Beyond Moderation: Emerging Research in Online Governance

The Justice Collaboratory's **Social Media Governance Initiative** 2023 Convening presents three essays on **Emerging Research on Online Governance**

THE JUSTICE COLLABORATORY

SOCIAL MEDIA GOVERNANCE INITIATIVE

YALE LAW SCHOOL

table of contents

Introduction

Matt Katsaros
Sudhir Venkatesh

In 2019, the Justice Collaboratory launched a new initiative—the Social Media Governance Initiative (SMGI). After observing many social media and technology platforms building out governance systems, we noticed that they were replicating many of the same outcomes (successes and missteps) in our offline criminal legal system. Given that the community of Justice Collaboratory scholars has been working to understand and bring evidence-backed transformation to the offline world, we saw an opportunity to apply these learnings to the relatively new, emerging space of online governance.

The need for effective and responsible online governance is considerable. To date, the industry-wide standard for content moderation has been the embrace of a single, uniform model in which platforms hire thousands, even tens of thousands, of contractors to review endless queues of text, photos, and videos. This material may either be reported by users or, increasingly, algorithmically identified, which the platform then decides whether those individual pieces of content violate an ever-changing set of rules created by the platforms.[1] Over the past few years, the financial investment in these complex moderation apparatuses has ballooned, as has the number of content moderators performing these jobs, with TikTok recently reporting having an incredible "40,000 safety professionals dedicated to keeping TikTok safe."[2] Despite such significant personnel and financial investments, it is not clear that the return on these investments is translating to an increased sense of safety across platforms or a mitigation of harms for their respective communities. Pew reports across three surveys in 2014, 2017, and 2021 that more severe forms of harassment, such as physical threats and stalking, have risen over the years (Vogels, 2021).

At the Justice Collaboratory, we have produced research that engages directly with these existing content moderation apparatuses, and we are actively committed to supporting improvements (Bradford et al., 2019; Tyler et al., 2021; Katsaros et al., 2022a). However, we note that our vision from the start is to look beyond an approach focused exclusively on removing harmful posts after they have occurred. Instead, we believe it is necessary to intervene further "upstream" (Katsaros 2022b), where online behavior and user interaction start to form, with the ultimate goal of promoting community vitality (Badiei et al., 2020; Kim et al., 2022).

To this end, in March of 2023, with the help of generous grants from the Stavros Niarchos Foundation and the Oscar M. Ruebhausen Fund at Yale Law School, we brought together eighty individuals approximately evenly split between research scholars and industry practitioners for two days of presentations and discussions. The scholars who joined us spanned many disciplines, including legal, HCI, sociologists, and psychologists. On the industry side, we were joined by a variety of data scientists, product managers, policy managers, and UX researchers from organizations including Jigsaw, Meta, Niantic Labs, Reddit, Roblox, Snap, Spectrum Labs, Spotify, TikTok, Tinder, Twitch, and Wikimedia. As an organizing theme, we called this convening "Beyond Moderation" to acknowledge a growing recognition in the limitations of current moderation approaches and the possibility of imagining a future of alternative approaches to building a vital and healthy online community.

[1] There are many scholarly writings describing various aspects of this content moderation apparatus. One of the earliest writings on this topic that has aged very well is Sarah T. Robert's "Behind the Screen", which was recently updated in a version published in 2021.

[2] https://www.tiktok.com/transparency/en-us/content-moderation/

To begin the convening, we asked attendees to write down a thought, question, idea, or provocation that they were bringing to the two days. What emerged from these notecards were a few consistent themes:

> Democratic and Community Governance — Who gets to define what problems, issues, and harms to even address? And who gets to decide how and who implements solutions to address these? How can individuals and communities be put at the center of these discussions to ensure that governance structures online are actually serving the communities for which they are designed?
>
> Economic Incentives — How do we better understand the relationship between a platform's economic incentives and the behaviors emerging on the platform? How do we align economic incentives with community vitality to promote more prosocial engagement on platforms?
>
> Organizational Design —What is the organizing structure of teams within these technology companies? How do these organizational structures promote (or not) safety on these platforms? What is the role of regulation in prescribing particular organizational structures?
>
> Academic-Industry Bridging — How do we, as research scholars, improve upon both our methods and our impact on platforms? To whom should we be talking, and whom should we be involving in our research? How can we better translate our research into practice? Conversely, how can we, as practitioners, better learn from and implement the knowledge being generated from within the academy?
>
> Comparative Perspectives —What can we learn by looking outside our everyday work? To which other structures and systems, offline or online, can we look for inspiration, evidence, and guidance in our work?

Over the two days, we engaged in a number of discussions on topics that felt familiar to those of us working in this online governance space. One theme that stood out across these conversations was a recognition that we are in the midst of a number of different shifts that are clearly impacting our work and, in so doing, creating new challenges while also opening up opportunities to reimagine and rethink approaches that previously felt hardened.

SHIFTING LANDSCAPE IN TRUST & SAFETY

Addressing harms online is as old as the internet itself. While even the earliest technology companies had individuals or teams devoted to mitigating spam and addressing online harms, the way that work was done was very inconsistent. Sometimes these teams were mostly security engineers focused exclusively on spam fighting, while at other times, they consisted of individuals reporting into customer service departments who would occasionally work in a queue reviewing graphic photos. In recent years, as platforms have invested heavily in these teams, there has been a maturing and professionalization of the individuals doing this work across the industry. One of the clearest and most notable examples of this professionalization is workers coalescing to form the Trust & Safety Professional Association (TSPA). The TSPA and groups like it offer those within the industry a more consistent opportunity and space to discuss their work, share best practices, and learn from one another. The TSPA has begun developing a curriculum and organizing virtual events and hosted its first in-person conference—TrustCon—in 2022. While this maturation presents an opportunity to harden and solidify existing dominant ideas, we see a greater opportunity to showcase the diversity of approaches across platforms both large and small and to enable people to learn more quickly from one another about new ideas and research.

This significant financial investment from major platforms in online safety is part of a growing economy of online trust and safety. The economy of content moderation that Roberts (2019) described so cogently in the first edition of her book "Behind the Screen" almost ten years ago is now prologue for a burgeoning industry of startups providing trust and safety services and software to platforms that has received significant capital investments. Companies such as Active Fence, Cinder, Sift, and Spectrum Labs provide platforms with services ranging from classifier development and AI models to graphic content detection and moderation and surveillance tools. Y Combinator, a technology startup accelerator, recently hosted a "demo day" that featured the largest number of cybersecurity, privacy, and trust companies in the history of Y Combinator.[3] The four companies listed above alone have raised over $200M in venture capital funding since 2020 (having raised a lifetime total of $315.6M). A large sum, from which investors will undoubtedly be expecting significant returns, raising questions about the financial incentives for these companies. What we know about these firms is limited to their marketing websites; for example, the mission of Active Fence includes "enabling a better, safer world by preventing online evil at scale."[4] While these services can be particularly important for smaller platforms that are unable to build and maintain sophisticated AI models for proactive detection, there are many unanswered questions about their motivations, approaches, economic incentives, and regulatory requirements.

Our conference discussion addressed one of the most apparent shifts underway, namely, regulatory changes that arise in the wake of the passing and implementation of the EU's Digital Services Act (DSA). These discussions also included the UK's Online Safety Bill, sundry attempts by US states to police or regulate online activity, and even firm-specific campaigns such as fierce efforts at both the US state and federal levels to ban TikTok. The industry is at a tipping point regarding online safety regulation, and we should expect to see more such initiatives. Whether efforts such as the DSA will contribute to building safety or simply stifle innovation is an open question. One does wonder whether such regulations will essentially prescribe what a platform views as the bare minimum to address safety, investing no more than required by law. While there are currently product oriented teams filled with engineers building novel AI models, data scientists identifying new opportunities, and designers building creative UI solutions to online safety, these regulations may shift a firm to have these teams look more like compliance teams seeking to check the required boxes.

Of course, at the top of mind for many was the economic environment and the reverberations across the technology industry. Lagging consumer growth, declining stock prices and a wave of layoffs resulting from bullish growth over the previous two years has not been matched with the kind of leadership and oversight that one would expect in a maturing industry. Some of our presenters had themselves been hit by layoffs–or, in the case of Meta, were waiting pending a round of preannounced layoffs. What this means for teams working at these firms was clear—Trust and Safety teams that are already underresourced are now being required to work with smaller budgets and personnel. All of this forces teams to think much more creatively about their problems. Where previously some systems and assumptions were hardened, we are in a new regime where prior assumptions can be more readily challenged. Teams are looking for ways to work much further upstream to address their problems, an approach for which we at the Justice Collaboratory have been advocating since we began the SMGI.

While we did devote an entire panel to this idea, one concept that presented itself multiple times over the two days was focusing on the design and architecture of platforms to promote community vitality. In the past, much of the focus was on individual pieces of content, what platform rules are (or are not), and the complex technical and operational processes through which those rules are enforced. Instead, discussions centered on the design and architecture of the platforms as a way to address harmful behavior and promote more prosocial interactions. In one panel, Design and Architecture of Healthy Online Spaces, we heard three presentations on this topic. Julia Kamin presented a framework for digital prosocial interventions organized by the Prosocial Design Network (Grüning et al., 2023), while Delia Mocanu talked about

[3] https://securityboulevard.com/2022/04/y-combinators-winter-2022-cybersecurity-privacy-and-trust-startups/

[4] https://www.activefence.com/about/

her data science work at Facebook and Twitter's Birdwatch (Wojcik et al., 2022), exploring this idea, followed by a presentation from Jen Weedon discussing how to build safety by design within an organization. In the discussion following these presentations, there was a healthy debate trying to actually define what prosocial behaviors even are. This brought to light the imbalance that exists with almost all of the attention focusing on antisocial behaviors, where we have many clearly defined harms (and many others not so clearly defined) but comparatively little focus on defining what sorts of positive behaviors platforms seek to promote. In another panel, Radical Futures for Social Media, we heard three incredible presentations from young scholars. Jane Im provided insight into how affirmative consent can be used as a framework to change the design of privacy and safety settings (Im et al., 2021),[5] while Ishita Chordia demonstrated the way that some platforms leverage deceptive design patterns to elicit and profit off of fear (Chordia et al., 2023), followed by Shamika Klassen presenting some of her design fiction work on *The Stoop,* a speculative social media platform created by a Black woman predominantly designed and developed by a Black team (Klassen & Fiesler, 2023). Each of these presentations engaged directly with the design, structure, and architecture of a platform in helping to shape the behaviors of users toward the promotion of safety and community vitality.

EMERGING RESEARCH IN ONLINE GOVERNANCE

As discussed above, one of the goals of the Beyond Moderation convening was to reimagine approaches to promoting community vitality online. To this end, we thought that it was critical to bring in fresh voices to the conversation as both presenters and attendees. One of our panels aimed at giving space to hear from three emerging scholars about the research they are doing around online governance. These three PhD students were gracious enough to share a written version of their presentations, which the Justice Collaboratory is very pleased to publish here.

In the first essay, "The Teenaged Adults in the Room: Understanding And Supporting Young Online Community Moderators," Jina Yoon draws our attention to the role that teenagers and young adults play in shaping our social internet. Building on an existing body of research exploring the way that communities can play a more central role in their own governance online, through a series of interviews, Yoon examines how teenagers are participating as volunteer moderators online. As expected, many social online spaces geared toward young adults also include volunteer moderators who are themselves young adults. Perhaps more unexpectedly, Yoon finds that many of the teenagers she interviews are also acting as moderators of communities that are made up of young and old adults alike. What we most appreciate about this work is the way in which Yoon connects these experiences described in the interviews toward a broader idea of promoting and supporting community vitality. She describes the way in which teenagers engaging as volunteer moderators online can be a way for these individuals to learn and grow. These young adults are obtaining hands-on experience with conflict resolution and community building and are leveraging the experience for work opportunities. She provides clear recommendations to support these individuals to thrive and grow through volunteer moderation, not simply act as an agent to remove unwanted content from a platform. In doing so, she reframes the dominant narrative that is limited to protecting teens from harm online towards a narrative that seeks to recognize and enhance their agency.

Our second emerging scholar featured, Monika Yadav, shares results of an amazing field experiment conducted with a colleague in her essay "Learning to Resist Misinformation: A Field Experiment." In this study, Yadav and her colleague conduct a rigorous, longitudinal field experiment in India providing participants weekly digests of factual information related to viral misinformation spread across the country over the past week. Here, we have a clear example of an approach that moves beyond the moderation status quo. Instead

[5] https://consentful.systems/

of seeking to simply remove or downrank misinformation, this approach aims to more broadly uplift individuals by providing education and support in an attempt to build individuals' resiliency to false and misleading narratives. Rather than being corrective after the proliferation and spread of misinformation, proactive approaches like the one presented here work further upstream to provide individuals with resources to resist false narratives on their own.

Our third emerging scholar, Adina Gitomer, and her coauthors explore activism online in her essay "Stop scrolling!: Youth activism and political remix on TikTok." Gitomer conducts timely research focusing on the specific design and structure of TikTok that is leveraged for political activism on the platform. Through a beautifully mixed methods approach, Gitomer and her colleagues look for the ways that younger and older users of TikTok are using the platform to create and spread political messages. One of the themes that courses through our work at the SMGI is trying to draw attention to the way that the design and architecture of platforms play a critical role in shaping the behaviors that unfold on these platforms (Kim, 2022). In her essay, Gitomer and colleagues directly engage with this idea. They show that TikTok's design and algorithm lend themselves particularly well to participatory and collective political action on the platform, especially among younger users.

We want to extend a thank you to all of those who helped make our Beyond Moderation convening possible. First and foremost, we highlight the generosity of the Stavros Niarchos Foundation and the Oscar M. Ruebhausen Fund at Yale Law School, which, importantly, helped support the attendance of our speakers in addition to that of many young scholars and PhD students who otherwise would not have been able to attend. Next, the convening would not have been possible without the hard work and support of the Justice Collaboratory's Program Coordinator, Molly Aunger. We would also like to thank some of the individuals who helped to shape and organize the programming of the convening: Lindsay Blackwell, Ravi Iyer, Caroline Nobo, Sarita Schoenebeck, and Joseph Seering. And of course, we extend our thanks to all of our speakers and presenters for their generosity in sharing their work and perspectives with us. Last, we thank the three authors of the essays in this publication, Adina Gitomer, Monika Yadav, and Jina Yoon, for presenting their work at our convening and allowing the Justice Collaboratory to publish this innovative work.

REFERENCES

Badiei, F., Meares, T., & Tyler, T. (2020). Community Vitality as a Theory of Governance for Online Interaction. *Yale JL & Tech., 23*, 15.

Bradford, B., Grisel, F., Meares, T. L., Owens, E., Pineda, B. L., Shapiro, J., ... & Peterman, D. E. (2019). Report of the Facebook data transparency advisory group. *Yale Justice Collaboratory*.

Chordia, I., Tran, L. P., Tayebi, T. J., Parrish, E., Erete, S., Yip, J., & Hiniker, A. (2023, April). Deceptive Design Patterns in Safety Technologies: A Case Study of the Citizen App. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-18).

Grüning, D. J., Kamin, J., Panizza, F., Katsaros, M., Lorenz-Spreen, P., Network, P. D., & Grüning, D. J. (2023). A framework of digital interventions for online prosocial behavior.

Im, J., Dimond, J., Berton, M., Lee, U., Mustelier, K., Ackerman, M. S., & Gilbert, E. (2021, May). Yes: Affirmative consent as a theoretical framework for understanding and imagining social platforms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-18).

Katsaros, M., Tyler, T., Kim, J., & Meares, T. (2022a). Procedural Justice and Self Governance on Twitter: Unpacking the Experience of Rule Breaking on Twitter. *Journal of Online Trust and Safety, 1*(3).

Katsaros, M., Yang, K., & Fratamico, L. (2022b, May). Reconsidering tweets: Intervening during tweet creation decreases offensive content. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 16, pp. 477-487).

Kim, J., McDonald, C., Meosky, P., Katsaros, M., & Tyler, T. (2022). Promoting Online Civility Through Platform Architecture. *Journal of Online Trust and Safety, 1*(4).

Klassen, S., & Fiesler, C. (2023). The Stoop: Speculation on Positive Futures of Black Digital Spaces. *Proceedings of the ACM on Human-Computer Interaction, 7*(GROUP), 1-24.

Roberts, S. T. (2019). Behind the screen. *Yale University Press*.

Tyler, T., Katsaros, M., Meares, T., & Venkatesh, S. (2021). Social media governance: can social media companies motivate voluntary rule following behavior among their users?. *Journal of experimental criminology, 17,* 109-127.

Vogels, E. A. (2021). The state of online harassment. *Pew Research Center, 13*, 625.

Wojcik, S., Hilgard, S., Judd, N., Mocanu, D., Ragain, S., Hunzaker, M. B., ... & Baxter, J. (2022). Birdwatch: Crowd Wisdom and Bridging Algorithms can Inform Understanding and Reduce the Spread of Misinformation. *arXiv preprint arXiv:2210.15723*.

The Teenaged Adults in the Room: Understanding and Supporting Young Online Community Moderators

Jina Yoon

Thousands of teens are already moderating their own online communities on platforms such as Discord, Reddit, and Twitch, helping to build healthy spaces on the internet and developing important interpersonal skills in the process. We interviewed 13 of these teen online community moderators and several of their mentors regarding their reasons for becoming moderators, how they learned to moderate, and the ways that they handle some of the hardest trust and safety challenges that we know of today. This essay highlights the contributions made by these teens and showcases the ways that online community moderation can benefit their personal growth. We also provide recommendations for how policy-makers, platforms, and parents or guardians might better support teen mods moving forward.

INTRODUCTION

Alex is a charismatic leader who owns an educational nonprofit with over 50,000 members from all over the US. She spends hours every day managing a staff of dozens, scaling up technical infrastructure, and monitoring town hall discussions. She is also only 15 years old. The organization that she leads is actually a Discord server registered as a 501(c)(3) nonprofit where high school students can earn volunteer hours for remotely tutoring peers in subjects such as algebra, French, and chemistry. Teens like Alex[1] are the backbone of thousands of communities that thrive on platforms including Discord, Reddit, and Twitch, where users of all ages gather around common interests online. Although the actual numbers of communities moderated by youth are not yet known, more than 20% of Reddit users are teenagers[2], and almost 25% of Discord users are under the age of 24[3]. Teen online community moderators, or "teen mods" for short, hold significant influence over these online spaces and their systems of governance, cultural norms, and content moderation practices. They are, essentially, operating as digital leaders and role models (Seering et al., 2017) for the next generation—and we need to understand these young moderators now more than ever.

Public interest in the effects of social media on adolescent development has increased significantly in the last few years, as 97% of teens in the US today report almost constantly being on the internet (Vogels et al., 2022). Recently, the US Surgeon General released an official advisory concerning the impacts of social media on youth mental health[4]. The state of Montana just passed legislation to ban TikTok[5], while Utah has set a strict 10 PM curfew for Instagram and TikTok to "combat poor mental health outcomes".[6] These types of actions, which are intended to protect youth, often end up being paternalistic at best and harmful at

[1] Alex is a persona based on one of our 15 semistructured, 60 minute interviews with young online community moderators of large Discord servers. The participants were aged 13-17 and had to moderate at least one server with over 1,000 members in order to be eligible. Most interviewees were from the US, UK or MENA regions. Interviews were held via Discord voice or text. Participants received $15 as compensation for participating. We recruited interviewees from the Moderator Mentorship Community, which is an official server for young moderators officially run by Discord, and similar moderation-related discussion servers.

[2] https://www.statista.com/statistics/1125159/reddit-us-app-users-age/

[3] https://www.statista.com/statistics/1327674/discord-user-age-worldwide/

[4] https://www.hhs.gov/sites/default/files/sg-youth-mental-health-social-media-advisory.pdf

[5] https://www.npr.org/2023/05/18/1176805559/montana-tiktok-ban

[6] https://fortune.com/2023/03/23/utah-nighttime-tiktok-instagram-curfew-teens/

worst. For example, the California Age-Appropriate Design Code Act (ADCA)'s age-assurance requirements are designed to minimize the corporate collection of children's data[7]. To comply, platforms will need to estimate the ages of all users prior to granting access, which critics say counteract the efforts exerted to protect children (Goldman, 2023; Wang et al., 2022). Decision-makers therefore must understand the nuances of these technologies and the ways that teens are actually using them to design effective regulations. Like any other tool, social media can be used for positive or negative purposes; certain features can actually improve psychological well-being, but these effects all depend on the pertinent details (Burke & Kraut, 2016; Blackwell et al., 2017). This work uncovers details regarding how one unique population of teens is connecting online in meaningful ways to jointly build, grow, and lead thriving communities.

WHY DO TEENS BECOME ONLINE COMMUNITY MODERATORS?

Being an online community moderator involves tasks such as growing the community, resolving conflicts between members, removing content, and developing automated tools (Seering et al., 2019). Although this can seem like a chore, it can be extremely rewarding and fun as a teenager.

> *"I think the appeal for it as a young person is you have responsibility and what you do matters. [If I make a change on a Discord server], I can feel that impact on how the community operates and the vibe and the personality that it has."* —MENTOR #1

This testimony is consistent with research that suggests that youth enjoy and greatly benefit from peer-based, networked learning (Ito et al., 2010), especially when the community centers around something that they love. Most interviewees were moderators of servers related to gaming, given Discord's roots in the industry[8], while others participated in communities focused on topics such as self-improvement, software development, music production, and digital illustration.

Another reason that these youths may enjoy moderating is that teens, like other moderators, are often highly motivated by a strong sense of justice, fairness, and altruism (Seering et al., 2019).

> *"It's a job that I enjoy because [it] helps others. It ensures that online communities are safe, that everybody's being treated respectfully and that there's no cyberbullying or online harassment of any kind."* —TEEN MOD #3

A few homeschooled interviewees also expressed that platforms such as Discord let them make friends with people of different perspectives and participate in groups that they would not encounter locally. Similarly, one international participant said they enjoyed the global cultural exchanges they made online.

> *"I don't live in, like, a city or a town so I can't really socialize as much … I just don't really have many opportunities to see people in real life (IRL)."* —TEEN MOD #11

HOW DO TEENS LEARN TO MODERATE ONLINE COMMUNITIES?

Many teen respondents emphasized the importance of learning from mentors and peers through "meta"-moderation servers where they discuss moderator philosophies, tips, and tools.

> *"I didn't want to learn from just looking at other moderators. I wanted to learn from talking and discussing with other moderators. I got what I came for, and I certainly believe I'm a lot better at moderating than I was before."* —TEEN MOD #6

[7] **https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202120220AB2273**

[8] **https://discord.com/company**

Teen mods also learn on the job through reading documentation and observing other mods. Some, however, were particularly inspired by negative examples of leadership. Teen mods may be more motivated than adults to challenge traditional ways of leadership and to seek new styles of governance.

> *"What led me to start pondering moderation and trying to educate myself regarding it was that I would see how moderators in other communities acted. Sometimes I as a user wouldn't like the way that they moderated. Perhaps they were too biased or they were enforcing the rules while also actively breaking them themselves because of their position of power. I strive to one day become a moderator that moderates in the way that I as a user would like to see moderators behave."* —TEEN MOD #4

In particular, many teen mods practice methods of procedural or restorative justice, although interviewees did not name them as such.

> *"If you're jumping in [and picking sides], it can lead to moderation disagreements and bad relationships [between members and moderators]. It can really divide a moderation team if you do. My tactic is to resolve as much as you can in DMs, scaling down the conflict from public chat … trying to find out how you can resolve these conflicts, making sure that they don't continue … I also talk to [the members involved in the conflict] after to make sure that they're feeling welcome in the server, and that it's not because of the server that they're acting out."* —TEEN MOD #3

Online community moderation is an inherently social job shaped by the participation of other people, and it requires significant emotional labor (Dosono & Semaan, 2019). Through these interactions, however, teen mods are learning advanced conflict resolution techniques and strong socioemotional skills. One example is Teen Mod #12, who remarked, "I have a lot more empathy for others than I used to".

Teen mods react to rule-breaking by following restorative and procedural justice methods.

HOW CAN BEING AN ONLINE COMMUNITY MODERATOR BENEFIT TEENS?

In addition to strengthening teens' interpersonal skills, moderating naturally increases social capital (Ito et al., 2010), which opens doors to future opportunities. Several interviewees were moderators of official servers maintained by popular streamers, musicians, and game studios, and a few had even been paid for temporary jobs. One teen who moderates for a partnered esports organization reported that some of the mods on their team had been flown out to in-person events, and a few had eventually been hired by the company as full-time employees.

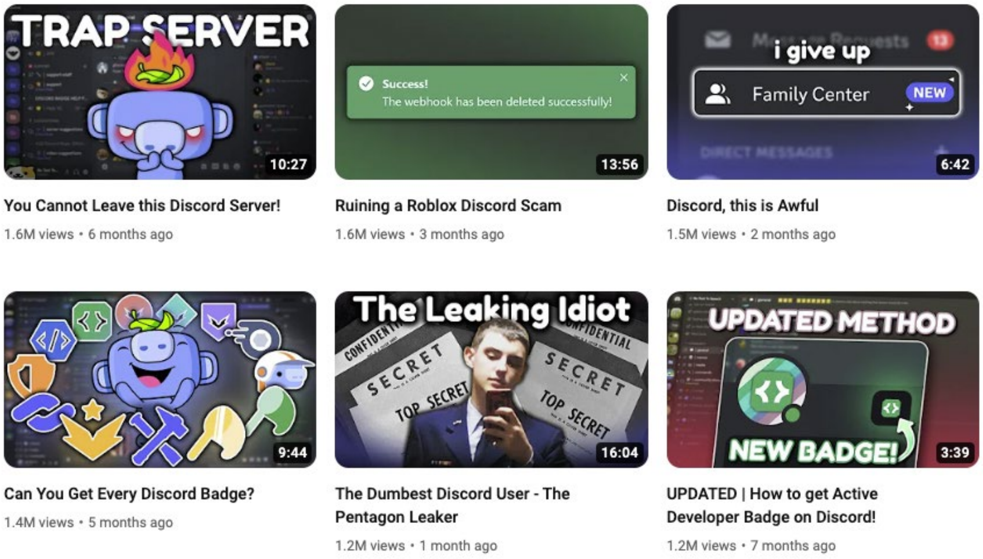
Particularly enthusiastic teens are already taking the initiative to pursue online community moderation as a long-term career. Teen Mod #12, who is sixteen years old, maintains a public resume on their profile that lists their extensive mod experience and consulting gigs. Teen Mod #2 likened the nonprofit educational Discord community that they served in as work experience:

> *"Today, I just got a new position. So I'm still gaining experience, learning new things from it, and it really is a very professional environment, at least the staff team is. It really gives us students a feel of what it's like to work at a big company and get these positions, network, all of that."* —TEEN MOD #2

HOW DO TEEN ONLINE COMMUNITY MODERATORS HANDLE CHALLENGING TRUST AND SAFETY INCIDENTS?

Being an online community moderator is not without challenges (Lapidot-Lefler & Barak, 2012). Teen mods must develop creative ways of handling minor incidents, such as customizing bots that automatically derank or "level up" user roles based on participation[9,10]. In more serious cases, however, such as threats of physical harm, teens often look to authority and official platform guidance (Freed et al., 2023) for help, and these resources are not always sufficient. According to interviewees, responses from official staff can take one to two business days at best, and report systems are also designed to react to incidents that have already occurred rather than proactively preventing future ones. To compensate for what platforms lack, teen mods have devised ways of learning, processing, and preventing through developing trust and discussing safety issues with each other.

When official information about issues such as account security or phishing scams is not available, teen mods may turn to third-party content for news. Multiple interviewees mentioned No Text to Speech, an independent YouTube channel that uploads videos about the latest online safety news that accrue millions of views on platforms such as Discord, Roblox, and Minecraft.[11]



You Cannot Leave this Discord Server!
1.6M views • 6 months ago

Ruining a Roblox Discord Scam
1.6M views • 3 months ago

Discord, this is Awful
1.5M views • 2 months ago

Can You Get Every Discord Badge?
1.4M views • 5 months ago

The Dumbest Discord User - The Pentagon Leaker
1.2M views • 1 month ago

UPDATED | How to get Active Developer Badge on Discord!
1.2M views • 7 months ago

Figure 1: The six most popular videos uploaded by No Text to Speech, a YouTube channel that focuses on Discord safety news.

When asked about child safety incidents such as sexual abuse or predatory behavior, the interviewees responded that they felt as if these issue do not happen as often as people think (Boyd, 2014) and that they usually just report and ban suspicious users. They also often consult friends and mentors about the situation afterward, making sense of what happened through a socially informed process (Hassoun et al., 2023). This act spreads awareness and inoculates them from future threats as they learn to recognize similar threats over time (Roozenbeek et al., 2022).

> *"I'm not saying you should experience it, but if you talk to someone else who's experienced it, that's pretty much all you'll need to know … I've seen a lot of people, if they do encounter these things on*

[9] https://mee6.xyz/en

[10] https://probot.io/

[11] https://www.youtube.com/@NoTextToSpeech/

> *Discord, they just kind of contact a friend, like, is this something I should do? But if you don't really have [that support], then you're more likely to be scammed because of that. So, it's kind of more of a thing with friends keeping friends safe."* —TEEN MOD #11

Reporting users or content is not always the first line of action, however, especially in times of crisis. Mentors explained that having a trusted adult or experienced mentor to talk to was the most important factor during these types of emergencies because they require immediate human support, emotional resilience, and the ability to distance oneself if needed. One interviewee recounted such a situation with one of their community members (warning—the following quotation makes a reference to suicide):

> *"[There was] a guy around my age, he'd always message me about wanting to kill himself and I would help him a lot. But it got to a point where… You know, I have my own life and problems as well. So when it became like a crisis situation, I didn't really know what to do. I don't know him in real life. I can't do anything except report it to Discord. But I was told from a friend that if you report it to Discord, they would just terminate their account. And if Discord is the only place that person can feel safe, then yeah that would be bad so I didn't do it."* —TEEN MOD #12

Teen mods have created many tools, processes, and networks to support each other online, but these workarounds also reveal gaps in official resources. Platforms and policy-makers must find ways to relieve some of these trust and safety burdens from teen mods without undermining their autonomy (Wei et al., 2023).

HOW MIGHT WE BETTER SUPPORT TEEN ONLINE COMMUNITY MODERATORS?

The best way to learn how to support teen mods is by actually talking to them.

First and foremost, we recommend the creation of youth online trust and safety councils, surveys, and other forms of participatory research to ensure that the voices of teens and teen online community moderators are heard. These insights were only made possible by meeting teen mods where they are (Druin, 2002; Yip et al., 2013). Decision-makers must speak directly with the teen mods of their own platforms and communities to address the actual needs of their group.

> **POLICYMAKERS** must devise legislative approaches that assess and distinguish between healthy and unhealthy teen behaviors on social media. Rather than imposing generalized blanket policies for all services, policy-makers need to target the specific features and behaviors that result in negative outcomes to ensure that they are not doing more harm than good for youth.
>
> **PLATFORMS** should offer professional certifications and career programs to recognize and legitimize online community moderation as a form of skilled voluntary labor. Teen mods spend countless hours voluntarily keeping their communities safe, learning technical tools, and exercising interpersonal skills, all of which are invaluable skills in the future workplace. Programs such as the (now discontinued) Discord Moderator Academy Exam[12] and the Reddit Mod Certification 201[13] are examples that, if professionally recognized, could be a huge boost for the future careers of these teen moderators (Cunningham, 2019).

[12] http://discord.com/blog/announcing-the-discord-moderator-academy-exam

[13] http://modeducation.reddithelp.com/reddit-mod-certification-201

PARENTS AND GUARDIANS need to show openness toward social media to earn teens' trust. Having access to a trusted adult is critical for teens (Meltzer et al., 2018), especially teen mods. However, several interviewees reported that their parents had wholly negative attitudes toward social media, compelling them to never disclose or discuss their mod activities. Expressing curiosity or acceptance rather than disapproval or overprotectiveness enables opportunities for candid discussions with teens (Ungar, 2009) about topics such as digital literacy, online safety, and social dynamics.

CONCLUSION

Further research must be conducted with teens to distinguish between the positive and negative behaviors of youth on social media. Teen mods are just one out of many examples of young people making the most of the internet, contributing to and benefiting from their global communities. They are the leaders of tomorrow—both online and offline—and we must see them as equal stakeholders in the content moderation ecosystem if we are to truly make the internet a better place for future generations.

ACKNOWLEDGMENTS

Many thanks to my collaborator Joseph Seering, my advisor Amy X. Zhang, and my friend Inyoung Cheong for their support and mentorship concerning this project. This work was indirectly sponsored by scholarships from the NSF and the ARCS Foundation.

REFERENCES

Blackwell, L., Dimond, J., Schoenebeck, S., & Lampe, C. (2017). Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW), 1-19.

Boyd, D. (2014). *It's complicated: The social lives of networked teens*. Yale University Press.

Burke, M., & Kraut, R. E. (2016). The relationship between Facebook use and well-being depends on communication type and tie strength. *Journal of computer-mediated communication, 21*(4), 265-281.

Cunningham, E. (2019). Professional certifications and occupational licenses. *Monthly Labor Review*, 1-38.

Dosono, B., & Semaan, B. (2019, May). Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1-13).

Druin, A. (2002). The role of children in the design of new technology. *Behaviour and information technology, 21*(1), 1-25.

Freed, D., Bazarova, N. N., Consolvo, S., Han, E. J., Kelley, P. G., Thomas, K., & Cosley, D. (2023, April). Understanding Digital-Safety Experiences of Youth in the US. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).

Goldman, E. (2023). Amicus Brief on the Constitutionality of the California Age-Appropriate Design Code's Age Assurance Requirement (NetChoice v. Bonta). *Available at SSRN 4369900*.

Hassoun, A., Beacock, I., Consolvo, S., Goldberg, B., Kelley, P. G., & Russell, D. M. (2023, April). Practicing Information Sensibility: How Gen Z Engages with Online Information. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-17).

Ito, M., Baumer, S., Bittanti, M., Boyd, D., Cody, R., Herr-Stephenson, B., ... & Tripp, L. (2010). *Hanging out, messing around, and geeking out*. Cambridge, MA: MIT Press.

Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in human behavior*, 28(2), 434-443.

Meltzer, A., Muir, K., & Craig, L. (2018). The role of trusted adults in young people's social and economic lives. *Youth & Society, 50*(5), 575-592.

Roozenbeek, J., Van Der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science advances, 8*(34), eabo6254.

Seering, J., Kraut, R., & Dabbish, L. (2017, February). Shaping pro and anti-social behavior on twitch through moderation and example-setting. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing* (pp. 111-125).

Seering, J., Wang, T., Yoon, J., & Kaufman, G. (2019). Moderator engagement and community development in the age of algorithms. *New Media & Society, 21*(7), 1417-1443.

Ungar, M. (2009). Overprotective parenting: Helping parents provide children the right amount of risk and responsibility. *The American Journal of Family Therapy, 37*(3), 258-271.

Vogels, E. A., Gelles-Watnick, R., & Massarat, N. (2022). Teens, social media and technology 2022.

Wang, G., Zhao, J., Van Kleek, M., & Shadbolt, N. (2022). 'Don't make assumptions about me!': Understanding Children's Perception of Datafication Online. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW2), 1-24.

Wei, M., Consolvo, S., Kelley, P. G., Kohno, T., Roesner, F., & Thomas, K. (2023, April). "There's so much responsibility on users right now:" Expert Advice for Staying Safer From Hate and Harassment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-17).

Yip, J. C., Foss, E., Bonsignore, E., Guha, M. L., Norooz, L., Rhodes, E., ... & Druin, A. (2013, June). Children initiating and leading cooperative inquiry sessions. In *Proceedings of the 12th International Conference on Interaction Design and Children* (pp. 293-296).

Learning to Resist Misinformation: A Field Experiment

Monika Yadav
Naman Garg

A well-informed electorate is vital to the functioning of democracies. However, in recent years, the information environment of voters has been drastically transformed due to the structural changes in news media markets caused by the internet. While the internet has played a vital role in expanding public access to information, it has also resulted in a decline in the quality of available news (Chen and Suen, 2022; Acemoglu et al., 2021). For instance, social media, today, is a major source of news for many people, but it also exposes them to a substantial amount of false and misleading information (Allcott and Gentzkow, 2017). This has led to growing concerns about the difficulty of discerning the veracity of online information.

A surge in the levels of exposure to misinformation combined with people's inability to spot false stories drives misperceptions about various political issues, subsequently influencing election outcomes, intensifying prejudice against minorities, fueling hate crimes (Müller and Schwarz, 2020, 2021) and ethnic violence, and shaping the global response to the COVID-19 pandemic. In light of these consequences, scholars across disciplines have argued that misinformation poses a significant challenge for human cognition and social interaction, calling for a broader work to identify and evaluate measures that can potentially help counter misinformation and mitigate its impact on attitudes and behavior.

For this paper, we conduct a large preregistered field experiment (N = 1301) in India to test an intervention aimed at improving people's ability to discern the veracity of the information encountered online and reducing their misperceptions about minorities. India is categorized as a 'Tier Zero' country by Facebook, making it the company's top priority group along with the US and Brazil in regard to the circulation of false or misleading content. Online misinformation has significantly impacted social and political discourse in the country, resulting in an increase in affective polarization. India also serves as a compelling case study illustrating the alarming consequences that can result when misinformed beliefs are allowed to shape policy opinions and even influence the very content of legislation.

Our intervention consisted of delivering weekly digests containing a compilation of fact checks of viral misinformation.[1] Since most misinformation revolves around certain issues and follows predictable patterns and manipulation techniques,[2] we reason that familiarity with these fact checks can help people internalize the heuristics needed to evaluate similar content that they might encounter in the future. Alternatively, we can characterize this dynamic using a machine learning analogy: by providing people with digests, we offer them a training dataset that helps them acquire the skills necessary to correctly classify content. Thus, the main objective of the intervention was to guide the participants through the landscape of online misinformation, including the issues that attract false news stories, the context of these issues, their typology, and the patterns of various types of viral misinformation, to better equip them to judge the veracity of content on their own.

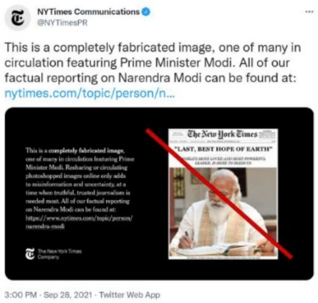
[1] These fact-checks were conducted by fact-checkers certified by the International Fact-Checking Network (IFCN). This certification is also the criteria used by Facebook for its third-party fact-checking program (https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking).

[2] Banaji et al. (2019) analyze a large amount of such misinformation that circulates on WhatsApp in India and categorize it into a concise typology.

Repeating Patterns

Issues and techniques of manipulation

Fake image of *New York Times* cover



NYTimes Communications
@NYTimesPR

This is a completely fabricated image, one of many in circulation featuring Prime Minister Modi. All of our factual reporting on Narendra Modi can be found at: nytimes.com/topic/person/n...

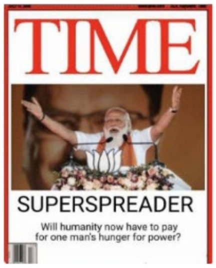
This is a completely fabricated image, one of many in circulation featuring Prime Minister Modi. Resharing or circulating photoshopped images online only adds to misinformation and uncertainty, at a time when truthful, trusted journalism is needed most. All of our factual reporting on Narendra Modi can be found at: https://www.nytimes.com/topic/person/narendra-modi

The New York Times Company

3:00 PM · Sep 28, 2021 · Twitter Web App

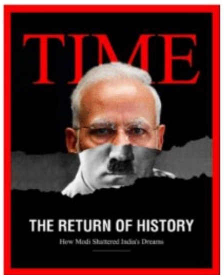
Went viral in September 2021
This was covered in the digest.

Fake *Time* magazine covers



TIME
SUPERSPREADER
Will humanity now have to pay for one man's hunger for power?

Went viral
in April 2021



Went viral
in June 2022

The issue of Modi's standing as a world leader continues to see a lot of misinformation, even including same manipulation technique of Photoshopping covers of prominent international media outlets.

Figure 1: Shows example of predictable patterns in misinformation in terms of issues and techniques of manipulation. Familiarity with fact checks can help people learn these patterns to become better at identifying misinformation.

In addition to the above fact-check summaries, in two of our digests, we included narrative explainers incorporating the relevant context of issues that are politically salient and serve as consistent targets of false stories. Specifically, we covered misperceptions against Muslims—a religious minority group—that are being increasingly fueled by a surge in exposure to misinformation. In addition to collating the different strands of misinformation regarding these issues, we also provided a detailed background regarding them, including not only numbers and statistics but also anecdotes and narratives of individuals who were impacted by these issues and the laws pertaining to them, listed instances of cases when mainstream media reporting on these issues was later discovered to be inaccurate or outright deceptive, and findings from investigations conducted both by law enforcement agencies and by independent media organizations.

We recruited participants using Facebook ads. The experiment lasted for ten weeks, from mid-August to October 2021, during which the treated individuals received nine digests through a custom-built mobile app. Participants completed a brief screening survey during the recruitment process, followed by four surveys in total, namely, a baseline survey, two subsequent surveys, and a final survey, that were each spaced approximately three weeks apart. This longitudinal experiment design enabled us to observe the changes in effects over time.

We focused on two sets of outcomes. First, we asked questions designed to measure the respondents' ability to accurately assess the veracity of statements related to ongoing events in the sociopolitical discourse. To this end, we showed some statements to our study participants and asked them whether they were familiar with the statement, i.e., whether they had seen or heard about it somewhere; whether they thought that the statement was true; and what their confidence level in their assessment of the statement's veracity was. In each of the follow-up surveys, eight headlines were presented that were varied across two dimensions: accuracy (mainstream/true versus false) and political valence (right versus left). All the true headlines were published by mainstream news sources within one month before the respective survey. All the false headlines circulated on social media within one month before the respective survey and were factchecked and labeled as false by at least one third-party fact-checking website. Additionally, it is important to note that we did not use any of the false news headlines that we had covered in the weekly digests in our follow-up surveys; hence, the estimated impacts of intervention were focused on individuals' ability to assess the veracity of information in general rather than on their ability to recall the information that had been provided in digests.

Second, we measured the change in misperceptions against Muslims, including factual beliefs and knowledge about minorities and any consequent changes in policy preferences or behavior.

Our findings show that the intervention increased the ability to detect misinformation by eleven percentage points. There was also a minor decline in the belief in true news, which fell by approximately four percentage points. By estimating a structural model, we find that the impact mechanism was driven by an increase in both truth discernment (the intervention increases people's ability to discern true versus false information as they become more familiar with the patterns of misinformation on social media) and skepticism (the intervention might lead people to update their prior assessments about the prevalence of misinformation on social media and become less credulous overall). The trends of these effects over time indicate that the intervention quickly boosts skepticism; however, it requires a considerable amount of time for people to improve their ability to discern the accuracy of statements.

Intervention increasese both skepticism and truch discernment

While skepticism increases immediately, it takes more time to become better at discerning the truth.

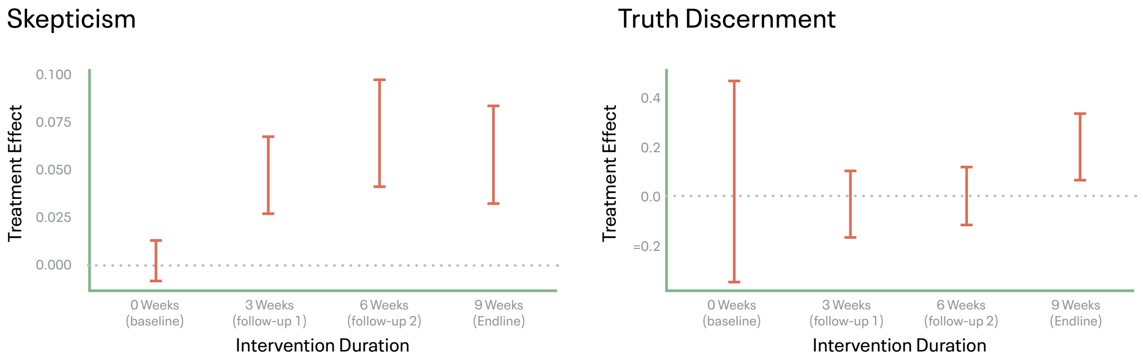


Figure 2: Shows structural estimates across survey rounds for the treatment effects on the parameters of skepticism and truth discernment.

The intervention also improved people's factual beliefs about minorities and led to changes in their policy preferences and behavior. Treated individuals are four percentage points less likely to support discriminatory policies and laws. They are also, on average, willing to pay more for initiatives aimed at curbing the harassment of minority communities.

We contribute to two main strands of research on misinformation. First, our study adds to the expanding body of research aimed at evaluating the effectiveness of interventions meant to mitigate the impact of misinformation. Nyhan (2020) classified these interventions based on their timing relative to the exposure to misinformation. Those interventions performed after exposure, such as providing fact checks of misinformation that individuals had been exposed to; those performed during the exposure, such as the appropriate tagging of false and misleading information; and those performed before exposure to inoculate against misinformation by teaching people to identify false or misleading content.

Our intervention aims to inoculate individuals against the effects of false news prior to their exposure to such misinformation. We show that being acquainted with the recognizable patterns and manipulative tactics used in misinformation can help individuals to develop a cognitive framework for discerning falsehoods. In other words, it allows us to act preemptively and help people build resistance to misinformation relatively broadly (van der Linden et al., 2017). It is also more suited to the combating of misinformation, as encrypted messaging apps have become the primary communication channels worldwide. The unobservability of information flows across these private networks precludes intervening, such as the removal or flagging of unreliable posts, either during or after the exposure.

Furthermore, our intervention is more robust and scalable than other inoculation strategies that, thus far, have focused on providing digital media literacy training around specific methods for the identification of misinformation (Guess et al., 2020; Lewandowsky and van der Linden, 2021; Roozenbeek et al., 2022; Badrinathan, 2021). Most of these existing digital media literacy interventions generally require a great deal of effort on the part of individuals. Moreover, the evidence regarding their effectiveness is mixed.[3] Our intervention does not require the leveraging of a large amount of cognitive resources for it to be effective; rather, the digests enable individuals to gain tacit knowledge of the typology and patterns of false stories, allowing them to develop helpful heuristics for identifying false narratives. It is also easier to scale across contexts, unlike other digital media literacy programs that must be tailored to fit the manipulation patterns and techniques that are prevalent in a particular region.

This study also provides a rigorous evaluation of recent similar initiatives that have been undertaken by various news outlets[4] and fact-checking organizations by distributing newsletters, conducting briefings to summarize fact-checks or dedicating sections of their websites to the debunking of viral online misinformation.[5]

Our study contributes to a second area of research that focuses on measuring the extent of people's misperceptions about various political issues and analyzes the effect of correcting these misperceptions on policy opinions and behavior. More specifically, our research explores the misperceptions about outgroups and minorities (see Bursztyn and Yang (2022) for a meta-analysis). A common observation in current studies is that despite changing people's factual beliefs, corrective information does not always lead to changes in policy attitudes, especially for politically salient and charged topics. For instance, Barrera et al. (2020) conducted a survey experiment in the context of French presidential elections focusing on the misinformation spread by the extreme-right candidate Marine Le Pen. They find that providing corrective information about the unemployment rate and gender ratio of immigrants leads people to update their beliefs about these facts but fails to reverse the effect of the original misinformation on people's resistance toward immigration. During the 2016 US presidential campaign, Nyhan et al. (2020) conducted a series of experiments in which respondents were randomly assigned to view various versions of a journalistic fact-check regarding then candidate Trump. The findings indicate that exposure to this information led to a

[3] Guess et al. (2020) train people in strategies such as checking the original source of information or comparing other news on same issues, which presumably require high levels of conscious effort. They find that even though the intervention was effective in the US, this effect did not last in India. Similarly, Badrinathan (2021) does not find any effect on truth discernment by an intensive hour-long in-person digital literacy intervention in India aimed at training people to do reverse image searching to verify the authenticity of images.

[4] For instance, during the 2020 US elections, the New York Times had a dedicated series of articles aimed at debunking and providing information about viral online misinformation (https://www.nytimes.com/live/2020/2020-election-misinformationdistortions). It also now has a section dedicated to tracking viral misinformation (https://www.nytimes.com/spotlight/disinformation)

[5] Even if major news organizations supply such fact-checking summaries, the level of demand for such products is still an open question. See Chopra et al. (2022) for such a demand estimation. They find that the demand for newsletters increases when the newsletter includes fact-checking.

reduction in the level of misperceptions concerning factual issues, specifically, changes in the prevalence of crime. However, there was no observable effect on the level of support for the candidate.

These findings lead many to posit that information delivered in a more narrative form might be more effective in affecting policy opinions. We test this hypothesis by supplementing corrective information about minority issues with narrative explainers that provide more background and list stories and anecdotes of individuals who have been impacted by misinformation. We find that such an intervention is indeed effective in changing people's policy attitudes and behavior in a setting characterized by high levels of affective polarization.

REFERENCES

Acemoglu, D., Ozdaglar, A., & Siderius, J. (2021). A model of online misinformation (No. w28884). *National Bureau of Economic Research*.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives, 31*(2), 211-236.

Badrinathan, S. (2021). Educative interventions to combat misinformation: Evidence from a field experiment in India. *American Political Science Review, 115*(4), 1325-1341.

Banaji, S., Bhat, R., Agarwal, A., Passanha, N., & Sadhana Pravin, M. (2019). WhatsApp vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India.

Barrera, O., Guriev, S., Henry, E., & Zhuravskaya, E. (2020). Facts, alternative facts, and fact checking in times of post-truth politics. *Journal of public economics, 182*, 104123.

Bursztyn, L., & Yang, D. Y. (2022). Misperceptions about others. *Annual Review of Economics, 14*, 425-452.

Chen, H., & Suen, W. (2019). Competition for attention and news quality. *American Economic Journal: Microeconomics*, forthcoming.

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences, 117*(27), 15536-15545.

Lewandowsky, S., & Van Der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology, 32*(2), 348-384.

Müller, K., & Schwarz, C. (2020). From hashtag to hate crime: Twitter and anti-minority sentiment. A*vailable at SSRN 3149103*.

Müller, K., & Schwarz, C. (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association, 19*(4), 2131-2167.

Nyhan, B. (2020). Facts and myths about misperceptions. Journal of Economic Perspectives, 34(3), 220-236.

Nyhan, B., Porter, E., Reifler, J., & Wood, T. J. (2020). Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior, 42*, 939-960.

Roozenbeek, J., Van Der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science advances, 8*(34), eabo6254.

van der Linden, S., Maibach, E., Cook, J., Leiserowitz, A., and Lewandowsky, S. (2017). Inoculating against misinformation. *Science, 358*(6367):1141–1142

Stop Scrolling!: Youth Activism and Political Remix on TikTok

Adina Gitomer
Julia Atienza-Barthelemy
Brooke Foucault Welles

The social video platform TikTok has recently exploded in popularity (Guinaudeau et al., 2022), leading to a great deal of talk and speculation about it. It has earned a strong reputation for frivolity, "known for serving up short videos on everything from dance moves to fashion tips, cooking tutorials to funny skits."[1] Simultaneously, it is seen as "a prominent venue for ideological formation, political activism and trolling"[2] and has been linked to successful instances of collective action (Bandy & Diakopoulos, 2020). While these perceptions occasionally intersect, as in the case of a makeup tutorial that simultaneously protested China's detention of Uighur Muslims[3], the former is often made out to discredit the latter[4]. As a result, TikTok's political potential is up for debate.

TikTok stands out from other platforms like Twitter, whose political implications have been widely studied, for being video-based, remix-based, and centered around algorithmically-curated content over personal networks. It also stands out for being dominated by youth (Vogels et al., 2022), who are turning to TikTok over other platforms for political engagement (Literat & Kligler-Vilenchik, 2019; Choi et al., 2022). Unsurprisingly then, the debate around TikTok's political strength is reflective of the broader debate around youth political participation. Youth tend to favor noninstitutional forms of political participation that leverage new media technologies and emphasize personal empowerment in political action (Bennett et al., 2011; Cohen & Kahne, 2012). While this nontraditional civic style continues to prompt charges of youth political depravity and disinterest (Flanagan & Levine, 2010; Weiss, 2020), it is increasingly examined, validated, and shown to be effective (Jenkins et al., 2016).

With all of that in mind, there is no shortage of examples of young people using TikTok to meaningfully impact political processes. To name a few: they have rallied against union-busting corporations[5], intervened in the recent assault on abortion rights in America[6], and fomented urgency around the growing threat of climate change (Hautea et al., 2021). It is therefore difficult to dismiss the political force behind TikTok and its young user base. Here, we take a closer look at how this force is built in an effort to better understand its power, and in turn, clarify the murky narratives surrounding TikTok activism, youth political participation, and the intersection of the two. In particular, we characterize activist remix strategies on TikTok, with special attention to those deployed by youth.

DIGITAL ACTION NETWORKS

We turn to the literature on digital action networks to determine the categories under which activist strategies on TikTok might fall. According to Bennett and Segerberg (2012), "two broad organizational patterns characterize … digitally enabled action networks." The first is "collective action," the more conventional pattern in which individuals rally around a shared "we" by way of organizational brokerage.

[1] https://www.nytimes.com/2021/03/20/books/booktok-tiktok-video.html

[2] https://www.nytimes.com/2020/06/28/style/tiktok-teen-politics-gen-z.html

[3] https://www.theguardian.com/technology/2019/nov/27/tiktok-makeup-tutorial-conceals-call-to-action-on-chinas-treatment-of-uighurs

[4] https://www.youtube.com/watch?v=D1J8t97gPW4

[5] https://www.wired.com/story/tiktok-army-union-busters-amazon/

[6] https://www.politico.com/news/magazine/2022/03/27/progressive-gen-z-for-change-tik-tok-00020624

Under collective action, individuals must adhere to a singular prescribed identity in order to be legible as participants in a given movement. The second pattern is "connective action," whereby individuals leverage digital media affordances to organize themselves, without relying on established organizations for coordination. Importantly, the logic of connective action replaces the shared singular identity of collective action with "personal action frames," wherein individuals take part in a larger movement by sharing personal narratives and experiences. Following this theory, we ask: are youth activist strategies on TikTok more collective or more personalized? Before presenting our hypothesis, we establish the central features of TikTok's design that inform it.

TIKTOK: WHAT TO KNOW

TikTok stands out from other platforms for its remix-based tools and their emphasis, which encourage users to creatively reimagine the content of other users. For example, the "duet" feature enables TikTokers to put their own video alongside that of another user in split screen fashion, and the "stitch" feature enables them to combine their own video with that of another user in successive fashion; meanwhile, there is no feature that directly allows users to repost someone else's video without remixing it first. Remix is thus the primary way that content circulates on the app. The wide array of sophisticated yet intuitive video editing tools (e.g., text, stickers, trimming) means these remixes and their original videos are often high quality. TikTok also has a unique focus on audio. The "sounds" that accompany videos are their own discrete entities, and users can either create original sounds or borrow them from other users. One can also search for videos based on their sound. Finally, TikTok's main page—the "For You Page" (FYP)—consists of an endless stream of content recommended by a black-box algorithm that is different for every user and not entirely dependent on who they follow. Because of this, every video has a potential audience (Guinaudeau et al., 2022), and users are incentivized to behave in ways that they believe the algorithm will reward. The prominence of the FYP also orients the app around content rather than around personal networks.

DIGITAL ACTION NETWORKS ▯ TIKTOK

TikTok's design and the theories surrounding it suggest that it would best support personalized activist strategies, and clear examples of connective action have been identified on the app (Becker, 2021; Sadler, 2022). To begin, it is built for customization[7]: the video format gives way to highly embodied communication (Raun, 2012), and remix—the core form of sharing on TikTok—requires the incorporation of personal taste (Church, 2017). Compounding this, scholars have emphasized the affective qualities of TikTok, which personalized expression thrives on. In the context of activism, Hautea et al. (2021) theorize the app as an "affective public," arguing that users' experience of watching others express concern about a given cause mobilizes their own concern and general support. Similarly, Haslem (2022) contends that TikTok creates an "affective loop," wherein users take in the emotions of others and respond with their own. Personalized activist strategies are also associated with young people, who represent TikTok's main user base. This is largely because young people have grown up surrounded by network technologies that facilitate self-organization, and are attributed with more self-expressive modes of political participation (Vromen et al., 2016). For these reasons, we expect to see TikTok users—and younger users in particular—engaging personalized strategies in their in-app activism.

[7] https://www.nytimes.com/2020/06/28/style/tiktok-teen-politics-gen-z.html

DATA

To test this hypothesis, we created a purposive sample of TikTok data by honing in on the American nonprofit organization Gen-Z for Change. Gen-Z for Change is aimed at "leveraging social media to promote civil discourse and political action" and identifies itself as a coalition of over 500 creators and activists. Run by a core team of 18 Gen Z organizers, the group has over 540 million followers and 1.5 billion monthly views on TikTok, which is their primary digital organizing tool. Currently, they use it to engage in a variety of large-scale collective organizing and progressive political education campaigns. For example, they crashed an anti-abortion whistleblower site with false tips and flooded a Starbucks application portal after the company fired workers for union organizing.[8]

Because TikTok did not have an official API at the time of data collection (5/2022), we were severely limited in terms of the data we could collect. Under this constraint, we collected all TikTok videos that included "#genzforchange" in the caption (N = 979) using an unofficial scraper[9]. We then retained only those that contained an original sound (N=464), meaning the author of the video matched the author of the sound. The resulting set represents our original activist videos. Next, using a different scraper[10], we collected all of the TikTok videos that borrowed a sound from one of those original videos (N = 1648). This set represents our activist remixes.

Finally, we filtered the original videos based on two conditions. First, the video had to be relevant—that is, it had to be explicitly aligned with Gen-Z for Change's progressive political agenda. Second, the video had to be remixed at least twice, so that we could compare strategies used across remixes of the same content. This process, combined with acute data attrition, left us with 60 original videos and 681 remixes for analysis.

METHODS

We hand coded every remix for 25 different attributes related to collective versus personalized strategies (see the Appendix for the full codebook). For example, we noted whenever information in the original video was repeated in the remix (more collective), as well as whenever a new idea was inserted (more personalized). Importantly, by relying almost exclusively on the ages and/or generations that users listed in their bios, we managed to code 213 of the remixes as to whether they were posted by a member of Gen Z (aged 10-25 at the time of coding)—which is how we define "youth" in this project—or by an older user. All coding was conducted by two coders, who achieved a percent agreement ranging from 94.3-100, a Cohen's Kappa ranging from 0.78-1, and a Krippendorf's Alpha ranging from 0.79-1 on a 10% reliability sample.

RESULTS

We were able to categorize the full set of remixes into just a handful of categories, which were taken up differently by Gen Z versus older users. The bulk of Gen Z users' remixes (69%) fell into one of the following categories: (a) a duet in which half the screen displays the original content, while the other half is completely black; (b) a format similar to that of (a), but rather than a completely black screen, the side not containing original content might include colors, patterns, and/or slight movement; and (c) a duet resembling (a) or (b), except that the alternate side incorporates text and/or images that engage with the original content displayed next to it. The top panel of Figure 1 shows an example of each category from left to right. Gen Z's association with these three kinds of remixes indicates an instinct for collective strategies over personalized ones, since they do not redirect any attention to their own identity or experience; instead,

[8] **https://www.politico.com/news/magazine/2022/03/27/progressive-gen-z-for-change-tik-tok-00020624**

[9] **https://github.com/bellingcat/tiktok-hashtag-analysis**

[10] **https://github.com/davidteather/TikTok-Api**

the added visuals and text serve only to keep the viewer's eyes and interest on the original content and its corresponding message.

To complicate that slightly, however, we note that there is a distinction between category (a) and a repost. TikTok does not have a button for directly reposting a video, but a repost can be accomplished by using the green screen feature to turn the original video into a backdrop, and then not layering anything on top of it. However, reposts were rare in the data, while instances of category (a) were common. Despite contributing almost nothing to the original—making it similar to a repost—the remixer declares their presence in these videos by taking up half of the screen. Thus, these remixes demonstrate an interest in collective strategies, but with a participatory edge. Furthermore, the remixer's presence makes them meaningfully different from reposts in the context of the FYP, where most users dwell. If one encounters such a remix on their FYP, it is obvious that at least two individuals endorsed the affiliated content: the original poster and the remixer suggested by the black screen; meanwhile, in a slew of algorithmically-recommended content, a proper repost could easily register as the original content, effectively erasing the additional endorsement for the viewer.

While younger users employed collective yet participatory strategies, older users employed distinctly personalized ones. The majority of their remixes (65%) consisted of duets with the remixer in the frame, reacting to the original content in some way, as exemplified in the bottom panel of Figure 1. As such, older users inserted themselves into the messaging and directed at least part of the viewer's attention to their own perspectives and experiences.

In terms of engagement, we find that Gen Z users' remixes and their collective strategies were disproportionately amplified compared with those of older users: While Gen Z remixes made up 61% of the set, they accounted for 73% of all likes, 82% of all comments, and (a whopping!) 97% of all shares.

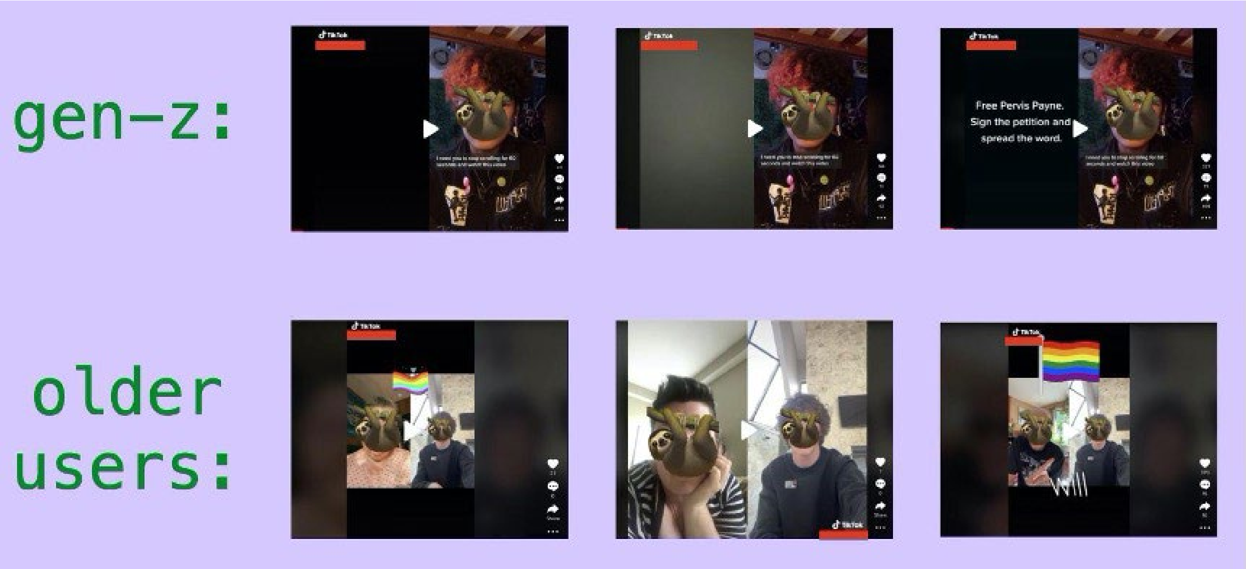


Figure 1: The top panel shows an example of each of the three types of remixes most popular among Gen Z users (categories (a) through (c) from left to right). The bottom panel shows three examples of the type of remix most popular among older users. The sloth emoji was inserted to protect user privacy.

CONCLUSION

The results presented above, though noncomprehensive, illustrate well our central finding: In spite of what the relevant theories might suggest, in spite of TikTok's design, and in spite of how older users behave, TikTok's main user base of young people employs deeply collective strategies to participate in activism on the app. Moreover, their collective strategies have a decidedly participatory bent, and they are successful insofar as they generate an outsized amount of engagement.

Social media, both in its relation to youth political participation and otherwise, has been noted for its ability to support personalized expression. Yet, we observe members of Gen Z articulating a desire for—and demonstrating the efficacy of—collective forms of participation and communication, at least in activist contexts. Therefore, to support their role in social change—if not simply to satisfy their consumer base—platforms might consider introducing more collective-oriented features, while maintaining the participatory nature fundamental to personalized expression.

REFERENCES

Bandy, J., & Diakopoulos, N. (2020). #TulsaFlop: A Case Study of Algorithmically-Influenced Collective Action on TikTok. *ArXiv:2012.07716 [Cs]*. http://arxiv.org/abs/2012.07716

Becker, A. B. (2021). Getting Out the Vote on Twitter With Mandy Patinkin: Celebrity Authenticity, TikTok, and the Couple You Actually Want at Thanksgiving Dinner . . . or Your Passover Seder. *International Journal of Communication, 15*(0), Article 0.

Bennett, W. L., Wells, C., & Freelon, D. (2011). Communicating Civic Engagement: Contrasting Models of Citizenship in the Youth Web Sphere. *Journal of Communication, 61*(5), 835–856.

Bennett, W. L., & Segerberg, A. (2012). The Logic of Connective Action. *Information, Communication & Society, 15*(5), 739–768.

Choi, A., D'Ignazio, C., Foucault Welles, B., Parker, A.G. (2022). Social Media as a Critical Pedagogical Tool: Examining the Relationship Between Youth's Political Engagement on Social Media and Their Critical Consciousness [Manuscript accepted for publication]. *CHI 23: CHI Conference on Human Factors in Computing Systems*.

Church, S. H. (2017). Amplificatio, Diminutio, and the Art of Making a Political Remix Video: What Classical Rhetoric Teaches Us About Contemporary Remix. *Journal of Contemporary Rhetoric*, 7(2/3), 158–173.

Cohen, C. J., & Kahne, J. (2012). *Participatory Politics: New Media and Youth Political Action*. Youth and Participatory Politics Research Network.

Flanagan, C., & Levine, P. (2010). Civic Engagement and the Transition to Adulthood. *The Future of Children, 20*(1), 159–179.

Guinaudeau, B., Munger, K., & Votta, F. (2022). Fifteen Seconds of Fame: TikTok and the Supply Side of Social Video. *Computational Communication Research, 4*(2), 463–485.

Haslem, B. (2022). TikTok as a Digital Activism Space: Social Justice Under Algorithmic Control. *Institute for the Humanities Theses*.

Hautea, S., Parks, P., Takahashi, B., & Zeng, J. (2021). Showing They Care (Or Don't): Affective Publics and Ambivalent Climate Activism on TikTok. *Social Media + Society, 7*(2), 20563051211012344.

Jenkins, H., Shresthova, S., Gamber-Thompson, L., Kligler-Vilenchik, N., & Zimmerman, A. M. (2016). *By Any Media Necessary: The New Youth Activism*. New York University Press.

Literat, I., & Kligler-Vilenchik, N. (2019). Youth collective political expression on social media: The role of affordances and memetic dimensions for voicing political views. *New Media & Society, 21*(9), 1988–2009.

Raun, T. (2012). DIY Therapy: Exploring Affective Self-Representations in Trans Video Blogs on YouTube. In A. Karatzogianni & A. Kuntsman (Eds.), *Digital Cultures and the Politics of Emotion: Feelings, Affect and Technological Change* (pp. 165–180). Palgrave Macmillan UK. https://doi.org/10.1057/9780230391345_10

Sadler, O. (2022). Defiant Amplification or Decontextualized Commercialization? Protest Music, TikTok, and Social Movements. *Social Media + Society, 8*(2), 20563051221094770. https://doi.org/10.1177/20563051221094769

Vogels, E. a, Gelles-Watnick, R., & Massarat, N. (2022, August 10). Teens, Social Media and Technology 2022. *Pew Research Center: Internet, Science & Tech*.

Vromen, A., Loader, B. D., Xenos, M. A., & Bailo, F. (2016). Everyday Making through Facebook Engagement: Young Citizens' Political Interactions in Australia, the United Kingdom and the United States. *Political Studies, 64*(3), 513–533.

Weiss, J. (2020). What Is Youth Political Participation? Literature Review on Youth Political Participation and Political Attitudes. *Frontiers in Political Science, 2*.

Stop Scrolling! **Appendix**

GEN-Z FOR CHANGE DUET CODEBOOK

TASK Taxonomize and analyze the #genzforchange duet space.

NOTE 1 Coding process should take into account both the video itself and the caption, unless otherwise specified.

NOTE 2 When coding duets of duets, code for the *outer duet,* or that posted by the user in question. In these cases, the 'original content' includes the original video plus all of the *inner duets*—i.e., everything the outer duet reacts to. If it is unclear what's part of the inner duet versus the outer duet, make sure to click on the inner duet and check.

INITIAL CODE		
CODE	**DESCRIPTION**	**TYPE**
REL	Marked as 1 if the video is relevant, and 0 if the video is irrelevant. Irrelevant videos are those that use the sound of an original video, but have no connection to the original content beyond that / no intent of amplifying the original message. This probably happens most often when the original content's sound is a song or noise that is not tied to its messaging.	0/1

IF REL = 1 ...		
CODE	**DESCRIPTION**	**TYPE**
CAT	**repost**: a duet is categorized as a repost if it is just the original content reposted, with no additions at all. The reposted video may be of lower quality; what matters is that the content of the video is identical to the original. **black**: a duet is categorized as black if the added visual just consists of a black screen and nothing else. A completely white or gray screen also counts, but a more eye-catching color can be marked as 'blank'. One way to tell the difference between 'black' and 'blank' is whether it makes more sense to leave **DISTR_VIS** empty (in which case, 'black') or code it as 'detract' (in which case, 'blank'). **blank**: a duet is categorized as blank if the added visual is largely static & does not seem to contribute anything to the *message* of the original. This is very similar to the 'black' category, but slightly more involved, as it can include patterns, colors, movements, [mostly] static images, etc. **blank+**: a duet is categorized as blank+ if it falls into the 'black' or 'blank' category, but incorporates stickers, text, and/or other mostly static material that is relevant to the message of the original content/its amplification. These duets should keep the original video largely intact. **self**: a duet is categorized as self if it consists of the duetter's face or body, where the face/body is engaged with/reacting to the original content to some extent, and the original sound is kept intact. Includes duets where the duetter is engaged in a task or activity unrelated to the original content, as long as they occasionally look up from what they're doing. It does not include duets where the duetter is technically in the frame, but they are cut off or there for no apparent purpose. **remix**: a duet is categorized as remix if it goes beyond the categories above, creatively refashioning the original content. Includes duets that use the original sound to transmit the original message, while showcasing something else to the viewer. **other**: a duet is categorized as other if it does not fit into any of the categories above.	string (options: 'repost', 'black', 'blank', 'blank+', 'self', 'remix', 'other')
TYPE	For this variable, it doesn't matter what feature the user technically used —we are interested in what the output looks like to the audience. So for instance, if the user technically used the green screen feature but the duet looks like a repost, code as 'repost.' **repost**: coded as repost if the video reads as a repost of the original content. **green**: coded as green if the video uses TikTok's green screen feature. **duet**: coded as duet if the video uses TikTok's duet feature. **duet+**: coded as duet+ if the video is a duet of 1 or more duets. **react**: coded as react if the video uses TikTok's react feature. **sound**: coded as sound if the video is completely new, but uses the sound of the original content.	string (options: 'repost', 'green', 'duet', 'duet+', 'react', 'sound')

IF REL = 1 ...		
CODE	**DESCRIPTION**	**TYPE**
GENZ	Coded as 1 if the duetter is/appears to be a member of Gen-Z (i.e., aged 10-25). Coded as 0 if the duetter is not a member of Gen-Z. Use the duetter's bio to determine their age/generation; if no information is given in the bio and they do not appear especially old, leave the field blank.	0/1 (or blank)
DESC	Short description (1-2 sentences) of the duet. Does not need to include a description of the original content, as we already have that. Can be seen as elaborating on the duet's category (**CAT**).	string
NOTES	Anything else to note, especially any manipulation of the original content that is not accounted for by the other codes. Try to use this field sparingly, so that we know that if there is a note associated with a video, it is probably important to read.	string (or blank)
IDEA	Coded as 1 if the duetter introduces a new idea, point, opinion, or cohering phrase that is not present in the original content, but is relevant to it. If we think of the original content as an essay, ask yourself if the potential new idea would be incorporated as a new sentence or a new paragraph; if not, it should not be considered as a new idea, but more of a repetition. The new idea/point/opinion/phrase should be introduced either in the video itself or in plaintext in the caption; it should not be introduced in hashtags in the caption, except for special cases where the hashtags in the caption are clearly and strongly being used to push a new idea/point/opinion/phrase. Coded as 0 otherwise.	0/1
ASK	Coded as 1 if the duetter introduces a new [action-oriented, concrete] ask or call to action that is not present in the original content. The ask may or may not include an attendant resource (e.g., link, contact info, etc). The ask should go beyond the TikTok itself—that is, it should not be related to sharing/liking/commenting on/etc the video. Coded as 0 otherwise.	0/1
AGENT	Coded as 1 if the duetter explicitly inserts themself into the original message/makes it somewhat about them, or makes themself an agent of the original message, whether or not they physically appear in the duet. For example, writing "I signed this petition!" when duetting a video calling on viewers to sign a petition, or writing "link in *my* bio" in reference to the original poster's link. Also includes when the duetter shows themselves partaking in the relevant action or message. Does not include language of "us" or "we" - should be personal to the duetter, and must include their agency beyond the fact that they duetted the video. Coded as 0 otherwise.	0/1
POINT	Coded as 1 if the duetter points to the original content in agreement, whether with their hand, an arrow, a sticker, etc. Also includes gestures in the spirit of pointing, such as dramatically turning to look toward the original content. Coded as 0 otherwise.	0/1
GESTURE	Coded as 1 if the duetter uses any hand or body motions to signal agreement or enhance the original message. Includes when people cover or wipe their eyes and other such gestures that are attached to the expression of a certain emotion like crying. Coded as 0 otherwise.	0/1

IF REL = 1 ...		
CODE	**DESCRIPTION**	**TYPE**
EMOTE	Coded as 0 if the duetter does not display any emotion; this includes when the duetter assumes a blank face, and when they are not in the frame at all. Coded as 1 if the duetter **subtly** emotes in agreement with the original content—that is, there is little movement, they are not super expressive, and the emotion is a bit more natural and subdued. Coded as 2 if the duetter **exaggeratedly** emotes in agreement with the original content—that is, there is a high degree of movement, they are very expressive, and attention is drawn toward their emotion. Includes all forms of crying. Note that nodding counts as a form of emoting.	0/1/2
MOUTH	Coded as 1 if the duetter uses lip syncing to agree with the original content. Coded as 0 otherwise.	string
VOL	Coded as 1 if you can hear the duetter speaking on top of the original sound. Coded as 0 otherwise.	string (or blank)
RE_RESOURCE	Coded as 1 if the duetter repeats (in the same or similar language) the **resource** offered in the original content as part of a call to action (if one is present)—this can be through text, sticker, lip-syncing, or something else. Some examples of resources include links, logistical information about protests or other events, and contact information (e.g., for a politician). Includes cases where the duetter references a resource, but doesn't necessarily repost it, as long as they provide some information as to where said resource can be found (e.g., "go sign the petition in their bio!"). The repetition should not appear in the form of a hashtag in the caption; it should appear in plaintext or in the video itself. Coded as 0 otherwise.	0/1
RE_ASK	Coded as 1 if the duetter repeats (in the same or similar language) the **ask** made in the original content (if one is present)—this can be through text, sticker, lip-syncing, or something else. The repetition should not be vague (e.g., "go help!"), but should instead repeat the specific ask made in the original video. The repetition should not appear in the form of a hashtag in the caption; it should appear in plaintext or in the video itself. Coded as 0 otherwise.	0/1
RE_ARG	Coded as 1 if the duetter repeats an **argument/point of information/idea/opinion/etc.** given in the original content (if one is present)—this can be through text, sticker, lip-syncing, or something else. The repetition should not appear in the form of a hashtag in the caption; it should appear in plaintext or in the video itself. Coded as 0 otherwise.	0/1

IF REL = 1 ...		
CODE	**DESCRIPTION**	**TYPE**
RE_PHRASE	Coded as 1 if the duetter repeats (in the same or similar language) a **cohering phrase** (e.g., “justice for Amir Locke”, or something that could reasonably work as a hashtag) referenced in the original content (if one is present)—this can be through text, sticker, lip-syncing, or something else. The phrase should be present in the video of the original post (i.e., not just in the caption), but the repetition can be in either the video or the caption (including hashtags). The repetition should be an exact repetition of the phrase. Only includes phrases related to the message/content of the original post, and not more generic phrases like “stop scrolling.” Coded as 0 otherwise.	0/1
ACCESS	Coded as 1 if the duetter does something to make the original content more accessible to the viewer, such as adding closed captioning, a voiceover of on-screen text, some sort of informative header, or a trigger warning. Coded as 0 otherwise.	0/1
ALG_UP	Coded as 1 if the duetter explicitly references doing something to game TikTok’s recommendation algorithm in order to boost the original content—can be through text, stickers, caption, etc. Also coded as 1 if the duetter includes some variation of ‘#foryoupage’ in the caption, since that is clearly for the sake of getting the algorithm’s attention (so to speak). Coded as 0 otherwise.	0/1
PLS_SHARE	Coded as 1 if the duetter calls on others to share/boost the original content. Boosting can be in the form of liking or commenting on the original content. Includes when the duetter says things like “spread awareness” or “spread the word,” as sharing the video is implied. Coded as 0 otherwise. While distinct, not mutually exclusive with **SUPPORT**. .	0/1
SUPPORT	Coded as 1 if the duetter expresses support or praise for the original speaker or content. This includes all variations of “please watch,” “stop scrolling,” “this is important,” “boosting this,” etc. Also includes non-specific repetitions of asks made in the original content, such as “go help!” or “go support!” Support can be included in either the video or the caption (or both!), but shouldn’t exclusively be in the form of a hashtag in the caption. Includes less straightforward forms of support like “yesss” and “I hear that.” Includes when the duetter tells the viewer to follow the original poster or to watch the original video. Pointing alone does not count as an act of support. Coded as 0 otherwise. While distinct, not mutually exclusive with **PLS_SHARE**.	0/1

IF REL = 1 ...		
CODE	**DESCRIPTION**	**TYPE**
DISTR_VIS	The following options apply to anything one can view in the video, including text. For 'black' duets and 'reposts', leave blank: **detract**: a duet is coded as detract if the visuals contribute nothing to the original content, and are distracting. This mainly includes visuals (often of poor quality) that do not seem to serve any purpose; for example, a duet with a shaky video of the floor alongside the original content. If the footage of the video detracts but includes other visuals that engage, code as 'engage'. **expand**: a duet is coded as expand if the visuals do not contribute to the message of the original content, are not particularly compelling to watch/look at, but seem to be aimed at expanding the content's audience. This includes, for example, duets with pictures of popular celebrities alongside the original content. **engage**: a duet is coded as engage if the visuals directly engage with and/or enhance the original content. Think of 'engaging' with the original content as acknowledging the original content, paying attention to the original content, and/or interacting with the original content. Includes videos in the 'self' category, even if they are not emoting. **compel**: a duet is coded as compel if the visuals are not relevant to the original content, but seem to be included because they are compelling to watch. This includes, for example, a video of someone drawing a beautiful picture with the original content's sound playing in the background. **other**: a duet is coded as other if there are visuals that are not described by any of the categories listed above.	string (options: 'detract', 'expand', 'engage', 'compel', 'other') (or blank)
DISTR_HASH	If the duetter includes hashtags *in the caption* (i.e., they are searchable) that are irrelevant to the original content, code as 1 and 0 otherwise. If unsure whether a hashtag is relevant or not, code as 1; however, if the hashtag is 'on theme' (e.g., tagging #women in a video about reproductive rights), it should be considered relevant. Do not count generic TikTok hashtags (amplification-oriented or otherwise) like #fyp, #boost, #greenscreen, #duet, etc. Do not count hashtags that are part of relevant plaintext sentences (e.g., I am #pissed). If there are no such hashtags, leave blank.	0/1
DISTR_OTHER	Coded as 1 if the duetter brings up an idea, opinion, concept, etc. that is not directly relevant to the original content, or does not work to enhance its message. Also includes when the duetter redirects attention to their business or product, even if the business/product is aligned with the original content somehow. Coded as 0 otherwise.	0/1
DISAGREE	Coded as 0 if the duetter does not disagree (i.e., agrees) with the original content. Coded as 1 if the duetter constructively disagrees with some part of the original content—that is, offers an alternative or note, etc. Coded as 2 if the duetter plainly and antagonistically disagrees with the original content.	0/1/2

ALG_DOWN	Coded as 1 if the duet does something aimed at avoiding suppression by TikTok's recommendation algorithm. This usually comes in the form of using 'algospeak,' such as writing '@b0rti0n' instead of 'abortion.' It may also be along the lines of sarcastically suggesting *not* to do something, when the intention is clearly to get the viewer to do that thing. The suppression avoidance tactic can be deployed in either the video or the caption. Coded as 0 otherwise.	0/1