation: Process Cards, Run Reports, and more

---

Process Card: Global Brain Deliberative Consensus

What is a process card? Process Card Template

Process Overview

Basic information

- **Process developer:** Social Protocols (Team: Jonathan Warden, Felix Dietze, Johannes Nakayama)
- **Process design dates**: May 2024
- **Process design version**: 0.2
- **Process card version**: 0.2
- **More information**: https://github.com/social-protocols/jabble https://social-protocols.org/deliberative-consensus-protocols/
- **Maturity**: prototype in the process of real-world evaluation
- **Tools and methods used**: Bayesian Statistics, Game Theory, Information Theory, Machine Learning
- **Where to send questions**: mail@social-protocols.org

Intended Use

- **Primary intended use**: Fact checking based on analyzing discussion trees. Estimate the **informed**, **de-biased** judgment of a group about the fact.
- **Primary intended users**: Fact-checking websites. Self-governing organizations. Any organization looking to aggregate knowledge of members.
- **Primary intended context**: A decision needs to be made based on the opinion of a group, but users are concerned that:
    - Participants may not be well-informed
    - Participants may be biased (e.g. ideologically motivated)
- **Out-of-scope use cases**:
    - Analysis of content without vote data
    - Cases where vote data cannot be obtained or cannot be anonymous
    - Argument mapping, formalized debate approaches
    - Currently: questions of **preference** and not **fact**

    - This is a small building block, and is based on a very generic model of argument trees and votes, and therefore can be part of virtually any kind of deliberative system.

Structure

Inputs

- A .jsonl file with vote events. Example:
  `{"user_id":"100","parent_id":null,"post_id":1,"vote":“up”,"vote_event_time":1708772663570,"vote_event_id":1}`
  `{"user_id":"101","parent_id":1,"post_id":2,"vote":“down”,"vote_event_time":1708772663573,"vote_event_id":2}`

- This file provides all the information our algorithm needs. It doesn’t need to know anything about the content of the posts. It works using *post IDs* and *user IDs* supplied by the deliberation platform integrating this algorithm, and *parent IDs* to encode hierarchical structure.

  A “post” can be any scorable entity (a claim, an argument, a pro/con, a question, a social media post, a reply). The algorithm then estimates how children of a post (e.g. comments and arguments) affect the probability of upvote on the parent.

Outputs

- Every vote event results in one or more “score” or “effect” events. Example:
  `{"vote_event_id":1,"vote_event_time":1708772663570,"score":{"post_id":1,"critical_thread_id":null,"o":0.9129,"o_count":1,"o_size":1,"p":0.9129,"score":0.7928}}`
  `{"vote_event_id":2,"vote_event_time":1708772663573,"effect":{"target_id":1,"comment_id":2,"p":0.9129,"p_count":0,"p_size":0,"q":0.9129,"q_count":0,"q_size":0,"r":0.9129}}`

- A score event is just an update of the score data for a post.
- The score data includes a probability that describes how a hypothetical all-knowing participant would agree if they considered all the comments.

- The score includes additional information that can be used to rank the argument tree based on the scores (Thompson sampling)
- An effect event is an update of the measure of the *effect* that a comment (e.g. argument) has on a post.
- A critical thread
    - Detailed data about which arguments change minds and change minds back

Additional impacts (state changes)

What else happens to participants or others as a result of the process, beyond the direct outputs? E.g. People learn about the spread of opinion.

- Distributing the most critical information among participants
    - Differences in relevant knowledge among participants decrease
    - Participants aggregate opinion converges towards informed opinion
- Downrank content which doesn't measurably contribute information to a discussion (approvals, spam, hate speech, clickbait, …)
- Users receive reputation based on reasonability and honesty (game theory)
- Everyone understands what information influences participant's beliefs

Details

Principles & Rationale*

What are the guiding principles and rationale behind this approach and process?

- A "fair" decision is based on an **informed**, and **unbiased** opinion
- We can empirically measure what content influences opinion to derive the **informed** opinion
- Unstructured machine learning can be used to identify **unbiased** opinions (e.g. bridging based ranking with matrix factorization).

Benefits

What are the reasons to use this process or include it in a larger process? What are difficult challenges that it addresses?

- Addresses the lack of a way to reliably "score" claims made in debates, arguments, and deliberations
- Provides a measure of misinformation (information that moves users further away from the informed consensus). Misinformation is systematically neutralized

- Addresses the challenges of deliberate manipulation, ideological bias and ignorance.
- The formal requirements for the algorithm are simple enough to be usable in Social Media contexts

Current Challenges & Limitations

What are the current challenges and limitations of the process which may be improved in future versions or process runs?

- Validating the algorithm with real-world scenarios (though we already have simulations)
- Many people are uncomfortable with downvoting
- Requires anonymity, and for users to trust that they will remain anonymous, for game theoretical assumptions to hold.
- Requires a certain scale: we need multiple votes from users in order for the matrix factorization algorithm to work, and also for game-theoretical dynamics to play out over time.

Intentional Limitations

What are the limitations of the process which are expected by design?

- Can't be applied to non-tree structures like group chats
- Does not enforce formalism (e.g. formal "claims", pro and con arguments, etc.). But this formalism can exist in the platform integrating the algorithm. This also simplifies UI requirements.

Assumptions

What assumptions must be true for the process to be applicable and effective?

- Meaning of up/downvotes: A user upvoting intends to draw more attention to a piece of information

Explanation Overview

[optional diagrams]

The algorithm takes an input stream of votes on posts in a comment tree, and outputs:

1. An estimate of the **informed** upvote probability for each post: the probability that a user would upvote the post *given they were informed of every comment in that post's reply tree*.
2. An estimate of the **effect** of each comment: how much more or less likely users are to upvote the post, given they have voted on (e.g. considered) that comment.

3. A **score** that can be used for ranking posts and comments

Parameters

What are the things that might change across different runs versus stay the same, and what are the variables that you might toggle for different variants of the process?

No variants of the process. No configuration.
Our algorithm is mostly deterministic except for randomly initialized weights and learned hyperparameters. Since it is probabilistic in nature, it will converge to the same results in multiple runs.

Evaluation

Results of current evaluations

Agent-based simulations show that our algorithm produces the results we expect.

Suggested evaluations for assessing process runs

See if a group of agents can become more intelligent than single agents and can solve problems for which we have verifiable answers on their own. We can evaluate that with LLM agents or people.
Example problems:

- Chess Moves
- Math problems
- Misinformation
- Evaluating a scientific paper for which we know it has flaws