Columbia Journalism School TOW CENTER FOR DIGITAL JOURNALISM
USC Annenberg Innovation Lab

# Controlling the Conversation: The Ethics of Social Platforms and Content Moderation

April 2018

Priyanjana Bengani
With introduction by Mike Ananny and conclusion by Emily Bell

Platforms and Publishers: Policy Exchange Forum III
February 23, 2018 | University of Southern California Annenberg
Hosted by Tow Center for Digital Journalism at Columbia University and the
Annenberg Innovation Lab

*The Policy Exchange Forums are a critical component of the Tow Center's Platforms and Publishers research project. In these sessions, participants representing both the platforms and publishing sides of the news industry can engage on issues related to the ethical and civic values of journalism. The forum focuses on the relationships between technology, business, journalism, and ethics, and brings together diverse stakeholders to discuss current issues and surface potential new ones.*

*The project is underwritten by the John D. and Catherine T. MacArthur Foundation, with additional support by the John S. and James L. Knight Foundation, the Foundation to Promote Open Society, and The Abrams Foundation.*

# Executive Summary

With social platforms' prevailing dominance, there are numerous debates around who owns information, content, and the audience itself: the publisher, or the platform where the content is discovered—or not discovered, as the case may be. Platforms rely heavily on algorithms to decide what to surface to their users across the globe, and they also rely on algorithms to decide what content is taken down. Meanwhile, publishers are making similar decisions on a significantly smaller scale, and not necessarily algorithmically or quite as generically. But how are any of these decisions made? And what are the various factors taken into account to ensure that the decision-making is fair and ethical?

On February 23, 2018, the Tow Center for Digital Journalism at Columbia University and the Annenberg Innovation Lab at USC Annenberg School for Communication and Journalism hosted a Policy Exchange Forum followed by a conference on the topic of "Controlling the Conversation: The Ethics of Social Platforms and Content."

The Policy Exchange Forum was a closed-group discussion that followed the Chatham House Rule. The discussion broadly focused on three topics: "Ethics of Moderation", "Moderation Tools", and "Technological Challenges."

**Findings**

- The internet has allowed for an information landscape that transcends national boundaries, and while journalism has transcended these boundaries, there is no singular, global, cross-border code of journalism ethics. Similarly, platforms are mostly governed by a US-centric sense of values, which are enforced on their global user-base.
- Platforms have created the baseline standards in moderation, which publishers are compelled to adopt prior to setting their own baselines predominantly due to the amount of audience engagement that occurs on platforms.
- There is a fundamental difference in speech moderation between platforms and publishers. For publishers, the conversations are often two-way, as they try to build a sense of community with their audiences, address their users' concerns, and focus on issues important to their readers. For platforms, moderation efforts aren't interactive—when content is taken down, the creator often doesn't even know why.
- To aid moderation efforts, most organisations must rely on a combination of automated tools and human labor. Automated tools are swift, consistent, and can handle scale, but they struggle with applying context. Conversely, human labour is not scalable, but more reliable when it comes to handling nuance. The immediate future isn't entirely reliant on artificial intelligence alone or humans alone, but striking the right balance to create an environment where machine-assisted human moderation thrives.

**Recommendations**

- Audience engagement doesn't have to be "all or nothing"—there are cases where newsrooms have shut down their comments' sections as they mandated too many resources or were highly toxic. However, it is worth adopting an engagement strategy which does not commit the organisation to a level of unsustainable activity, but allows for experimentation and building audience relationships. This could entail opening comments on a subset of articles, having dedicated groups for the audience to discuss issues close to the community either on a publisher platform or on social media, or holding off-line events, such as town halls to which members of the community are invited.
- The emotional toll for human moderators is high, and there needs to be workflows in place to ensure the mental well-being of moderators. This could include allowing them time off if they need it, having counselors at hand, and ensuring that their schedules give them time to recover.
- Publishers and social platforms should be more transparent about their protocols and practices. From the platforms' perspective, this includes stating their principles and guidelines upfront, as well as more engagement from their moderators when content is taken down. From the publishers' perspective, this includes publishing transparency reports, including commentary on how and why certain editorial decisions were taken.
- While designing algorithms to help in moderation, it is imperative to get input from the human moderators early on, as they have a pulse on the minutiae of the moderation process.
- As much as possible, algorithms should be designed in a way that they are applicable globally, and do not risk imposing Western values on the rest of the world.

# Introduction

*By Mike Ananny, Assistant Professor of Communication and Journalism at University of Southern California Annenberg*

Journalists have always had complex relationships with audiences. Are audiences sources of revenue, accountability, relevance, or information? Are journalists supposed to meet audiences' expectations and treat them as customers, or challenge them with new and perhaps unwanted perspectives and see them as readers who can come and go? Should journalists collaborate with audiences and involve them in their work, or keep them at a safe and objective distance? The answer to all of these questions is a frustrating "it depends."

In earlier eras of the press, journalists met audiences through letters to editors, call-in shows, ratings summaries, aggregate demographics, or on-the-street interviews. To journalists, audiences were often after-thoughts or necessary distractions. Historian Robert Darnton[1] describes journalists' visceral fear of the general public, finding that reporters preferred to imagine their readers as their friends, families, and colleagues—never some abstract public. Communication scholar Karin Wahl-Jorgensen[2] similarly finds that editors of newspaper letters sections see people who write into the paper as "insane" and unrepresentative of who they think their real audiences are, or want them to be. Throughout history, journalists have had uneasy relationships with audiences: they are economically essential, core to the profession's public accountability, bellwethers of popular culture, and frequent sources for secret information—but they are also abstractions, masses that journalists can never really know. It's easier to turn them into clichés, ignore their existence, or pretend that they are either supportive or dismissible.

Today, the audience lives not only in journalists' attitudes or small-scale media like letters, call-in shows, and interviews. Contemporary journalists also meet audiences through complex social media infrastructures, third-party algorithms, and content moderation policies. News organizations are increasingly closing their own websites' comment sections, firing their public editors, or returning to older forms of editorial letters—assuming that some audiences are in a vague digital "elsewhere" of tweets, clicks, likes, shares, and web traffic that is beyond their control. But as news organizations outsource their public interactions to technology companies, those very platforms eschew their roles as public media outlets. The editorial judgments and cultures of public accountability that, ideally at least, drove newsrooms' relationships with audiences are being displaced by the language of social media companies: commercial content moderation[3], algorithmic filters, online community standards, and platform terms of service.

[1] Darnton, R. (1975). Writing news and telling stories. *Daedalus, 104*(2), 175-194.

[2] Wahl-Jorgensen, K. (2007). *Journalists and the public: Newsroom culture, letters to the editor, and democracy*. Cresskill, NJ: Hampton Press.

[3] Roberts, S. T. (2017, March 8, 2017). Social Media's Silent Filter. *The Atlantic*. Retrieved from https://www.theatlantic.com/technology/archive/2017/03/commercial-content-moderation/518796/

What does this new terrain between news organizations and platform companies look like? How does it understand audiences, free speech, and public accountability? What kind of new languages and standards need to be invented to help journalists see where their audiences now live, and help technologists better appreciate the public responsibilities they have, perhaps unknowingly, accepted? These were the kinds of questions underpinning the symposium this report describes.

While the report details well the meeting's discussions and debates, twelve key issues stood out to me:

1. How can this new terrain between platforms and publishers sustain multiple ideals of the public, and diverse notions of ethics? Instead of trying to create and police a single way of understanding social media audiences, how can moderation nurture diverse forms of public life without, as Bengani says in her executive summary, overprivileging the West?
2. How can platform-publisher intersections foreground the mental health, fair compensation, and compassionate treatment of the people tasked with moderation labor? New types of emotionally difficult and ethically taxing work are required to maintain these free speech systems, and it is incumbent upon both platforms and publishers to care for their workers.
3. How can platforms and publishers better appreciate the diverse types of moderation their intersections demand? Just as platforms are defined by no single platform culture, and news organizations are governed by no one set of editorial standards, content moderation at platform-publisher intersections needs to account for multiple cultures of expression.
4. How can the challenges of content moderation be understood as functions of platform business models? Better forms of content and moderation are only likely to come from direct engagement with how platforms commodify people and content.
5. When, how, and why should journalists engage with audiences? New spaces and traditions of audience engagement raise anew longstanding questions about the ethics of journalists' relationships to their readers and viewers, what they owe their audiences, and when they should remove themselves from public conversations.
6. How can content moderation tackle the core problem of scale underpinning many social media platforms? Are small-scale content moderation policies and practices possible anymore, or are news organizations beholden to the massive, global-scale spaces platforms' growth-driven business models require them to create?
7. What new tools and infrastructures are platforms creating to manage content moderation labor, and automate aspects of moderators' work? What kinds of ethics and success metrics drive the design and deployment of such tools, and what role might news organizations and moderators alike have in their invention?
8. What are the threat models underpinning content moderation best- and worst-case scenarios? What, exactly, do platforms and publishers fear comes with unmoderated content, what harms dominate these models and what harms are not being talked about enough?

9. How can platform-publisher content moderation break out of US-centric ways of thinking? Many debates fall back on the (US) First Amendment, section 230 of the (US) Communications Decency Act, and the values of Silicon Valley companies, either neglecting or attempting to standardize the variations that exist across global media systems and cultures of expression.
10. How might platform-publisher content moderation be seen as public infrastructure? Instead of leaving questions of moderation practices, policies, and tools as proprietary issues to be resolved within private companies, how might they instead be seen as collective concerns requiring public governance?
11. In thinking about the complex nature of the challenges of platform-publisher content moderation, how can researchers and practitioners see the landscape as a multi-faceted one requiring multiple levels of analysis? The answers to these challenges are not to be found in any one place, but a mix of technological, social, political, economic, and cultural interventions.
12. Finally, how can the community of moderation researchers, regulators, technologists, and civil society activists sustain themselves as an engaged, effective, and generous collective that can respond to both the near- and long-term challenges of creating ethically defensible platform-publisher content moderation?

# Discussion I: Ethics of Moderation

Drawn from a lightning talk by Stephen J.A. Ward (Distinguished Lecturer on Ethics / Founding Director - Center for Journalism Ethics at the University of Wisconsin-Madison)

For publishers and platforms, the challenges around online moderation are growing. The digital revolution has resulted in a global information system where it is almost impossible to reach a consensus on a universal code of ethics that can be applied to this new media landscape. The first session of the day focused on the need for a reboot in ethics, with participants discussing their experiences, challenges, and concerns around the ethics of online journalism and content moderation. The first discussion kicked off with a lightning talk from Stephen Ward, which set the stage for a conversation that explored:

1. Foundational Journalism Ethics versus Content Moderation Ethics
2. The Role of Publishers versus The Role of Platforms

# Ethics: A Short Introduction, or Democracy Without Dialogue

There are no concrete answers to the question, “What is ethics, anyway?” What can be said, however, is that ethics cannot exist without context or in the abstract. Instead, codes of ethics emerge as a response to the common or uncommon understanding of the problems of a place and time, where the response is driven by the moral interpretation of the said problems. In the case of media and journalism, there exists a large scope for disagreement as everyone comes in with their own widely different moral interpretations of the goals of journalism.

The digital revolution, amongst other things, has resulted in the globalization of the media industry, which has led to the new participants (and platforms) questioning the now-stale professional objective model that was constructed in the United States in the 1920s. This old model allowed for a parochial, nation-bound medium bound to the codes of ethics specific to its locale. Now, journalism extends far beyond a nation’s borders, but there is no unified sense of journalism ethics.

Consensus across borders isn’t the only dilemma confronting journalism ethics in the digital age. Looking inward at the United States, as an example, journalists disagreeing with the “Trump agenda” are considered partisan according to a large part of the population. Therefore, if *The New York Times* were to publish a big exposé, half the country might shrug off the scoop as untrustworthy because it came from the “liberal media.”

Online interactions among strangers with conflicting opinions typically end in a yelling match between the participants; there are few public fora that provide a space for respectful albeit frank conversations across the ideological factions. One way for newsrooms to address this is to adopt “democratically-engaged journalism”, i.e. journalism engaged with society that promotes a plural and egalitarian democracy while respecting facts and spreading media literacy. This mandates dialogic journalism, where journalists create bridges of understanding among the divided—those at the far ends of the political spectrum—in the communities served by the news organization. This helps contribute to a free and democratic press where open discussions are encouraged, thereby suppressing mob rule and promoting the kind of populism where ordinary people’s voices carry weight.

An example of one such news organization is the Canadian Broadcasting Corporation (CBC). By holding town-hall meetings with the objective of getting the cadence of truth and reconciliation, and giving people microphones to tell their side of the story on air, they are effectively creating bridges.

Setting static rules that are applied on a per-case basis is not the panacea for ethics in journalism and media. Instead, journalism ethics requires a reboot, with the focus on *invention*, i.e. constructing a new set of principles; *open-ended discussions*; *plurality* of perspectives within the journalism and ethics communities, even if it may lead to contention amongst those participating; and thinking about the *global* landscape to transcend borders and impact civilisation and culture beyond the United States.

With this in mind, newsrooms should ask themselves these ethically-tinged questions when it comes to interacting with their audience online:

1. Does mediation promote a dialogic public sphere medium while avoiding ideological confrontational debates captured by extreme voices?
2. Does conversation online lead to a fair representation of the different viewpoints without resorting to stereotypes?
3. How do we fight the misinformation in the public sphere?
4. Is mediation passive or active, i.e. are people coming together to discuss problems and come up with solutions, or are people simply expressing themselves in what could practically be a void.

## Journalism Ethics versus Content Moderation Ethics

In the Venn diagram composed of ethics in journalism and ethics in content moderation, what is the intersection and what are the differences? And, to what extent can the norms of foundational journalism ethics be translated to ethics in online moderation?

The consensus amongst the participants was that the norms around “engagement reporting”, for the most part, isn’t that different from traditional reporting. Engagement reporters cover stories that matter to the audience, are deeply reported and fact-checked, and in which the purpose of the interaction between the newsroom and the audience is made clear to all participants at the outset. Further, prior to embarking on an engagement-driven piece of work, there are pre-existing rules put in place to ensure the participation from the community doesn't steer off track and the focus remains on the purpose stated at the outset., if and when the conversation goes astray, the rules allow moderators to rein it in. The dilemma arises when users misunderstand the role of journalists, and attempt to use the journalists to promote personal agendas. In some cases where these users have been wronged or are swayed towards advocacy, they expect news organization to side with them or circulate grievances on official channels. But, it is up to the journalists to draw those lines and minimise the fuzzy areas in these interactions.

Drawing these clear, bright lines isn't a trivial exercise, though: One member at the roundtable asked, "where do we even start? Do we build an edifice to objectivity akin to what we had in [legacy] journalism?" Objectivity can be "pernicious," according to another participant, as it has the potential to delegitimize voices and discussions, whereas the online moderators who enjoy the most social capital are the ones who invite everyone into their context and understand where people are coming from.

A participant, who has spoken to news organizations across the globe (including Europe, Mexico, Brazil, United States, and Canada), said there are similarities in how foundational ethics are discussed, and even as some of these ethics are evolving, it is still possible to find solid, concrete ethics across all kinds of journalism. These foundational ethics include being accurate and truthful, being fair, avoiding harm, and avoiding conflict of interest.

So, now, assuming a set of universal journalistic ethics exist, how can they be shifted to these new roles of online moderation? And, if we are looking at automated tools to contribute to the moderation efforts, how do we consider nuance and context—something that is essential if we were to be compliant with the foundational journalism ethics?

## The Role of Publishers versus The Role of Platforms

Journalists aren't the only publishers, and often, the website of each news organization is not even the primary medium used to interact and engage with their reporting. Instead, social media sites, i.e. platforms, provide a common way for readers to access the content posted online. Additionally, a lot of the engagement efforts take place on these platforms and are subject to their terms and conditions, which makes platforms an integral part of this conversation.

Every website, blog, and platform—along with their content moderators—has a set of rules that it tries to abide by. But, the role of content moderators differs widely between platforms and publishers. A "participant moderator" or "engagement reporter" actively takes part in the conversation, be it via commenting, shaping discussions, or enforcing norms. However, moderation on platforms is a completely different animal: mass-scale with minimal (if any) proactive participation from the moderators.

And, yet, the sheer volume of activity (user attention and engagement, groups, platform-specific content) on platforms has made them critical to publishers, who, consequently, have to abide by the rules of each platform. This means that, perhaps even without realizing it, platforms have created the baseline standards in the moderation space. Subcommittees from news organizations on these platforms, tasked with ensuring productive constructive discussion spaces for their journalistic efforts, are compelled to create guidelines within the frameworks of what platforms deem acceptable. However, at no point did all the parties involved get on board and agree with the principles laid down by the platforms. And, so, we are in a situation where it is unclear if the public knows what it wants, or is even aware of the debates taking place on the moderation front. Meanwhile, publishers are acting on suppositions of what platforms want, while the platforms themselves are still thinking through what their moderation strategy should be, with all the nuances and the complications that arise with contextual newsworthy information, (e.g. a police killing video—should it stay or should it be taken down?) while insisting that they do not want to be the arbiters of truth.

Moderating content from news publications isn't the only role that platforms play in mediating between factions of the public. The rise of "state-run social media" has resulted in journalists no longer being the gatekeepers. Instead, this role has been extended to police departments and heads of states, where with a tweet or a post, they can control the narrative as they build their audiences. Some police departments have audiences larger than local newspapers in rural areas at this point. Journalists cannot avoid interacting with this state-run media, but there are questions around how journalists should engage with the "new publisher in town". The possible questions, as posited by one participant, included: "are we going to amplify the messages from these accounts and build their audience?", "will we let them control the narrative and be stenographers while they upload police killings on YouTube?", and "are we going to fight with the President of the United States?"

## Key Challenges and Opportunities

Multiple participants brought up the design implications of looking for technical solutions to social problems. The social problems, in this context, are twofold: the need for localization in a global space and the problem of building ethical and effective tools.

### Localization in a Global Space

One of the participants said that it continues to "confuse me, concern me, perturb me" that all these conversations are predicated on American legal norms and precedents as complete fait accompli. This is especially troublesome when platforms' values are inherently Western even though most of their traffic is not from the United States or Europe. Accounting for different environments and cultures is essential, but since the media ecosystem varies in each country, capturing all the nuances specific to all countries is a non-trivial endeavour. For example, in countries where the government has influence over a public broadcaster, it is hard to sift through good journalism for what could possibly be state-sponsored disinformation campaigns.

**Building Ethical Tools**

People should design for democracy, and bake ethics into tools as they are being built. But how does one actually go about doing that? And how much time do we have?

Attempting to solve this problem without a higher macro-level understanding of what we are designing for—what the values, priorities and norms are—is futile. Throwing money at engineering to solve this problem is also likely to be futile without solving the design problem first.

For example, the ideology of extremes might be a norm that should be tempered or excluded from conversations. But should there be a difference between political views that are fact-based and political views that are not, political views that advocate harm and political views that advocate liberation? Or should all extreme voices be treated just the same?

Addressing social quandaries like this has to be a priority in this current pressing moment, because it is harder to roll out a fix (unlike, say, a website patch) if something goes awry. Things need to change sooner rather than later, but it also takes time to deliberate on the various aspects of the debate and conduct a thorough analysis before reaching implementable conclusions, as evidenced by Wikipedia taking up to a couple of years to decide how they would handle the controversy surrounding depicting Muhammad. Rushing to deliver tools or conflating designing and engineering could lead to suboptimal solutions that perpetuate the toxic state of the current public sphere.

But do we have two or more years at hand to make the necessary design decisions?

## Discussion II: Moderation Tools

Drawn from a lightning talk by Anika Gupta (Senior Product Manager at The Atlantic)

After discussing the broad strokes of global ethics for media in the information age, the second session turned to explore the current environment of moderation and its shortcomings in an attempt to identify how to improve the situation to ensure good information public health, which has spent the past 12-18 months in the metaphorical emergency room.

The discussion kicked off with a lightning talk from Anika Gupta, which set the stage for a far-reaching conversation that covered:

1. Objectives of Different Stakeholders: Newsrooms, Platforms, and Users
2. Automated Moderation versus Human Moderation

## Objectives of Stakeholders: Newsrooms, Platforms, Advertisers, Users

The online ecosystems that encourage user engagement have different priorities, which are reflected in the moderation practices of these organizations—be it platforms, publishers, or publishers on platforms.

There are two angles to moderation in this context:

- what moderation entails, i.e. what content is allowed, what comments are permitted, and what interactions are acceptable
- what these norms are driven by—business outcomes, democratic outcomes, or something else

Different groups within the same organization can have conflicting priorities, too. For example, in a newsroom, the metrics that business executives focus on may not align with the goals the editorial team cares about. Business executives might want to maximise eyeballs on a page, but this can be at the expense of mental health of the human moderators who have to manually deal with the plethora of comments. This can be downright difficult for them, especially if there are no protections in place for the moderation team and the comments are hostile.

One of the ways in which publishers build close ties with their communities is by interacting with their readers, and providing their readers with a space to interact with them. This can take many forms, including creating Facebook groups or activating comments. Right now, though, if the online comments get hard to handle or are consistently not in accordance with the community moderation guidelines, the reaction of some media companies is to shut the comments down.

There are, however, news organizations committed to keeping their comments open, as it allows them to push the conversation further along and ensure that all sides of the debate exist on their pages. This, in certain cases, even permits extreme views provided that the comments themselves do not violate the rules. After all, as one participant said, if society is going downhill, it will do so with or without the media's approval.

And, yet, one of the problems is that just the presence of an extreme voice can have a chilling effect—large swathes of people, especially those from marginalised communities, are reluctant to contribute to a contentious conversation, and it is the job of the moderators to encourage those silent voices to speak up on their platforms vis-à-vis extreme content.

This problem is not exclusive to news websites—it's common on platforms as well. However, a participant said that platforms aren't dedicating enough capacity or building infrastructure necessary for effective moderation work where the moderators have two-way conversations with the users and address their concerns. Further, as multiple other participants pointed out, the business imperative for platforms is predicated on the intangible sense that everyone needs to be online expressing themselves all the time, either by engaging with existing content or uploading new content, both of which can be monetized by the platform company. And it is a documented fact that emotive content leads to higher engagement than more even-tempered views, which means that it's in the platforms' best interests to promote these extreme or emotive voices irrespective of their accuracy.

Recent events, though, have compelled platforms to revisit their approach to moderation to avoid being in "clean up the mess" mode. These market pressures include big brands pulling their advertisements from YouTube, daily average usage on Facebook slipping in the United States and Canada for the first time, and journalists and academics flagging subpar activities on the platforms regularly. As a result, platforms are taking direct action to improve their ecosystem, be it via hiring more content moderators or by partnering with fact-checking organizations.

These efforts are likely to sift out extreme views and blatantly false stories, but there are two other areas which both platforms and publishers should address for the benefit of their users: transparency and education.

More *transparency* from platforms and publishers will mandate that guidelines and principles are clearly communicated to the user-base, leaving little—if any—room for ambiguity. One participant suggested platforms should communicate more with the world on their principles, what they're doing, and have visible external guidelines. Currently, there is a complete lack of engagement from the platforms; users hear "a whole lot of nothing" when they attempt to interact with platform moderators because a piece of their content has been taken down, or they have reported a questionable piece of content, or they have been suspended from the platform for a week. But, this need for transparency isn't specific to just the platforms, as another participant pointed out. Using the example of the Weinstein story that NBC killed, the participant suggested that media outlets also have a responsibility to publish transparency reports so that, amongst other things, readers are aware of how many stories are killed.

*Educating* users is a hard and costly problem that cannot be solved with a 1-3 hour online video course. There needs to be an core curriculum that's regularly updated, and platforms and publishers need to promote media literacy to their users via this core curriculum on an ongoing basis. We can create tools to guide users, but there needs to be a minimum baseline of knowledge amongst the users, or else the tools will be rendered useless.

## Automated Moderation versus Human Moderation

The deluge of user-generated content being uploaded every minute means that it is highly unlikely that a human being will be able to approve every single item manually. This necessitates designing machine learning algorithms to aid in this task. But machine learning isn't the panacea that it's often made out to be, and it is important to think about when these algorithms will work and when they won't, and the fallbacks that are put in place when the algorithms fail.

Designing these algorithms also requires tackling and codifying some difficult (if not impossible) questions, starting with what kinds of content can be moderated using algorithms. For child abuse, it is easier to create algorithms as there is a concrete definition. Misinformation is a harder problem to solve, as even the human detectors struggle with making the right call.

Among other questions: Who gets to define what "extreme" is, and what the parameters of extreme are? Or, for that matter, what hate speech is? This is especially hard in the case of platforms, which can have over a billion daily users. Simply using outliers is one way of imagining extremist politics, but jumping on this kind of a statistical approach means that the two poles of extremism will inevitably get conflated into one—a point also made during the first discussion of the day.

Finally, creating algorithms that are generic enough to work in a global system is important, or else western values are imposed on other cultures, which is suboptimal.

These factors illustrate that there is a lot of ambiguity when it comes to creating effective all-purpose machine learning algorithms, and consequently a lack of a clear path. That's not to say that there isn't a social science component to this. Norms like civility, elimination of extreme views, and role representation can be encoded with corresponding—potentially democratic or editorial—outcomes. We can then use this as a basis to design technology that maximises these outcomes. But, again, identifying and accurately codifying all the norms is a non-trivial exercise.

And, so, if the "techno-utopian dream of magical robots" solving all problems is not to be realized, perhaps the answer lies in building tools to make the lives of human moderators easier, instead of trying to replace them. One such tool could be designed to use past moderation data as the training data input into an algorithm to help classify future decisions. This could, in theory, help moderators get through their queues faster.

For this to be truly effective, human moderators need to be involved—co-creators even—in the design process, since they are the domain experts; moreso, even, than designers and data scientists. They are aware of the nuances of everyday moderating as well as the emotional toll and the trauma that goes into the job, all of which should be factored into the design of any moderation system.

Human moderators spend hours at a time looking at some of the most abhorrent content online, be it child pornography or violent extremism. Moderating this type of content is typically black-and-white, but other content that needs to be moderated—including misinformation campaigns and news stories—are harder to crack, as mentioned earlier. This is especially true when moderation efforts are outsourced to firms in developing countries where the labor is cheaper and will allow platforms to capitalize on the economies of scale. Too often, moderators living thousands of miles away from the United States are expected to figure out the nuances of the American political system, and deal with the misinformation present therein.

This highly stressful job also comes with a non-disclosure agreement meaning that the workers cannot rely on the traditional support system of family and friends. This puts the onus on the companies to ensure the wellness of the employees, be it by allowing them to tap out on a given day or avoid content that is triggering for them. Some companies hire contractors and mental health professionals to create the emotional space needed by their moderators, but, even so, the rate of attrition is high.

## Key Challenges and Opportunities

In considering the sheer scale of content creation, a monumental question arises: How moderators can handle the public expectation that they will create a rule-of-the-law environment where undesirable content is dealt with quickly and in a consistent fashion, while accounting for the intricacies in each specific case? Provided a set of rules, machine-driven moderation can ensure consistency and speed, but it fails on the nuances. On the other hand, humans are better at contextualizing each case (with caveats for instances where offshore moderators have to handle misinformation cases in the United States), but they are unable to do so at scale. A combination of these two things yields an environment that can appear riddled with double-standards and faux pas.

While technology-assisted human moderation is one of the immediate solutions that platforms and publishers are turning to, there are still kinks to be worked out. The inconsistencies and possible mistakes—false positives and false negatives bubbling up to the surface—means that the onus is on the companies to ensure their users are aware of the processes and practices in place, and to allow for a dialogue to exist between the users and moderators.

When creating these algorithmic systems, companies often use metrics that seem easy to gauge or monitor (e.g. click-through rates, time on page, satisfaction levels), but what about metrics that are harder to quantify? For example, one participant pointed out that transparency is a desired trait, but is there a way to be able to quantify transparency in order to determine the impact it does or doesn't have?

Looking inwardly, companies need to recognize the importance of new workflows to help their moderators deal with the emotional toll that comes with the job. Otherwise, there is a risk that companies will default to doing nothing for the mental well-being of their moderators, possibly causing them severe long-term harm.

# Discussion III: Technological Challenges

Drawn from a lightning talk by Andrew Losowsky (Director of The Coral Project)

Turning to technology to solve societal problems comes with its own set of ethical dilemmas. The promise of machine learning is to use a one-size-fits-all solution to reach incredibly consequential decisions. However, the algorithms are only as good as the data they are trained on, and one of the fundamental problems of machine learning is that biases are the rule, not the exception.

The final session for the day got underway with a lightning talk from Andrew Losowsky about The Coral Project (a Mozilla Foundation project that brings journalists and communities they serve closer together), and the following discussion explored:

1. Designing and Implementing a Technology-Driven Solution
2. The Legal Side of Platforms as Editors

## Designing and Implementing a Technology-Driven Solution

Before newsrooms build automation tools or adopt third-party products to aid their moderation efforts, they should undergo an exercise in introspection to help determine what their strategy around engagement should be. There is a whole spectrum of engagement, and if a particular newsroom doesn't have the resources or "moderation power" to support comments across the board, it doesn't mean they are out of engagement options. Answering questions around what a newsroom hopes to achieve with a comments section given its mission, what its community needs, and how its site or group is different can allow them to tackle the harder questions around when and how to engage with the audience.

From the audience's perspective, there are four main areas around interaction: **clarity** in terms of what the news organization is doing and why; **utility** in terms of what the commenter is getting from this interaction; **safety**, i.e. is the commenter at any kind of risk for initiating or participating in a conversation; and **recognition**, i.e. are their efforts at engaging with the news organization and their contributions being appreciated?

Talk, a commenting platform created by The Coral Project, is one open-source tool that allows news organizations to experiment with their commenting strategy, allowing sites to activate comments on a single article or across the board. Activating comments doesn't have to mean opening the floodgates: Moderation can start before the comment is posted, or by disallowing comments that consist of undesirable language, be it from a blacklist of words and phrases or from a list of hidden hate speech codes.

*Talk* supports this functionality. In addition, it also hooks into Google Jigsaw's Perspective API, a toxicity-detector that provides a toxicity score for a comment.

A participant pointed out that this toxicity metric is not rooted in truth, but in harm, i.e. the premise is to rephrase the commenter's truth in a less offensive way. And, different organizations might want to tweak their thresholds or parameters for toxicity, or even explore how the toxicity metric evolves over time.

Breaching the acceptable toxicity threshold prompts users to revisit their comment (lest the comment was made in a single "bad moment"), but only once to minimise the risk of users trying to game the system. If users decide to submit the comment anyway, human interaction will be needed prior to the comment being published onto the site. This system aims to minimize the number of comments that make their way to human moderators, thereby helping tackle the problem of sheer volume.

This computer-assisted human moderation requires ongoing testing and analysis, to ensure that it is not being gamed, that the false positives (i.e. good comments mistaken as toxic) are addressed — an established problem with Jigsaw's Perspective API —, and the threat model is well-defined. Adding, tweaking, or removing features should never be done without revisiting the threat model to determine if users are safe from abuse, harassment, or being silenced as a result of these new changes. What happens, though, when hate groups start using code words like Skype and Yahoo in place of racial slurs? And, while we have some ideas on how to deal with text, what happens when faked (and deepfaked) photos and videos take on an even bigger role in our conversations? One participant pointed out that we are almost there, and we don't have a clue—technologically or as a society—on how to deal with that problem.

Codifying ambiguity is another tremendously difficult problem. Using the Trolley Problem as an example, one of the participants raised a question: How do we express that accurately and at scale in an algorithm?

Further, even with ongoing testing, threat model analysis, and research, there is the evergreen issue when it comes to algorithms—that they are only as good as the training data and biases in the data along with the biases of the people writing the algorithms. One concerning example that was provided was a classifier that had been built to label clickbait appropriately, but it ended up manifesting the biases of the raters and disproportionately affected a marginalised group. This illustrates that we cannot take the tools for granted or buy into the Pollyannish assumption that this sort of bias is merely an exception (which has been proven false over and over again).

Platforms have to address these problems as well. Further, platforms err on the side of being neutral, not wanting to be arbiters of truth, which means that they have an additional problem: clamping down on misinformation.

Journalists, in theory, can help the platforms in a meaningful way to battle misinformation, an idea that is under-explored. However, this model needs to be developed such that there is a real value proposition for the journalists, rather than doing the platforms' work for free while the platforms absolve themselves of responsibility.

## Platforms, Editorial Oversight, and the Laws

While discussing moderation efforts on platforms, multiple participants brought up the platforms' reluctance to accept their role as editors, even though editorial activity is rampant on the platforms. As one participant said, moderation is fundamentally an editorial decision, as is coming up with a counter-messaging strategy for terrorism — both of which platforms uncomplainingly do.

This reluctance of platforms to accept their role as publishers, to an extent, stems from their desire to be perceived as politically neutral, not discriminating against any user or post based on ideology. There is political pressure on this front as well, as witnessed by the public outcry when allegations emerged that Facebook's trending topics suppressed conservative views.

But there is no real-world legal liability here. Section 230 indemnifies platforms almost completely for all user-generated content in the United States without requiring them to be politically neutral. Section 230 is a U.S. statute that says, "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."

We see this disconnect in terms of how platforms talk about the law versus what the reality is. For example, the First Amendment doesn't apply to private actors like the platform companies. It is important to challenge these inaccurate assumptions and public perceptions about the law that platforms leverage to evade responsibility.

So what will make platform companies take ownership of the problems living in their domains?

While nothing holds any platform's feet to the fire at the moment, the regulatory climate ~~for platforms~~ in Europe is changing. Recently, a U.K. parliamentary committee hearing was held in Washington D.C., where representatives from platforms were asked point-blank why their businesses shouldn't be considered publishers. The European Union's Electronic Commerce Directive's liability of intermediaries specifies the notion of a neutral entity or mere conduit, and this is very different from the United States, and brings with it drastic legal challenges for the platforms.

One participant said that we are looking to Europe to show leadership the subject of neutrality and moderation on platforms, but in the United States, we should look into how to motivate the companies to go beyond Section 230. This includes the ideas of general social responsibility, corporate citizenship, and corporate social responsibility—be it inter-company (as we saw with brands threatening to pull their advertisements from YouTube) or in conjunction with social activism. Activists have elevated the discussion around the major issues on these platforms, and the platforms have sometimes taken action—be it the Twitter bot purge or removing verification checks from the accounts of white supremacists.

## Key Challenges and Opportunities

Using technology to help solve the moderation problem is an attractive proposition for all parties involved, but it has its share of problems that platforms and publishers need to take into account.

Finding the right balance between human moderation and automated moderation will be key, but there need to be safeguards and checks in place to ensure that the errors in the system don't end up silencing voices or pushing them out of the conversation. It is also important to ensure that the error rates in the algorithms don't disproportionately affect minority or marginalised groups.

Operationalizing values, especially those that are hard to define, is challenging, and can result in inadvertently delegitimizing the experiences of some members of the communities that platforms and journalists are trying to build, and among whom they are trying to instil trust.

It is also worth exploring other ways in which media companies can build trust and distinguish themselves from the platforms (or other news organizations). A lot of the commentary and conversation takes place on platforms today, and one concern is that users conflate platforms and publishers, and are unable to disambiguate between decisions made by platforms and decisions made by publishers, despite different vastly different value systems. One participant, who had spoken to dozens of people in terms of what they trust in the news ecosystem, said that there are other technology-driven solutions that could work independently or co-exist with the commenting platform including annotation tools, in-person events, and letting users vote on quality of a website and its content.

The Silicon Valley mantra has always reflected Facebook's recently abandoned—and somewhat embarrassing in hindsight—motto: “move fast and break things.” In light of recent events, perhaps its ethos ought to shift from building and launching products first and cleaning up the mess later to doing due diligence and research prior to designing, developing and releasing products.

# Conclusion

*By Emily Bell, founding director of the Tow Center for Digital Journalism at Columbia's Graduate School of Journalism*

Moderation at scale is perhaps the most challenging aspect of the rapid increase in social publishing. The involvement of ever larger numbers of citizens in creating material which they share with others, the automated amplification of speech by bots and coordinated online campaigns, and the growth of use in social media platforms has made moderation the central challenge for Facebook, YouTube, Twitter and their smaller peers.

For legacy publishers even the challenge of maintaining comment threads on their own websites has proven too much, and they have either closed down public access to below-the-copy commentary or outsourced it to social platforms.

As we have seen over the course of 2017 and 2018, a failure to be clear and consistent about moderation policies, or an inadequacy of resource or understanding, can have profound consequences for individuals and societies. The 'fake news' epidemic of the 2016 US Presidential election cycle and the subsequent revelations about the propaganda campaign from the Internet Research Agency based in Russia, have brought about wide recognition of what academics and journalists have long identified as a problem in the information ecosystem.

Examples like the viral hate campaigns against Rohingya Muslims in Myanmar and the proliferation of "digital pharmacies" selling opioid drugs through Facebook have been blamed on the irresponsibility of Facebook in particular in not monitoring what material might be circulating on its network.

For journalistic organizations and for social platforms, the business of moderation is inherently uncomfortable and expensive. Platforms are reluctant to see themselves as exercising editorial judgment over content, and news organizations often prioritize resources into reporting their own material rather than moderating commentary by readers. The result of this is a large neglected delta of content which contains everything from hate speech through to videos of torture and death, which have to be removed from the system by both algorithmic sorting and by humans.

There are also many finer, more nuanced calls which cannot be decided by machines alone. When is an individual being hateful? Is a meme political manipulation? When is 'fake news' really fake? To some extent the problems wrought by a globalized unregulated market for speech and content will never be 'solved' by moderation tools or even human intervention. Social norms, literacy in the new systems of creation and distribution, new types of accountability rules and organizations will all need to develop alongside more sophisticated and transparent systems of human and mechanistic editing and intervention.