

IN DEPTH

Bridging Algorithms as Practical Tools for Depolarisation

Luke Thorburn

Research fellow at the AI & Democracy Foundation

It is widely recognised that by optimising for “engagement,” social media algorithms tend to disproportionately surface the more extreme, inflammatory voices^[1], raising the emotional temperature of online conversations and contributing to political polarisation^[2]. But could we design ranking algorithms that instead help build common ground? This is the motivation behind an emerging area of research and practice around “bridging algorithms” and “bridging-based ranking.” In this piece, I give a concise introduction to bridging algorithms, orient you to where and how they have been used so far by both civil society organisations and social media platforms, and summarise their current limitations and challenges.

The term bridging-based ranking can describe any method for ranking alternatives – be they social media posts, policy proposals, survey responses, or candidates for elected office – that helps build mutual understanding and trust across divides^[3]. This qualitative goal can be operationalised in many different ways. I will explain the most common approaches, but to help ground this, imagine that you are convening – in whatever community or region you are most familiar with – an online process in which members of groups that have been in conflict with one another submit written proposals for what should happen next. As the facilitator of the process, you need to determine which of the submissions – of which there could be thousands – to make most visible when you report the results back to the community, and you want to do this in a way that is bridging, that is, helps build mutual understanding and trust.

“ Social media algorithms tend to disproportionately surface the more extreme, inflammatory voices. Could we design ranking algorithms that instead help build common ground? ”

The most common approach is to identify proposals that represent some form of common ground or “diverse approval,” that is, those approved by people who usually disagree with each other. If you have some vote data from people in the community about whether or not they support each proposal, then it’s possible to quantify the level of support for each proposal within each group, and then identify those proposals which have a high level of support in multiple groups. Finding such proposals and making them visible has at least two important effects. First, it demonstrates that there is, in fact, common ground –often quite a lot of it –which is not widely accepted in conflict settings. Second, I believe that directing attention to ideas at the intersections of different groups, even if such ideas are not universally supported, helps reduce sorting [4]– that is, the homogeneity of opinion clusters, which is a known risk factor for conflict escalation. Here are some examples:

This approach has been used recently by the Alliance for Middle East Peace (ALLMEP) in a sequence of online dialogues^[5] to help build common ground between Israeli and Palestinian peacebuilders, which resulted in a set of five demands for world leaders and the international community that all had over 90% support on each side.

It has been implemented at scale in Community Notes, a feature on social media platform X (formerly Twitter), where people can propose notes that add important context to tweets that they believe to be misleading, but where these notes are only displayed publicly if an algorithm determines they are likely to be rated as helpful by people on both ends of the political spectrum. This algorithm is open source, and the feature is also being trialled on Meta’s platforms (Facebook, Instagram, Threads), TikTok, and YouTube.

Polis, an open-source tool that facilitates online collective dialogue and has been widely used in civil society and peacebuilding contexts, makes this ranking very accessible. In the standard report from a Polis conversation, scroll down to the “All Statements” section and select to sort the statements by “group-informed consensus.”

“ The term bridging-based ranking can describe any method for ranking alternatives – be they social media posts or policy proposals – that helps build mutual understanding and trust across divides ”

A second approach is to reflect a set of proposals that is representative of the community’s opinions, by ensuring that diverse proposals are included and that there is sufficient context about the level of support for those proposals and where that support is coming from. Such representation and social context are important because in polarised contexts, there are often significant perception gaps^[6] – we tend to think the other side holds more extreme or intolerable views than they actually do. The social context needed to debunk these false perceptions is also usually missing from social media^[7], which both acts as a “prism”^[8] by disproportionately surfacing the most inflammatory perspectives. It only provides raw counts of likes, shares, etc., with no information about the degree to which a post resonates with the population as a whole, or in which sub-communities that feeling is strongest. Because of this, it can be crucial for processes run by civil society to help provide such a social context.

This approach is also implemented in Polis, the tool mentioned above, which provides, in its standard report, both a visualisation of the most prominent opinion clusters and the extent to which each cluster agrees with each statement.

It is also one of the core rationales behind the use of more traditional public opinion polling in peace processes^[9], such as those conducted by Colin Irwin^[10] in Northern

Ireland, Kashmir, the Balkans, Sri Lanka, and elsewhere, which aim – in part – to reduce help people realize that many more people share their views than they realize, that is, to reduce “pluralistic ignorance”.

“ Bridging algorithms are just one of a set of features often present in an emerging ecosystem of online “deliberative tools” that can facilitate deliberation in polarised contexts ”

A third class of approaches uses automated classifiers to identify and uprank proposals that exhibit hallmarks of being written in good faith. Such classifiers are usually more “black box” than the above two approaches, meaning that it is not always possible to know why a given text was scored a certain way by the classifier. Because of this, the use of such classifiers might be less appropriate in processes where transparency and procedural fairness are critical. But they can be helpful in more informal online fora—for example, prompting people to reflect on how their contributions might land with others and empowering them to filter for the kinds of contributions they consider valuable.

For example, Google Jigsaw has built automated classifiers, free to use, that can score text comments in an online forum based on the degree to which they exhibit “bridging attributes” like compassion, curiosity, nuance, or respect.

Challenges and limitations

Each of these approaches has limitations, and the design of bridging algorithms remains an active research area. Three challenges, in particular, deserve attention.

First, ranking items of content is inherently zero-sum, and optimising for anything – be that common ground, representativeness, nuance, or anything else – will mean that you face trade-offs with other goals you have as a facilitator of the process. As an example, consider the goal of common ground. It’s much easier for people to agree on

vague platitudes about creating a more harmonious future than it is for them to decide on, say, precisely worded clauses that could go into a substantive peace agreement. And many peacebuilders will be aware of the diplomatic “move” in which, to allow dialogue to continue despite an irresolvable disagreement, one progressively makes the language of agreement more abstract or procedural. Resulting, for example, not in an agreement about how benefits and burdens will be distributed, but in an agreement merely that they should be distributed “fairly”, or that a particular process will be followed for deliberating on how they will be distributed.

While sometimes the only way forward is possible, such vagueness kicks the can down the road – the “can” being the disagreement which will inevitably have to be resolved – by leaving ambiguous the language – and thus the substance – of what is agreed upon. In this way, the goal of maximising common ground is in tension with the goal of minimising ambiguity, and so ranking by diverse approval can result in relatively vague or ambiguous statements rising to the top. You can get a sense of this in the results from the ALLMEP process mentioned earlier, where some of the most bridging “values” and “visions” for the future have this flavour. I am currently collaborating on an ongoing project that aims to develop a method for identifying statements that are both bridging and unambiguous.

“ To allow dialogue to continue, one progressively makes the language of agreement more abstract or procedural; the goal of maximising common ground is in tension with the goal of minimising ambiguity ”

Second, just because you can use an algorithm to recognise common ground or nuance, that doesn’t guarantee there will be much of it to find. The most successful applications of these methods have been in contexts where substantial effort is put into creating conditions in which people engage in good faith and contribute ideas they believe have the potential to resonate broadly. A widely acknowledged limitation of

Community Notes,^[11] for example, is that only a relatively small percentage – under 10% – of all the notes written are found to be bridging and are displayed publicly. Its effectiveness as a check on falsehoods in the public sphere is thus significantly limited by the thoughtfulness of the note contributors. Still, it is challenging to build such a culture at the scale of a modern social media platform. In contrast, when bridging algorithms are used in processes with additional onboarding and trust-building, with time for deliberation, and in communities small enough for people to have a shared sense of responsibility – for example because they live in the same place and recognise that they have to build some form of everyday life together – they will likely be more successful because there will be more bridging ideas to be found.

Third, it is essential to acknowledge that the accessibility of bridging-based ranking as a tool for facilitators and peacebuilders is not yet where it should be. Perhaps the most accessible implementation is the open-source tool Polis. If you are comfortable using a spreadsheet, it is also straightforward to use bridging-based ranking to curate the results of processes run on Remesh – the tool used by ALLMEP – or of votes collected using any standard survey tool. Bridging algorithms are just one of a set of features often present in an emerging ecosystem of online “deliberative tools” that can facilitate deliberation in polarised contexts. Financial support is available to access these tools if needed^[12].

“ The impact of bridging algorithms is likely to come from their strategic use by peacebuilding and civil society organisations in polarised contexts ”

Where to from here

Bridging algorithms remain an active area of research^[13]. In the social media context, a large academic study recently tested several bridging algorithms with roughly 6,000 participants for 4 months across multiple platforms (Facebook, X, and Reddit), and we

will know soon whether it was successful in reducing affective polarisation. A smaller recent study^[14] testing a form of bridging algorithm on X has already shown that this is possible. But while much of the focus has been on social media, the impact of bridging algorithms is – at least in the short term – likely to come from their strategic use by peacebuilding and civil society organisations in polarised contexts at a more local or regional scale.

*If you are interested in using these methods but are not sure where to begin, please feel free to reach out to the author or to the Deliberative Tech hub of the [Council on Technology and Social Cohesion](#), an emerging community of practice for civil society organisations using these tools.

[1] Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao & Anca D. Dragan. March 2025. “Engagement, user satisfaction, and the amplification of divisive content on social media”. In PNAS Nexus, Volume 4, Issue 3.

[2] Tiziano Piccardi, Martin Saveski, Chenyan Jia, Jeffrey T. Hancock, Jeanne L. Tsai, & Michael Bernstein. 2024. “[Social Media Algorithms Can Shape Affective Polarization via Exposure to Antidemocratic Attitudes and Partisan Animosity](#)”. Preprint.

[3] Aviv Ovadya & Luke Thorburn. 2023. “[Bridging Systems: Open problems for countering destructive divisiveness across ranking, recommenders, and governance](#)”, ArXiv.

[4] Luke Thorburn, Maria Polukarov & Carmine Ventre. 2024. “Societal Sorting as a Systemic Risk of Recommenders”. In 18th ACM Conference on Recommender Systems (RecSys ’24), New York: Association for Computing Machinery, 2024, p. 951–956.

[5] Andrew Konya, Luke Thorburn, Wasim Almasri, Oded Adomi Leshem, Ariel Procaccia, Lisa Schirch, & Michiel Bakker. 2025. “[Using collective dialogues and AI to find common ground between Israeli and Palestinian peacebuilders](#)”. In The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’25), New York: Association for Computing Machinery, 2025, p. 312–333.

[6] Jonathan Stray. 2025. “[Why Do They Think We’re Extreme?](#)”. In *Better Conflict Bulletin*.

- [7] E. Glen Weyl, Luke Thorburn, Emillie de Keulenaar, Jacob Mchangama, Divya Siddarth & Audrey Tang. 2025. “Prosocial Media.” *ArXiv*.
- [8] Chris Bail. 2021. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Prince University Press.
- [9] Colin Irwin. 2025. “The People’s Peace Third AI and UN Edition. Public Opinion, Public Diplomacy and World Peace”. *Peace Polls Publication*.
- [10] See the [Peace Polls](#) site.
- [11] Madison Czopek. 2023. “Why Community Notes mostly fails to combat misinformation”. In *Poynter*.
- [12] Specifically, the AI & Democracy Foundation makes deliberative tools freely available to try under its [Deliberative Tools Access Program](#). If this is of interest, please do feel free to fill out the linked form.
- [13] Madon Revel, Smitha Milli, Tyler Lu, Jamelle Watson-Daniels & Max Nivkel. 2025. “Representative Ranking for Deliberation in the Public Sphere”, *ArXiv*.
- [14] Tiziano Piccardi, Martin Saaveski, Chenyan Jia, Jeffrey T. Hancock, Jeanne L. Tsai & Michael Bernstein. 2024. “[Social Media Algorithms Can Shape Affective Polarization via Exposure to Antidemocratic Attitudes and Partisan Animosity](#)”, *ArXiv*.

About the author

Luke Thorburn

Research fellow at the AI & Democracy Foundation and a final year PhD student in computer science at King’s College London. His research currently focuses on how social media recommender systems can be designed to mitigate conflict risks, an idea sometimes called “bridging-based ranking”. He coauthors the Understanding

Recommenders project of the Center for Human-Compatible AI at UC Berkeley, is a member of the GETTING-Plurality Research Network in the Allen Lab for Democracy Renovation at Harvard University, and a research affiliate of the Machine Intelligence and Normative Theory Lab at Australian National University. Previously, Luke has worked with or advised Ofcom, the newDemocracy Foundation, and Meta. His background is in probability and statistics.

Photography

Symbolic representation of algorithms and their role in polarization in the digital sphere. Autor: Achira22 (Shutterstock).