The Myth of "The Algorithm": A system-level view of algorithmic amplification

ANONYMOUS AUTHOR(S)

With the ever-changing social media landscape and the spread of algorithmic systems that recommend content, there has been an increasing interest in the concept of "algorithmic amplification" in order to understand how attention to content is allocated online by recommendation systems. The typical approach is to compare an algorithm-based approach to some definition of a "neutral" baseline. In this position paper, we elucidate two key points about the concept of algorithmic amplification. First, we enumerate the typical components of a social media system design, illustrating in the process that it is extremely difficult to isolate one single "algorithm" that is responsible for how content is amplified. Second, we show, through simulation studies, that there is no neutral baseline to compare against in amplification studies. The most commonly used baseline, a reverse chronological display of content, itself contains biases and disparities that were present in other system components of the past. Our overall conclusion is that practitioners studying amplification need to think carefully about the choices they make and what assumptions those choices carry about the overall system being studied.

ACM Reference Format:
Anonymous Author(s). 2023. The Myth of "The Algorithm": A system-level view of algorithmic amplification. In *EAAMO '23*. ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Algorithmic recommender systems underlie much of the technology we interface with today, influencing everything from what products we buy, to what news sources or political viewpoints we're exposed to, to how much income we earn from content that we create. In the context of social media, recent attention has turned to how these algorithmic systems can increase the reach of certain types of content relative to the reach it would have gotten under some other neutral baseline[1]– a phenomenon called "algorithmic amplification." While the concept of algorithmic amplification could be applied to any type of content– for example, we could study the excess reach of cat memes on a platform– by and large, existing efforts to study, measure, or regulate algorithmic amplification have sprung from concern that algorithmic amplification acts as a vector for specific societal problems or harms.

One societal problem that has been central to the discourse and scholarship around algorithmic amplification is the ability of algorithmic systems to expose users to overtly harmful content, such as extremist or radicalizing content [29] and misinformation [9]. For example, Representatives Anna G. Eshoo of California and Tom Malinowski of New Jersey introduced the Protecting Americans from Dangerous Algorithms Act, which is intended to hold companies "liable if their algorithms amplify misinformation that leads to offline violence," according to a press release from Representative Eshoo's office [2]. More recently, the Supreme Court heard a pair of cases, Gonzalez v. Google and Twitter v. Taamneh, which considered whether social media companies could be held liable for allowing extremist content from ISIS on their platforms [10].

[1] This definition is based on that given by Dean Eckles in testimony to the U.S. Congress. [5]

When it comes to overtly harmful content, amplification is undesirable in an absolute sense— any amplification is unwanted. By contrast, the second concern about algorithmic amplification is relative; it hinges on disproportionalities or unfairness in amplification, especially as it relates to disparities in amplification among groups. In this case, the views on what amplification is justified can vary; some argue that amplification is not an issue if it aligns with user interests, while others might believe that amplification should act as a positive social force to amplify traditionally marginalized voices. The interest in disproportionalities in amplification is founded in concern that social media platforms unfairly allocate more influence to some types of people than others or tip the scales of public opinion by amplifying certain viewpoints over others, for example the U.S. political left over the political right [13]. Additionally, as people increasingly gain economic opportunities via their social media presence, disparities in algorithmic amplification could unfairly allocate the economic spoils that follow from having a robust social media presence to already privileged groups.

Although the topic of algorithmic amplification has been prevalent in the public discourse, the concept remains murky. In particular, although algorithmic amplification as an abstract concept is clear, operationalizing this definition to detect and measure algorithmic amplification remains challenging. In our view, this largely stems both from imprecision about which models and systems are included in “the algorithm” and a lack of clarity about an appropriate “neutral baseline” against which to measure amplification. In this work, we will disentangle how various components of a typical social media platform beyond what is commonly considered “the algorithm” can contribute to amplification and simultaneously confound its measurement. In doing so, we enumerate the various components that go into producing curated feeds of content on social media beyond just the ranking algorithm itself, including content moderation, account recommendation, search, business rules, auxiliary models, and user behavior. We then challenge the notion of a “neutral baseline” and illustrate how, in practice, the most common choice of baseline fundamentally depends on the state of some components of the system. By conditioning on the state of any system component in constructing a baseline, we assume away the effects of past bias and amplification that brought the system component to its current state. This leads to unsatisfying and potentially misleading conclusions about whether and to what extent certain content is amplified. We conclude by returning to the underlying concerns spurring the discussion of algorithmic amplification and propose concepts and measures we believe may more actionably address those concerns because they do not rely on the existence of a neutral baseline against which to compare.

2 ALGORITHMIC AMPLIFICATION

In order to study algorithmic amplification, we must first define it. We adopt the working definition: “the extent to which the algorithm gives certain content greater reach than it would have gotten with some other neutral baseline algorithm.” Operationalizing this definition requires us to further define “the algorithm” and the neutral baseline against which it is measured. We take these in turn.

2.1 Defining “the algorithm"

Algorithmic amplification is ultimately interested in the different levels of attention and exposure content receives. Thus, we begin our tour of “the algorithm” by defining the surfaces through which the algorithm delivers content to users and the different technical components that drive each of the surfaces. The pieces we define here are not specific to a particular social media platform, but are instead amalgamations of features we have observed or have been publicly documented across the industry. This general framework can be compared to the more technical documentation recently released by Twitter regarding the structure of their system.

2.1.1 Surfaces of exposure. When envisioning a typical social media system, we often first think of the "main content feed" – for example, Twitter's For you timeline, Facebook's Feed, TikTok's For You Page [7, 20, 23]. This is usually the first surface that users encounter when they log in, and it contains content curated to their tastes. This feed can include content from accounts that the user has specifically opted into (which we will call "followed" or "in-network" accounts), as well as content from accounts that the user has not opted to follow (which we will call "unfollowed" or "out-of-network" accounts). The policies as to what kind of out-of-network content can be included in a user's feed vary by platform, with some platforms displaying mostly "in-network" content and others relying almost entirely on machine learning algorithms to infer what— from the universe of all content– a user would most like to see [17].

In addition to the main feed, users have other ways of finding content. Most platforms include some space primarily intended for exploration of out-of-network content. For example, Twitter and Instagram both have an "Explore" page [12, 21]. These pages contain trending content, but are often still curated to a user's inferred preferences. Another path to a particular account's content is through the search bar. By searching for a particular topic or a particular account name, users can discover content relevant to their search query; Twitter and TikTok both have examples of this functionality [19, 25]. Additionally, content creators often collect the links to their social media pages on a single page – for example, linktr.ee provides this capability, and external websites often link to specific authors or content when citing sources [1]. As such, it is possible for a user to find a creator's content through a direct profile view or directly landing on a particular piece of content as well. Finally, there are some app-specific features, such as notifications, curated lists and topics, hashtags, etc. that give users pathways to different kinds of content [26, 27].

2.1.2 Primary ranking models. The surfaces enumerated above are all driven to some extent by recommendation and ranking models - machine learning models that decide how relevant a particular piece of content is to a user. This model takes in a set of candidates and scores them according to how likely the content is to be relevant to the user. The candidate pieces of content are then sorted by their score and displayed. Because it would take an infeasible amount of computation to score every piece of content for every user, the ranking model is usually fed a smaller set of candidates by a computationally cheaper model or set of heuristics. This first step is called the candidate generation step. When a recommendation system works in this way, it is called a two-stage recommender system [28]. In practice, the rules and criteria used to generate the set of candidates can have a great impact on what content is ultimately seen by the user [4]. Also, depending on the design of the system, different surfaces may have different two-stage recommenders trained for them, depending on the objective of the surface. For example, an Explore page likely has a different model driving it than the main content feed. This ultimately depends on what each surface is optimized for and how much model reuse can happen. It is a technical design choice for the platform.

2.1.3 Peripheral ranking models. In addition to the models deciding how to order content on different surfaces of exposure, there are other models that drive various interactions on the platform. One prime example is the account recommendation model, which suggests other users that you may be interested in following [8, 22]. Another is the search ranking algorithm, which will rank both content and user accounts based on their relevance to a user's query. Your past activity, inferred interests, and existing connections can all factor into what these models rank highly. Generally speaking, a peripheral model is any model which is ranking content of some kind that does not appear on one of the main exposure surfaces. This could include ad ranking, trending news, and many other types of content.

2.1.4 Auxiliary models. On top of models that do ranking, there are also many machine learning models that compute scores or vector representations for content or accounts. The most prominent of these are related to content moderation, or the process of monitoring potentially harmful content on platforms. This monitoring can include both content that explicitly violates platform terms of service or rules and content that is not clearly violative of terms of service but does detract from a healthy platform, often referred to as "toxic."

Models are used at many different stages of content moderation. Some are used to decide what content to refer to human reviewers [15]. Other models assign a score to determine how toxic or marginally abusive a particular piece of content is, and this score may be used in determining overall ranking scores [3, 30]. In other cases, users might even be prompted to reconsider their content before posting it based on a model's toxicity score [14].

In addition to content moderation scoring, there are often models which seek to extract representations of items, like content or accounts, that can be fed into downstream models. Some approaches include extracting joint "embeddings" - dense vector representations - for accounts, advertisements, and posts [6]. Others label content as pertaining to different topics or interests that are relevant to a user.

In this work, we refer to these models as auxiliary models. While they are usually not directly responsible for ranking or suggesting content, they do greatly affect what content is seen, removed, or amplified on the platform. These are also auxiliary in the sense that the scores or vectors these models produce are often used as inputs to other models.

2.1.5 Manual components or "business logic". In addition to all of the data-driven pieces of the system, there are often also manually coded policy and business decisions that influence and moderate the display of content. For example, "brand safety" policies might be implemented to ensure that certain types of content are not shown next to particular brands' advertisements [24]. Other heuristics might be used to ensure a user's feed is not too repetitive, such as limiting the number of times ads or posts from the same user can be shown consecutively. Other business rules might decide to highlight a particular new feature by displaying it prominently in the content feed. While the existence of such policies is often opaque to the end user, these too have a profound effect on what content gets shown and where.

2.1.6 How everything is connected. When it comes to algorithmic amplification, which part of this complex system is "the algorithm"? Typically, the content ranking model underlying the main feed is thought of as "the algorithm". However, the content that is ultimately shown to users is impacted by every part of this system, either directly or indirectly. For example, Figure 1 presents a bird's eye view of the interconnected "model spaghetti monster" that is a social media platform. This shows the different surfaces of exposure and various models that drive them, along with other encapsulations of the state of the system such as the follow network. Gray arrows show how various models' and algorithms' feed into one another to ultimately deliver content to a user. Colored errors show pathways through which a user's interactions with that content they are shown impacts system components, either by serving as new training data for an underlying model or directly altering the state of the follow network. One obvious pathway by which all models ultimately contribute to the main content ranking model that populates the main feed is via the follow network. All system components ultimately are involved in displaying content to users in one way or another. The content the user is shown (or not shown) all ultimately impacts what they choose to follow. The follow network then serves as an input to the main content ranking models, thereby creating an indirect effect of every system component on what appears in the main feed. Because of the inter-connected way in which all of the system components can ultimately impact what content is served to a user, "the algorithm" is, in fact, the entire system.

2.2 Defining the baseline

Per our working definition, amplification is defined relative to some "neutral baseline" algorithm. More broadly, this baseline could be defined as exposure that would have been obtained without the existence of the platform at all, though

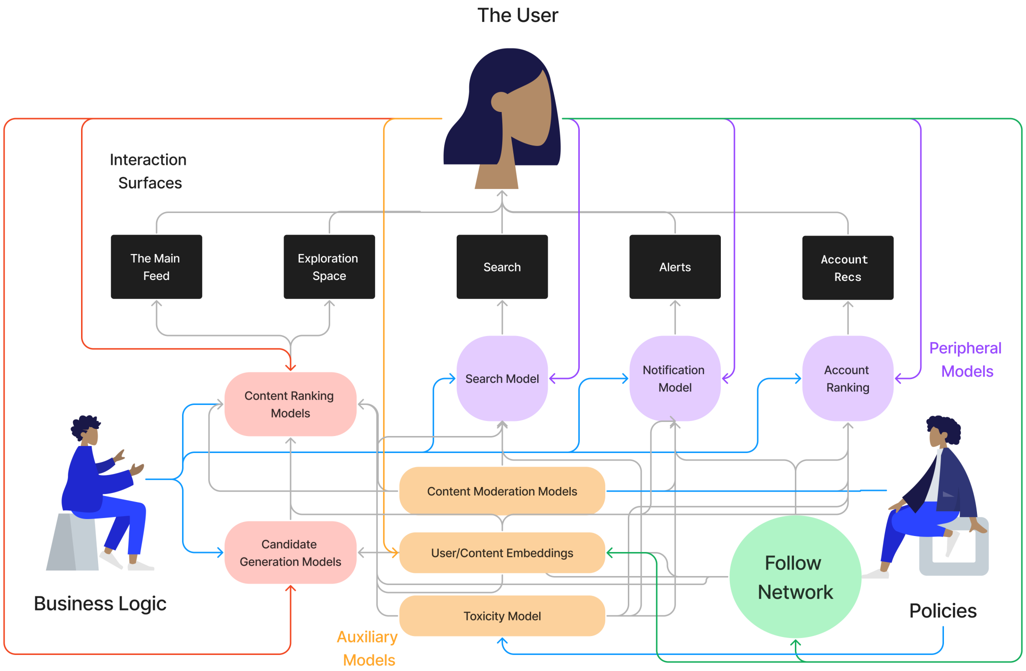


Fig. 1. A schematic diagram of a typical social media platform

in practice, the neutral baseline is typically thought of as the outcome that would have occurred if some other, more "neutral" algorithm had been used on the platform as it otherwise exists. What, then, is a reasonable neutral baseline against which to compare?

Recently proposed legislation also points towards some candidates for what could be considered a "neutral baseline." While the previously mentioned Protecting Americans from Dangerous Algorithms Act did not venture into technical definitions of algorithmic amplification, it does specify what types of systems the proposed bill would and would not apply to, strongly implying what types of algorithms they might consider neutral by exclusion. Specifically, this bill designates that the proposed law would apply to any interactive computer service that "used an algorithm, model, or other computational process to rank, order, promote, recommend, amplify, or similarly alter the delivery or display of information" and specifically excludes systems that are easily understood by humans, such as displaying content in reverse chronological order, in order of overall popularity, or cases in which a user explicitly seeks out the content. While this does not specify how amplification ought to be measured, by stating what types of rank ordering are ineligible, it provides useful guidance on what the authors might consider a neutral baseline against which to measure excess transmission.

Of those options, the most commonly proposed baseline algorithm is "reverse chron", i.e. an algorithm that displays in-network posts in reverse chronological order. This baseline was recently used in [11] to compare amplification of

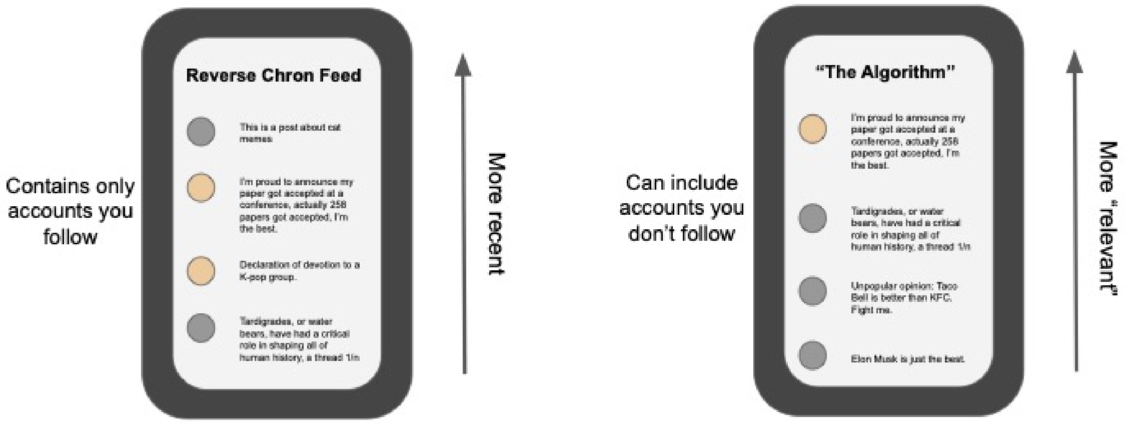

Contains only accounts you follow
Reverse Chron Feed
This is a post about cat memes
I'm proud to announce my paper got accepted at a conference, actually 258 papers got accepted. I'm the best.
Declaration of devotion to a K-pop group.
Tardigrades, or water bears, have had a critical role in shaping all of human history, a thread 1/n
More recent
Can include accounts you don't follow
"The Algorithm"
I'm proud to announce my paper got accepted at a conference, actually 258 papers got accepted. I'm the best.
Tardigrades, or water bears, have had a critical role in shaping all of human history, a thread 1/n
Unpopular opinion: Taco Bell is better than KFC. Fight me.
Elon Musk is just the best.
More "relevant"

Fig. 2. Example comparing a reverse chronological feed (left) to an algorithm-driven feed (right)

different elected officials on Twitter in what is– in our opinion– the best example to date of measuring algorithmic amplification on a real social media platform. In light of this, going forward we focus on analyzing amplification with reverse chron as the baseline and the implications of this choice of baseline on our measurement of algorithmic amplification. Then, in mathematical terms, our working definition of amplification becomes

$$\text{Amplification}(x) = \frac{\text{Impressions to x under current systme}}{\text{Impressions to x under reverse chronological}} \tag{1}$$

where x defines the group of users or content type for which we are calculating algorithmic amplification and an impression is the event of a user viewing a piece of content[2]. In practice, this value could be calculated counterfactually by keeping track of how many impressions would have gone to x under each algorithm, regardless of which was actually used. Or, it could be calculated by comparing the number of impressions to x by users using the current system to the number using reverse chronand normalizing appropriately for the relative number of users in each condition.

The implicit assumption in adopting reverse chron as a neutral baseline is that the accounts that a user follows are a neutral representation of the content that user wants to see. But, is this necessarily the case? Suppose the peripheral account recommendation model preferentially recommends accounts of a certain type. Or, similarly, suppose that the search function preferentially places certain types of accounts near the top of the returned results, making it easier for users to find some accounts than others. This amplification of the "preferred" accounts by search or account recommendation would cause those accounts to be over-represented among the accounts users follow relative to other accounts that received no such amplification from peripheral models. Thus any "algorithmic amplification" that would be detected by comparing the exposure the preferred accounts received to the exposure those accounts would have gotten under reverse chron would only account for the marginal amplification due to the ranking model alone. The amplification due to the biased search or account recommendation would be subsumed by the "neutral baseline" and subtracted away.

[2] This definition is very similar to that used in [11] to calculate political amplification for different politicians. In their analyses x was either individual politicians or groups of politicians defined by their political party.

Extending this argument, adopting a reverse chronological baseline bakes in the assumption that the past behavior of the system, including the ranking model, was neutral. For example, suppose the system has historically preferentially placed content produced by certain accounts near the top of the ranking, allowing those accounts more opportunities to amass followers. If we adopt a baseline built upon the follow graph, we bake the historical advantage enjoyed by some accounts into the calculation, thus under-estimating the advantage they currently receive relative to what we would have calculated if the past had truly been neutral. By conditioning on the state of any system component in constructing a baseline, we assume away the effects of past bias and amplification that brought the system component to its current state.

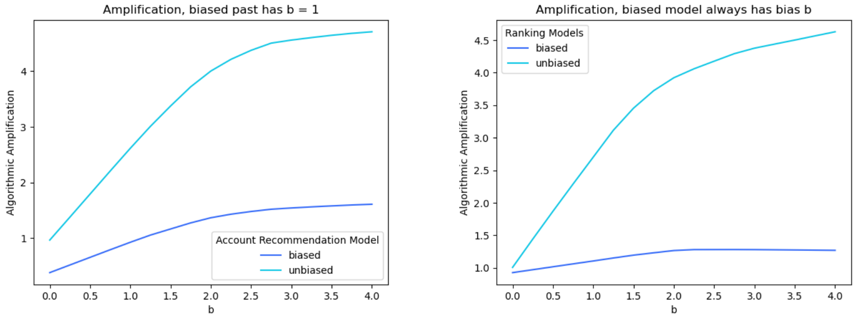
2.2.1 An illustrative simulation. To concretize the dynamics described above, we present an illustrative simulation. First, suppose we have a population of n = 100 accounts, all of which are "identical" in the sense that their interests are all drawn from the same distribution. That is, at each time point t, each account draws a value from a vit N(0,1) which represents the interests of user i at time t. They broadcast this value in a post. Under our model, posts that are most similar to a user's posted value at each time point are the most relevant to that user. Specifically, for dijt = |vit - vjt | the absolute difference between the value user i posted and the value user j posted at time t, we define the most relevant posts for user i at time t to be those for which dijt is smallest.

In this simulation, we are interested in calculating the extent of algorithmic amplification of a set of m users that "the algorithm" treats preferentially. We select the users to be treated preferentially by the system to be the first m users— the ordering is arbitrary so this is essentially a random selection.

In this case, content is ranked as follows. For each i, we sort the values dijt - bIj from smallest to largest and display the first KT. Here, b is the amount by which the advantaged accounts are artificially boosted, and Ij is an indicator of whether the jth account is one of the accounts that get the unfair benefit. In a nutshell, under our biased ranking algorithm, each user is shown the accounts that are most relevant to them, with the exception that some accounts get a boost in the ranking that is unrelated to the relevance of the content they produced to the other users. Note that when b=0, the algorithm does not preferentially amplify any accounts. At each time point, each user then elects to follow each of the accounts it was shown with probability p = 0.05. After an initial period of 50 iterations, we then calculate algorithmic amplification over the subsequent five iterations by dividing the number of times the advantaged accounts appeared on the generated timelines using "the algorithm" by the number of times the advantaged accounts appeared in timelines using reverse chron. Referring back to our mathematical definition of algorithmic amplification, x in that equation refers to the advantaged users in this simulation.

We study this under two scenarios. In the first, we set the bias to be b=1 during the initial 50 iterations. Then, we calculate algorithmic amplification for several values of b. This is shown by the dark blue line in the left figure. For convenience, we also provide a comparison to what the calculation of algorithmic amplification would have been if in the past, the algorithm had not shown preferential treatment to the advantaged accounts, i.e. what we would have calculated if our "neutral baseline" were, in fact, "neutral". This is shown in light blue.

Two things stand out in this figure. First, for small values of b, we estimate amplification with a value less than one, meaning de-amplification. This measurement would imply that the preferred accounts are actually getting less exposure than they should under the chosen neutral baseline. Not only would we conclude that the preferred accounts did not get amplified by the system, we may infer that they are being unfairly treated. Even for very large values of b (i.e. when the preferred accounts have a very large advantage), we still estimate amplification of those accounts as barely greater than one, indicating negligible amounts of amplification of those accounts. The second notable facet of



(a) $b = 1$ in the past

(b) Current b equal to past b

Fig. 3. Simulations of amplification measurements with different configruations of biased pasts

this figure is that as the preferred accounts' advantage grows (i.e. as b gets larger) , so does the gap between the true level of amplification under an unbiased past and what we actually measure: for the largest amounts of algorithmic amplification, we underestimate it the most.

In the second scenario, we allow "the algorithm" to maintain the same level of bias in the initial 50 iterations as it has in the subsequent iterations during which amplification is calculated. The results of this simulation are shown in the right figure. Here, we see that across all values of b, we measure no amplification. This is because the follow network on which the baseline (reverse chron) is based is imbued with exactly the same bias as "the algorithm". This fundamentally limits our ability to test for algorithmic amplification. By comparison, if we had not had a biased content recommendation algorithm in the past (light blue), we would have correctly been able to infer that the advantaged accounts enjoy a significant amount of amplification for large values of b— the algorithm allocates about four times as many impressions to the advantaged accounts as does reverse chron.

This simulation clearly falls short of representing the complexity of link formation and content delivery on social media platforms. For example, in reality people would not follow accounts they are shown at random, but rather, would likely select those they find most interesting or relevant to them. Similarly, in the real world all users are not identically distributed and heterogeneity of preference complicates the measurement of algorithmic amplification even further. We have chosen this very simple setting because it is only in an extremely simplified universe can we easily develop a non-controversial truly neutral baseline against which to compare measures of algorithmic amplification.

3 PATHS FORWARD

We've given some examples of how measuring algorithmic amplification is complicated by the realities of a complex system that evolves over time. In the absence of a neutral baseline that has not been influenced by the past behavior of the system, directly measuring algorithmic amplification is difficult if not impossible. In light of this, what can we do?

It is useful to return to the underlying concerns motivating the desire to study and address algorithmic amplification in the first place. First, when it comes to overtly harmful content, any exposure to the harmful content is undesirable. Perhaps the question is less about how many times people using different algorithms for sorting content see the harmful content, but rather, how many people see it at all, regardless of whether they follow its creator or not. This suggests it may be useful to track impressions on harmful content without baselining to any algorithmic comparator or neutral baseline.

When it comes to the unfair allocation of impressions, the underlying concern is that algorithmic amplification causes some groups or individuals to receive an unfair amount of exposure relative to others. We can further break down what is meant by an "unfair amount of exposure". One interpretation of that might be that there is some component of the system that is unjustifiably biased towards or against some group or individual. Or, reversing that, we might say that the system is not unfair if all of its components are operating in such a way that no group or individual is unjustifiably advantaged by any component. This suggests that an audit be performed to ensure that each of the components of the system is behaving "fairly"– that is to say, does not exhibit predictive bias towards or against any socially salient group or individual. Accompanying metrics might be things like measures of group-wise model performance disparities for each system component. Mitigating unjustified disparities in amplification under this approach would then be equivalent to minimizing each system component's performance disparities, however defined.

While this approach is appealing, it fails to account for the ways in which all of the system components interact. In a complex system, such as a social media system, it is possible that the constituent models, policies, and heuristic rules could interact in unpredictable ways leading to "unfair" allocation, even if each of the components is "fair" in isolation. In light of this, one potential alternative definition of unfair allocation of exposure is if content from some groups or individuals is given higher levels of exposure relative to the size of that content's receptive audience than is given to content from other groups or individuals. In some sense, this is already at the heart of our operationalized definition of algorithmic amplification— where we compare the exposure that was given by the algorithm (the numerator) to a baseline (reverse chron) that serves as a proxy for the size of the receptive audience (the denominator). Even in this case, however, it is clear that aligning to the receptive audience is not a neutral baseline, because the measure of the audience (the "follow graph") is itself influenced by the algorithmic system.

Conceptually, comparing algorithmic amplification across groups or individuals is a way of considering disparities in exposure (via the numerators) while simultaneously accounting for whether the content is reaching a receptive audience (via the denominator). To address the same desiderata, we could instead turn to measures of exposure inequality across socially salient groups– a venture akin to comparing the numerators directly. Separately, we can use a counter-balancing metric that is designed to account for how satisfied the consumers are with the content they have been presented to account for the competing need to ensure that flattening the distribution of impressions does not come at the expense of delivering readers content they actually enjoy. Several methods for calculating inequality of exposure or groupwise disparities in exposure have been proposed recently, and there exist many metrics for inferring the quality of the content that has been displayed to users [16, 18]. This could be implemented via experimentation or A/B testing by tracking both inequality metrics and metrics for reader-side satisfaction to try to address the problem of the system allocating disproportionate influence to some users relative to the size of the audience that is receptive to their content.

Connection to algorithmic amplification aside, this approach to addressing the concerns underlying discussion of disparities in algorithmic amplification may be beneficial in its own right. Reducing inequality can create a platform where more voices can be heard and can help avoid concentrating influence and exposure in fewer people and perspectives. Given the increasing reliance on social media for building professional networks and marketing, reducing algorithmic amplification could also be a useful tool for flattening real world access to economic and career opportunities.

4 CONCLUSION

Defining algorithmic amplification to account for all of the ways in which a social media system can increase exposure to certain types of content requires us to expand our definition of "the algorithm" from simply a single recommender system to a complex web of interacting models, policies, and actors. In doing so, we are forced to grapple with the fact that commonly proposed baselines against which to measure algorithmic amplification are not as neutral as they first appear– they condition on the state of system components that have been influenced by the past behavior of "the algorithm" or could arguably be considered part of "the algorithm" themselves. Indeed, any neutral baseline that conditions on the state of the system will suffer from this same issue, confounding our ability to measure amplification as it is currently defined.

At its heart, the term algorithmic amplification is used in relation to concern about two related but distinct issues: exposure to overtly harmful content and unjustifiable disparities in exposure between groups and individuals. In the interest of clarifying what is already a complicated and politically fraught issue, going forward it may be useful to discuss "algorithmic exposure" in the context of overtly harmful content. In this case, any algorithmic exposure is undesirable, whether the algorithm amplified it or not. In the context of disparities in exposure on a platform, "algorithmic inequality" may more appropriately address the underlying concern that these systems disproportionally allocate influence and benefits. Semantics aside, further careful study and regulation of algorithmic amplification are critical for ensuring equitable benefits from online platforms and reducing the damage that can result from exposing and promoting harmful content.

REFERENCES

[1] [n. d.]. What is Linktree? | Linktree Help Center. https://help.linktr.ee/en/articles/5434130-what-is-linktree

[2] 2021. Reps. Eshoo and Malinowski Reintroduce Bill to Hold Tech Platforms Accountable for Algorithmic Promotion of Extremism. http://eshoo.house.gov/media/press-releases/reps-eshoo-and-malinowski-reintroduce-bill-hold-tech-platforms-accountable

[3] Jack Bandy and Tomo Lazovich. 2022. Exposure to Marginally Abusive Content on Twitter. https://doi.org/10.2139/ssrn.4175612

[4] Amanda Bower, Kristian Lum, Tomo Lazovich, Kyra Yee, and Luca Belli. 2022. Random Isn't Always Fair: Candidate Set Imbalance and Exposure Inequality in Recommender Systems. (2022). https://doi.org/10.48550/ARXIV.2209.05000

[5] Dean Eckles. [n. d.]. Algorithmic transparency and assessing effects of algorithmic ranking. https://www.commerce.senate.gov/services/files/62102355-DC26-4909-BF90-8FB068145F18

[6] Ahmed El-Kishky, Thomas Markovich, Serim Park, Chetan Verma, Baekjin Kim, Ramy Eskander, Yury Malkov, Frank Portman, Sofía Samaniego, Ying Xiao, and Aria Haghighi. 2022. TwHIN: Embedding the Twitter Heterogeneous Information Network for Personalized Recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, Washington DC USA, 2842–2850. https://doi.org/10.1145/3534678.3539080

[7] Facebook. [n. d.]. How Feed WorksFacebook Help Center. https://perma.cc/CZ9G-FDBN

[8] Facebook. [n. d.]. People You May Know. https://www.facebook.com/help/336320879782850

[9] Miriam Fernández, Alejandro Bellogín, and Iván Cantador. 2021. Analysing the Effect of Recommendation Algorithms on the Amplification of Misinformation. https://doi.org/10.48550/arXiv.2103.14748 arXiv:2103.14748 [cs].

[10] Jennifer Stisa Granick. 2023. Is This the End of the Internet As We Know It? | ACLU. https://www.aclu.org/news/free-speech/section-230-is-this-the-end-of-the-internet-as-we-know-it

[11] Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. 2022. Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences* 119, 1 (Jan. 2022), e2025334119. https://doi.org/10.1073/pnas.2025334119

[12] Instagram. [n. d.]. Explore Tab. https://perma.cc/BK2R-YY2N

[13] Cecilia Kang and Sheera Frenkel. 2018. Republicans Accuse Twitter of Bias Against Conservatives. *The New York Times* (Sept. 2018). https://www.nytimes.com/2018/09/05/technology/lawmakers-facebook-twitter-foreign-influence-hearing.html

[14] Matthew Katsaros, Kathy Yang, and Lauren Fratamico. 2022. Reconsidering Tweets: Intervening during Tweet Creation Decreases Offensive Content. *Proceedings of the International AAAI Conference on Web and Social Media* 16 (May 2022), 477–487. https://doi.org/10.1609/icwsm.v16i1.19308

[15] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–18. https://doi.org/10.1145/3491102.3501999

[16] Tomo Lazovich, Luca Belli, Aaron Gonzales, Amanda Bower, Uthaipon Tantipongpipat, Kristian Lum, Ferenc Huszár, and Rumman Chowdhury. 2022. Measuring disparate outcomes of content recommendation algorithms with distributional inequality metrics. *Patterns* 3, 8 (Aug. 2022), 100568. https://doi.org/10.1016/j.patter.2022.100568

[17] Arvind Narayanan. [n. d.]. Understanding Social Media Recommendation Algorithms. http://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms

[18] Guillaume Saint-Jacques, Amir Sepehri, Nicole Li, and Igor Perisic. 2020. Fairness through Experimentation: Inequality in A/B testing as an approach to responsible design. https://doi.org/10.48550/arXiv.2002.05819 arXiv:2002.05819 [cs, econ].

[19] TikTok. [n. d.]. Discover and search | TikTok Help Center. https://perma.cc/3K28-UPS4

[20] TikTok. [n. d.]. For You - TikTok Help Center. https://perma.cc/V763-MXM5

[21] Twitter. [n. d.]. About our approach to recommendations. https://perma.cc/557T-8QG4

[22] Twitter. [n. d.]. About Twitter's account suggestions. https://help.twitter.com/en/using-twitter/account-suggestions

[23] Twitter. [n. d.]. About your For you timeline on Twitter. https://perma.cc/4ZCC-MSMF

[24] Twitter. [n. d.]. Brand Safety @Twitter. https://business.twitter.com/en/help/ads-policies/brand-safety.html

[25] Twitter. [n. d.]. How to use Twitter search – search Tweets, people, and more. https://perma.cc/NZF5-ND65

[26] Twitter. [n. d.]. Topics on Twitter | Twitter Help. https://perma.cc/Q9SC-H8FK

[27] Twitter. [n. d.]. Twitter Notifications timeline and quality filters. https://perma.cc/RAD5-S42X

[28] Lequn Wang and Thorsten Joachims. 2023. Uncertainty Quantification for Fairness in Two-Stage Recommender Systems. https://doi.org/10.48550/arXiv.2205.15436 arXiv:2205.15436 [cs].

[29] Joe Whittaker, Seán Looney, Alastair Reed, and Fabio Votta. 2021. Recommender systems and the amplification of extremist content. *Internet Policy Review* 10, 2 (June 2021). https://policyreview.info/articles/analysis/recommender-systems-and-amplification-extremist-content

[30] Kyra Yee, Alice Schoenauer Sebag, Olivia Redfield, Emily Sheng, Matthias Eck, and Luca Belli. 2022. A Keyword Based Approach to Understanding the Overpenalization of Marginalized Groups by English Marginal Abuse Models on Twitter. https://doi.org/10.48550/arXiv.2210.06351 arXiv:2210.06351 [cs].