Designing recommender systems to depolarize

by Jonathan Stray

Abstract
Polarization is implicated in the erosion of democracy and the progression to violence, which makes the polarization properties of large algorithmic content selection systems (recommender systems) a matter of concern for peace and security. While algorithm-driven social media do not seem to be a primary driver of polarization at the country level, they could be a useful intervention point in polarized societies. This paper examines algorithmic depolarization interventions aimed at transforming conflict: not suppressing or eliminating conflict, but making it more constructive. Algorithmic intervention is considered at three stages: what content is available (moderation), how content is selected and personalized (ranking), and content presentation and controls (user interface). Empirical studies of online conflict suggest that not only could the exposure-diversity intervention proposed as an antidote to ‘filter bubbles’ be improved: under some conditions, it can even worsen polarization. Using civility metrics in conjunction with diversity in content selection may be more effective. However, diversity-based interventions have not been tested at scale, and may not work in the diverse and dynamic contexts of real platforms. Instead, intervening in platform polarization dynamics will likely require continuous monitoring of polarization metrics, such as the widely used ‘feeling thermometer’. These metrics can be used to evaluate product features, and can potentially be engineered as algorithmic objectives. While using any metric as an optimization target may have harmful consequences, to prevent optimization processes from creating conflict as a side effect it may prove necessary to include polarization measures in the objective function of recommender algorithms.

Contents

1. Introduction
2. Depolarization as conflict transformation
3. Measuring polarization
4. Algorithmic depolarization interventions
5. Learning to depolarize
6. Conclusion
1. Introduction

Polarization is a condition where myriad differences in society fuse and harden into a single axis of identity and conflict (Iyengar and Westwood, 2015), and it has been increasing for decades in several democracies (Boxell, et al., 2022; Draca and Schwarz, 2018). Comparative studies that examine polarization across countries argue that increasing polarization is a factor contributing to the democratic erosion seen in many of them, including Venezuela, Hungary, Turkey, and the United States (McCoy, et al., 2018; Somer and McCoy, 2019). Polarization produces a feedback loop where diverging identities lead to less intergroup contact which in turn leads to increased polarization, culminating in a hardened us-versus-them mentality that can contribute to the deterioration of democratic norms. Most conflict escalation models consider polarization a key element in the feedback dynamics that lead to violent conflict (Collins, 2012). Peace and security demand that we address situations of increasing polarization, which is why the international peacebuilding community concerns itself with the issue (Ramsbotham, et al., 2016).

Scholars have long studied the relationship between media and conflict, a tradition that now includes digital media (Hofstetter, 2021; Tellidis and Kappler, 2016), much of which is algorithmically selected and personalized. The algorithms that choose which items are shown to each user are called recommender systems, and all major news aggregators and social media platforms have such a system at their core. Modern recommender systems select content based on a variety of information sources such as the content of each item, a user’s expressed preferences, their past consumption behavior, the behavior of similar users, user survey responses, fairness considerations, and more (Aggarwal, 2016). Note that ‘recommender’ is a computer-science term of art that covers all algorithmic content selected on the basis of implicit information, i.e., not as the result of a search query. This content might be presented as ‘recommended for you’, or labelled as ‘news’ or ‘trends’, or might appear as a feed or timeline.

There has been intense interest in the question of whether recommender systems affect large-scale conflict dynamics. Most of the work on recommenders and polarization has taken place within the ‘filter bubble’ paradigm and has therefore explored the idea of exposure diversity (Helberger, et al., 2018). Selective exposure is the idea that individuals will preferentially choose news sources and articles that are ideologically aligned (Prior, 2013). Because recommender systems respond to user interests, there is the possibility of a feedback loop in which both recommendations and user interests progressively narrow. Indeed, simulations have demonstrated such polarization-increasing effects in stylized settings (Jiang, et al., 2019; Krueger, et al., 2020; Rychwalska and Roszczyńska-Kurasińska, 2018; Stoica and Chaintreau, 2019).

However, the available evidence mostly fails to support the hypothesis that recommender systems are driving polarization through selective exposure, a.k.a., ‘filter bubbles’ or ‘echo chambers’ (Bruns, 2019; Zuiderveen Borgesius, et al., 2016). Algorithmically personalized news seems to be quite similar for all users (Guess, et al., 2018), and is typically no less diverse than selections by human editors (Möller, et al., 2018), while social media users consume a more diverse range of news sources than non-users (Fletcher and Nielsen, 2018). Most recently, Feezell, et al. (2021) find no difference in affective polarization scores between Americans who get their news from conventional sources as opposed to social media.

Non-news personalized content could still be polarizing. Lelkes, et al. (2017) compare the introduction of broadband access across U.S. states from 2004 to 2008 and find a small causal increase in affective polarization. Yet polarization began increasing in the U.S. decades before social media, and is increasing faster among individuals aged 65 and up, a demographic with low Internet usage (Boxell, et al., 2017). An analysis across different countries shows no clear relationship between polarization and increasing Internet usage, as many OECD countries with high internet usage such as Britain, Sweden, Norway, and Germany show decreasing affective polarization (Boxell, et al., 2022).

Direct experimental intervention is probably the best way to study the causality of recommender systems. Allcott, et al. (2020) paid U.S. users to stay off Facebook for a month and found that an index of polarization measures decreased by 0.16 SD (standard deviation). This may have been due to a decrease in exposure to polarizing posts, comments, and discussions, but this intervention also decreased the time spent
on news by 15 percent, and news consumption can itself be polarizing (Martin and Yurukoglu, 2017; Melki and Sekeris, 2019). By contrast, a similar study of users in Bosnia and Herzegovina who deactivated Facebook during a genocide remembrance week showed greater polarization, a 0.24 SD increase on an index of ethnic polarization (Asimovic, et al., 2021). The increase was smaller for users who had a more ethnically diverse off-line social group, suggesting that Facebook was in this case providing depolarizing diversity. While these studies do suggest causation, the effects are not unidirectional or straightforward.

Rather than asking if social media is driving polarization, it may be more productive to ask if social media interventions can decrease polarization. The main contribution of this paper is to propose several methods for building recommender systems that actively reduce polarization.

Note that polarization is conceptually distinct from radicalization. Polarization is a process that “defines other groups in the social and political arena as allies or opponents”, while radicalization involves people who “become separated from the mainstream norms and values of their society” and may engage in violence (van Stekelenburg, 2014). There is a growing body of work studying the connection between recommender systems and radicalization (Baugut and Neumann, 2020; Hosseinmardi, et al., 2020; Ledwich and Zaitsev, 2020; Munger and Phillips, 2022; Ribeiro, et al., 2020), but this is methodologically challenging, and has not yet established a robust causal link. While social media may plausibly be involved in radicalization processes, the nature of this connection is complex and poorly understood. This work concerns polarization only, arguing that polarization itself is a bad outcome and a precursor to more extreme conflict.

In this paper I first make the moral argument for attempting to reduce polarization through recommender systems, framing it as a conflict transformation intervention. I then review definitions and metrics of polarization before considering depolarization interventions at three stages: which content is available (moderation), how content is selected and personalized (ranking), and content presentation and controls (user interface). The depolarization intervention most commonly proposed is exposure to ideologically diverse content: this may not be effective, however, because mere exposure does not necessarily depolarize, and sometimes it polarizes further. While there are other promising approaches, such as exposure to civil counter-ideological content, these may not be sufficiently robust to withstand the incredibly diverse conditions of real-world platforms. Instead, I propose continuously monitoring the measures used for surveying affective polarization so as to drive recommender outcomes in a feedback loop. Polarization metrics can be used both at the managerial level and at the algorithmic level, potentially through reinforcement learning.

2. Depolarization as conflict transformation

There are complicated questions around intervening in societal conflicts through media, and additional concerns around the use of AI for this purpose. At worst, algorithmically suppressing disagreements could amount to authoritarian pacification. The Chinese social media censorship regime is an instructive example of democratically questionable interventions in the name of harmony (Creemers, 2017; G. King, et al., 2017). I therefore frame the goal of depolarization as conflict transformation: not eliminating or resolving conflict, but making that conflict better in some way, e.g., less prone to violence and more likely to lead to justice (Jeong, 2019).

Indeed, it is not clear that platform users want to be ‘depolarized’, and in any mass conflict situation there will be people who argue for escalation in the strongest moral terms. There is a corresponding line of argument that polarization is beneficial. Political theorists have argued that it reduces corruption by increasing accountability (Melki and Pickering, 2020), and that it generally helps differentiate between political parties in a way that offers voters a meaningful choice. In the mid-twentieth century, mainstream political scientists worried that America wasn’t polarized enough (American Political Science Association,
1950). Importantly, fights for justice or accountability can also increase polarization, such as the American civil rights movement of the 1960s (D.S. King and Smith, 2008). There are parallels with the idea of a just war.

Yet polarization also has serious downsides. At the elite level it causes a ‘gridlock’ that makes effective governance difficult (F.E. Lee, 2015), but contemporary polarization reaches far beyond lawmakers. The politicization of all spheres of society destroys social bonds at the family, community, and national levels (A.H.-Y. Lee, 2021). By some measures, cross-partisan dislike in the U.S. is now considerably stronger than racial resentment, and has widespread effects on social choices such as hiring, university admissions, dating, family relations, friendships, and purchasing decisions (Iyengar, et al., 2019). Polarization erodes the norms that constrain conflict escalation, leading to ‘morally outrageous’ behavior on all sides (Deutsch, 1969), and it is a key precursor to violence (Collins, 2012). Ultimately, polarization appears to be a causal factor in the destruction of democracies (McCoy, et al., 2018; Somer and McCoy, 2019).

There is a tension between peace and justice. Actions that promote peace may make justice harder, and vice versa. Yet a democracy requires both — an observation that leads to the concept of a just peace (Fixdal, 2012). Instead of trying to eliminate conflict, we can try to understand what makes it good or bad. In an agonistic theory of democracy it is considered normal for political adversaries to be engaged in “opposing hegemonic projects”, and conflict is to be not eliminated but “tamed” (Mouffe, 2002). Perhaps the most sophisticated understandings of conflict come from the peacebuilding tradition, which came into its own as an applied discipline after World War II. Over fifty years ago, Deutsch described the difference between “constructive” and “destructive” conflict, paying particular attention to the dynamics of escalation:

Paralleling the expansion of the scope of conflict there is an increasing reliance upon a strategy of power and upon the tactics of threat, coercion, and deception. Correspondingly, there is a shift away from a strategy of persuasion and from the tactics of conciliation, minimizing differences, and enhancing mutual understanding and good-will. And within each of the conflicting parties, there is increasing pressure for uniformity of opinion and a tendency for leadership and control to be taken away from those elements that are more conciliatory and invested in those who are militantly organized for waging conflict through combat.

[...]

It leads to a suspicious, hostile attitude which increases the sensitivity to differences and threats, while minimizing the awareness of similarities. This, in turn, makes the usually accepted norms of conduct and morality which govern one’s behavior toward others who are similar to oneself less applicable. Hence, it permits behavior toward the other which would be considered outrageous if directed toward someone like oneself. (Deutsch, 1969)

On the other hand, Lederach (2014) describes how conflict is necessary for positive social change and how conflict transformation leads to better conflict processes:

A transformational approach recognizes that conflict is a normal and continuous dynamic within human relationships. Moreover, conflict brings with it the potential for constructive change. Positive change does not always happen, of course. As we all know too well, many times conflict results in long-
standing cycles of hurt and destruction. But the key to transformation is a proactive bias toward seeing conflict as a potential catalyst for growth.

[...]

A transformational approach seeks to understand the particular episode of conflict not in isolation, but as embedded in the greater pattern. Change is understood both at the level of immediate presenting issues and that of broader patterns and issues. (Lederach, 2014)

Or as Ripley (2021) puts it:

The challenge of our time is to mobilize great masses of people to make change without dehumanizing one another. Not just because it’s morally right but because it works. [1]

Polarization is potentially an important intervention point in conflict dynamics because it features in multiple pathways to escalation. It can be exploited for political mobilization through us-versus-them rhetoric, as has long been understood by activists (Layman, et al., 2010) and other “political entrepreneurs” (Sommer and McCoy, 2019) — and as demonstrated by the fact that the most politically engaged citizens are found at the ideological extremes (Pew Research Center, 2014). However, this kind of exploitation further increases polarization. Indeed, polarization features in a variety of pernicious feedback loops: polarization leads to less intergroup contact, which causes polarization (A.H-Y. Lee, 2021); polarization is a precursor to violence, which causes polarization (Collins, 2012); polarization leads to selective information exposure, which causes polarization (Kim, 2015), and so on. These causal dynamics suggest that, in conflict escalation, polarization could indeed be an important intervention point.

Conflicts that involve democratic erosion or violence are deeply troubling, to the point where conflict-transforming interventions may be warranted on human rights grounds. In the U.S., support for violence as a means to political ends is increasing on both the left and the right (Diamond, et al., 2020). In short, partisans are willing to violate democratic norms when polarization is high. A recent review concluded that “the goal of these [depolarizing] interventions is to move toward a system in which the public forcefully debates political ideals and policies while resisting tendencies that undermine democracy and human rights” (Finkel, et al., 2020).

3. Measuring polarization

Quantitative measures are needed for evaluating polarization at scale. This is a problem not merely of measurement, but of definition. Polarization has been studied through differences in legislative voting patterns (Hare and Poole, 2014), and in the language used in U.S. Congressional speech (Gentzkow, et al., 2017). At the population level it has been operationalized as an increasing correlation of policy preferences on multiple issues (Draca and Schwarz, 2018; Kiley, 2017) and increasing animosity toward a political outgroup, known as affective polarization (Iyengar and Westwood, 2015). All of these indicators show that polarization has been increasing in the U.S. over the last 40 years. Globally the results are more mixed, with some OECD countries experiencing increasing polarization and others showing flat or decreasing trends (Boxell, et al., 2022; Draca and Schwarz, 2018).

Affective polarization has become a key concept in the analysis of American politics, as “ordinary Americans increasingly dislike and distrust those from the other party” (Iyengar, et al., 2019). Affective
Designing recommender systems to depolarize is a consequence of partisan identity, which is a better model of contemporary political conflict than differences of preferred policy on various issues (Finkel, et al., 2020). It also has the advantage of being operationalizable through straightforward survey measures, such as the feeling thermometer, which is one of the oldest and most widely used polarization measures. This method asks respondents to rate their feeling about each political party on a scale from 0 (cold) to 100 (warm). The difference in scores, the net feeling thermometer, is taken to be a measure of affective polarization. This question has been asked in the American National Election Study since the 1970s, and is frequently used in studies of polarization and social media (Feezell, et al., 2021; Levy, 2021; Suhay, et al., 2018). While there are different measures of affective polarization, they are mostly highly correlated (Druckman and Levendusky, 2019).

Affective polarization — negative feelings about the ‘other side’ — has serious interpersonal consequences. Tellingly, 13 percent of Americans reported that they had ended a relationship with a family member or close friend after the 2016 election (Whitesides, 2017). Affective polarization correlates with dehumanization, “a significant step toward depriving individuals who belong to certain groups or categories of individual-level depth or complexity of feelings, motivation, or personality” (Martherus, et al., 2021). It leads to the destruction of social bonds and increased outgroup prejudice across all facets of social and political life (Iyengar, et al., 2019; A.H.-Y. Lee, 2021; Somer and McCoy, 2019). In short, affective polarization now strongly colors the experience of daily life and relationships in multiple countries and has potentially grim consequences for democracy.

4. Algorithmic depolarization interventions

Recommender-based systems such as social media and news aggregators are more than just ‘algorithms’, and an analysis of the polarization effects of this wide array of products and platforms could potentially be very broad. To narrow the scope, I will consider three key places where changes to recommender systems might be used for depolarization:

- **What content is available (moderation).** Much previous work on polarization has concerned itself with what content is allowed on a platform. For example, hate speech and incitement to violence are routinely removed through a combination of human moderators, machine-learning classifiers, and user flagging.

- **How content is selected (ranking).** Algorithmic content selection is essentially a prioritization problem, and all contemporary recommendation systems score each item based on a number of criteria. An intervention in content ranking addresses the core question of who sees what. Most of the approaches considered in this paper are modifications of content ranking.

- **How content is presented (interface).** Selected items are presented in some way to the user, who can interact with the recommender system through predefined controls. Different presentations or different controls may lead to a better or worse conflict.

It should immediately be said that there are many possible non-algorithmic social media depolarization interventions, such as community moderation (Jhaver, et al., 2017). There are also hybrid approaches, such as The Commons, which uses automated messages (social media bots) to find people who want to engage in depolarizing conversations and then connects them to human facilitators (Build Up, 2019). There is also a wide variety of depolarization strategies entirely outside of algorithmic media, such as approaches based in journalism, politics, or education, any of which may prove to be more effective. Nevertheless, this paper considers only algorithmic interventions in recommender systems, because algorithmic content selection has been a central topic of concern, automation offers a means of scaling interventions, and the polarization properties of recommender algorithms are important in any case.
### 4.1. Removing polarizing content

Many kinds of content are now removed from platforms, including spam, misinformation, hate speech, sexual material, criminal activity, and so on (Halevy, *et al.*, 2022). While the removal of violent material and incitements to violence may be particularly important in the context of an active conflict (Schirch, 2020), the removal of less extreme material is a blunt approach that may not be justified as a mass depolarization intervention.

This kind of content removal is often called ‘moderation’, but it is important to distinguish between community moderation and algorithm-assisted moderation at scale. At the level of an online community or discussion group, volunteer moderators are able to set and enforce norms that lead to a productive discussion of polarized topics, as a study of the r/ChangeMyView subreddit shows (Jhaver, *et al.*, 2017). Such studies of the micro-dynamics of conflict provide important clues for potential depolarization interventions. Moderators remove posts and suspend accounts, but they also state the reasons for their actions, take part in discussions about appropriate policy, and consider appeals.

Platform moderation, by contrast, operates at a vast scale to identify unwanted content through a combination of paid moderators and machine-learning models. It is acontextual, impersonal, and difficult to appeal against (York and Zuckerman, 2019). The low rates of offending content mean that true positives (material correctly removed) may be vastly outnumbered by false positives (material incorrectly removed) unless automated classifiers can be made unrealistically accurate (Duarte and Llansó, 2017). Moreover, content removal is concerning from a freedom of expression perspective, and the standards for removal are widely contested (Keller, 2018). Facebook alone is “most certainly the world’s largest censorship body” (York and Zuckerman, 2019).

Given these concerns, there should be a high bar for automated content removal as a mass depolarization intervention. What should be the standard for unacceptably polarizing material? We could algorithmically remove all angry political comments, but do we want to? Removing all material that might intensify conflict would leave the public sphere arid, authoritarian and devoid of any real politics.

### 4.2. Increasing exposure diversity

Most prior work on the relationship between polarization and social media has been based on the concept of exposure diversity. The most frequently proposed fix is to increase the diversity of social media users’ feeds algorithmically (Bozdag and van den Hoven, 2015; Elisa Celis, *et al.*, 2019; Helberger, *et al.*, 2018), and a variety of recommender diversification algorithms have been developed (Castells, *et al.*, 2015). This is intuitively appealing, as inter-group contact has been demonstrated to reduce prejudice (Pettigrew and Tropp, 2006).

This approach presupposes, however, that a lack of diversity in online media content is causing polarization — which is questionable, as discussed above. ‘Diversity’ is also poorly defined, and may refer to source diversity, topic diversity, author diversity, audience diversity, and more. A review of media diversity by Loecherbach, *et al.* (2020) notes that “research on this topic has been held back by the lack of conceptual clarity about media diversity and by a slow adoption of methods to measure and analyze it”. Furthermore, the causal connection between exposure diversity and polarization is complex, and under some conditions exposure to outgroup content can actually increase polarization (Bail, *et al.*, 2018; Paolini, *et al.*, 2010; Rychwalska and Roszczyńska-Kurasińska, 2018; Taber and Lodge, 2006).

Yet increasing exposure diversity can work, at least to some extent. One experiment tested the effect of asking U.S. Facebook users to subscribe to (to ‘like’) up to four liberal or conservative news outlets, measuring the changes in the users’ affective polarization through a survey two weeks later. This level of exposure to outgroup information decreased affective polarization by about one point on a 100-point scale (Levy, 2020). By comparison, the rate of increase in affective polarization in the U.S. since 1975 is estimated at 0.6 points per year (Finkel, *et al.*, 2020). Rescaled to the same 100-point scale, the previously discussed experiment of leaving Facebook for a month resulted in a decrease of about two points [2],
although only on issue-based rather than affective measures. All of these estimates should be considered quite rough.

This demonstrates that increased exposure diversity can be a useful intervention point for depolarization, although its effect so far has been modest. Are different or better approaches possible? For example, Levy (2021) tested only news diversity, meaning professional journalism; polarization may, however, turn out to be more sensitive to non-news content or to user comments.

4.3. Recommending civil arguments

Several studies have attempted to determine the conditions under which polarization and depolarization occur. Kim and Kim (2019) found that those who read uncivil comments arguing for an opposing view rated themselves as being closer to ideological extremes on a post-exposure survey than those who did not. Civility may not be depolarizing per se, but incivility does seem to be polarizing. Suhay, et al. (2018) similarly show that comments that negatively describe political identities (e.g., ‘Liberals are ignorant’) increase polarization as measured by the feeling thermometer question. This effect also appears in the context of partisan media sources (e.g., MSNBC, Fox) where the “incivility [of] out-party sources affectively polarizes the audience” (Druckman, et al., 2019).

It seems likely that ‘civility’ and ‘partisan criticism’ can be algorithmically scored through existing natural language processing techniques, drawing on previous work classifying hate speech and harassment. All are conceptually close to the ‘toxicity’ operationalized by contemporary comment classification models (Noever, 2018). While these models are mostly used for moderation — that is, removing offending comments — they could also provide a ‘civility’ signal that is incorporated into recommender item ranking. Twitter has experimented with this idea (Wagner, 2019), but I am not aware of any production recommender that incorporates a civility signal in content ranking (as opposed to content moderation).

In addition to demoting uncivil content, it is possible to promote civil content. Experimental evidence shows that ranking high-quality comments at the top can have a positive effect on the tone of subsequent discussion (Berry and Taylor, 2017). In effect, this intervention hopes to model respectful disagreement — something that may not work if there are not many natural examples of productive inter-group conversation. In particular, there may be a lack of journalistic content that takes a depolarizing approach to reporting on controversial issues (Hautakangas and Ahva, 2018; Prior and Stroud, 2015; Ripley, 2018).

Of course, uncivil language can be necessary and important. We certainly don’t want an algorithmic media system that redirects attention away from anyone raising their voice. Indeed, several theories of democracy require confrontation of this kind, such as critical approaches (Helberger, 2019) or agonistic models (Mouffe, 2002). Hence there is a tension between encouraging expression and intervening to make the conversation more productive — this is the art of (algorithmic) mediation.

4.4. Priming better interactions

Given a particular set of items selected for a user, it may be possible to present them in a way that encourages more productive conflict. Language seems particularly important in political disagreements. Intriguingly, replacing the usual ‘like’ button with a ‘respect’ button increased the number of clicks on counter-ideological comments — that is, people were more likely to ‘respect’ something they disagreed with than to ‘like’ it (Stroud, et al., 2017).

While civility norms have been shown to contribute to successful online discussions of polarized topics (Jhaver, et al., 2017), it is difficult to automate the promulgation and enforcement of such norms. One intriguing possibility, however, is to change the content of automated messages, such as the one welcoming someone to a group. In a large-scale experiment on r/science on Reddit, adding a short note explaining what types of posts will be removed and noting that “our 1,200 moderators encourage respectful discussion” greatly reduced the rate at which newcomers violated community norms (Matias, 2019).
In a sense, changing user behavior is the strongest depolarization intervention. Accomplishing this is by no means easy, but these studies demonstrate that simple changes in the user interface can have profound effects.

5. Learning to depolarize

The approaches discussed above are justified on the basis of sociological theory, from results in laboratory settings, or through modest platform experiments. Real platforms are enormous, diverse, dynamic environments, and ecological validity is a serious problem for the development of social media interventions (Griffioen, et al., 2020). It is likely to be difficult to predict which depolarization interventions will succeed, and the best approach will vary between subgroups, in different contexts, and over time.

The effective management of polarization will therefore depend on a continual monitoring of polarization outcomes by platform operators. Affective polarization measures may prove to be the most useful category of metrics, in part because they do not depend on the type of content that drives polarization. More cognitive measures of polarization, such as issue position surveys (Draca and Schwarz, 2018; Kiley, 2017), may be less diagnostic for social media, where many interactions will not involve discussions of substantial policy preferences.

Platforms already monitor various non-engagement measures and incorporate them into recommender design and ranking (Stray, 2020). Facebook asks users whether specific posts led to a meaningful social interaction on or off the platform. This is a construct from social psychology that appears to be similarly interpretable across cultures (Litt, et al., 2020). YouTube likewise incorporates user-satisfaction ratings obtained by asking users what they thought of specific recommendations (Zhao, et al., 2019). Such metrics are used to drive product choices at the managerial level by selectively deploying changes, a form of A/B testing. They are also incorporated directly into the predictive models underlying item ranking, as the next section describes, but the first and most fundamental depolarization intervention is simply to monitor for actual polarization outcomes, rather than betting on theory.

5.1. Optimizing for depolarization

Survey responses can be used to train recommender ranking algorithms, for example by building a model that predicts whether an item is going to lead to a positive survey answer for a particular user in a particular context. This is, technically speaking, very similar to predicting which items will result in a click. Optimizing for predicted survey responses is an important technique in the nascent field of recommender alignment — the practice of getting recommender systems to enact human values (Stray, 2021; Stray, et al., 2020).

The feeling thermometer has been used experimentally to evaluate the polarizing effect of seeing a post by taking the difference between treatment and control groups (Kim and Kim, 2019; Suhay, et al., 2018). If it proves possible to know whether individual posts or conversations are polarizing, it should be possible to build a model to predict the polarization effect of showing novel posts. Similar classifiers are already in use to detect misinformation, hate speech, bullying, etc. One plausible technique is the TIES (Temporal Interaction EmbeddingS) model, which takes into account not only the text and image content of a specific post but the sequence of interactions around it, including discussions in comments, likes, shares, etc. (Noorshams, et al., 2020). In the context of an online discussion, the goal would be to determine whether users are having a productive exchange of views or a divisive argument, so the history of interactions carries significant information.

Alternatively, affective polarization measures could be used longitudinally, perhaps by asking a panel of users to respond to a feeling thermometer question daily or weekly, thereby measuring attitudes over time.
When compared to a control group, this amounts to a difference-in-differences design which gives robust causal estimates under certain assumptions [3]. That is, it should be possible to learn the actual polarizing effects of selecting different distributions of items. However, using longitudinal data to drive recommendation systems toward selecting depolarizing content is technically challenging, as it takes much longer and entails a higher level of abstraction than feedback on individual items.

Reinforcement learning (RL) algorithms may be the most general and powerful approach to learning patterns of recommendation that optimize long-term outcomes (Ie, et al., 2019; Mladenov, et al., 2019). In principle, affective polarization survey measures could be used as a reward signal for reinforcement learning-based recommenders. However, this sort of learning from sparse survey feedback has not yet been demonstrated. Additional algorithmic development will be necessary before longitudinal polarization measures can be incorporated into content selection algorithms, but the necessary technical research is underway because other sparse, long-term signals, such as user subscriptions, have immediate business value.

In other words, the same methods that make it possible to predict what movies to show someone to get them to subscribe may also make it possible to learn which patterns of interaction increase or reduce polarization.

5.2. Unintended consequences and the necessity of specification

The effective use of sociological metrics is complicated and can fail in a number of ways, regardless of whether the metric is used by people or algorithms. Using reinforcement learning to attempt large-scale political intervention should be regarded as a particularly alarming prospect. While there is a strong moral case for designing recommender systems to depolarize, unintended consequences could swamp any positive effects.

A metric is an operationalization of some theoretical construct, and might be an invalid measure for a variety of reasons (Jacobs and Wallach, 2021). Even a well-constructed metric almost never represents what we really care about: clickbait lies entirely in the difference between ‘click’ and ‘interest’. When used as targets, metrics suffer from a number of problems involving gaming and spurious correlations, which can be understood in causal terms as variations of Goodhart’s law (Manheim and Garrabrant, 2018). It is particularly important to use ongoing qualitative methods and undertake user research, to see whether current metrics are adequately tracking the intended goals — and to find out whatever else may be happening.

Metrics often fail when used in management contexts because they are irrelevant, illegitimate, or gamed, or are not updated as the context changes (Jackson, 2005). Using metrics to train a powerful optimizing system introduces further concerns (Thomas and Uminsky, 2020). Different effects for different subgroups may be a particular problem for recommender systems, which typically optimize average scores (Li, et al., 2021). While it is always useful to monitor for slippage between a metric’s intent and what it is actually measuring, this is particularly important when a measure becomes the target of society-wide AI optimization (Stray, 2020). If we choose to apply reinforcement learning to polarization metrics, those metrics will require continuous evaluation.

On the other hand, not using polarization measures in algorithmic content selection may be far worse. Optimization algorithms that do not penalize polarization measures might learn, as humans do, that polarization can be exploited for engagement. Or they might merely increase conflict as an agnostic side effect, which is no better. In general, under-specification is a serious hazard in the creation of machine learning models (D’Amour, et al., 2020). If we do not specify the intended effect of a recommender system on polarization, we should not be surprised to find unexpected outcomes.
6. Conclusion

Polarization is a hardening division of society into ‘us’ versus ‘them’. It interacts with a number of conflict feedback processes and eventually leads to democratic erosion and violence (McCoy, et al., 2018; Somer and McCoy, 2019). The goal of a depolarization intervention is not to suppress conflict, but to produce a better conflict that moves toward constructive societal change (Deutsch, 1969; Jeong, 2019; Lederach, 2014; Ripley, 2021). While all societies face complex tensions between peace and justice, depolarization interventions may ultimately be justified on human rights grounds (Finkel, et al., 2020), just as other peacebuilding interventions are.

The available evidence suggests that social media usage is not driving increases in polarization at the country level (Boxell, et al., 2022, 2017). In particular, there is little empirical support for the idea that personalization is reducing exposure to diverse information (Guess, et al., 2018; Zuiderveen Borgesius, et al., 2016). Nevertheless, there is some evidence that social media-based interventions can reduce polarization among users. In a recent experimental test, increasing news diversity led to a small decrease in polarization (Levy, 2021). And paying users to stay off Facebook for a month produced small decreases in issue polarization, although not affective polarization (Allcott, et al., 2020).

Moderation — the removal of unwanted content — can be important especially in the context of a violent conflict (Schirch, 2020), but is probably too blunt an instrument for depolarization. Content ranking defines what each user sees, and is the most general intervention point. While exposure to diverse perspectives can actually increase polarization (Bail, et al., 2018), increased exposure diversity does depolarize in some contexts (Levy, 2021; Pettigrew and Tropp, 2006). Recommenders could augment diversity by de-prioritizing content that has been shown to be polarizing, including uncivil presentations of outgroup opinions (Kim and Kim, 2019) and criticism of partisan identities (Suhay, et al., 2018). Content presentation and user interface may also have depolarizing effects, as has been shown in experiments changing ‘like’ to ‘respect’ (Stroud, et al., 2017) and adding a message reminding users of community norms (Matias, 2019).

Yet none of the above approaches directly targets the outcome of interest. Any depolarization method based on selecting content according to a pre-existing theory may prove unable to cope with the radically diverse and dynamic contexts of a real recommender system. The solution is to measure polarization outcomes directly and continuously.

Existing polarization measures — in particular, affective polarization measures — have been used to evaluate the effect of encountering different types of comments on news articles (Kim and Kim, 2019), and the same methods should apply to other types of items including user posts, discussion threads, and so on. Such survey data can be used to evaluate recommender system changes and make deployment decisions. It can also be used to train polarization prediction models, much as existing recommender models predict meaningful social interactions and other survey results (Stray, 2020). Ultimately, polarization survey feedback could be used as a reward signal for reinforcement learning-based recommendation algorithms. This powerful emerging approach has the potential to learn what actually depolarizes, and to adapt continuously to changes. Optimizing for such a signal may have unintended harmful consequences, so such a system would need to be continuously monitored in other ways — through qualitative studies, for instance. In any case, it may prove necessary to incorporate polarization measures into recommender systems to prevent the creation of conflict as a side effect of optimization (D’Amour, et al., 2020).

It is not known whether this sort of feedback-driven intervention would succeed in reducing the average dislike of an outgroup, whether it would be better than doing nothing — or, more broadly, whether intervening in platform recommenders will be an effective depolarization strategy within the complex and dynamic media ecosystem of any particular community — but there is reason to suspect this is possible. At the very least, the collection of individual-level affective polarization survey data provides a managerial incentive in the direction of depolarization. In any case, the use of affective polarization survey data to drive platform recommender systems is a theoretically grounded, technically feasible, and potentially robust
Designing recommender systems to depolarize strategy for a social media depolarization intervention, and deserves further study.

About the author

Jonathan Stray is Senior Scientist at the Center for Human-Compatible AI at the University of California, Berkeley, where he works on the design of recommender systems for better personalized news and information. He previously taught the dual master’s degree in computer science and journalism at Columbia University, built several pieces of software for investigative journalism, worked as an editor at the Associated Press, and developed graphics algorithms for Adobe. He holds an M.Sc. in computer science from the University of Toronto and M.A. in journalism from the University of Hong Kong.
E-mail: jstray [at] berkeley [dot] edu

Notes


References


Designing recommender systems to depolarize


Designing recommender systems to depolarize


Designing recommender systems to depolarize


doi: [https://doi.org/10.1080/1369118X.2016.1271900](https://doi.org/10.1080/1369118X.2016.1271900), accessed 11 April 2022.


doi: [https://doi.org/10.1111/ajps.12152](https://doi.org/10.1111/ajps.12152), accessed 11 April 2022.

doi: [https://doi.org/10.1080/0300393042000332837](https://doi.org/10.1080/0300393042000332837), accessed 11 April 2022.

doi: [https://doi.org/10.1145/3442188.3445901](https://doi.org/10.1145/3442188.3445901), accessed 11 April 2022.

doi: [https://doi.org/10.4324/9781315182070-2](https://doi.org/10.4324/9781315182070-2), accessed 11 April 2022.


doi: [https://doi.org/10.1145/3306618.3314288](https://doi.org/10.1145/3306618.3314288), accessed 11 April 2022.


J.N. Matias, 2019. “Preventing harassment and increasing group participation through social norms in 2,190 online science discussions,” *Proceedings of the National Academy of Sciences*, volume 116, number 20 (29 April), pp. 9,785–9,789. doi: [https://doi.org/10.1073/pnas.1813486116](https://doi.org/10.1073/pnas.1813486116), accessed 11 April 2022.


Designing recommender systems to depolarize


---

**Editorial history**

Received 26 March 2022; accepted 11 April 2022.

---

This paper is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Designing recommender systems to depolarize

by Jonathan Stray.

*First Monday*, volume 27, number 5 (May 2022).

doi: [https://dx.doi.org/10.5210/fm.v27i5.12604](https://dx.doi.org/10.5210/fm.v27i5.12604)