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 around 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, software for conducting civic forums, 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, cross-cutting, polarization, depolarization, deliberative technology, facilitation, recommender system, ranking, artificial intelligence*

*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.

Contents

1. Introduction **3**
- 1.1. Background . . . 5
- 1.2. Contribution . . . 7

2. Attention-Allocation Systems **8**
- 2.1. Allocation Process . . . 9
- 2.2. Learning Process . . . 9
- 2.3. Optimization . . . 11
- 2.4. Bridging as a Property of Attention-Allocators . . . 12

3. Representation **13**
- 3.1. Data . . . 13
- 3.2. Relation Model . . . 14

4. Quantification **18**
- 4.1. Relation Metrics . . . 18
  - 4.1.1. Quantifying the Degree to which Relations are Good . . . 18
- 4.2. Bridging Metrics . . . 19
  - 4.2.1. Quantifying the Degree to which Attention Events are Bridging . . . 19
  - 4.2.2. Quantifying the Degree to which Attention-Allocators are Bridging . . . 22

5. Evaluation **23**
- 5.1. Validity . . . 23
- 5.2. Reliability . . . 23

6. Discussion **24**
- 6.1. Implementation . . . 24
- 6.2. Challenges, Limitations, and Risks . . . 24

7. Conclusion **26**

A. Appendix **32**
- A.1. Glossary . . . 32
- A.2. System Diagrams . . . 34
- A.3. Roadmap (for this document) . . . 36

1. Introduction

> Imagine a platform that gave people status not for clever takedowns of political opponents but for producing content with bipartisan appeal. ... Instead of boosting content that is controversial or divisive, such a platform could improve the rank of messages that resonate with different audiences simultaneously.
>
> **Chris Bail**, *Breaking the Social Media Prism* [7]

Division impacts cooperation and conflict. We face compounding global challenges including climate change, pandemics, and transformative artificial intelligence, all of which are likely to require significant cooperation to navigate. At the same time, there is significant public concern around increasing societal division [64, 48] and the resulting increase in *destructive conflict* [28]—which can increase the likelihood of large-scale political violence [73]. Destructive conflict and the violence that can result from it not only harms innumerable lives directly—it may also make addressing those global challenges exceedingly difficult [54].

The systems that allocate our attention can impact division. Increases in societal division may be related to the incentives of the systems which guide people's attention—what we call *attention-allocators*. We are all attention-allocators in that we have some agency over how we allocate our own limited attention. But before we can even choose among what to attend to, *upstream* attention-allocators such as the recommender systems on social media platforms, search engines, news media, and even human facilitation have already done much of that allocation for us. From a firehose of potential information, they choose a much smaller set of items for us to attend to in our limited time.

In this paper we focus on these upstream attention-allocators, as they shape the incentives of our attention economy by direction attention to some kinds of behavior over others. Because attention can translate to money and power, such systems help determine what kinds of behaviors are rewarded in many spheres of life.

Systems that directly reward attention may have "bias toward division". Many attention-allocators reward behavior that seeks to maximize attention toward themselves. Recommender systems, for example, largely seek to maximize measures of attention (what is commonly referred to as *engagement-based ranking*), and most news media entities need to attract attention for their work out of financial necessity. Similarly, politicians often aim to attract attention in order to win elections. These incentives reward *engagement-bait*—content and ideas intended to generate engagement, which are often misleadingly sensational and hyperbolic. In practice, engagement-bait can crowd out good faith efforts to communicate across divides, decrease understanding and trust among people of diverse viewpoints, and thus increase division [43, 15].

Two common ways of defining division are *ideological polarization* (differences in policy positions) and *affective polarization* (emotional dislike of those from the other party). Measurements of affective polarization, for example, show increases in many parts of the world [14]. This increase appears to be correlated with the increasing adoption of ubiquitous digital communication, but the extent to which this relationship is causal remains uncertain [43, 39]. Recent work has also moved towards using multiple measures of division, including "resistance to cross-partisan collaboration", "resistance to interpersonal contact with outpartisans", and "willingness to use violent tactics against outpartisans" [72].

This paper explores how to incentivize *bridging*. The goal of bridging is to increase mutual understanding and trust across divides, creating space for productive conflict, deliberation, or cooperation. Every system that involves human attention—from social media recommendations, to search engine ranking, to governance processes—will, to some extent, reward or punish bridging. We explore two core questions: *What do systems that reward bridging look like? How might they be designed?*

Our goal is to clearly articulate an interdisciplinary field that can inform the development and accelerate the adoption of bridging systems across domains. In other words, just as there are fields devoted to other biases we aim to articulate a domain focused on overcoming the "bias toward division."

Bridging is *not* about eliminating conflict or creating homogeneity. The goal is to check the default tendencies of many environments (including much of social media), which can potentially push us toward

extremes when combined with our psychological predilections.[1] As articulated in Stray [63], the intent is "conflict transformation" [42, 13]: not to remove divisions or interfere with the substance of civic debates, but to "[make] conflict better in some way". 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 [74].[2] Figure 1 provides a speculative causal loop diagram [23] illustrating the potential ideal impacts of bridging systems.

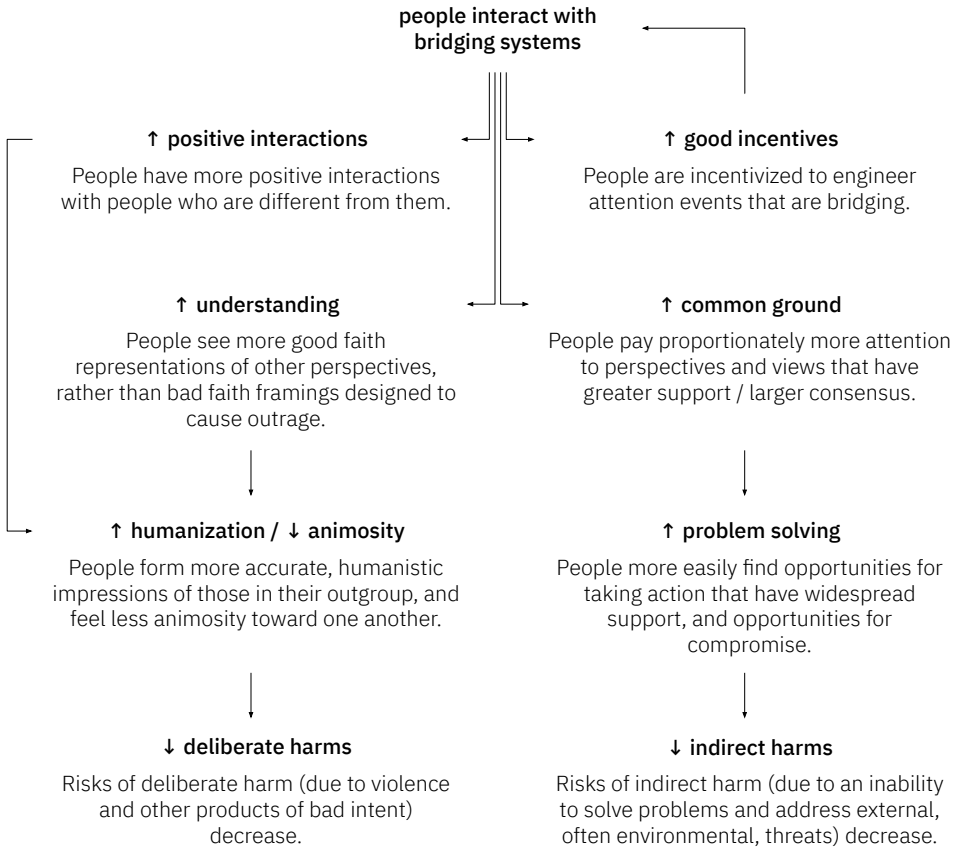

Figure 1: **A causal loop diagram illustrating how bridging systems might impact society.** The goal of this diagram is not to make strong, precise claims about causality, but simply to provide intuition on how a proliferation of bridging systems could have important and beneficial societal consequences 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 [53] 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" [77]. In this way, beliefs about *other people's beliefs* are changing, while personal beliefs are staying fixed.

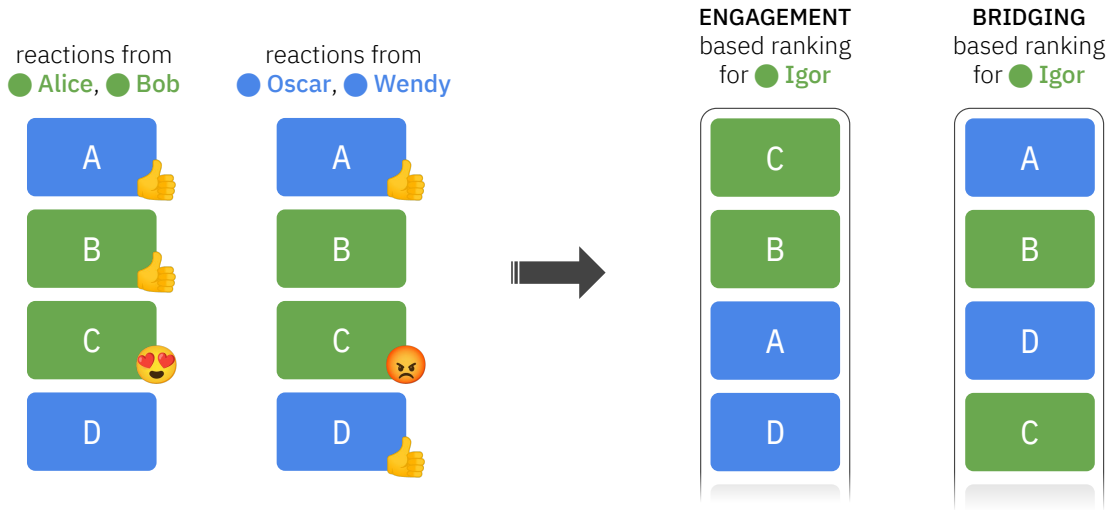

Figure 2: **A simple example of bridging-based ranking.** We see on the left how Alice, Bob, Oscar, and Wendy reacted (or did not react) to posts A, B, C, and D. The color of the post represents the party affiliation of the original poster. On the right, we can see the resulting ranking for the posts for Igor (also green party) under bridging-based vs. engagement-based ranking (posts toward the top are ranked higher). With bridging, the post ranked the highest does not originate from a green party member but is the only post actively liked by both parties. The lowest-ranked post is the one that is the most divisive.

1.1. Background

Throughout this paper we will use examples from three domains to illustrate the concept of a bridging system: (1) recommender systems on social media platforms, (2) software for supporting large-scale civic discussion, and (3) facilitated, in-person deliberation and mediation. Each of these domains is briefly introduced below.

Recommender System: Overview *Example 1*

A recommender system is an algorithm that selects which items of content, from a large pool of available items, should be shown to a user [2]. Recommender systems are commonly used on social media platforms—prominent examples include the algorithms behind the Twitter timeline, the Facebook news feed, and the YouTube homepage.

Their implementations vary, but most recommender systems on social media operate using the same basic logic: they "optimize for engagement", selecting the items of content that are most likely to elicit clicks, likes, reactions, comments, reshares, and other behaviors that the platform can measure [65]. Such behaviors are often an effective proxy by which to select what people want [68], but can also incentivize the creation of content that is misleading, sensational, outrageous, or addictive [11].

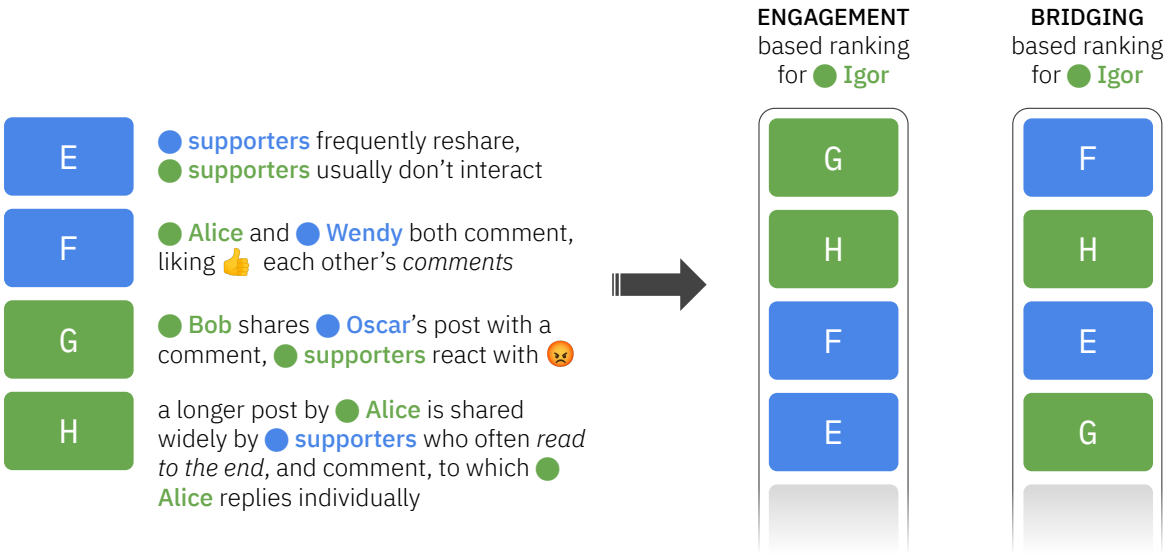

Figure 3: **A more sophisticated example of bridging-based ranking.** This example shows how interaction patterns might be used to determine a bridging-based ranking, and contrasts this with an engagement-based ranking. Note that this is just one way in which bridging-based ranking could be implemented—the exact ranking outcomes will depend on the particular algorithm used.

Civic Forum: Overview *Example 2*

Governments and other civil society organizations that act on behalf of large groups of people often need to elicit the views of their constituents and stakeholders. A variety of tools for coordinating such large-scale dialogues have been created, including some that explicitly seek to elevate common ground. In this document we call these *civic forums*, and focus on two illustrative examples.

The first example, YourView [76], was a not-for-profit civic forum that operated in Australia during the lead-up to the 2013 federal election [61, 35, 36]. YourView provided concise explanations of policy proposals, under which participants could contribute comments for or against the proposal, as well as vote on the comments of other participants, and on the issue overall. From this voting data, YourView derived two measures of public support for the proposal: the raw percentage of participants in favor, and a "Public Wisdom" percentage in which the participants with the highest "credibility score" were given the most weight.

The second example, Polis [56], is an open-source civic forum that has been used to conduct public consultations informing digital policy in Taiwan, among other applications [40, 62]. Polis instances are context-specific forums where participants can contribute comments and vote on the comments of others. From this voting data, Polis clusters participants according to their voting patterns and generates visualizations of the support for each comment among the participants in each cluster.

Human Facilitation: Overview *Example 3*

While there is a very broad range of potential examples of human facilitation, here we focus specifically on mini-publics [31]. They involve the convening of a diverse group of people to deliberate over a particular policy issue, guided by impartial facilitators. The group is selected through sortition, providing a representative random sample (roughly analogous to the selection of jury candidates in the criminal justice system, though more similar in process to representative polling). Increasingly, such mini-publics are being used to make progress on contested political issues, including abortion in Ireland [26] and climate change in France [38]. There is an emerging profession of facilitators who are skilled at convening and coordinating such groups [29, 19]. Mini-public facilitation is of particular interest because, unlike many multi-stakeholder forums, participants are not just elites and represent a wide range of personal experiences and viewpoints in a single facilitation environment.

More broadly, the design of bridging systems draws on work from many disciplines that inform our understanding of sociotechnical systems.[3] This includes work on:

- *articulating what "good" public discourse looks like* for attention-allocators generally, from domains including political theory, philosophy (e.g., epistemology), economics (e.g., social choice theory), communication, anthropology, sociology, rhetoric, and science & technology studies.
- *the collection of accurate social data*, from domains across the behavioral and social sciences including, quantitative history, measurement theory, and survey design;
- *modeling social phenomena*, from domains including computational social science, opinion dynamics, game theory, and political science;
- *the design of interfaces and environments*, from domains including human-computer interaction (e.g., data visualization), economics (e.g., mechanism design, choice architecture), science & technology studies, organizational design, facilitation design, urban planning, and architecture.

We relatedly build on work specifically around existing systems and practices that satisfy our "bridging" definition, including:

- models of deliberative democracy, facilitation, conflict mediation, and peace-building;
- depolarizing recommender systems [63, 4, 6, 5, 78, 20, 58];
- automation of aspects of facilitated deliberation, including deliberative quality analysis [34]; and
- argument-mining, e.g., to enable quantification of the "convergence or polarization" impacts of a discussion [71];

Even this broad set of disciplines is far from comprehensive, and we do not attempt a thorough review of all relevant work here.[4] However, in many cases there are direct precursors to particular bridging system components. We will cite many of these in the examples discussed throughout the paper.

1.2. Contribution

Our goal in this working paper is to articulate a research and practice direction around bridging systems. We provide an overview of the components of attention-allocators, using the examples of recommender systems, collective dialogue systems, and human-facilitated deliberations. We then provide a detailed account of how each component is structured, and how it might support bridging. Finally, we discuss several potential implementation approaches for technical bridging systems. Throughout, we propose intellectually compelling and societally impactful open problems, to help support both cross-domain collaboration and rapid beneficial deployment.

The appendix includes a glossary that provides concise summaries of key terms and concepts, and our current roadmap of how this working paper might be expanded upon. Our goal with this version of the working paper is not to be comprehensive, but to be illustrative and generative—to prompt discussion and further inclusion of what is missing.

[3] There are also many associated non-academic bodies of knowledge around attention-allocators generally and bridging specifically which can be learned from across industry, government, and civil society.

[4] As this is a working paper, we are most definitely open to additional suggestions of work that should be included in this overview.

2. Attention-Allocation Systems

> If a tree falls in a forest and no one is around to hear it, does it make a sound?
> If content is distributed and no attention is paid to it, does it matter?[5]

In this section, we formally define attention-allocation systems (or attention-allocators) and frame bridging as a property of them.[6] Individual people ultimately allocate their own limited attention (and are therefore themselves attention-allocators), but they do not do this alone—much of that allocation work is first partially delegated to upstream social and technological systems.[7] Our three examples—recommender systems, civic forums, and human-facilitated mini-publics—all help allocate this limited human attention and can thus be understood as attention-allocators.

We first introduce the idea of attention events, then build on that definition to more formally define attention-allocation, and finally introduce the bridging property.

Attention Events An *attention event* is the filling of an *attention slot* with an *attention object*. An attention object is simply anything that can be attended to, and an attention slot is a container that can contain an attention object. Examples of attention slots include discrete positions within a recommender feed, or continuous intervals of first-person experience.

The simplest form of attention event occurs when one attention slot is filled with one attention object. We describe these simple attention events as *atomic*. Atomic attention events are represented as

(`slot`, `object`, `properties`).

The third element, `properties`, is a catch-all for data that describe the nature, qualities, and context of the attention event, and which are formalized differently in different contexts. For example, if the attention event represents a real person focusing on something, `properties` might contain how they were focusing, whether they were giving their full attention, other objects which were competing for the same attention slot, etc. More complex attention events can be described as sets of atomic events. For example, an attention event where a thousand people attend to the same object would be represented as the set of atomic events describing each individual attending to that object.

Recommender System: Attention Events *Example 4*

Recommender systems (and more broadly, the user interfaces of social media platforms) create atomic attention events each time a particular item of content (an `object`) is used to fill a particular position within a recommender feed (a `slot`). The `properties` include engagement data such as how long the user subsequently paused on the item (dwell time), but also the social context such as reactions and comment counts which were shown alongside the content.

Civic Forum: Attention Events *Example 5*

Civic forums often have two kinds of attention events. The first occurs when users are shown items for evaluation (e.g., dialogue, voting, etc.), and the second occurs when the most widely supported items (e.g., the results of the votes) are presented to everyone.

Human Facilitation: Attention Events *Example 6*

In facilitated mini-publics, attention events occur fluidly as people engage in dialogue. Here, the attention `object` is an idea being expressed by another person, and the additional properties of the attention event include the tone, facial expressions, and body language with which it is communicated.

A *potential* attention event is one that has not yet happened (and may or may not happen), and a *realized* attention event is one that has already happened.

[5] Content intended for humans might also only be consumed by the "attention" of a machine learning system, and this may have downstream impacts on content that humans do actually consume.

[6] We define "attention-allocation" in an idealized manner which abstracts away many of the complexities of human attention.

[7] This delegation is a result of a number of factors, including the vast extent of potential information to attend to, and power structure, economic incentives, etc.

Attention-Allocation As we formally define it, attention-allocation is a process that determines which of many potential attention events will be realized. An *attention-allocation system* (or *attention-allocator*) takes as input a set of potential attention events and outputs a set of realized attention events.

Often, multiple attention-allocators are required to meaningfully model a given situation. For example, consider a person browsing a social media feed. In this scenario, there are at least two attention-allocators involved. The recommender system is an upstream attention-allocator, algorithmically determining which of many potential attention events (that is, positions of content within the feed) to realize. But the positioning of content within a recommender feed does not wholly determine what the person pays attention to. The person has agency too, and acts as their own, personal attention-allocator deciding which of the many possible ways of allocating their continuous, first-person experience to ultimately enact or realize. We can say that the individual partially delegates their attention-allocation to the recommender. Or, more technically, that the recommender (as an attention-allocator) influences the set of potential attention events that are available as inputs for the person's own, downstream attention-allocator.[8]

While it is perhaps more common to speak of systems for "content distribution" or (algorithmic) "amplification", we believe that "attention-allocation" is often a more useful frame and term because it acknowledges that attention slots are scarce. In particular, human attention is finite and any given individual will only be able to attend to a small fraction of the information to which they have access. For this reason, it is important to focus on what ultimately matters—whether content is actually attended to.

Attention-allocators consist of an *allocation process* and, optionally[9], a *learning process*.

2.1. Allocation Process

The allocation process is the core of an attention-allocator. Its purpose is to determine which of a set of potential attention events to realize—that is, which attention objects (e.g., posts, people, etc.) to use to fill a finite set of available attention slots. The process takes as inputs a set of potential attention events and predicts how each event would, if realized, change the state of the world along several dimensions. These predicted impacts are aggregated into a measure of the "attention-worthiness" of each attention event using a normative *value model*[10]. Finally, the most valuable attention events (according to the value model) are selected and realized.

2.2. Learning Process

The purpose of the learning process is to improve the predictions generated by the allocation process. When a trigger indicates that models need updating, relevant data is retrieved from storage, collected, or elicited from the people interacting with the system[11]. This data is used to update (or finetune) both *state models*—static descriptions of the current state of the world—and *predictive models*—models that predict the impacts of attention events, optionally conditioned on the current state. These updated models are then substituted into the allocation process.

All three of our examples can be viewed as attention-allocators.

[8] This phenomenon of attention-allocators feeding into each other can be called an *attention stack* or *attention delegation network* depending on the structure.

[9] The learning process is optional because some systems (e.g., chronological recommender systems) do not require learning.

[10] The value model formally defines what it means for an attention event to be "attention-worthy". For example, a recommender system might define value to be a weighted sum of different measures of engagement, while a human facilitator will have a qualitative value model that balances the needs of group members to be heard with the need for the group to deliver on its remit.

[11] Going forward, we will simply use "data collection" as a catch-all term.

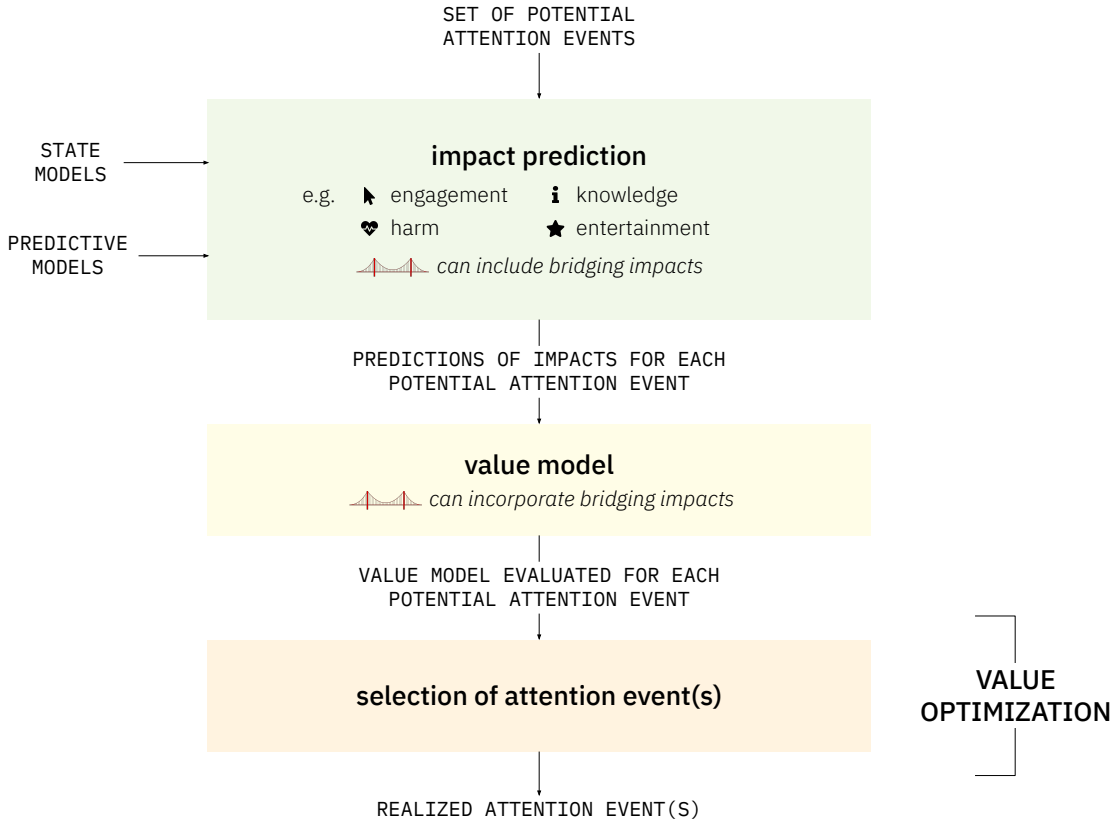

Figure 4: **The allocation process in an attention-allocator.** A bridge icon indicates where bridging can be incorporated. Not shown are the ways in which the system itself is optimized.

Recommender System: Attention-Allocator *Example 7*

For a given social media account, every item of content corresponds to a potential attention event or, more accurately, to multiple potential attention events, one for each context and position in which the content could appear in the user interface[12]. A set of impacts of each of these potential events is predicted. Commonly considered impacts include engagement behaviors (e.g., clicks, comments, shares), whether the user will be entertained (e.g., what they would rate the content), or whether they are likely to be harmed (e.g., whether the content is a financial scam). The predictions are then aggregated using a formal value model [65], and the resulting scores are the primary factor that determines how items are ranked within the recommender feed, influencing which attention events take place. For the most part, recommenders currently direct attention to optimize engagement [65], but could target other goals [63].

[12] In practice, most items will be removed from consideration during a *candidate generation* phase, to reduce computational cost [65].

Civic Forum: Attention-Allocator *Example 8*

As mentioned, civic forums often have two kinds of attention events. For example, Polis only presents one item at a time for voting, and it tries to choose which item to show in order to learn as much as possible about the overall structure of the views and perspectives present in a population. Thus the main "predicted impact" in the allocation process is the information about participants' views provided by the attention event.

Polis then uses that information to create a non-personalized ranking and visualization, showing which items are agreed with the most across those divides—which it calls "group aware consensus". The allocation process here allocates the most prominent attention slots to items that have the most agreement across divides, in order to highlight those items and encourage people to riff on them, suggesting variations that attract yet broader support [62].

Human Facilitation: Attention-Allocator *Example 9*

Facilitators can promote or discourage certain kinds of attention events through the way they structure the deliberations—such as by giving certain people the floor at certain times—with goals (a qualitative value model) including the maintenance of baseline civility and ensuring the group delivers on its remit.

2.3. Optimization

To think clearly about attention-allocators, it is important to recognize that they involve multiple levels of optimization (Figure 5)—what we call an *optimization stack*.

The core attention-allocator described above is an example of bilevel optimization [24], consisting of *accuracy optimization*[13] during the learning process and *value optimization* during the allocation process. But in algorithmic attention-allocators (such as recommender systems), there are at least two other levels of optimization. Downstream of value optimization is the fact that individuals—both producers and consumers of (potential) attention events—will strategically optimize their behavior to further their own goals. Upstream of accuracy optimization are decisions made about the design of the learning and allocation processes. What impacts should be predicted? How should they be weighted in the value model? What data should be collected? Each of these questions will be answered to "optimize" some (perhaps qualitative) measure of the value of the system as a whole.[14]

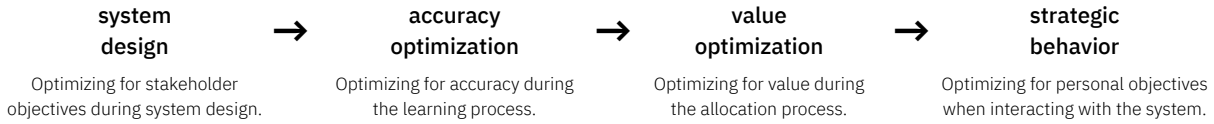

Figure 5: **The attention-allocation optimization stack.** The four levels of optimization that take place within (or adjacent to) an algorithmic attention-allocator, such as a recommender system.

At all levels of optimization, *metrics* are used in a number of ways to quantify the degree to which optimization efforts are successful. As metrics are a significant focus of the remainder of this paper, we have included a list below which summarizes the most common ways in which metrics are used for optimization. Readers with a technical background may wish to skip to Section 2.4.

1. **Ranking.** If the value of the metric is predictable or known for each of a set of alternatives, the alternatives can be directly ranked from most favorable to least favorable according to the metric,

[13] Note that we can only observe phenomena like bridging-ness or entertaining-ness indirectly via measurable outcomes, such as the diversity of people who comment on a post or the watch time on a YouTube video. Thus, whenever we use the word "accuracy", we mean accuracy as measured against these measurable outcomes, which are proxies for the phenomena of interest. In many cases, there may not be a literal ground truth against which to measure accuracy, as it may not be possible to know the true impacts depending on what is predicted (e.g., we can't know true bridging-ness or entertaining-ness, we rely on proxies for them).

[14] *Reward Reports* provide an approach for understanding the components and interactions within this optimization stack [37].

and the most favorable alternative(s) chosen. For example, value optimization may consist of evaluating the value model for each potential attention event, and then promoting those events which are deemed most valuable.

2. **A/B Testing.** If the value of the metric is not easily predictable, then experiments or A/B tests can be performed to produce estimates of the causal effects of each alternative intervention on the value of the metric. For example, system design often consists of conducting a large number of A/B tests to inform which design changes are implemented.

3. **Machine Learning.** If the process being optimized is a machine learning model, then metrics can be included as part of the loss or reward function on which that model is trained. For example, a reinforcement learning-based value optimization process could be trained to maximize the value (according to the value model) of the attention events it selects.

All of these have human facilitation analogs. A facilitator may directly rank topics to prioritize for discussion, may experiment with different methods of structured deliberation to see which work best in a given context, and will be constantly learning over time what facilitation strategies to use, according to their qualitative, internal "metrics" that measure the degree to which deliberation is successful.

2.4. Bridging as a Property of Attention-Allocators

An attention-allocator is *bridging* to the extent that it realizes attention events which increase mutual understanding and trust across divides, creating space for productive conflict, deliberation, or cooperation (i.e., the "bridging goal").

Bridging is a property of attention-allocators, rather than a distinct kind of system. In general, attention-allocators may either facilitate or discourage such events to varying degrees, so there is a spectrum of attention-allocators that are more or less bridging. Bridging systems differ in degree, not in kind, from non-bridging systems.

In principle, one way to promote bridging would be to cause people to value bridging impacts more within their own, personal attention-allocators. We suspect this approach, requiring people to unilaterally change their behavior, is likely quite ineffective. In the remainder of this paper, we focus on algorithmic attention-allocators such as recommender systems and civic forums. In algorithmic systems, the qualitative bridging goal can be formalized using models and metrics which are intended to capture how near a population is to the goal, or the degree to which a given attention event moves a population nearer to the goal. These formalisms are introduced in Sections 3 and 4.

There are several specific places in which bridging can be incorporated into an attention-allocators, some of which are indicated by bridge icons in Figure 4 and 11. For example, the system can model divisions present in a population and can predict, in the *impact prediction* stage of the allocation process, whether divisions will improve or deteriorate. These predictions can then be included in the value model [63]. *Bridging metrics* could also be considered at the level of system design, and be used to adjudicate which of a set of potential interventions should be implemented. In Sections 3-4, we formalize the notion of relation metrics and bridging metrics, and describe how attention-allocators can become bridging.

3. Representation

Above, we stated that an attention-allocator is bridging to the extent that it facilitates attention events which support the "bridging goal": increasing mutual understanding and trust across divides, creating space for productive conflict, deliberation, or cooperation. This qualitative property must be formalized if we are to extend the insights from offline bridging practices that have been developed over millennia into digital attention-allocator such as recommender systems and civic forums.

In the next two sections, we describe how the notion of bridging can be formally represented and then quantified. In particular, we introduce the concept of a *relation model*—a representation of the relationships or affinities between individuals in a given population. This model is used along with relation metrics that quantify the extent to which a relation model is "good" (the state of the relation model), and bridging metrics that quantify how a relation model is "improving" (the dynamics of the relation model).

3.1. Data

To model bridging, you need relevant data. Different systems use different approaches for collecting or eliciting such information, and the data can take different forms. Examples are given below from our three example domains.

Recommender System: Data Elicitation *Example 10*

Users of online platforms generate engagement data such as likes, reactions, comments, shares, dwell time, tagging, direct messages, and so on. Recommender systems use this data to learn about the preferences and perspectives of users. Implicitly, this data reveals how users differ, and the degree of affinity between them.

Civic Forum: Data Elicitation *Example 11*

Each instance of a civic forum is focused around an issue, prompt or question. Participants can either provide new responses or vote on existing ones. The user interfaces generally differ from those of conventional forums such as Reddit. For example in Polis, people are shown a succession of responses one by one, and can choose to agree, disagree, or pass in response to each statement. Polis' algorithm selects which statement to show next using a number of criteria, one of which is to maximize learning about the relations between participants, responses and, implicitly, each other [62].

Human Facilitation: Data Elicitation *Example 12*

Facilitators "reading a room" identify a host of subtle cues over the course of deliberation. Common examples include flared nostrils, changes in breathing, and hushed silences. Structured exercises may also be used to sort the room into groups of perspectives [19].

These data contain information about people's preferences, perspectives, opinions, affiliations or worldviews. Implicitly, they thus contain information about the relationships and affinities that exist in a population.

In many cases, these "data" would have been "collected" anyway. For example, an effective facilitator will intuitively gather such information in the course of interacting with a group. Similarly, a recommender system where users interact with each other will implicitly be eliciting data useful for relation modeling.

Note, however, that the data elicited by such systems is very contextual—it is highly dependent on system/process design, affordances, culture, the environment, existing divisiveness and many other factors. In all cases, respect for privacy is valuable both ethically and for gathering accurate data.

Open Questions: Eliciting Data for Bridging *Set 1*

1.1. What affordances provide the most useful information about relationships in a population?

1.2. Can elicitation methods from one domain (say, mini-publics) be translated to another (say, recommender systems)?

1.3. How can we mitigate and account for the fact that the act of eliciting information about relationships can itself influence those relationships?

3.2. Relation Model

A *relation model* is a formal representation of the relationships between people in a population, which can be learned or inferred based on the data available. These relationships could be explicit (e.g., friendships) or implied (e.g., the affinities between people who have similar preferences or worldviews). The relation model is a "state model", describing a snapshot of the world at a given point in time.

Formally, a relation model—as we define it—can be decomposed into three components: `people` (the people who interact with the system), `items` (the alternative objects to which `people` may attend), and `relations` (the one-to-one relations between `people` and `items`, intended to capture goodwill, agreement, affinity, reactions, or similar). This general framework highlights a common structure shared by our three examples, as summarised in Table 1.

	Recommender System	**Civic Forum**	**Human Facilitation**
`people`	users	participants	participants
`items`	posts / content	comments	claims / positions
`relations`	revealed preferences	agree, disagree, or pass	qualitative opinions

Table 1: Recommender systems, civic forums, and human facilitation all share a common structure.

The three components will be modeled differently in different contexts. Most comprehensively, the relation model may simply be the totality of all available data available about the population. More structured models can also be used, two common examples of which are *graph-based* models and *space-based* models.

Graph-based Graph-based models represent `people` as nodes in a (mathematical) graph or network. The edges in the graph characterize the relationship between people, and can represent explicit, active communication channels, or be weighted to represent more fine-grained and abstract types of affiliation. A simple example of a graph-based model is given in Figure 6(a).

Space-based Space-based models represent `people` as locations within an ambient "opinion space". The similarity between people's opinions, preferences or viewpoints is characterized by how close to one another they are in this space. For example, people may be modeled as a location on the left-right political spectrum (a one-dimensional space), or assigned a position on a political compass (a two-dimensional space). A simple example of a space-based model is given in Figure 6(b). In general, the space might have hundreds or thousands of dimensions. The `items` may also be represented as points in the same space, in which case the proximity of a person to an item represents the degree to goodwill, agreement, or other "favorable" relationship between them.

These two approaches to relation modeling are not exhaustive. For example, it may make sense to model relations at a higher level of abstraction, because it is not necessary to model individuals to represent useful information about divisions in a population. For example, a very simple relation model might be: {70% of people like *The Beatles*, 50% like *Adele*, 10% like *Nickelback*}[15]. This tells us that there is overlap between people who like The Beatles and Adele, that both are fairly popular, and that Nickelback is comparatively not. Such *aggregate models* could incorporate item classifications (e.g., Nickelback might

[15]We use musical artists as an example, but these could equivalently be levels of support for different political positions.

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

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 [27, 59].

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 [59].

Civic Forum: Relation Model *Example 14*

Civic forums 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 [62]. An example of the YourView space-based model is shown in Figure 7.

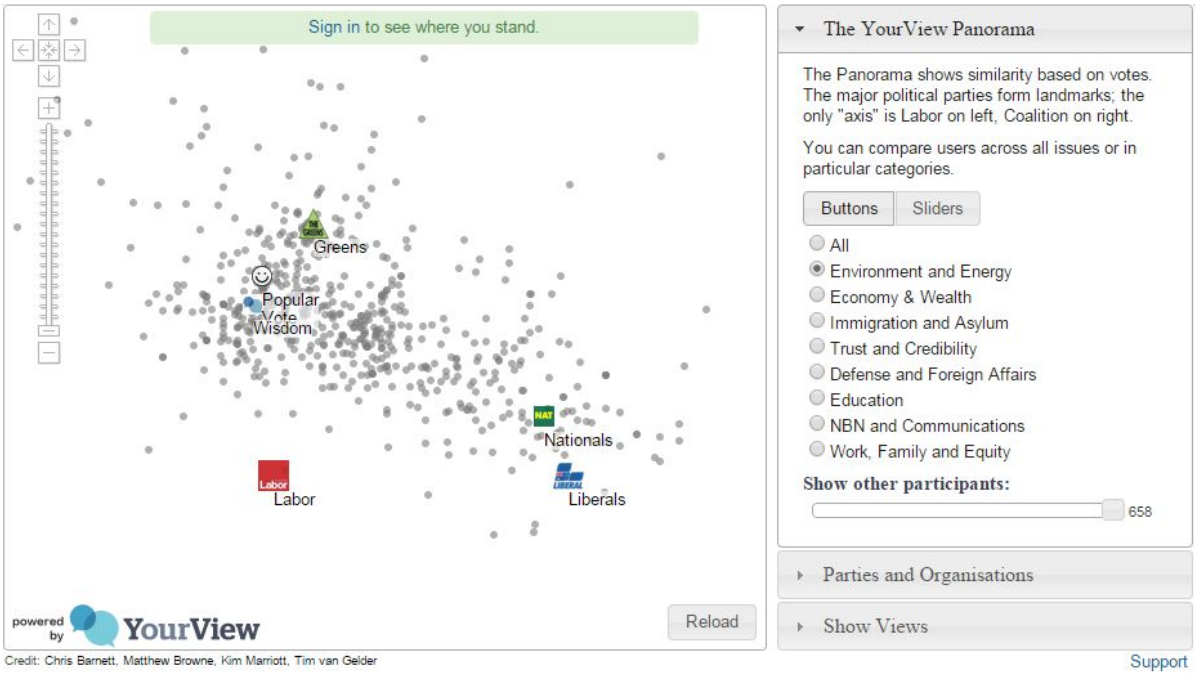

Figure 7: The YourView "Panorama"—a two-dimensional space-based model.

Human Facilitation: Relation Model *Example 15*

Facilitators pay attention to how people relate to each other, and where they fall along the salient axes of disagreement within the groups that they are facilitating, thus intuitively applying qualitative versions of both graph-based and space-based models.

By introducing the term "relation model" we are not aiming to prescribe a particular kind of model or claiming to have invented one, but merely aiming to describe a *role* that certain models can play within an attention-allocator. Models from many existing fields (computational social choice, opinion dynamics, recommender systems, etc.) could be used as a relation model. The new term also helps when talking about the commonalities across recommender systems, civic forums like YourView and Polis, and human-facilitated deliberations.

How can information about the quality of relations be extracted from space- or graph-based models? In both cases, the models can be used to identify clusters or groups of people who think similarly. Divisions can be thought of as the spaces between these groups: how far apart they are, how they think about each other, and how they interact. Annotating the models with such additional structure can provide insight into the nature and strength of divisions that exist in a population. Figure 8 presents examples of this sort of clustering in both graph-based and space-based models.

Civic Forum: Clustering *Example 16*

In Polis, `people` are then clustered into two to five distinct groups, an example of which is shown in Figure 8(b). Clustering is performed directly on the two-dimensional projection using the k-means clustering algorithm, and a goodness-of-fit statistic is used to determine which number of clusters best fits the vote data [62].

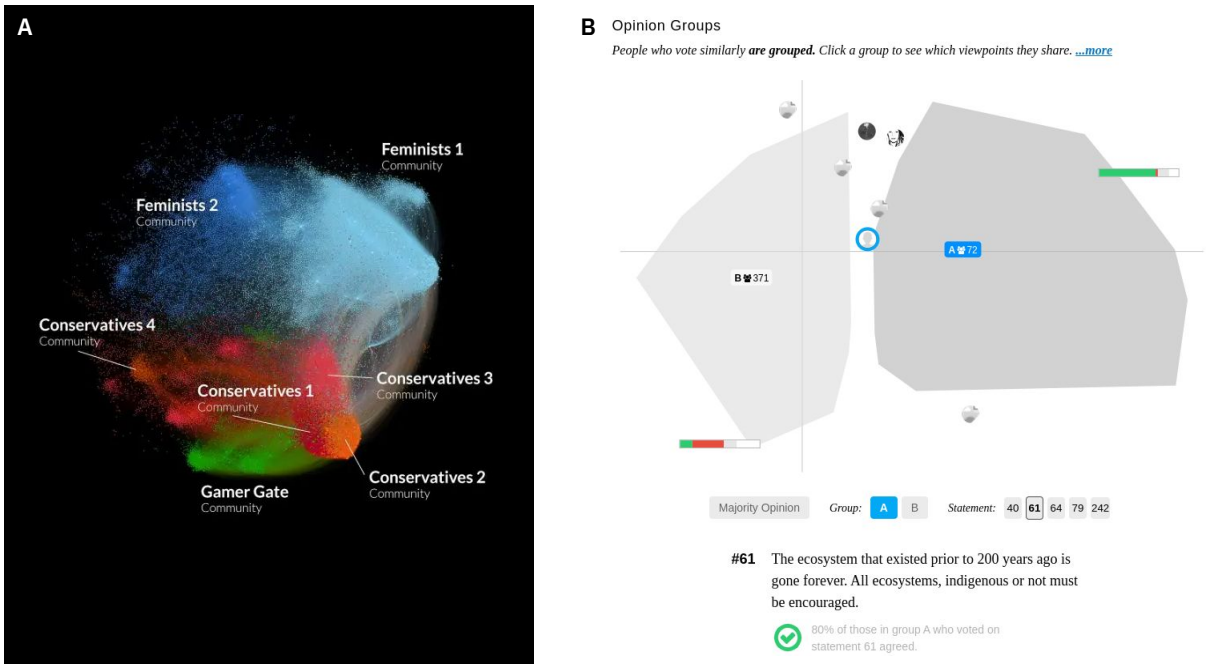

Figure 8: **Examples of clustering in relation models.** Figure (a) is a graph-based model of Twitter users, clustered by their stance regarding the legal status of abortion. Figure (b) is a space-based model generated by Polis, where participants in a civic forum conducted by a New Zealand newspaper expressed their views on protecting biodiversity. *Image credits: Clifton [21] and Scoop [60].*

Open Questions: Relation Models *Set 2*

2.1. What kinds of representations are appropriate for a given context? For example, discrete clusters (Polis) versus continuums (Twitter Community Notes).

2.2. How can these representations be best operationalized in that context? For example, Polis currently uses PCA to project the vote space into two dimensions, sets k for k-means clustering to 2-5 groups, and chooses the best k using silhouette coefficient to identify clusters.

2.3. Is the data collected and relation models of existing attention-allocators, such as recommender systems, sufficient to perform bridging-based ranking?

4. Quantification

Relation models represent the structure of opinions and divisions within a population. However, they do not capture normative judgments about how "good" that structure is, so it is difficult to articulate in their terms what constitutes an improvement or deterioration. To incorporate bridging into automated attention-allocators, we need metrics that formalize the bridging goal—that is, metrics that encode normative values about what it means for the relation model to be good, bad, improving or deteriorating.

We focus here on two kinds of metrics: *relation metrics* and *bridging metrics*. Relation metrics capture the quality or "health" of relations at a given point in time (as represented by the relation model). In contrast, bridging metrics capture the degree to which attention events or attention-allocators are associated with improvements in the relation model, over time.

While one way of defining bridging metrics is to simply calculate differences between relation metrics, this may not be sufficiently practical or computationally efficient for many use cases. We thus include bridging metrics as a distinct kind of metric (Section 4.2) to also capture other, more heuristic methods of computing them.

4.1. Relation Metrics

Relation metrics are summary statistics that characterize the *state* of a relation model at a given point in time. They can be used for monitoring the quality of relations over time, and as the foundation for bridging metrics, which are described in Section 4.2.

4.1.1. Quantifying the Degree to which Relations are Good

To be useful, relation metrics need a clear normative interpretation. Within a given context, we should be able to say that higher metrics reflect success at achieving the bridging goal, or that certain configurations of multiple relation metrics are better than others.

Candidates for relation metrics include existing measures of polarization, which have been proposed for both space-based [30, 10, 44] and graph-based models [51, 47]. However, there are many different kinds of polarization [16], and it is not clear whether existing polarization measures are appropriate targets for optimization. We believe the space of relation metrics is significantly underexplored.

Recommender System: Relation Metrics *Example 17*

In the context of recommender systems and social media, relation metrics have mostly been proposed as summaries of graph-based models, such as follow networks on Twitter or the hyperlink networks of online news publications. These include metrics based on *homophily* (the degree to which nodes are connected to nodes that are similar to themselves), *modularity* (the number of intra-group connections relative to the number of inter-group connections), *random walk controversy* (a measure of how likely you are to cross between groups when randomly traversing the network), and *balance theory* (which measures how consistent the network is with properties such as "my friend's friend is my friend" and "my friend's enemy is my enemy"). See Interian et al. [41] for a comprehensive review.

Open Questions: Relation Metrics *Set 3*

3.1. For a given impactful system, e.g., a search engine ranking system, what kinds of relation metrics are meaningful?

3.2. How does the answer differ by system, type of relation model, and context? How can we evaluate the appropriateness of relation metrics?

3.3. What can be done to minimize negative side effects if relation metrics are used as targets for optimization?

3.4. What specific technological approaches are most applicable to enabling relation metrics at scale, what foundational work can help accelerate that, and what are the potential limitations and risks of those approaches?

4.2. Bridging Metrics

While relation metrics quantify the quality of relations at a given point in time, *bridging metrics* quantify the degree to which relations improve or deteriorate over time. We consider two types of bridging metrics: *event-level* and *system-level*.

4.2.1. Quantifying the Degree to which Attention Events are Bridging

To promote bridging attention events, we need to be able to identify them. For example, we want to be able to quantify the extent to which recommending a particular Facebook post, or presenting a particular claim in Polis, will lead to an improvement or deterioration in the relation model.

An *event-level bridging metric* is a summary statistic that characterizes the effect of a specific attention event on the relation model. For a given attention event, a positive bridging metric should correspond to an improvement in the relation metrics. Conversely, a negative bridging metric should correspond to a deterioration in relation metrics.

Event-level bridging metrics can be observed or predicted, depending on whether the attention event has already happened. When predicted, they can be included among the predicted impacts of the attention event in the allocation process of an attention-allocator (Figure 4), and included in the value model by which attention events are selected. When observed (retrospectively), they can be used for monitoring, and compiled into datasets to train, in a supervised fashion, the models used for impact prediction.

Event-level bridging metrics can be either framed in terms of the relation model, or be based on a heuristic. We describe each of these approaches below.

Relation Model-based In principle, an attention-allocator could be designed to first test every single attention event by showing (or not showing) the relevant item to small treatment and control groups of similar individuals in a paired study design (e.g., via A/B tests). After observing how the relation model changes in each group, it would be possible to produce estimates for the changes in relation metrics attributable to the attention event. These estimates could be used to predict how similar events would impact relation metrics. The estimates from the tests, and the predicted impacts are examples of event-level bridging metrics.

Such bridging metrics are explicitly grounded in the relation model and the associated relation metrics, so represent the (academically) "ideal" way of quantifying the degree to which attention events are bridging. However, there are a number of practical challenges to computing bridging metrics in this way.

- It is likely computationally intractable to perform such experiments and update the entire relation model for every potential attention event.
- Subjecting people to attention events—some of which will be dividing—for the sake of quantification is ethically dubious and would hamper efforts to improve the relation model.
- Changes in relation metrics that summarize the entire relation model caused by atomic attention events may be too small to observe or meaningfully compare (particularly given the noise of other events).
- Relation metrics describe the quality of a relation model across the complete population of users, so it may be difficult experimentally isolate relation metrics for the treatment and control groups.

For these reasons, the most immediately implementable bridging metrics will use heuristics, rather than be explicitly defined in terms of, or derived from, the relation model.

Open Questions: Bridging Metrics *Set 4*

4.1. How can model-based, event-level bridging metrics be made tractable? (Perhaps by considering "local" effects or small subsets of the relation model?)

Heuristic Heuristic rules can be used to identify attention events that belong to a broader class, the prevalence of which is thought (or known) to cause relation metrics to improve.[16] These classes of attention events are characterized by particular patterns of activity or interaction that can be easily observed.

Visual intuition for four previously-proposed heuristics is given in Figure 9. We describe each in more detail below.

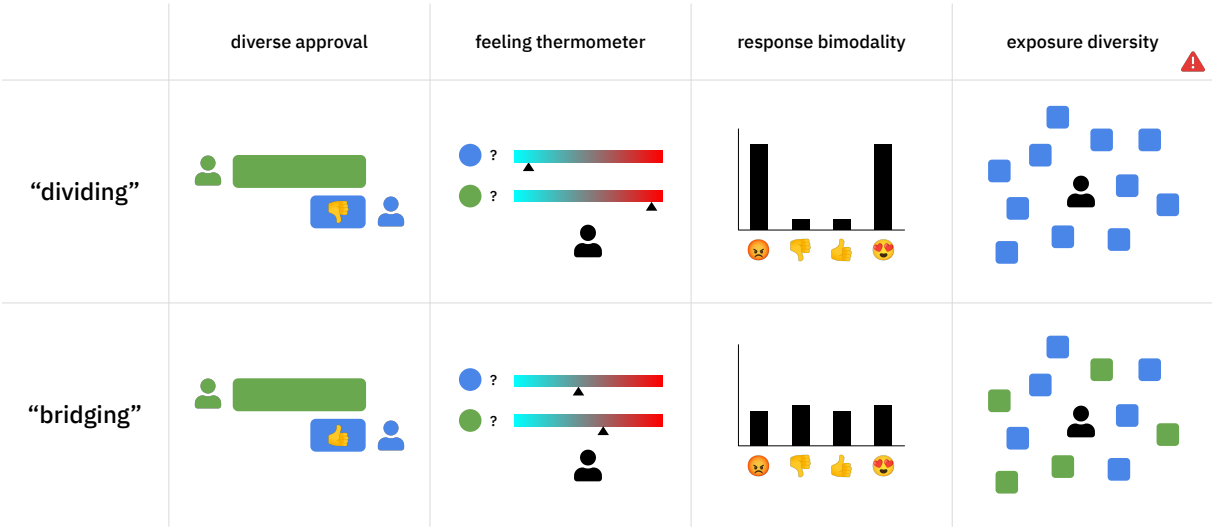

Figure 9: **Visual intuition for four heuristic approaches to bridging.** Note that evidence for the validity of all four heuristics is limited, and in particular, use of the exposure diversity heuristic has been shown to worsen divisions in some contexts (see main text).

- **Diverse approval** The first heuristic, *diverse approval*, states that bridging occurs when a person creates an item that is approved, supported, or otherwise endorsed by people who would normally disagree with them.[17] A slightly stronger version states that bridging occurs when an item is endorsed by people from diverse viewpoints. Versions of this heuristic have been operationalized with success in at least two bridging systems.

Civic Forum: Outgroup Approval *Example 18*

In YourView, each participant had a "credibility" rating—a type of reputation score that quantified, primarily, the degree to which the comments they contributed on any particular issue attracted support from people who disagreed with them about that issue. Thus grounded, the definition of credibility could be extended recursively: YourView assumed that people were credible if they were respected or trusted by credible people they disagreed with. This circular definition is akin to the logic of the PageRank algorithm or the concept of "eigentrust" [1].

Recommender System: Diverse Approval *Example 19*

Community Notes (formerly known as Birdwatch) is a feature Twitter recently launched that allows users to "collaboratively add helpful notes to Tweets that might be misleading" [75]. In its current form, Community Notes contributors can also rate notes contributed by others as "helpful", "somewhat helpful" or "not helpful". Notes are awarded a high helpfulness score (and hence displayed publicly) only if they have been "rated helpful by raters with a diversity of viewpoints". This is achieved using a version of the matrix factorization algorithm that is commonly used in recommender systems [69].

[16] While there is the possibility of demonstrating causality at the class level, such as via A/B tests, causality can not be shown at the level of individual items at scale, so we refer to all these approaches as "heuristics".

[17] It is theoretically straightforward to determine when there is a divide between two people; this can be quantified by a metric between their respective vertices or positions in graph- or space-based relation models. Such divides are already inferred automatically by many existing attention-allocators.

- **Feeling thermometer** The second heuristic, the *feeling thermometer*, is a short survey instrument commonly used to measure affective polarization. In its simplest form, it consists of asking people to rate their feelings towards their country's two largest political parties on a scale from 0 (cold) to 100 (warm). The difference between these two scores indicates the degree of affective polarization. Stray [63] proposed incorporating this short survey instrument into social media platforms. Survey responses could be linked with items recommended to the user shortly prior to their completing the survey. In this way, the degree to which responses trend towards less affective polarization could be used to quantify the degree to which recent attention events were bridging. However, as it places an additional burden on the user to periodically complete short surveys, the data collected will likely be sparse.

- **Response bimodality** Under the third heuristic, *response bimodality*, attention events are said to be (relatively) bridging if the distribution of ratings or reactions elicited by the relevant item is not polarized. Intuitively, a distribution is most polarized if it is markedly bimodal (that is, it has a "U" shape). One line of research on the use of recommender systems for depolarization relies on this heuristic [4, 6, 5]. They quantify the degree of polarization in a rating distribution by training a binary classifier that takes in features computed from a histogram of item ratings and outputs whether or not the distribution is polarized, as trained on "human expert" classifications.

- **Exposure diversity** The fourth heuristic, *exposure diversity*, states that bridging occurs when people attend to items from diverse sources, particularly from sources they don't normally see. This heuristic suggests that to facilitate bridging attention events, people should be shown items from outside their "filter bubble" or "echo chamber". However, a number of studies have failed to find evidence of algorithmic filter bubbles [18], and other research suggests that indiscriminately showing people posts from their outgroup may cause relations to deteriorate [8]. For these reasons, the validity of the exposure diversity heuristic is questionable.

Depending on whether the metrics are predicted or observed, event-level bridging metrics based on this heuristic would either be (a) the probability that a potential attention event results in the heuristic interaction pattern, or (b) the degree to which the items being attended to already take part in instances of the heuristic interaction pattern.

The interaction patterns captured by the above four heuristics are conceptually simple.[18] We can also define more complex and subtle heuristics that, for example, model longer patterns of interaction, or incorporate more nuanced information about the people and items involved. Below, we give an example of a class of heuristic, *positive interactions across divides*, which generalizes the idea of "diverse approval".

Recommender System: Positive Interactions Across Divides *Example 20*

Consider a hypothetical social media platform named BridgeTok. BridgeTok built a dataset of posts with their subsequent social exchanges (reactions, comments, shares, etc.), along with human ratings of the degree to which each exchange was "positive". They then trained a model to predict the positivity of any arbitrary interaction pattern. These predicted positivity scores, combined with existing user embeddings, are used to quantify the degree to which new interactions represent "positive interactions across divides". The most bridging attention events, according to this heuristic, are then selected for in the BridgeTok recommender system. Concrete examples of what such *bridging-based ranking* might look like are given in Figures 2 and 3.

Note there is a precedent for this kind of modeling at Facebook: among the Facebook Papers provided by Frances Haugen there is mention of experiments using certain interaction patterns (referred to as "motifs") to identify positive "conversations" and civic interactions [32, 33].

The focus on patterns of interaction, without consideration of the meaning of natural language content, increases the extent to which these heuristics can be implemented internationally, across languages. That said, the effectiveness of a given "structural" heuristic for identifying bridging behaviors may depend on cultural and social context, and in some settings content-based heuristics can also be useful. For example, one can use a machine learning model to estimate the degree to which content is dehumanizing or contemptuous. The inverse of this estimate can then be used as a heuristic bridging metric for attention events involving that item of content.

[18] Though in most cases, their computational complexity remains an open question.

There is currently limited evidence linking heuristics to improved relation metrics, so they should be used with caution. The *Strengthening Democracy Challenge* [72] provides one example of what such evidence might look like. Both model-derived and heuristic approaches can be evaluated within the framework of *measurement theory* (Section 5). Heuristics may also be derived from more expensive deliberative signals [49].

Open Questions: Bridging Metrics *Set 5*

5.1. How can we evaluate the quality of bridging heuristics?

5.2. What are the relative strengths and weaknesses of bridging heuristics? Which should we use?

5.3. How common are examples of outgroup approval in practice (e.g., on modern social media platforms)? Are there enough examples to make a meaningful impact if they are featured more prominently by recommender systems?

5.4. To what extent does giving heuristic bridging metrics more weight in the value model actually incentivize bridging behavior? In economic terms, what is the "value model elasticity of bridging"?

5.5. How computationally complex are different bridging heuristics?

5.6. To what extent are particular bridging heuristics valid across international contexts?

4.2.2. Quantifying the Degree to which Attention-Allocators are Bridging

A *system-level bridging metric* is a summary statistic that characterizes the effect an entire attention-allocator has on relations in a population, over time. The bridging metrics for an attention-allocator should be positive if the system leads to an improvement in relation metrics over time, and negative if it leads to a deterioration in relation metrics over time. Such metrics can be used by system designers when evaluating the impacts of potential new features (as measured by A/B tests) and deciding which new features should be widely implemented. They can also be used by external auditors when assessing the impact of the system.

There are two main approaches to defining system-level bridging metrics. The first is to consider how relation metrics have changed over time. This may be simply as the system has been in place, or specifically the difference in relation metrics pre and post some change in the system design. The second approach is to aggregate event-level bridging metrics. For example, the average value of an event-level bridging metric across all recent attention events would constitute a system-level bridging metric. Other possible aggregation methods include taking the proportion of recent attention events that have positive event-level bridging metrics, or the absolute frequency of such events.

Open Questions: Bridging Metrics (system-level) *Set 6*

6.1. Given the challenges of measuring the effects of online platforms [67], how can we produce system-level bridging metrics that can be interpreted as causal estimates?

6.2. How might bridging metrics be misaligned, or be affected by the different versions of Goodhart's Law [45]?

6.3. How could bridging metrics be gamed or abused? What would 'bridging-bait' content look like?

5. Evaluation

The metrics described above are intended to be optimized as part of the value model in an attention-allocator. For this reason, it is important that these formalisms are closely connected to the underlying bridging goal (an increase in mutual understanding and trust across divides, creating space for productive conflict, deliberation, or cooperation). If the metrics do not capture the aspects of reality we care about, then efforts to optimize them will be at best ineffective, and at worst harmful.

Fortunately, there is a mature literature in the field of *measurement theory* that provides a framework for thinking about the quality of relation metrics and bridging metrics. In this section we discuss two properties that it is good for metrics to have—validity and reliability—and consider what they mean in the context of bridging systems.

5.1. Validity

Validity refers to whether a metric actually measures what it claims to measure. More precisely, validity is the "degree to which evidence and theory support the interpretations of [metrics] for proposed uses of [those metrics]" [3].

Importance To see the importance of validity for relation and bridging metrics, consider the consequences of optimizing for a metric that evidence suggests has poor validity. For example, a platform with good intentions may decide to reward exposure diversity in their recommender system—that is, to specifically show their users content from political leanings they disagree with. They might be "successful" according to this measure, and interpret the increasing diversity of content viewed as indicative of increasing mutual understanding and trust across divides. However, it is possible that increased exposure to contrary viewpoints would actually cause people to double down on their existing beliefs and hence cause relations to deteriorate [8]. The reason for the discrepancy between the platform's good intentions and the regrettable outcome is that exposure diversity has poor validity as a measure of bridging.

Relation and bridging metrics may lack validity for several reasons. They may be based on misconceptions (as in the case of exposure diversity). They may track the quality of relationships in a narrow scope observable by the platform, but not correlate with the quality of relationships in broader society. They may identify content that is bridging if shown to one person, but not bridging when shown to 10 million (due to second-order effects). They may be based on relation models that are not expressive enough to capture the underlying plurality of perspectives (this may be the case with space-based models [12, 66]). They may succumb to Goodhart's Law [45]—that is, they may be inflated by actions that do not respect the original spirit of the metric. For example, promoting puppy videos may create more cases of diverse approval, but not have any effect on the quality of political discourse.

5.2. Reliability

Reliability is the degree to which a metric takes on similar values when calculated in similar contexts. A reliable relation metric will give similar values if calculated multiple times in quick succession, if independently calculated by different teams within a company, or if calculated on multiple random samples of a population of platform users.

Importance Reliability is important because it is a prerequisite for validity. If a metric is too noisy or cannot be replicated, then it cannot be valid as an optimization target.

Open Questions: Evaluation *Set 7*

7.1. What gold standard measures or benchmarks should we use to evaluate relation models, relation metrics and bridging metrics?

7.2. Which measures are context dependent, and in what contexts are they appropriate?

7.3. What are the limitations of graph- and space-based relation models (see e.g., [66])? How significant are they?

6. Discussion

In this paper we have emphasized connections and open questions across three domains—recommender systems, civic forums, and human facilitation—to help formalize an interdisciplinary research area focused on bridging.

6.1. Implementation

Table 2 provides a summary of concrete steps that can be taken by practitioners to incorporate bridging into attention-allocators.

Allocation Process	Include bridging impact(s) as a predicted impact
	Give meaningful weight to bridging impact(s) in the value model
Learning Process	Collect / retrieve / elicit data on relations
	Maintain and update a relation model
	Maintain and update model(s) that predict the impact of an attention event on the relation model
System Design	Monitor both relation and bridging metrics over time
	Consider bridging metrics when interpreting the outcomes of A/B tests
	Collect data that will help you do bridging better in future

Table 2: Examples of how to incorporate bridging into attention-allocators.

6.2. Challenges, Limitations, and Risks

The idea of bridging systems and the interdisciplinary research we aim to articulate are not free from challenges or controversy, and should not be considered a panacea. Here we explore these potential challenges and related limitations and risks. After grounding the discussion in terminology and common misunderstandings we articulate challenges to the goal of bridging itself, to our formalization of bridging, and to the implementation of bridging.

Intuition versus formalism We defined the intuition for the bridging as follows: *an activity or relationship is said to be bridging to the extent that it leads to increased mutual understanding and trust across divides, e.g., creating space for productive conflict, deliberation, and/or cooperation.* It is common when going from intuitions to formalisms that only a small part of the intuition is captured, sometimes with significant negative societal implications. We attempt to navigate that dilemma by making the intuition explicit, and emphasizing it as the ultimate goal to measure success (and formalisms) against. We relatedly don't attempt to make the formalism fully capture this intuition, as no single formalism understandable by a human is likely to do so. We expect that the best outcomes are likely to result from relying on a broad set of bridging metrics that address different aspects of the intuition.[19]

Common misunderstandings This notion of bridging is commonly misunderstood (and may be incorrectly formalized) in ways that would lead to outcomes that fall outside of the definition. Bridging does *not* mean bringing everyone to a homogeneous center or eliminating conflict. Pluralism and productive conflict are valuable. Bridging also does *not* mean just showing content from across divides; research suggests that this can sometimes *increase* division [7].

Challenges to the goal of bridging As the bridging intuition is imprecise, it can have different meanings at different levels of abstraction, formalism, and implementation. This makes the notion of bridging and the formalisms used to measure or apply it contestable.

[19] We also recognize that this level of investment will not always be feasible. Even creating a set of modular bridging metrics and algorithms appropriate to a large variety of contexts is a significant public goods challenge.

Many would argue that there should not even be an attempt to bridge some kinds of divides due to the abhorrence of particular groups, or their lack of respect for human rights [46]. This implies a key question: when is bridging appropriate, and when is it not [57]? There are also arguments that destructive conflict may be an important driver of social change, so reducing destructive conflict could limit the potential for such change [25, 50]. It has been argued that bridging could "interfere with important processes that may be necessary for a democracy to determine the best path forward" [22]. Finally, there may be differential benefits of bridging to particular groups. For example, some groups may be less willing or able to change than others, e.g., due to significant punishments for engaging outside of the group dogma or because some groups do much more "internal bridging work" than others.

Challenges to our formalization of bridging Formalizing bridging around attention-allocation does not account for material sources of division, such as physical violence, economic inequality, or historical injustices. Extensions of bridging to more general allocation problems (e.g., around physical resources) may perhaps be helpful in addressing some of these critiques, but bridging would still be far from a panacea. In a world with many divides, attention events that bridges some divides may exacerbate others, and formalisms may not be able to meaningfully balance those contradictory impacts.

Specific choices of relation models, relation metrics and bridging metrics may lack validity (see Section 5.1) or may not be aligned with other societal values and goals. For example, some approaches might cause unintended side effects or consequences, some of which may be significant, such as an undermining of human agency or manipulation of public opinion. Relatedly, some may argue that any formalism or optimization of concepts as complex and interpersonal as bridging is inherently inhumane or invalid.

Regardless of whether humane formalization of bridging is possible, it is true that any practical formalism will be a considerable simplification of the complex psychological and social changes that correspond to the bridging goal. Such simplifications may not reflect important aspects of bridging and would benefit from ongoing critique and improvement.[20]

Challenges to a technological implementation of bridging Adversaries are likely to try to use fake or manipulated accounts in order to influence bridging-based ranking algorithms. This is already true of existing algorithms and similar safeguards may apply, but it is possible that there are either more or less algorithmic vulnerabilities as a result of a bridging-based ranking approach. We may also not have the technological capacity to predict societal outcomes with sufficient accuracy to implement important aspects of the intuition for bridging [52]. Instead, we may have to rely on heuristics like diverse approval (Section 4.2.1), which then need to be either validated or justified as having impacts "close enough" to the bridging goal.

By including bridging in a value model, there is a trade-off between bridging and other goals, such as relevance or engagement. Including bridging may, in some cases, reduce the extent to which these other goals can be achieved. Bridging is less likely to be implemented by large social media and search engine companies unless it is either economically beneficial, neutral, or there is significant and sustained external pressure (or regulation). While bridging-based ranking may reduce moderation costs and increase engagement over time, it may have short term negative impacts on engagement and growth. Bridging may be particularly challenging to maintain in an environment with sustained competition between platforms for attention, and easiest to maintain in a monopoly. However, in spite of these challenges there have now been publicly disclosed implementations of bridging across multiple major platforms, albeit in limited areas thus far, and we have been informed of others which have not yet been made public [69, 32, 33].

Finally, there is considerable uncertainty over the impacts of a bridging implementation. It is possible that bridging attention events cannot be incentivized and remain scarce, or that a particular an implementation of bridging practically results in predominantly benign, widely acceptable things content such as cat videos being promoted. Some may argue that algorithmic attention-allocators are delegated too small a portion of human attention (relative to, say, interpersonal relationships or mainstream media organizations) for efforts incorporating bridging to have any meaningful impact. Bridging-based ranking may also be less impactful or less tractable to implement within distributed or decentralized social media environments, or in models of social media and mass communication that are yet to emerge and may not be based around ranking.

Overcoming these challenges Some of these challenges may be easily overcome, some are unavoidable facts, and some are speculative. A large part of the research required in this area is to better understand

[20] If AI advances enable simulation and automation of high quality human-like facilitation, that may allow us to move beyond formalisms, though at the risk of reduced understanding of what such systems are doing.

the extent to which these challenges can be avoided, mitigated, or managed. They must also be considered in the context of the status quo, key elements of which (e.g., optimizing for engagement in social media, political messaging and advertising) are rarely subject to the same level of caution or scrutiny before implementation that we are applying here. The risks of intervention must be balanced with the risks of inaction.

Open Questions: Challenges, Limitations, and Risks *Set 8*

8.1. How can we best overcome the most critical challenges, limitations, and risks?

8.2. In what contexts is incorporating bridging worse than the status quo?

8.3. To what extent is the bridging goal in tension with other goals such as engagement or relevance?

7. Conclusion

Our social spaces should not *default to divisive*. Bridging is a core part of healthy social fabrics and the systems that allocate attention within them, whether human or machine, implicit or explicit. Without sufficient bridging, destructive conflict may undermine our relationships and prevent us from cooperating effectively to respond to societal challenges.

While bridging may help counteract the "bias toward division", it is unlikely to address the myriad of other biases that social spaces may entrench. We see this work as complementary to the lines of research around other biases, and it is important that we also keep in mind (and measure where appropriate) these other forms of bias when designing for bridging. We should also caveat that designing systems for meaningful positive interactions is part art, and in this paper we have framed it as a science. We encourage work that can go beyond "shallow bridging heuristics" to "deep bridging"—from bonds around cat videos to bonds about our very different struggles in living complex lives. In this working paper we have also focused primarily on bridging as it relates to the selection of what to attend to among a set of options, but the intuitions and formalisms we have explored can also be adapted to the *synthesis* of such options—i.e., to generative systems and foundation models such as ChatGPT and GPT-3 [9, 17].

New kinds of social spaces, both online or offline, do not always incorporate the hard-won lessons from the old. Modern computational and communication technologies have changed the structure of civilization, bringing billions of people onto the internet. But as we moved online, bridging is also one of the elements of offline spaces that we have most under-resourced.

Belatedly, some online spaces are catching up and setting an example. The deployment of the Community Notes feature on Twitter [70, 75]—first to the US in October 2022, and then globally visible in December 2022—is the first publicly acknowledged large-scale implementation of bridging-based ranking. Since the publication of the original bridging-based ranking paper [53], we have heard from people interested in bridging across organizations of all sizes and kinds, from the largest tech companies, to recently incorporated startups, to the technology teams at traditional media organizations. Through this working paper, we have attempted to provide a framework for understanding how such bridging systems can be understood and developed, and laid out a set of research questions to support and accelerate their safe deployment.

Our omnipresent social media, search, and synthetic media systems were not designed to satisfy the bridging goal: to increase mutual understanding and trust across divides, creating space for productive conflict, deliberation, or cooperation. But these systems that help allocate our attention are not irredeemable—they have the potential to support a more deliberative, peaceful, and pluralistic future.[21]

[21] There remains the normative question of what extent bridging is appropriate for a given system, which is outside of the scope of this paper. Ideally, such decisions would be a collective choice of those impacted by the system. While such collective decision-making might seem impractical for e.g., platforms with billions of users, mini-publics (and some kinds of civic forums) provide a potentially viable mechanism for such democratic decision-making at global scale [55].

Acknowledgements

We thank Tim van Gelder and Colin Megill for sharing information and insights from their work on YourView and Polis, respectively. We would also like to thank Natania Antler, Priyanjana Bengani, Leisel Bogan, Joaquin Quiñonero Candela, Austin Clyde, Joe Edelman, Thomas Gilbert, Amritha Jayanti, Julia Kamin, Andrew Konya, David Krueger, Stephen Larrick, Jesse McCrosky, James Mickens, Kathy Pham, Maria Polukarov, Afsaneh Rigot, Bruce Schneier, Jonathan Stray, Ted Suzman, Carmine Ventre, Jessica Yu, Glen Weyl and Cathy Wu, among others, for helpful discussions and feedback. Any errors or limitations of this work remain those of the authors.

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.

References

[1] Scott Aaronson. "Eigenmorality". In: *Shtetl-Optimized* (June 18, 2014). URL: https://scottaaronson.blog/?p=1820 (visited on 05/30/2022).

[2] Charu Aggarwal. *Recommender Systems*. Vol. 1. Springer, 2016.

[3] American Educational Research Association et al. *Standards for educational and psychological testing*. American Educational Research Association, 2014.

[4] Mahsa Badami. "Peeking into the other half of the glass: handling polarization in recommender systems". PhD thesis. University of Louisville, 2017. DOI: 10.18297/etd/2693. URL: http://ir.library.louisville.edu/etd/2693 (visited on 01/04/2022).

[5] Mahsa Badami and Olfa Nasraoui. "PaRIS: Polarization-aware Recommender Interactive System". In: (Oct. 2, 2021), p. 8.

[6] Mahsa Badami, Olfa Nasraoui, and Patrick Shafto. "PrCP: Pre-recommendation Counter-Polarization". In: *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. 10th International Conference on Knowledge Discovery and Information Retrieval. Seville, Spain: SCITEPRESS - Science and Technology Publications, 2018, pp. 282–289. ISBN: 978-989-758-330-8. DOI: 10.5220/0006938702820289. URL: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0006938702820289 (visited on 01/04/2022).

[7] Chris Bail. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Princeton University Press, 2021.

[8] Christopher A Bail et al. "Exposure to opposing views on social media can increase political polarization". In: *Proceedings of the National Academy of Sciences* 115.37 (2018), pp. 9216–9221.

[9] Michiel A. Bakker et al. *Fine-Tuning Language Models to Find Agreement among Humans with Diverse Preferences*. Nov. 2022. DOI: 10.48550/arXiv.2211.15006. arXiv: 2211.15006 [cs].

[10] Fabian Baumann et al. "Emergence of polarized ideological opinions in multidimensional topic spaces". In: *Physical Review X* 11.1 (2021), p. 011012.

[11] Priyanjana Bengani, Jonathan Stray, and Luke Thorburn. "What's Right and What's Wrong with Optimizing for Engagement". In: *Understanding Recommenders* (Apr. 27, 2022). URL: https://medium.com/understanding-recommenders/whats-right-and-what-s-wrong-with-optimizing-for-engagement-5abaac021851 (visited on 05/27/2022).

[12] Anna Bogomolnaia and Jean-François Laslier. "Euclidean preferences". In: *Journal of Mathematical Economics* 43.2 (2007), pp. 87–98.

[13] Johannes Botes. "Conflict Transformation: A Debate over Semantics or a Crucial Shift in the Theory and Practice of Peace and Conflict Studies?" In: *International Journal of Peace Studies* 8.2 (2003), pp. 1–27. ISSN: 10857494. URL: http://www.jstor.org/stable/41852899 (visited on 01/08/2023).

[14] Levi Boxell, Matthew Gentzkow, and Jesse M. Shapiro. "Cross-Country Trends in Affective Polarization". In: *The Review of Economics and Statistics* (Jan. 2022), pp. 1–60. ISSN: 0034-6535. DOI: 10.1162/rest_a_01160. eprint: https://direct.mit.edu/rest/article-pdf/doi/10.1162/rest_a_01160/1986030/rest_a_01160.pdf. URL: https://doi.org/10.1162/rest%5C_a%5C_01160.

[15] William J Brady and Jay J Van Bavel. *Estimating the effect size of moral contagion in online networks: A pre-registered replication and meta-analysis*. 2021. DOI: 10.31219/osf.io/s4w2x. URL: osf.io/s4w2x.

[16] Aaron Bramson et al. "Understanding Polarization: Meanings, Measures, and Model Evaluation". In: *Philosophy of Science* 84.1 (2017), pp. 115–159. DOI: 10.1086/688938.

[17] Tom Brown et al. "Language Models Are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.

[18] Axel Bruns. *Are filter bubbles real?* John Wiley & Sons, 2019.

[19] Lyn Carson. *Facilitating Public Deliberations*. Podcast Series. newDemocracy Foundation, 2020. URL: https://facilitatingpublicdeliberation.libsyn.com/.

[20] L. Elisa Celis et al. "Controlling Polarization in Personalization: An Algorithmic Framework". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19: Conference on Fairness, Accountability, and Transparency. Atlanta GA USA: ACM, Jan. 29, 2019, pp. 160–169. ISBN: 978-1-4503-6125-5. DOI: 10.1145/3287560.3287601. URL: https://dl.acm.org/doi/10.1145/3287560.3287601 (visited on 01/04/2022).

[21] Brian Clifton. "How to tell whether a Twitter user is pro-choice or pro-life without reading any of their tweets". In: *Quartz* (Oct. 9, 2015). URL: https://qz.com/520309/how-to-tell-whether-a-twitter-user-is-pro-choice-or-pro-life-without-reading-any-of-their-tweets/ (visited on 05/28/2022).

[22] Austin Clyde. *Algorithmic Systems Designed to Reduce Polarization Could Hurt Democracy, Not Help It*. Tech Policy Press. Feb. 17, 2022. URL: https://techpolicy.press/algorithmic-systems-designed-to-reduce-polarization-could-hurt-democracy-not-help-it/ (visited on 02/18/2022).

[23] Peter T. Coleman, Larry S. Liebovitch, and Joshua Fisher. "Taking Complex Systems Seriously: Visualizing and Modeling the Dynamics of Sustainable Peace". In: *Global Policy* 10.S2 (2019), pp. 84–92. DOI: https://doi.org/10.1111/1758-5899.12680. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1758-5899.12680. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-5899.12680.

[24] Benoit Colson, Patrice Marcotte, and Gilles Savard. "An overview of bilevel optimization". In: *Annals of operations research* 153.1 (2007), pp. 235–256. DOI: 10.1007/s10479-007-0176-2.

[25] Lewis A Coser. *The Functions of Social Conflict*. Vol. 9. Routledge, 1998.

[26] Dimitri Courant. "Citizens' Assemblies for Referendums and Constitutional Reforms: Is There an "Irish Model" for Deliberative Democracy?" In: *Frontiers in Political Science* (2021). DOI: 10.3389/fpos.2020.591983.

[27] Paul Covington, Jay Adams, and Emre Sargin. "Deep Neural Networks for YouTube Recommendations". In: *Proceedings of the 10th ACM Conference on Recommender Systems*. RecSys '16. Boston, Massachusetts, USA: Association for Computing Machinery, 2016, pp. 191–198. ISBN: 9781450340359. DOI: 10.1145/2959100.2959190. URL: https://doi.org/10.1145/2959100.2959190.

[28] Morton Deutsch. "Conflicts: Productive and Destructive*". In: *Journal of Social Issues* 25.1 (1969), pp. 7–42. ISSN: 1540-4560. DOI: 10.1111/j.1540-4560.1969.tb02576.x.

[29] John S. Dryzek et al. "The crisis of democracy and the science of deliberation". In: *Science* 363.6432 (2019), pp. 1144–1146. DOI: 10.1126/science.aaw2694. eprint: https://www.science.org/doi/pdf/10.1126/science.aaw2694. URL: https://www.science.org/doi/abs/10.1126/science.aaw2694.

[30] Jean-Yves Duclos, Joan Esteban, and Debraj Ray. "Polarization: concepts, measurement, estimation". In: *Econometrica* 72.6 (2004), pp. 1737–1772.

[31] Oliver Escobar and Stephen Elstub. *Forms of mini-publics: An introduction to deliberative innovations in democratic practice*. Research and Development Note. May 8, 2017. URL: https://newdemocracy.com.au/wp-content/uploads/2017/05/docs_researchnotes_2017_May_nDF_RN_20170508_FormsOfMiniPublics.pdf.

[32] Facebook. "MSI Metric Changes for 2020 H1". In: *Facebook Papers Directory*. Ed. by Dell Cameron, Shoshana Wodinsky, and Mack DeGeurin. Gizmodo, 2022. URL: https://www.documentcloud.org/documents/21601827-tier2_rank_ro_0120.

[33] Facebook. "News Feed Research: Looking Back on H2 2020". In: *Facebook Papers Directory*. Ed. by Dell Cameron, Shoshana Wodinsky, and Mack DeGeurin. Gizmodo, 2022. URL: https://www.documentcloud.org/documents/21748428-tier1_news_ir_0221.

[34] Eleonore Fournier-Tombs and Michael K. MacKenzie. "Big Data and Democratic Speech: Predicting Deliberative Quality Using Machine Learning Techniques". In: *Methodological Innovations* 14.2 (May 2021), p. 20597991211010416. ISSN: 2059-7991. DOI: 10.1177/20597991211010416.

[35] Tim van Gelder. "Cultivating Deliberation for Democracy". In: *Journal of Deliberative Democracy* 8.1 (May 1, 2020).

[36] Tim van Gelder. "Public Wisdom". In: *2020: Vision for a Sustainable Society*. Ed. by Craig Pearson. Melbourne: Melbourne Sustainable Society Institute, 2012. Chap. 10, pp. 79–86. URL: https://sites.google.com/site/timvangelder/publications-1/public-wisdom.

[37] Thomas Krendl Gilbert et al. "Reward Reports for Reinforcement Learning". In: *arXiv:2204.10817 [cs]* (Apr. 2022). arXiv: 2204.10817 [cs].

[38] Louis-Gaëtan Giraudet et al. ""Co-construction" in Deliberative Democracy: Lessons from the French Citizens' Convention for Climate". Preprint. May 2022. URL: https://hal-enpc.archives-ouvertes.fr/hal-03119539.

[39] Jonathan Haidt and Christopher A. Bail. "Social Media and Political Dysfunction: A Review". New York University, Nov. 2021.

[40] Yu T Hsiao et al. "vTaiwan: An Empirical Study of Open Consultation Process in Taiwan". In: (Aug. 4, 2018). DOI: 10.31235/osf.io/xyhft. URL: osf.io/preprints/socarxiv/xyhft.

[41] Ruben Interian et al. *Network polarization, filter bubbles, and echo chambers: An annotated review of measures, models, and case studies*. 2022. DOI: 10.48550/ARXIV.2207.13799. URL: https://arxiv.org/abs/2207.13799.

[42] John Paul Lederach. *The Little Book of Conflict Transformation*. Simon and Schuster, 2015.

[43] Philipp Lorenz-Spreen et al. "A systematic review of worldwide causal and correlational evidence on digital media and democracy". In: *Nature Human Behaviour* (2022). ISSN: 2397-3374. DOI: 10.1038/s41562-022-01460-1. URL: https://doi.org/10.1038/s41562-022-01460-1.

[44] Michael W. Macy et al. "Polarization and tipping points". In: *Proceedings of the National Academy of Sciences* 118.50 (2021), e2102144118. DOI: 10.1073/pnas.2102144118. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2102144118. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2102144118.

[45] David Manheim and Scott Garrabrant. *Categorizing Variants of Goodhart's Law*. 2018. DOI: 10.48550/ARXIV.1803.04585. URL: https://arxiv.org/abs/1803.04585.

[46] Avishai Margalit. *On Compromise and Rotten Compromises*. Princeton: Princeton University Press, 2009. ISBN: 9781400831210. DOI: doi:10.1515/9781400831210. URL: https://doi.org/10.1515/9781400831210.

[47] Antonis Matakos, Evimaria Terzi, and Panayiotis Tsaparas. "Measuring and moderating opinion polarization in social networks". In: *Data Mining and Knowledge Discovery* 31.5 (Sept. 2017), pp. 1480–1505. ISSN: 1384-5810, 1573-756X. DOI: 10.1007/s10618-017-0527-9. URL: http://link.springer.com/10.1007/s10618-017-0527-9 (visited on 01/09/2022).

[48] Jennifer McCoy and Murat Somer. "Toward a Theory of Pernicious Polarization and How It Harms Democracies: Comparative Evidence and Possible Remedies". In: *The ANNALS of the American Academy of Political and Social Science* 681.1 (Jan. 1, 2019). Publisher: SAGE Publications Inc, pp. 234–271. ISSN: 0002-7162. DOI: 10.1177/0002716218818782. URL: https://doi.org/10.1177/0002716218818782 (visited on 02/18/2022).

[49] Smitha Milli, Luca Belli, and Moritz Hardt. "From Optimizing Engagement to Measuring Value". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Mar. 2021), pp. 714–722. DOI: 10.1145/3442188.3445933. arXiv: 2008.12623.

[50] Chantal Mouffe. "Deliberative Democracy or Agonistic Pluralism?" In: *Social Research* 66.3 (1999), pp. 745–758. ISSN: 0037783X. URL: http://www.jstor.org/stable/40971349 (visited on 01/07/2023).

[51] Christopher Musco et al. *How to Quantify Polarization in Models of Opinion Dynamics*. 2021. arXiv: 2110.11981.

[52] Arvind Narayanan. "How to recognize AI snake oil". Princeton University. 2019. URL: https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf.

[53] Aviv Ovadya. *Bridging-Based Ranking*. Tech. rep. Belfer Center for Science and International Affairs, Harvard Kennedy School, May 2022.

[54] Aviv Ovadya. *Holding Platforms Accountable Is Not Enough. We Need A 'Compass' For Social Technologies*. Nov. 2021.

[55] Aviv Ovadya. *Towards Platform Democracy: Policymaking Beyond Corporate CEOs and Partisan Pressure*. Oct. 2021.

[56] *Polis*. URL: https://pol.is/ (visited on 05/27/2022).

[57] John A Powell. "Overcoming Toxic Polarization: Lessons in Effective Bridging". In: *Minnesota Journal of Law & Inequality* (2022), pp. 247–247.

[58] Bashir Rastegarpanah, Krishna P. Gummadi, and Mark Crovella. "Fighting Fire with Fire: Using Antidote Data to Improve Polarization and Fairness of Recommender Systems". In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. WSDM '19: The Twelfth ACM International Conference on Web Search and Data Mining. Melbourne VIC Australia: ACM, Jan. 30, 2019, pp. 231–239. ISBN: 978-1-4503-5940-5. DOI: 10.1145/3289600.3291002. URL: https://dl.acm.org/doi/10.1145/3289600.3291002 (visited on 01/04/2022).

[59] Venu Satuluri et al. "SimClusters: Community-Based Representations for Heterogeneous Recommendations at Twitter". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '20. Virtual Event, CA, USA: Association for Computing Machinery, 2020, pp. 3183–3193. ISBN: 9781450379984. DOI: 10.1145/3394486.3403370. URL: https://doi.org/10.1145/3394486.3403370.

[60] Scoop. *Protecting and Restoring NZ's Biodiversity*. Sept. 22, 2019. URL: https://pol.is/3atycmhmer (visited on 09/06/2022).

[61] Michael Short and Tim van Gelder. "Full transcript: Tim van Gelder". In: *The Age* (Mar. 14, 2012). URL: https://www.theage.com.au/national/full-transcript-tim-van-gelder-20120313-1uxzr.html (visited on 05/27/2022).

[62] Christopher Small et al. "Polis: Scaling Deliberation by Mapping High Dimensional Opinion Spaces". In: *Recerca. Revista de Pensament i Anàlisi* 26.2 (2021), pp. 1–26. DOI: 0.6035/recerca.5516.

[63] Jonathan Stray. "Designing Recommender Systems to Depolarize". In: *arXiv:2107.04953 [cs]* (July 10, 2021). arXiv: 2107.04953. URL: http://arxiv.org/abs/2107.04953 (visited on 08/30/2021).

[64] Milan W Svolik. "Polarization versus democracy". In: *Journal of Democracy* 30.3 (2019), pp. 20–32.

[65] Luke Thorburn, Priyanjana Bengani, and Jonathan Stray. "How Platform Recommenders Work". In: *Understanding Recommenders* (Jan. 20, 2022). URL: https://medium.com/understanding-recommenders/how-platform-recommenders-work-15e260d9a15a (visited on 05/27/2022).

[66] Luke Thorburn, Maria Polukarov, and Carmine Ventre. "Information Loss in Euclidean Preference Models". In: *arXiv preprint arXiv:2208.08160* (2022).

[67] Luke Thorburn, Jonathan Stray, and Priyanjana Bengani. "How to Measure the Effects of Recommenders". In: *Understanding Recommenders* (July 20, 2022). URL: https://medium.com/understanding-recommenders/how-to-measure-the-causal-effects-of-recommenders-5e89b7363d57 (visited on 08/22/2022).

[68] Luke Thorburn, Jonathan Stray, and Priyanjana Bengani. "What Does it Mean to Give Someone What They Want? The Nature of Preferences in Recommender Systems". In: *Understanding Recommenders* (Mar. 11, 2022). URL: https://medium.com/understanding-recommenders/what-does-it-mean-to-give-someone-what-they-want-the-nature-of-preferences-in-recommender-systems-82b5a1559157 (visited on 05/27/2022).

[69] Twitter. "Note Ranking". In: *Community Notes Guide* (2022). URL: https://twitter.github.io/communitynotes/ranking-notes/ (visited on 01/09/2023).

[70] Twitter. "Overview". In: *Community Notes Guide* (2022). URL: https://twitter.github.io/communitynotes/ (visited on 01/09/2023).

[71] Eva Maria Vecchi et al. “Towards Argument Mining for Social Good: A Survey”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 1338–1352. DOI: 10.18653/v1/2021.acl-long.107.

[72] Jan G Voelkel et al. “Megastudy identifying successful interventions to strengthen Americans’ democratic attitudes”. In: (2022). URL: https://www.strengtheningdemocracychallenge.org/paper/.

[73] Barbara F. Walter. *How Civil Wars Start*. First Edition. New York: Crown, 2022. ISBN: 9780241988404.

[74] Glen Weyl. “Why I Am a Pluralist”. In: *RadicalxChange* (Feb. 10, 2022). URL: https://www.radicalxchange.org/media/blog/why-i-am-a-pluralist/ (visited on 09/08/2022).

[75] Stefan Wojcik et al. *Birdwatch: Crowd Wisdom and Bridging Algorithms can Inform Understanding and Reduce the Spread of Misinformation*. 2022. DOI: 10.48550/ARXIV.2210.15723. URL: https://arxiv.org/abs/2210.15723.

[76] *YourView*. URL: https://web.archive.org/web/20140226034818/https://yourview.org.au/ (visited on 02/26/2014).

[77] Daniel A Yudkin, Stephen Hawkins, and Tim Dixon. *The Perception Gap: How False Impressions are Pulling Americans Apart*. 2019. DOI: 10.31234/osf.io/r3h5q. URL: psyarxiv.com/r3h5q.

[78] Liwang Zhu and Zhongzhi Zhang. “Minimizing Polarization and Disagreement in Social Networks via Link Recommendation”. In: (Nov. 2021), p. 13.

A. Appendix

A.1. Glossary

Bridging

bridging goal (intuition)

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

bridging (formalism)

An improvement in *relation metrics* that corresponds to the bridging goal.

Note. We distinguish between the intuition and formalism of bridging. Our aim in doing this is to decrease the likelihood that the formalism (and its optimization) will overshadow the rich human experiences it is intended to make legible. Where formalism conflicts with intuition, intuition should generally be favored (in order to avoid Goodharting [45]).

Attention Events

attention object

Anything that can be attended to.

attention slot

A container to which an attention object can be allocated. Examples include discrete positions within a recommender feed, or continuous intervals of first-person experience.

atomic attention event

The allocation of an *attention object* to an *attention slot*. In general, it is modeled formally as (`slot`, `object`, `properties`). If the slot is an interval of someone's first-person experience, then we sometimes write this as (`person`, `object`, `properties`) to emphasize that a human is involved.

(general) attention event

A set of *atomic attention events*.

A **potential** attention event is one which has not (yet) happened (and may not happen), and a **realized** attention event is one which has already happened.

Attention-Allocation Systems

attention-allocation system / **attention-allocator**

A system that takes as input a set of *potential attention events* and outputs a set of *realized attention events*. It consists of an *allocation process* and, optionally, a *learning process*.

allocation process

A process (Figure 10) that takes as input a set of *potential attention events*, and outputs a set of *realized attention events*. It is the core component of an *attention-allocator*.

learning process

A process (Figure 11) that collects, retrieves or elicits data, and uses it to update the state and prediction models used in the *allocation process*. It is an optional component of an *attention-allocator*.

value model

A method for aggregating the multiple predicted impacts of a *potential attention event* into an overall measure of value. It is used in an *allocation process*.

Optimization Stack

optimization stack

The multiparty, multilevel optimization that occurs in an *attention-allocator*. At minimum, it consists of optimization for stakeholder objectives during *system design*, *accuracy optimization* during the learning process, *value optimization* during the allocation process, and the strategic behavior of users who optimize for personal objectives when interacting with the system.

system design

The process of making decisions about the design of an *attention-allocator*, that are made outside of the normal course of operating the system.

accuracy optimization

Optimizing for accuracy during the *learning process*.

value optimization

Optimizing for value during the *allocation process*.

Representation & Quantification

relation model

A model of the state of relationships among people in a population, consisting of the triple (`people`, `objects`, `relations`). Common examples include *graph-based* and *space-based* models.

relation metrics

Statistics summarizing the *state* of a relation model (or a subset of the relation model) at a given point in time.

Relation metrics should have a clear normative interpretation. Within a given context, we should be able to say that an increase in relation metrics is good (i.e., corresponds to the bridging goal), or that certain configurations of multiple relation metrics are better than others.

bridging metrics

Statistics summarizing a *change* in relation metrics.

Event-level bridging metrics capture the effect of an attention event on relation metrics. *System-level* bridging metrics capture the effect of an entire attention-allocator on relation metrics.

Bridging metrics can either be *model-based* (if they are framed or defined in terms of the relation model), or *heuristic* (if they are based on certain interaction patterns which are thought to cause relation metrics to improve).

interaction pattern

An archetypal pattern of interaction between people.

bridging heuristic

An interaction pattern that is correlated with improved bridging metrics.

System Properties

division bias / bias towards division

A property of attention-allocators if their *system-level bridging metrics* are negative.

A.2. System Diagrams

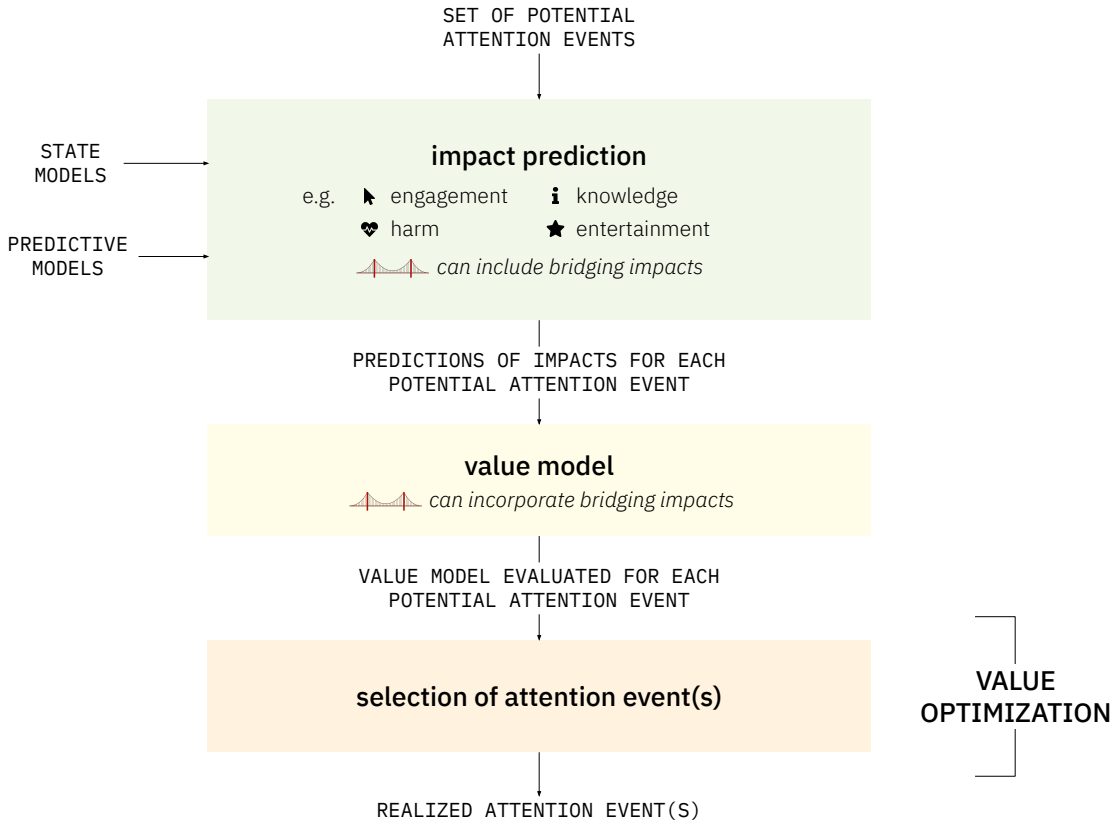

Figure 10: **The allocation process in an attention-allocation system.** A bridge icon indicates where bridging can be incorporated. Not shown are the ways in which the system itself is optimized.

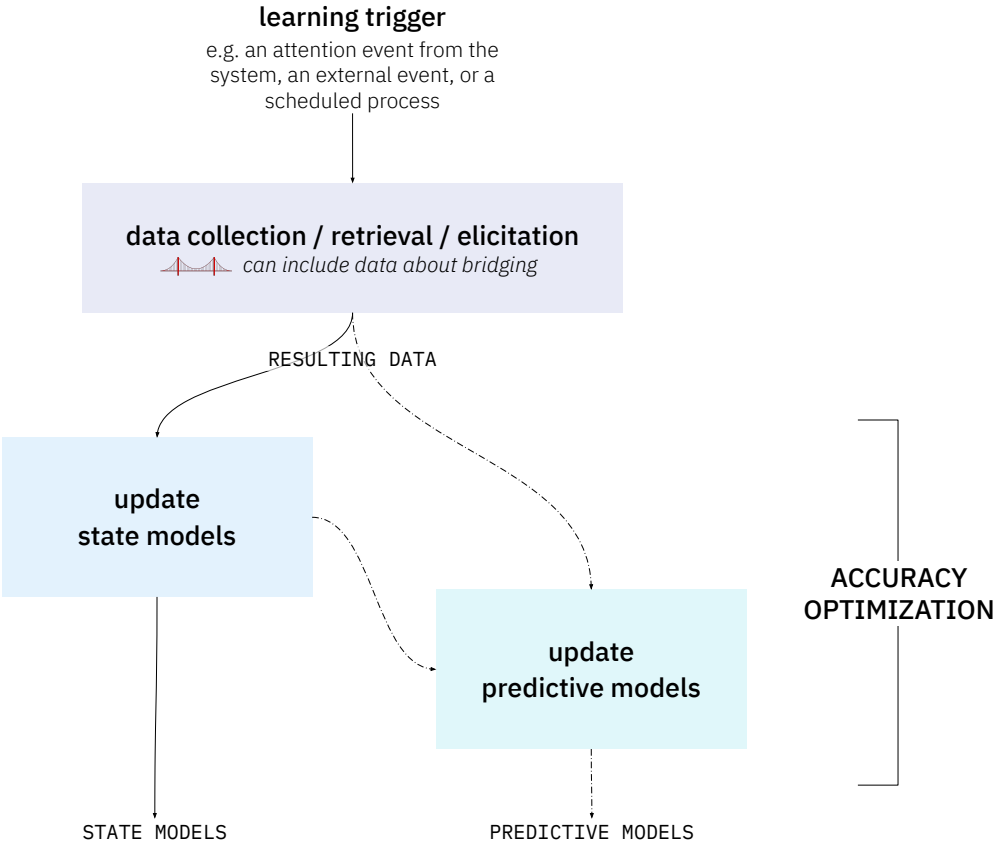

Figure 11: **The learning process in an attention-allocation system.** A bridge icon indicates where bridging can be incorporated. The dashed lines indicate that many technical systems will update state models after each attention event, but update predictive models less frequently. Human facilitators update their implicit state models as they facilitate (noticing how different events impact people), and also update their predictive models slightly as they work, but might do a much larger "update" of their predictive models (improving their facilitation skills) during a post-facilitation retrospective. Not shown are implicit inputs such as the previous models, or how the act of data elicitation can itself change the state of the world.

A.3. Roadmap (for this document)

This document is work-in-progress, and there are still substantial changes and additions we intend to make. We would like to:

- Expand the literature review with additional interdisciplinary domains and related work.
- Incorporate more insights from the study and practice of human-facilitated bridging (including mediation, organizational practices, peace studies, etc.).
- In Section 4, add a subsection on quantifying the degree to which attention events reduce uncertainty about the state of division.
- Add a suite of open problems around the social science of bridging systems, e.g. evaluating the theory of change from Figure 1.
- Provide example roadmaps for transitioning systems to support intentional bridging. For example, exploring questions such as "What are some paths for taking a recommender system or a search ranking system and implementing bridging metrics and optimization for them? How might this be different for systems with different existing processes and competencies; e.g. Google with its Search Rater Guidelines?"

We are putting together a working group to expand this paper and the broader work, adding more open problems and deeper explorations across disciplines. Please complete this form if you would be interested in contributing, or just kept in the loop on further developments.