Bridging-Based Ranking

What we know, What we need, Connection to deliberation

Aviv Ovadya + Luke Thorburn

June 2023

bridging An increase in mutual understanding and trust across divides, creating space for productive conflict, deliberation, or cooperation.

What we know

What we need

Connection to deliberation

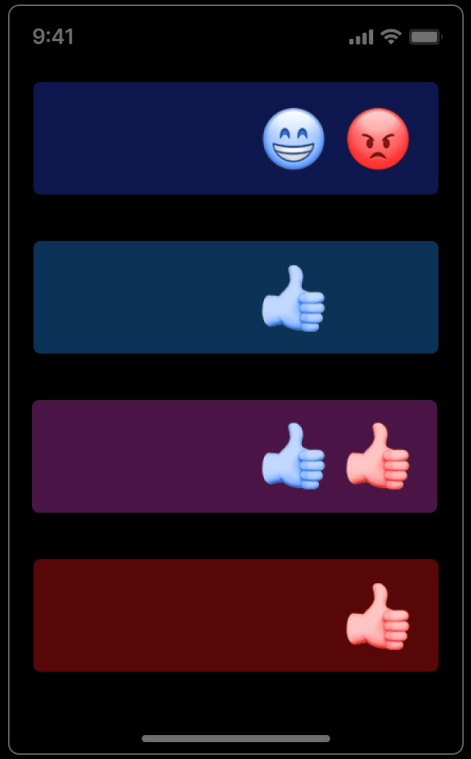
What we know

What we need

Connection to deliberation

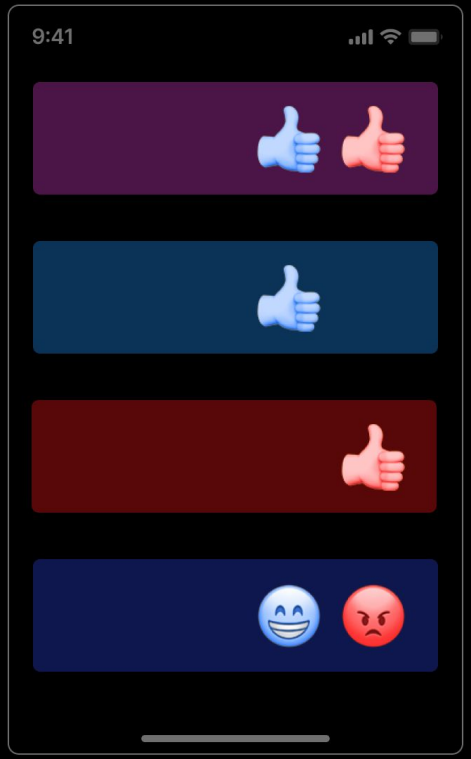
diverse approval A positive response from a diverse group of people.

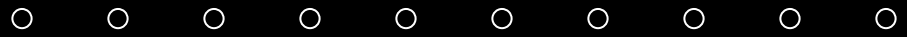
ENGAGEMENT
based ranking
for



BRIDGING
based ranking
for

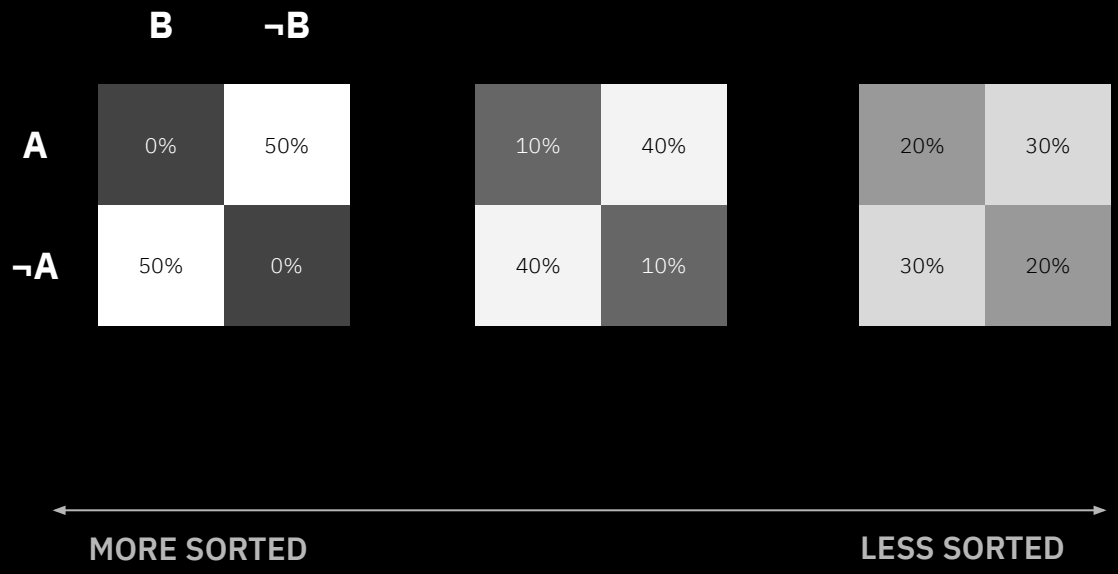
"diverse approval"





diverse approval ≈ reduced sorting

(political science, conflict studies)



We know that certain types of contact can help.

(contact theory)

A meta-analysis with **515 studies** and more than **250,000 subjects** demonstrates that intergroup contact typically reduces prejudice (**mean r = −.21**). Allport's original conditions for optimal contact – equal status, common goals, no intergroup competition, and authority sanction – facilitate the effect but are **not necessary conditions**. There are other positive outcomes of intergroup contact, such as **greater trust and forgiveness for past transgressions**. These contact effects occur **not only for ethnic groups but also for such other groups as homosexuals, the disabled and the mentally ill**. ... Moreover, these effects **typically generalize** beyond the immediate outgroup members in the situation to the whole outgroup, other situations, and even to other outgroups not involved in the contact. They also **appear to be universal** – across nations, genders, and age groups.

→ **Recent advances in intergroup contact theory**, Pettigrew et al. (2011)

diverse approval ⊃ surprising validation

(Cass Sunstein, Richard Reisman)

Instead of...

“how predictable and uninformative that someone like that would think something so evil and foolish”

...we respond:

“if someone like that disagrees with me, maybe I had better rethink”

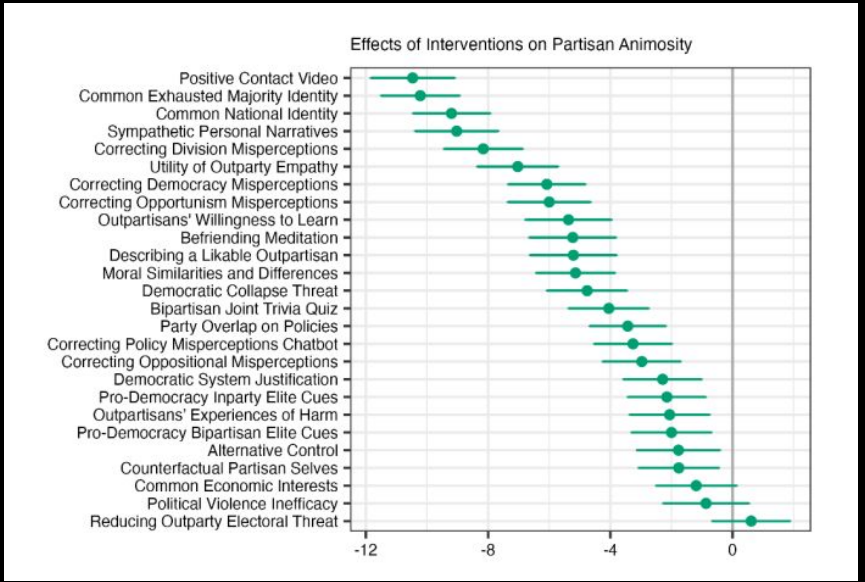
→ **Breaking Up the Echo**, Cass Sunstein (2012)

→ **Filtering for Serendipity — Extremism, “Filter Bubbles” and “Surprising Validators”**, Richard Reisman (2012)

Media Effects

(psychology; Strengthening Democracy Challenge)

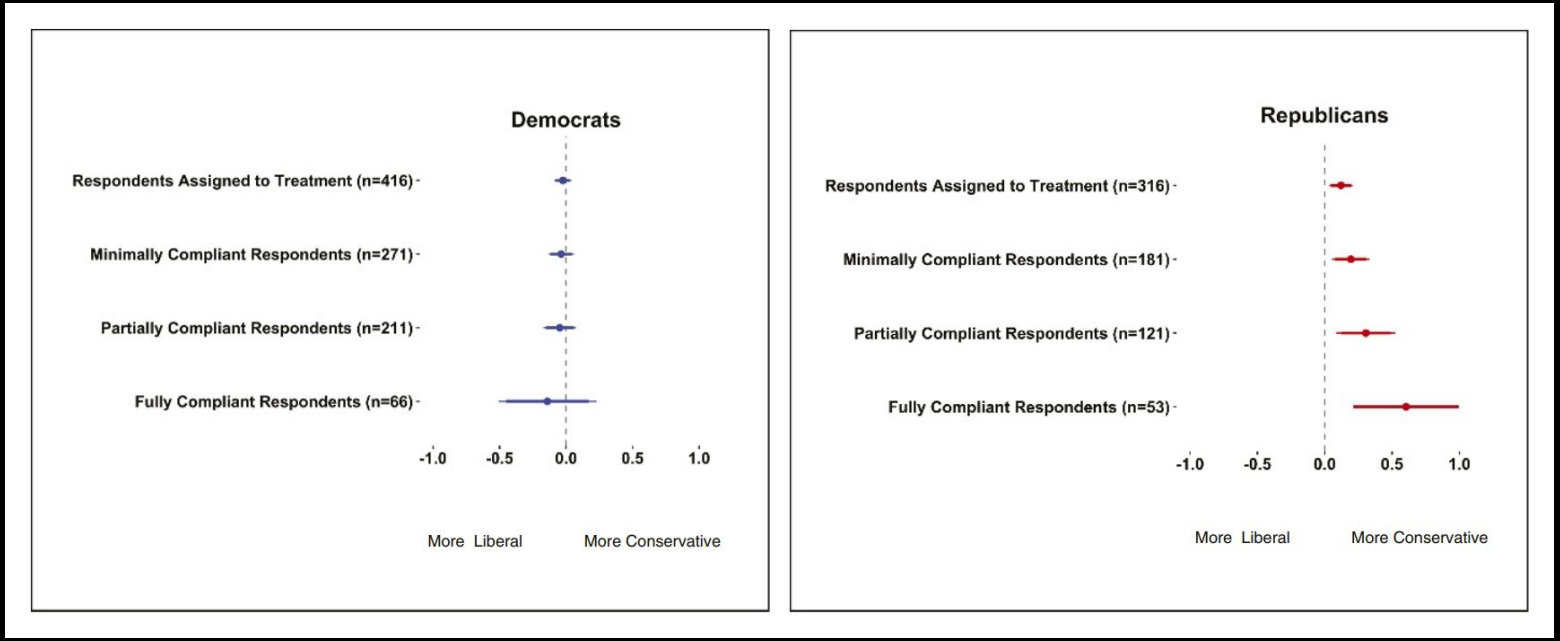
- mere-exposure effect
- illusory truth effect
- experience effects
- anchoring
- learning new facts
- discovering new interests
- pressure to conform



→ **Megastudy identifying effective interventions to strengthen Americans' democratic attitudes**, Voelkel et al. (2023)

‘Mere’ diversity is insufficient.

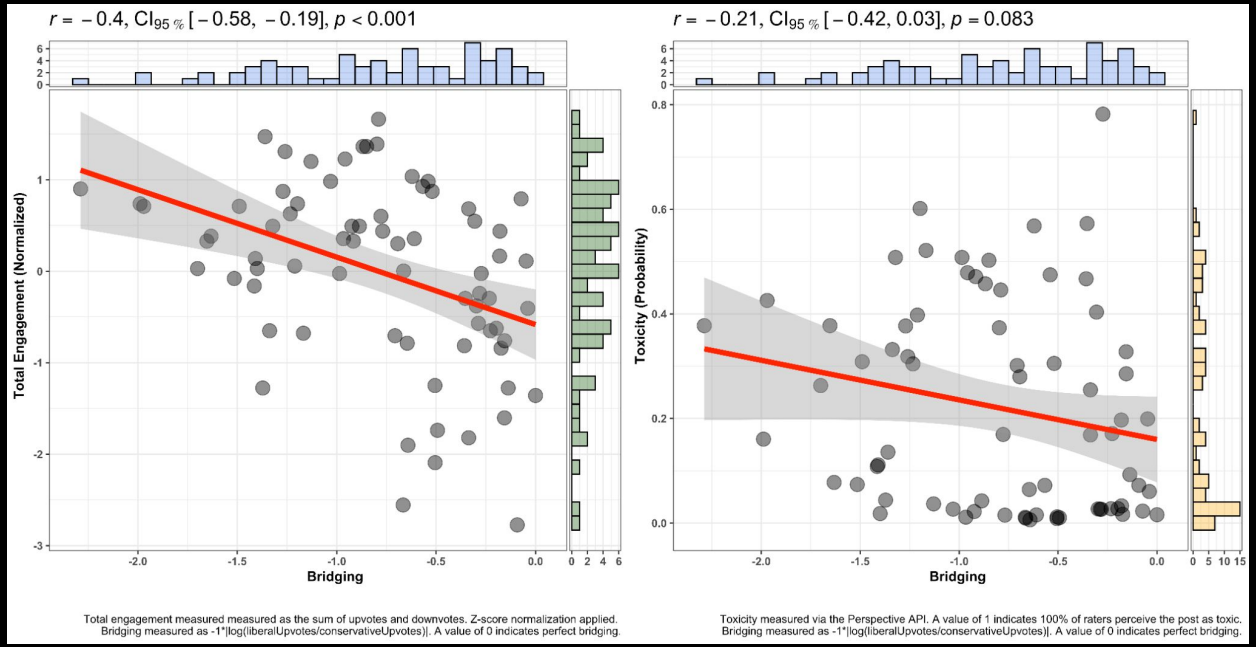
(Chris Bail)



→ **Exposure to opposing views on social media can increase political polarization**, Bail et al. (2018)

Lab Evidence

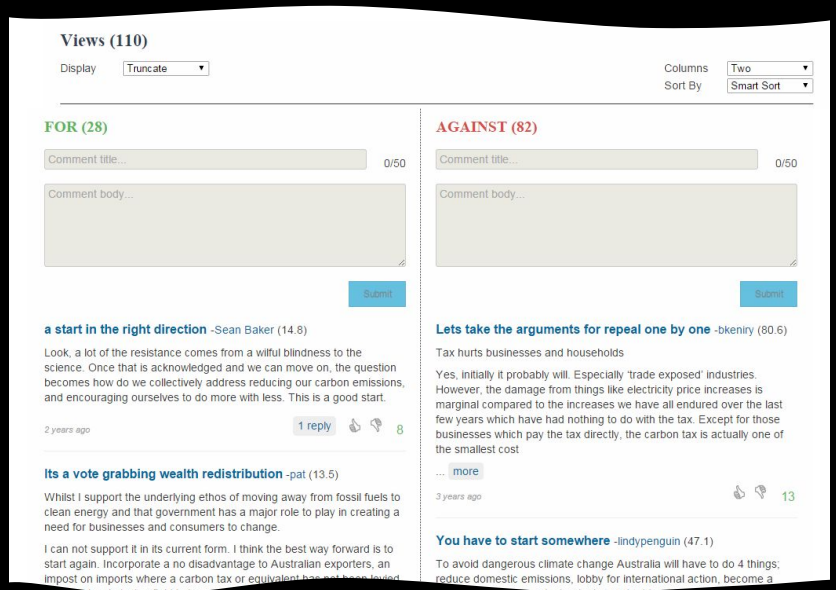
(Jason Burton, Max Planck Institute for Human Development)



→ Jason Burton (in progress)

YourView

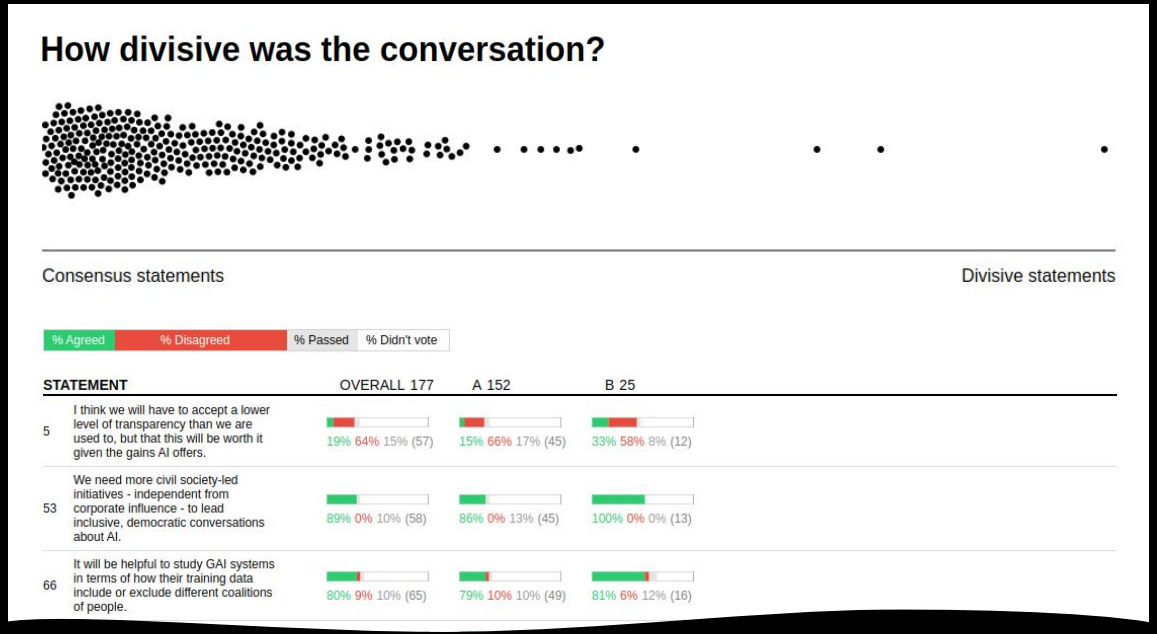
(Tim van Gelder)



→ **Cultivating Deliberation for Democracy**, Tim van Gelder (2020)
→ **Public Wisdom**, Tim van Gelder (2020)

Polis

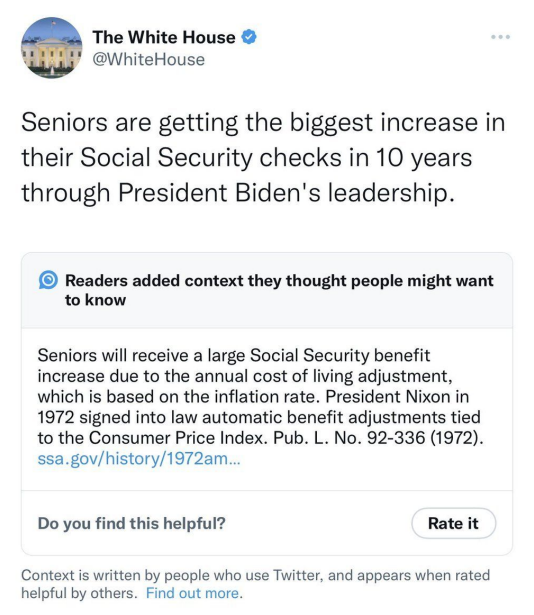
(Computational Democracy Project)



→ **Polis: Scaling deliberation by mapping high dimensional opinion spaces**, Small et al. (2021)

Twitter

(Community Notes)



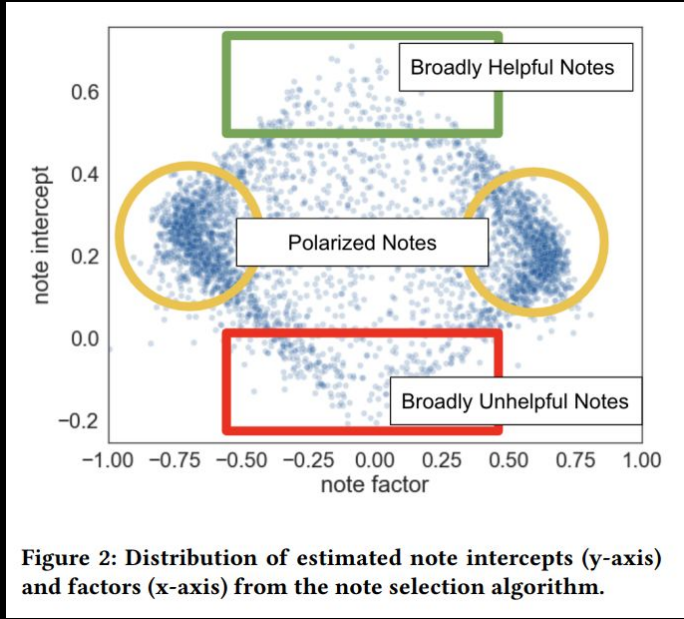
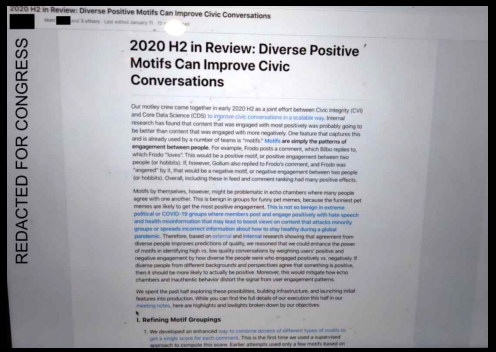


Figure 2: Distribution of estimated note intercepts (y-axis) and factors (x-axis) from the note selection algorithm.

→ **Birdwatch: Crowd Wisdom and Bridging Algorithms can Inform Understanding and Reduce the Spread of Misinformation**, Wojcik et al. (2022)

Facebook

(Facebook Papers)



Comments with high positive motifs from a highly diverse audience — relative to those with high positive motifs from a *non*-diverse audience — were:

- **10.9x less likely to be reported**
- **3.4-5.2x lower rates of violations than comments**

At the user-level, people with a reputation for getting positive motifs from more diverse audiences were:

- **much less likely to violate community standards,**
- **much less likely to be reported by other users for being fake, bullying, hate, or violence incitement**

“Thus, reputations for garnering positive motifs are especially informative of the likely value of content when enhanced by the diversity of people engaging with that content.”

→ **2020 H2 in Review: Diverse Positive Motifs Can Improve Civic Conversations**, Facebook Papers (2021)

What we know

What we need

Connection to deliberation

What we know

What we need

Connection to deliberation

Bridging Systems

Open Problems for Countering Destructive Divisiveness across Ranking, Recommenders, and Governance

Aviv Ovadya[*]
Harvard University
aviv@aviv.me

Luke Thorburn[†]
King's College London
luke.thorburn@kcl.ac.uk

Divisiveness appears to be increasing in much of the world, leading to concern about political violence and a decreasing capacity to collaboratively address large-scale societal challenges. In this working paper we aim to articulate an interdisciplinary research and practice area focused on what we call *bridging systems*: systems which increase mutual understanding and trust across divides, creating space for productive conflict, deliberation, or cooperation. We give examples of bridging systems across three domains: recommender systems on social media, collective response systems, and human-facilitated group deliberation. We argue that these examples can be more meaningfully understood as processes for *attention-allocation* (as opposed to "content distribution" or "amplification") and develop a corresponding framework to explore similarities—and opportunities for bridging—across these seemingly disparate domains. We focus particularly on the potential of *bridging-based ranking* to bring the benefits of offline bridging into spaces which are already governed by algorithms. Throughout, we suggest research directions that could improve our capacity to incorporate bridging into a world increasingly mediated by algorithms and artificial intelligence.

Keywords: *bridging, bridging-based ranking, cross-cutting, polarization, depolarization, deliberative technology, facilitation, recommender system, ranking, artificial intelligence*

This document is a draft. A revised version will be published with the Knight First Amendment Institute at Columbia University.

[*]Aviv Ovadya is an affiliate at the Berkman Klein Center for Internet & Society at Harvard University (at the Institute for Rebooting Social Media), and a visiting scholar at the Leverhulme Centre for the Future of Intelligence at Cambridge University. This work began while he was a Technology and Public Purpose Fellow at the Harvard Kennedy School's Belfer Center (2021-2022).

[†]Luke Thorburn is a researcher in the UKRI Centre for Doctoral Training in Safe and Trusted AI at King's College London.

extremes when combined with our psychological predilections.[1] As articulated by Stray [82], the intent is "conflict transformation" [51, 17]: not to remove divisions or interfere with the substance of civic debates, but to "[make] conflict better in some way". In other words, to mitigate the risks of "high conflict" [74] whilst supporting "healthy" or "constructive" conflict. When we use the terms bridging, or reducing division, we are thus *not* referring to "making everyone believe the same things"—these are just shorthand for "enabling mutual understanding and respect across divides"—or in other words, supporting pluralism [96].[2] Figure 1 provides a speculative causal loop diagram [29] illustrating the potential ideal impacts of bridging systems.

Figure 1: **A causal loop diagram illustrating how bridging systems might impact society.** The goal of this diagram is not to make strong, precise claims about causality, but simply to provide intuition on how a proliferation of bridging systems could have important and beneficial societal consequences and reduce both deliberate and indirect harms. Significant work is required to determine which causal relations hold (including those not drawn on this diagram), and under what conditions.

Moving from engagement-based ranking to *bridging-based ranking*. By definition, optimizing more for bridging means optimizing less for engagement, but the extent to which these two goals are in tension is an open question. It may be possible to include bridging impacts within the objective function of a recommender system without undermining financial sustainability. Figures 2 and 3 give examples of what such bridging-based ranking [63] might look like in the context of a recommender system on a social media platform.

[1]An economics framing of the ultimate goal might describe it as countering the "incentivization of divisive behavior" by "intentionally subsidizing bridging".

[2]For example, a system might bridge divides by facilitating understanding that one's existing values and beliefs are much more similar than expected to those of other people, closing what has been called the "perception gap" [103]. In this way, beliefs about *other people's beliefs* are changing, while personal beliefs are staying fixed.

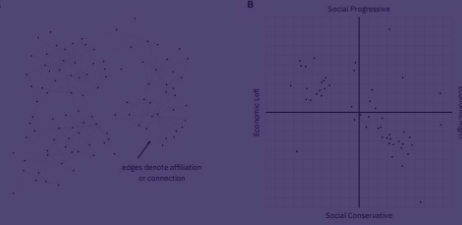


Figure 6: Simple examples of (a) a graph-based model and (b) a space-based model.

is comparatively not. Such *aggregate models* could incorporate item classifications (e.g., Nickelback might be labeled as non-bridging), or be broken down by subgroups. Note, however, that such models are likely less expressive than graph or space-based models, and depending on the percentages it will not always be possible to infer overlap.

Recommender System: Relation Model *Example 13*

The recommender systems on modern social media platforms are often built using deep neural networks that learn a numerical representation of each user and item. These vectors or "embeddings" are usually high-dimensional (hundreds or thousands of dimensions), and correspond to a position within a latent embedding space that characterizes each user's history of behavior on the platform and, by implication, their opinions, preferences, and worldview. Thus, these embeddings constitute a space-based model [33, 75].

Social networks also often have an underlying graph-based structure, such as graphs of friends on Facebook, or follow networks on Twitter. Such networks constitute graph-based models of `people`. There are other possible graph structures. For example, you could consider a single graph where both `people` and `items` are vertices, and edges are used to indicate interactions between `people` and `items`. In some cases, information from such graph-based models is translated to space-based models for use in recommender systems [75].

Collective Response System: Relation Model *Example 14*

Collective response systems like YourView or Polis involve `people` submitting `items` (called comments or responses), and voting on items shown via a recommendation algorithm. These votes are represented by a matrix where rows correspond to `people`, columns correspond to `items`, and the cell values indicate votes: "Agree" (encoded as +1), "Disagree" (−1), or "Pass" (0). The rows in the vote matrix can be viewed as the locations of people in a space-based model. So that it can be visualized, both YourView and Polis compress this relatively high-dimensional representation into a two-dimensional space-based model, positioning people closer together if their votes are more similar. Intuitively, this means that if two people voted the same way on every item, they will end up at the same point [81]. An example of the YourView space-based model is shown in Figure 7.

OPEN PROBLEMS

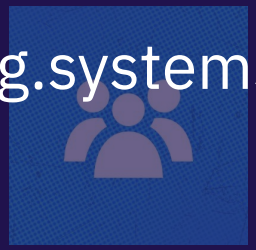
→ bridging.systems

1. Methods for implementation
2. Actual implementations
3. Metrics for evaluation

1. Methods for implementation
2. Actual implementations
3. Metrics for evaluation

CASE STUDIES

→ bridging.systems/case-studies/ (soon!)



EXPERIMENTS

→ talk to Jonathan, Ravi, Smitha, Micah, Luke ...

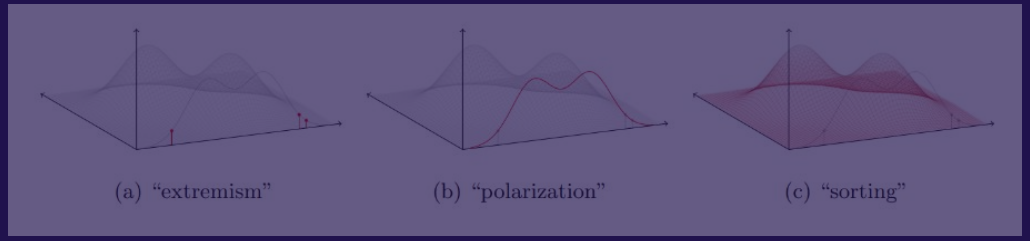
1. Methods for implementation
2. Actual implementations
3. Metrics for evaluation

SURVEY OF METRICS

→ bridging.systems/metrics/

METRIC	INTUITION	SCOPE	MODEL TYPE	STRUCTURE REQUIRED	SAFE TO OPTIMIZE?	FORMULA	REFERENCES
node-level homophily	The proportion of a vertex's neighbours which are in its own group.	individual	graph	groups	NO	$\frac{d_i(v)}{d(v)}$	Currarini et al. (2009), Interian + Ribeiro (2018), Reese et al. (2007), Interian et al. (2022)
group-level homophily	An average measure of the degree to which vertices in a group are connected to vertices in their own group, rather than others.	sub-group	graph	groups	NO	$\frac{\sum_{v \in G_i} d_i(v)}{\sum_{v \in G_i} d(v)}$	Lelkes (2016), Currarini et al. (2009), Interian et al. (2022)
population homophily	An average measure of the degree to which all vertices are connected to vertices in their own group, rather than others.	population	graph	groups	NO	$\frac{\sum_{v \in V} d_{s(v)}(v)}{\sum_{v \in V} d(v)}$	Interian et al. (2022)
modularity	The number of intra-group edges in the graph minus the expected number of intra-group edges in a graph with the same nodes, groups and degrees, but with edges placed at random. (Up to a multiplicative constant.)	population	graph	groups	NO	$\frac{1}{4\lvert E\rvert} \sum_{u,v \in V, u \neq v} \left[\mathbb{1}((u,v) \in E) - \frac{d(u)d(v)}{2\lvert E\rvert} \right] \mathbb{1}(u, v \text{ in same group})$	Newman (2006), Zhang et al. (2007), Wolfowicz et al. (2021), Garcia et al. (2015), Dal Maso et al. (2014), Interian et al. (2022)
E-I index	The difference between the proportions of edges that are inter- and intra-group.	population	graph	groups	NO	$\frac{\lvert\{(u,v) \in E : g(u) \neq g(v)\}\rvert - \lvert\{(u,v) \in E : g(u) = g(v)\}\rvert}{\lvert E\rvert}$	Krackhardt + Stern (1988), Interian et al. (2022)
random walk controversy	Given that two random walks ended in different groups, the difference between the probability that they started from those same groups and the [illegible] groups.	population	graph	groups	NO	[illegible] $\mathbb{P}(R_1(\text{start}) \in G_1, R_2(\text{start}) \in G_2 \mid R_1(\text{end}) \in G_1, R_2(\text{end}) \in G_2)$	Garimella et al. (2018), Garimella et al. (2016), Cossard et al. (2020), Rumshisky et al. (2017), Emamgholizadeh et al. (2020), Interian et al. (2022)
node-level random walk controversy	The probability that a random walk which ends in one [illegible] started in a vertex of interest relative to the same value for other groups.	individual	graph	groups	NO	$\frac{\mathbb{P}(R(\text{start}) = v \mid R(\text{end}) \in G_i)}{\sum_{j=1}^{M} \mathbb{P}(R(\text{start}) = v \mid R(\text{end}) \in G_j^*)}$	Garimella et al. (2018), Garimella et al. (2016), Cossard et al. (2020), Rumshisky et al. (2017), Emamgholizadeh et al. (2020), Interian et al. (2022)
degree of balance	Degree of consistency with properties such as "my friend's friend is my friend" and "my friend's enemy is my enemy".	population	graph	signed edges	MAYBE	$\frac{c_3(G)}{\binom{N}{3}}$	Harary (1959), Aref et al. (2020), Interian et al. (2022)
line index of balance	Minimum number of edge modifications that must be made to be perfectly consistent with properties such as "my friend's friend is my friend" and "my friend's enemy is my enemy".	population	graph	signed edges	MAYBE	Not easily notated—see references.	Harary (1959), Aref et al. (2020), Interian et al. (2022)
point index of balance	Minimum number of vertices which must be deleted to be perfectly consistent with properties such as "my friend's friend is my friend" and "my friend's enemy is my enemy".	population	graph	signed edges	MAYBE	Not easily notated—see references.	Harary (1959), Aref et al. (2020), Interian et al. (2022)
diameter	The maximum distance between any two points.	population	space		NO	$\max_{x,y \in \mathcal{X}} \lVert x - y \rVert$	Bramson et al. (2016), Bramson et al. (2017)
volume	The volume of the minimal convex polytope that includes all points.	population	space		NO	$\min\{\text{Volume}(S) \mid \mathcal{X} \subseteq S, S \text{ convex}\}$	Bramson et al. (2016), Bramson et al. (2017)
mean difference	Average pairwise distance between any two points.	population	space		NO	$\frac{1}{\binom{N}{2}} \sum_{x,y \in \mathcal{X}, x \neq y} \lVert x - y \rVert$	Bramson et al. (2016), Bramson et al. (2017)
average absolute deviation	Average distance between each point and the mean.	population	space		NO	$\frac{1}{N} \sum_{x \in \mathcal{X}} \lVert x - \bar{x} \rVert$	Bramson et al. (2016), Bramson et al. (2017)
standard deviation	Scalar standard deviation of the set of points.	population	space		NO	$\sqrt{\frac{1}{N} \sum_{x \in \mathcal{X}} \lVert x - \bar{x} \rVert^2}$	Bramson et al. (2016), Bramson et al. (2017)

EXIT



QUANTIFYING SORTEDNESS

→ talk to Luke

1. Methods for implementation
2. Actual implementations
3. Metrics for evaluation

What we know

What we need

Connection to deliberation

What we know

What we need

Connection to deliberation

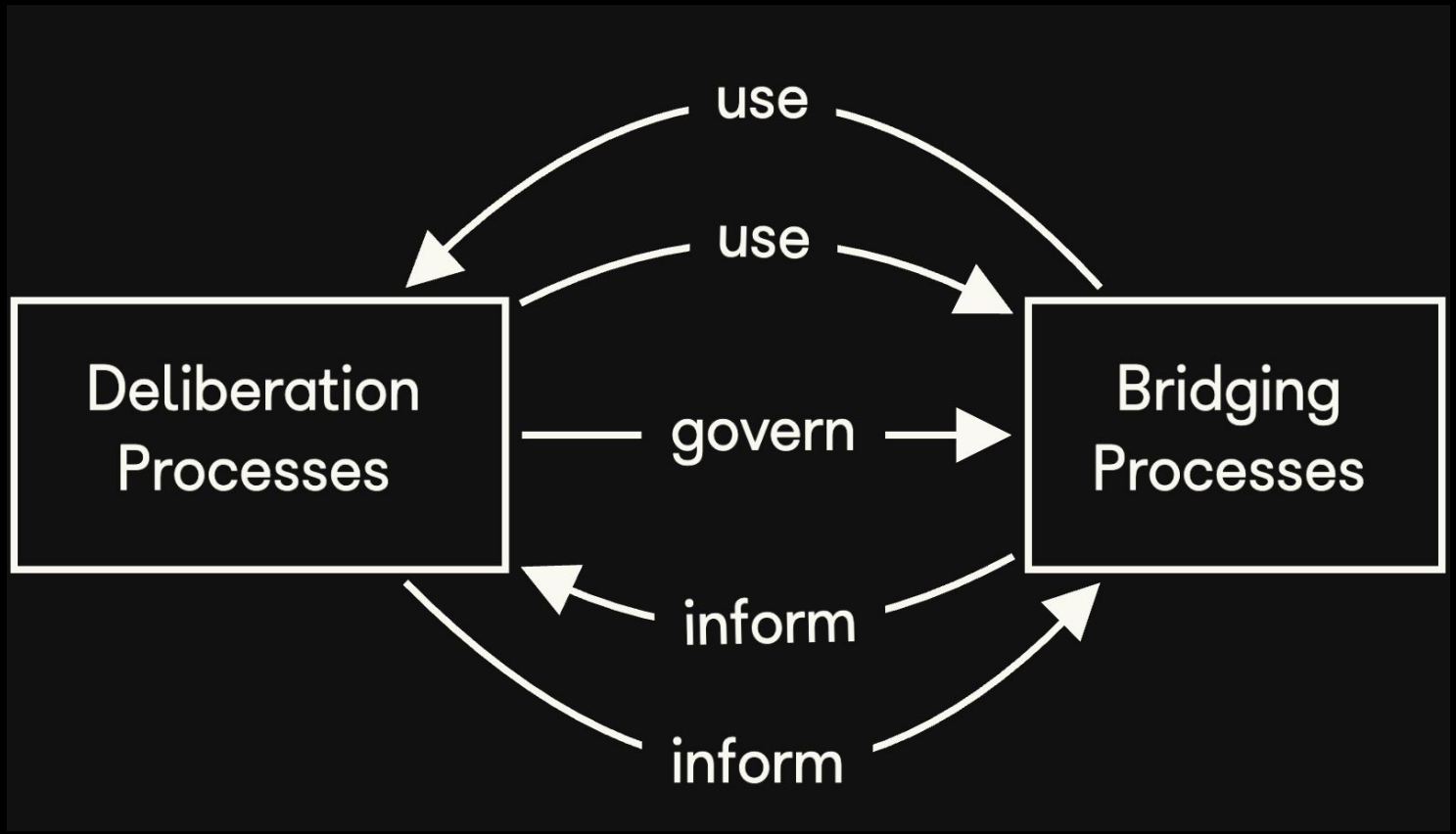
Bridging Processes × **Deliberation** Processes

Underlying ‘ultimate goals’

- **Bridging workstream**: Processes that ‘bridge’ (instead of divide).
- **Deliberation workstream**: Processes that:
  - Provide high quality & legitimate decisions
  - For issues where there may be significant conflict
  - At every scale, including transnationally and global
  - At the speed of technological change

  Why? Consider AI: Low trust, significant conflict, every scale, extremely fast-changing.

  Improved deliberation processes can improve AI governance.


use
use
Deliberation Processes
govern
Bridging Processes
inform
inform

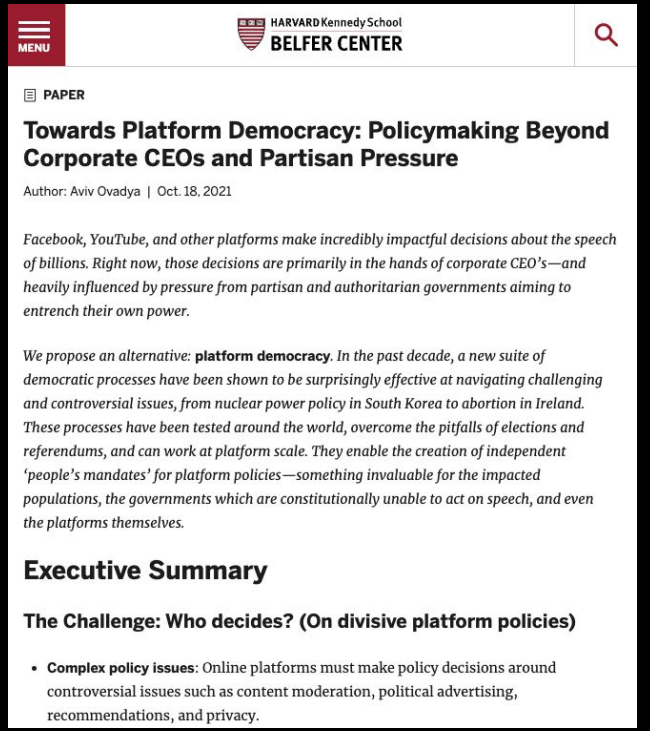
Where does bridging fit into deliberation?

1. Learning from effective deliberation processes—
e.g. Citizen assemblies, deliberative polls, etc.—
and related conflict resolution/transformation processes
(cold vs. hold “deliberations”)

Where does bridging fit into deliberation?

2. Governing the baseline 'bridging force'—determining the extent of bridging that is 'subsidized' by global & transnational infrastructure (e.g. online platforms)

- The extent to which our environment is bridging can influence the the level of pluralism across society.
  - Corporate incentives, partisan incentives, and decentralized infrastructure can all encourage less bridgingness.
- Democratic deliberative processes provide a mechanism for collective self-determination—for democratically *'tuning bridgingness'* across national boundaries, without paternalism.



MENU

HARVARD Kennedy School
BELFER CENTER

PAPER

Towards Platform Democracy: Policymaking Beyond Corporate CEOs and Partisan Pressure

Author: Aviv Ovadya | Oct. 18, 2021

Facebook, YouTube, and other platforms make incredibly impactful decisions about the speech of billions. Right now, those decisions are primarily in the hands of corporate CEO's—and heavily influenced by pressure from partisan and authoritarian governments aiming to entrench their own power.

We propose an alternative: **platform democracy**. *In the past decade, a new suite of democratic processes have been shown to be surprisingly effective at navigating challenging and controversial issues, from nuclear power policy in South Korea to abortion in Ireland. These processes have been tested around the world, overcome the pitfalls of elections and referendums, and can work at platform scale. They enable the creation of independent 'people's mandates' for platform policies—something invaluable for the impacted populations, the governments which are constitutionally unable to act on speech, and even the platforms themselves.*

Executive Summary

The Challenge: Who decides? (On divisive platform policies)

- **Complex policy issues**: Online platforms must make policy decisions around controversial issues such as content moderation, political advertising, recommendations, and privacy.

Where does bridging fit into deliberation?

3. Bridging processes exist within a deliberative process

- As critical components within a collective response system like Polis
- As part of the practice of facilitating a citizen’s assembly
- Through the development of semi-automated facilitation tools

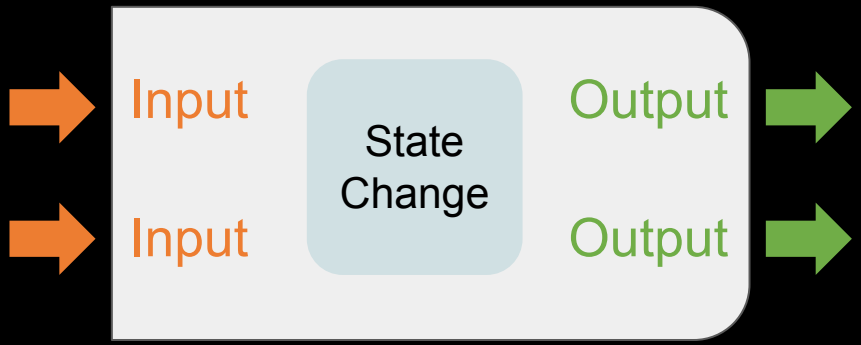
Current Projects

Bridging × Deliberation

Current Projects: Deliberation

What are the **‘type signatures’** and **key properties** of deliberative processes?

Type signature ≅ the structure of the inputs & outputs.



Collective Response Systems Criteria

Process & Structure

Key Properties

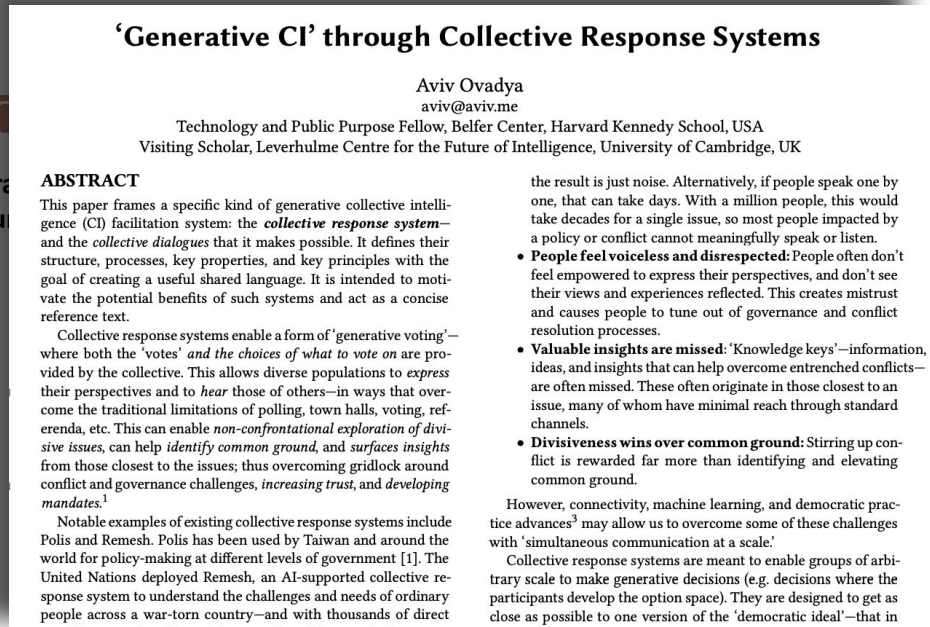
Participant Agency

Para Commu

Key Principles

Collective Oriented

Prompt Oriented



'Generative CI' through Collective Response Systems

Aviv Ovadya
aviv@aviv.me
Technology and Public Purpose Fellow, Belfer Center, Harvard Kennedy School, USA
Visiting Scholar, Leverhulme Centre for the Future of Intelligence, University of Cambridge, UK

ABSTRACT

This paper frames a specific kind of generative collective intelligence (CI) facilitation system: the ***collective response system***—and the *collective dialogues* that it makes possible. It defines their structure, processes, key properties, and key principles with the goal of creating a useful shared language. It is intended to motivate the potential benefits of such systems and act as a concise reference text.

Collective response systems enable a form of 'generative voting'—where both the 'votes' *and the choices of what to vote on* are provided by the collective. This allows diverse populations to *express* their perspectives and to *hear* those of others—in ways that overcome the traditional limitations of polling, town halls, voting, referenda, etc. This can enable *non-confrontational exploration of divisive issues*, can help *identify common ground*, and *surfaces insights* from those closest to the issues; thus overcoming gridlock around conflict and governance challenges, *increasing trust*, and *developing mandates*.[1]

Notable examples of existing collective response systems include Polis and Remesh. Polis has been used by Taiwan and around the world for policy-making at different levels of government [1]. The United Nations deployed Remesh, an AI-supported collective response system to understand the challenges and needs of ordinary people across a war-torn country—and with thousands of direct participants and a substantial proportion of the population engaged—the result is just noise. Alternatively, if people speak one by one, that can take days. With a million people, this would take decades for a single issue, so most people impacted by a policy or conflict cannot meaningfully speak or listen.

- **People feel voiceless and disrespected:** People often don't feel empowered to express their perspectives, and don't see their views and experiences reflected. This creates mistrust and causes people to tune out of governance and conflict resolution processes.
- **Valuable insights are missed**: 'Knowledge keys'—information, ideas, and insights that can help overcome entrenched conflicts—are often missed. These often originate in those closest to an issue, many of whom have minimal reach through standard channels.
- **Divisiveness wins over common ground:** Stirring up conflict is rewarded far more than identifying and elevating common ground.

However, connectivity, machine learning, and democratic practice advances[3] may allow us to overcome some of these challenges with 'simultaneous communication at a scale.'

Collective response systems are meant to enable groups of arbitrary scale to make generative decisions (e.g. decisions where the participants develop the option space). They are designed to get as close as possible to one version of the 'democratic ideal'—that in collective decisions:

“Collective Response” Process

Inputs:

group, prompt
(a question)

Response

All can respond (generatively)
(e.g. free text, not multiple choice)

Evaluation

All responses are listened to

Distillation

‘Best’/representative is selected /synthesized

Outputs:

representative distillation,

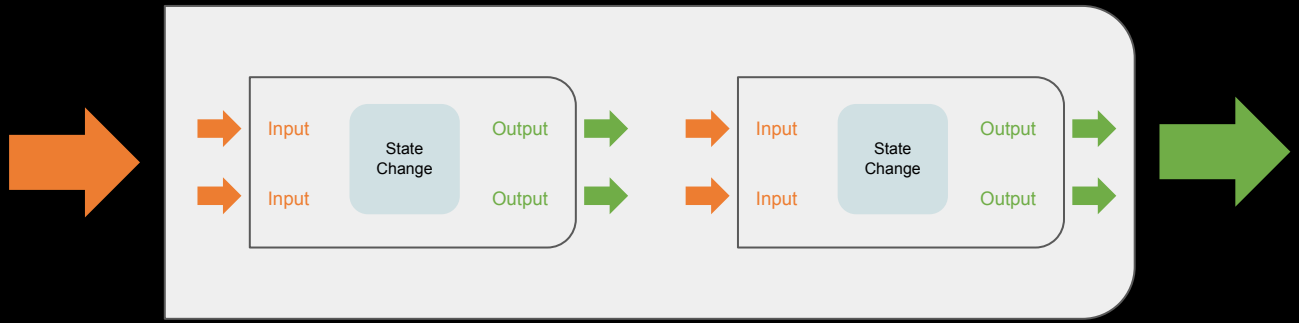
raw & derived data (e.g. response map)

Bridging-based ranking and Bridging systems with Luke Thorburn

Current Projects: Deliberation

What are the '**type signatures**' and **key properties** of deliberative processes?

Type signature ≅ the structure of the inputs & outputs.

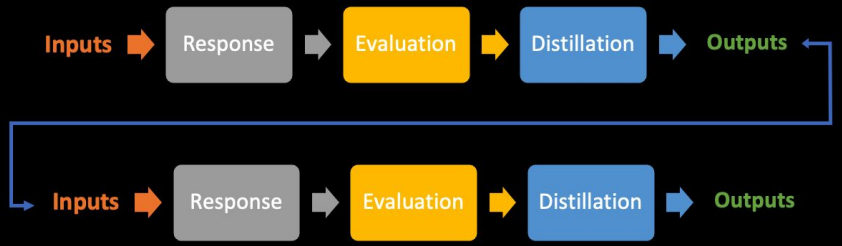


Current Projects: Deliberation

What are the '**type signatures**' and **key properties** of deliberative processes?

Type signature ≅ the structure of the inputs & outputs.

Combining via chaining



Integrating within a larger process

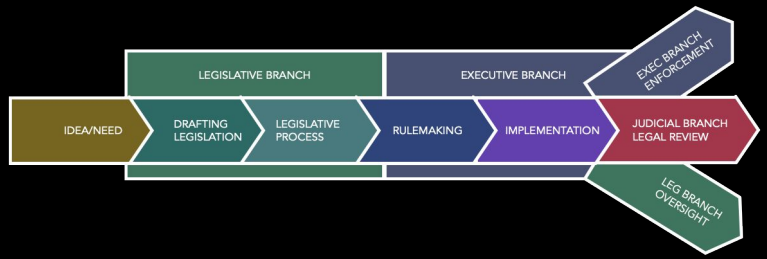
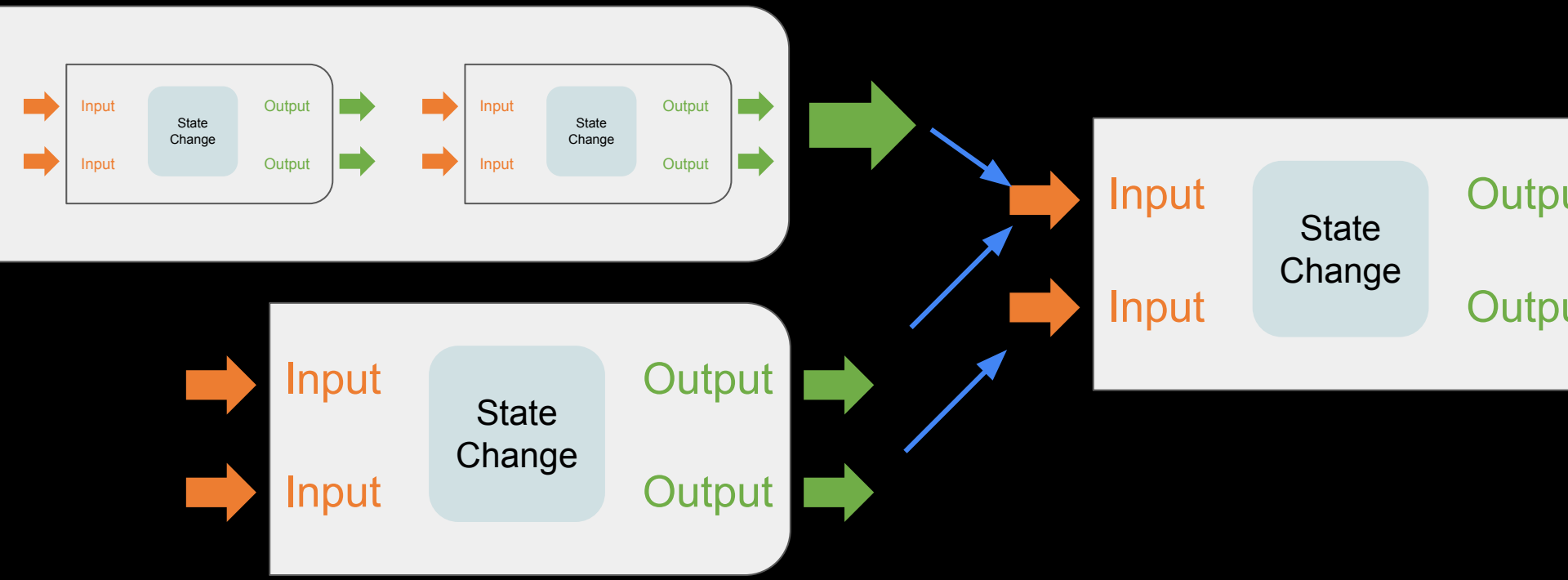


Diagram via Marci Harris from popvox



If you have frameworks, processes, examples, that you want to share—or if you are interested in doing this sort of mapping, contact me! aviv@aviv.me

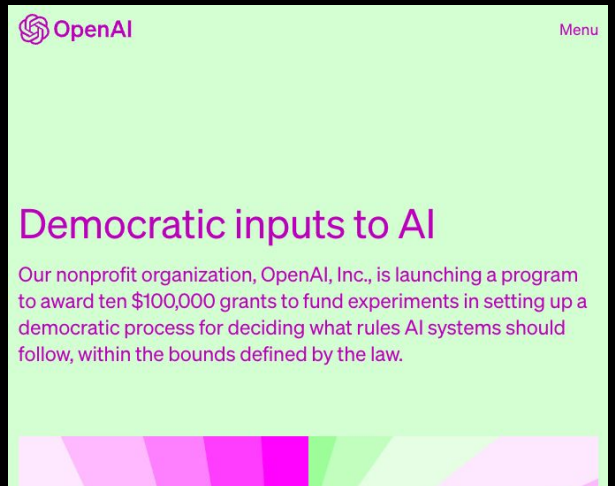
Current Projects: Deliberation

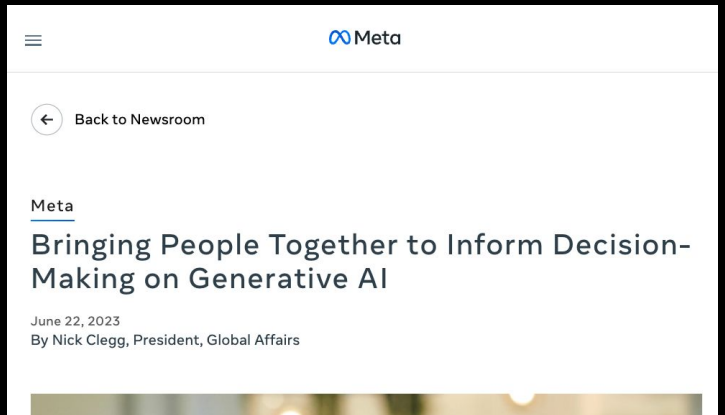
Critical Applications of **deliberative processes** for AI governance and alignment

Building on those types and properties

06.22.2023

Majority Leader Schumer Floor Remarks On Launching The SAFE Innovation Framework For AI And First Of Their Kind AI Insight Forums





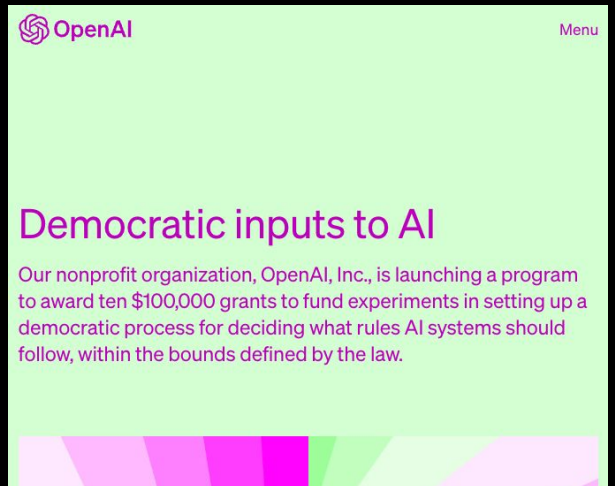
Current Projects: Deliberation

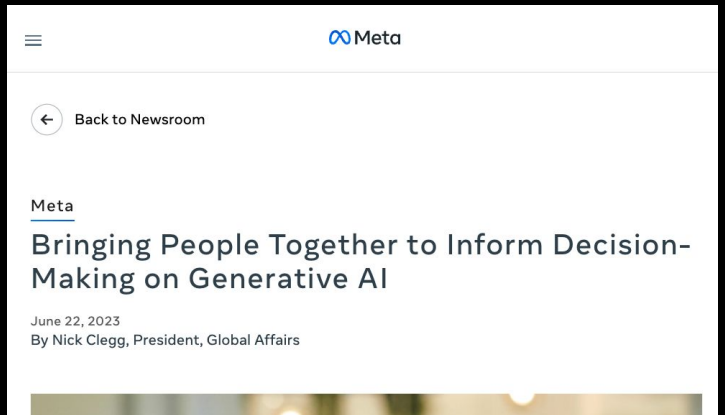
Critical Applications of **deliberative processes** for AI governance and alignment

Building on those types and properties

06.22.2023

Majority Leader Schumer Floor Remarks On Launching The SAFE Innovation Framework For AI And First Of Their Kind AI Insight Forums





bridging.systems

Aviv @metaviv av@aviv.me

Luke @LukeThorburn_ luke.thorburn@kcl.ac.uk

Thank you to the many collaborators & mentors who have contributed to or informed this work!

GETTING PLURALITY Research Network

This work is conducted in partnership with the GETTING-Plurality Research Network housed within the Allen Lab for Democracy Renovation at Harvard University.

Aviv Ovadya was supported in part by a Technology and Public Purpose Fellowship at the Belfer Center for Science and International Affairs, Harvard Kennedy School. Luke Thorburn was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (safeandtrustedai.org), King's College London.