High Impact Questions and Opportunities for Online Harassment Research and Action

August 2016 | Cambridge, MA

Contributors:

J. Nathan Matias, PhD Candidate, MIT Media Lab
Camille Francois, Google/Jigsaw
Amy Johnson, PhD Candidate, MIT History, Anthropology, Science, Technology, and Society
Emilie Reiser, Developer and Project Lead, MIT Center for Civic Media
Susan Benesch, Director, Dangerous Speech Project
Lindsay Blackwell, PhD student, University of Michigan School of Information; researcher with HeartMob
Amy Bruckman, Professor, Georgia Tech School of Interactive Computing
Jen Carter, Product Manager, Policy & Enforcement, YouTube/Google
Soraya Chemaly, Director, Women's Media Center Speech Project
Justin Cheng, PhD student, Stanford University Computer Science
Jason Coon, Engineering Manager, Xbox Product Services, Microsoft
Lucas Dixon, Chief Scientist, Jigsaw
Nicole Ellison, Professor in the School of Information, University of Michigan
Eric Gilbert, Associate Professor, Georgia Tech School of Interactive Computing
Rey Junco, Associate Professor in Education & Human Computer Interaction at Iowa State University
Cliff Lampe, Associate Professor, University of Michigan School of Information
Mariel Garcia M, incoming Master's student, MIT Comparative Media Studies
Merry Mou, M.Eng Student, MIT CSAIL
Katherine Lo, PhD student, University of California Irvine
Alice Marwick, Fellow, Data & Society Research Institute; Assistant Professor of Communication, University of North Carolina, Chapel Hill
Kevin Munger, PhD candidate, NYU department of politics
Sarah Otts, Software Engineer and Community Moderator, MIT Scratch
Derek Ruths, Associate Professor of Computer Science, McGill University
Andy Sellars, Director of BU/MIT Technology & Cyberlaw Clinic
Sarah Sobieraj, Associate Professor, Tufts University Department of Sociology and Director of the Digital Sexism Project
T.L. Taylor, Professor of Comparative Media Studies and Director of Research for AnyKey
Nithum Thain, Jigsaw Fellow
Ellery Wulczyn, Data Scientist, Wikimedia Foundation

High Impact Questions and Opportunities for Online Harassment Research and Action

Online harassment has been an enduring and evolving social concern for over 40 years, yet many of the most urgent empirical questions for public well-being and freedom remain unexplored. Nor can our answers evolve at the pace of socio-technical change. On August 17th and 18th, the MIT Media Lab and Jigsaw convened 35 researchers, advocates, and platform representatives to identify and advance high impact research about online harassment.

The report from this event is divided into the following sections:

Key Questions For Progress on Online Harassment

The workshop opened with a morning discussion of key questions essential for achieving impact on online harassment. While our discussion focused on needs, we acknowledge that many hundreds of valuable resources have been published by advocates and researchers in the last 40 years of attention on conflict, harassment, and bullying online. Our statement of questions is built on great work that has already been done.

What are the social costs of online harassment?

We urgently need to understand the social costs of online harassment to guide our attempts to address and prevent the harms involved. With 73% of US adult internet users reporting that they observed harassment and 40% experiencing it personally, online harassment has touched the lives of roughly 140 million people in the US and many more elsewhere.1 Yet we lack basic understandings of who harassment impacts, what the harms are, and how to measure those costs. We also struggle to know how to ask good questions when talking to affected parties.

Without robust ways to understand or estimate the costs, platforms rarely know how to prioritize the impact of harassment on their missions and business models. As new populations come online around the world, we were left asking what the impact of this work might be on people who choose to avoid the internet altogether. What might the costs be for people who adopt current flawed approaches to online conflict as the default for their digital experiences?

To address this question at the workshop, one group worked on a taxonomy of harms from online harassment. One early-stage project tests the effect of witnessing harassment on bystanders. Another group developed qualitative and statistical methods to estimate the possible negative effects of harassment on platform participation and people's exercise of their speech rights.

Why do people participate in online harassment?

Efforts to address online harassment are deeply hampered by fundamental gaps in our understanding of who participates in online harassment and why they do so. Many current approaches label people as trolls, harassers, and abusers without substantial, systematic evidence grounded in the experience of people who participate in those behaviors. Research is urgently needed on the motivations of people who harass others, the social contexts of harassment, and the role of changing networks and situations in fostering conflict online. Without the capacity to understand and differentiate the factors involved in online harassment, any efforts to reduce harms or resolve conflicts will be little better than guesswork.

---

[1] Duggan, M. (2014). Online Harassment. Pew Research Center.

The main roadblock to understanding people who engage in harassment seems to be ethical research access to people involved in harassing behaviors. Many platforms keep internal records of people accused of harassment, but those records are not accessible to researchers. At the workshop, one group further advanced an academic archive of harassment reports, which could manage legal and ethical concerns around access to platform reports. One qualitative study asks what motivates online harassing behavior. Another project is developing predictive models of who is most likely to non-consensually share intimate photos online. All of these projects need further data access to deliver the greatest possible benefits.

How can we prevent and respond to online harassment?

Across law, platforms, and the hundreds of thousands of peer responders to online harassment, fundamental questions remain unanswered about prevention and response.

In the absence of randomized trials and other causal studies, some design interventions and machine learning systems have backfired at large scales, exposing many thousands of people to new problems rather than reducing the problems they were designed to address.2 Our workshop offered support to research on the effect of peer responses to racist slurs and political harassment, testing theories about what kinds of people are likely to reduce their racist statements, depending on the nature of the response. Since a small number of individual experiments cannot keep pace with the scale of online harassment or its rate of evolution, we also discussed infrastructures for experimentation in moderation and platform design that could grow our knowledge at the scale of these social problems.

"a small number of individual experiments cannot keep pace with the scale of online harassment or its rate of evolution"

Very little public knowledge is available about people who do respond to reports of online harassment, from platforms to volunteer moderators, peer responders, and bystanders. Researchers have a tremendous opportunity to support these stakeholders. But support cannot be meaningfully offered without greater public knowledge on the processes that shape moderation work, the policy and training behind it, the ideas that drive it, and the technical infrastructures involved. While our workshop did not explore these questions in detail, several attendees are doing in-depth research with volunteer moderators and peer responders. Others are working to establish industry groups that could make advances in the policies, infrastructures, and labor of responding to harassment reports.

---

[2] Cheng, J., Danescu-Niculescu-Mizil, C. & Leskovec, J. (2014). How Community Feedback Shapes User Behavior. ICWSM 2014.

Halfaker, A., Geiger, R. S., Morgan, J. T., & Riedl, J. (2012). The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. American Behavioral Scientist

One urgent gap in public knowledge involves the experiences of people who are on the receiving end of responses to alleged harassment. If we know little about the effect of moderation on the level of harassment in a platform or community, we know even less about the outcomes for people whose participation is removed, whose accounts are disabled, and who are banned entirely from platforms. Nor do we have much evidence on the effects of alternative processes for conflict prevention and resolution. Two participants are interviewing people who have been blocked and banned by peer-generated "block lists," but research in this area is held back by limitations in access to the people who platforms and moderators have acted against.

"a focus on social change can direct our attention toward deeper causes of conflicts and societal interventions to reduce and prevent those conflicts"

Many perspectives on online harassment focus on individual people rather than the social structures, networks, and cultures that contribute to harassment. We may not easily be able to hold cultures and social movements responsible for specific actions. Yet a focus on social change can direct our attention toward deeper causes of conflicts and societal interventions to reduce and prevent those conflicts.[3] At the workshop, we discussed ways that qualitative research could inform network analyses and interventions for social change.

How can we improve definitions of online harassment and its responses?

Poor, fragmented definitions of online harassment disrupt the possibility of widespread collaboration, argued many at the workshop. Some communities affected by online harassment have developed helpful language for the problems they face. Yet among researchers, platforms, policymakers, and advocacy groups, many fundamental conversations are impeded by the lack of common language to describe harassing behaviors, potential harms, and platform or peer responses. Consequently, each new social platform and online community re-invents its own way of understanding harassment, without the chance to learn from the experiences and practices of others. Poor definitions also impede the role that machine learning can play in advancing our understanding of online harassment, even as projects like the Wiki Detox system for detecting personal attacks are advancing the state of the art.

At the workshop, we discussed which kinds of definitions would be most valuable. One group worked on a taxonomy of harms from online harassment. Others discussed the challenges of surveying people for information about experiences that people may not see as harmful, especially in cultures that normalize violence against women and other marginalized groups.

---

[3] Paluck, E. L. (2012). The dominance of the individual in intergroup relations research: Understanding social change requires psychological theories of collective and structural phenomena. Behavioral and Brain Sciences, 35(06), 443-444.

We also discussed the risks involved in taxonomy-creation. Taxonomies of harassment might pathologize kinds of behavior that are protected rights, and diagnostic tools can encode injustice as easily as promote justice. By differentiating online harassment from research on sexual harassment, conflict, and social deviance, we distance the conversation from important intellectual resources. The work of definition-making also risks imposing predominantly Western notions on structures of power that shape the digital, social lives of billions of people elsewhere in the world. Taxonomies can allow computer scientists to turn complex issues into computable questions open to machine learning, but those same taxonomies can encode harmful biases into the fundamental assumptions of those systems.

"Poor, fragmented definitions of online harassment disrupt the possibility of widespread collaboration on online harassment."

Overall, the workshop revealed that meaningful definition work can unlock powerful cooperation around online harassment, but it needs substantial engagement from all affected stakeholders. This high-yield area will also require substantial commitment and resources from those stakeholders to achieve meaningful impacts.

A Note on Research Ethics

Throughout the workshop, participants discussed the research ethics of our specific research projects.This report does not offer any recommendations or guidelines. At the same time, we all agreed that careful attention to ethics is especially important in research on online harassment.[4]

Infrastructures to support research

Online harassment is widespread and occurs in a wide range of cultures. The scope and scale of harassment online—and the continuing evolution of platforms and harassment strategies—outpaces ongoing research in this area.To address this challenge, we discussed infrastructures of knowledge that could help knowledge evolve at the scale of the problem.

Archive of Online Harassment Reports

(Amy Johnson and J. Nathan Matias)

[4] Some participants suggested the Association of Internet Researchers guideline: (2012) Ethical decision-making and Internet research 2.0: Recommendations from the AoIR ethics working committee

Workshop discussions revealed a common need for research access to reports of harassment if we are to make meaningful progress on the risks that people face online. Current challenges/ limitations to existing research include:

- Our definitions of online harassment are not fit for this purpose because harassing material is often removed, and because researchers lack data on what people consider harassment. As a result, we struggle to know what questions to ask when defining harms, designing algorithms, and imagining fair and just conflict resolution.
- Research on the effects of harassment is skewed. While some people are vocal about their experiences, those who have been silenced are difficult to access (Sarah Sobieraj)
- Studying motivations and cultures of harassment is difficult because it is hard to access samples of people who have been accused of it (Alice Marwick)
- Studies on harassment stay one-sided because very few datasets support access to alleged harassers (Amy Bruckman)
- We can't easily study the outcomes and backlash from being blocked or banned because those users are typically removed from the public record (Amy Bruckman)

To overcome these problems, the archive will offer a secure repository of harassment reports that have been submitted to platforms by their users, allowing a carefully vetted group of researchers, designers, and possibly journalists to study online harassment as reported by internet users. Designed around in-depth discussions with online platforms and legal advisors over the last two years, the archive will accept data in two ways: directly from harassment reporters, and directly from online platforms, with user consent. Anyone whose report is held by the archive will be able to access their reports and manage the privacy of the data it holds. Having developed this project in conversation with platforms over two years, Amy and Nathan are recruiting advisors, seeking a host institution, and seeking expressions of interest as they start fundraising in 2016.

At the workshop, participants made progress on:

- Outlining high impact questions that this archive could help answer
- Refining the value to reporters of harassment
- Further scoping the archive's legal and ethical considerations
- Discussing best approaches to move forward conversations with platforms
- Aggregating resources and providing access that would be valuable to individual reporters of harassment
- Finding possible research partners

Contact Amy Johnson and Nathan Matias for more information.

Taxonomy of Harms from Online Harassment

(Derek Ruths, Susan Benesch, and others)

This group worked toward a unified approach to defining and discussing harassment. Our definitions of harassment have profound implications for the well-being and rights of millions of people. Yet as we realized during discussions on the first day of the workshop, different arbiters of harassment (especially online platforms and the US government) all have different notions of exactly what kinds of content constitute harassment. Users, too, have varied understandings of harassment, in part because of cultural context.

Since no single definition of harassment will serve all ends, this group proposed using harm as a currency for universalizing discussions of harassment—these better definitions could harmonize research on effective methods to diminish harassment and the harms that it inflicts. This taxonomy would be designed for use by platforms, governments, educators, researchers, and algorithms that need a way to clearly explain their definitions of harassment in ways that others can understand.

"better definitions could harmonize research on effective methods to diminish harassment and the harms that it inflicts"

Next steps from the group include:

- Continued meetings to refine the taxonomy of harm.
- Participation in conferences and venues where the group can continue to workshop the taxonomy with wider communities.
- Compiling cases of harassment that they can use to "test" the validity of their taxonomy.

Contact Derek Ruths and Susan Benesch for more information.

Detecting Personal Attacks with Machine Learning

(Camille François, Lucas Dixon, Nithum Thain, Ellery Wulczyn)

The Wiki Detox Project is a research initiative that aims to use machine learning and statistics to understand how attacking or "toxic" language affects the contributor community on Wikipedia.[5] The focus of the analysis is initially on talk page comments that exhibit harassment, personal attacks and aggressive tone. The project aims to:

1. Build machine learning models to detect toxic comments.
2. Determine how prevalent toxic comments are, who they target and what impact they have on user retention and productivity.
3. Facilitate discussion and implementation of tools to help address toxicity in Wikipedia (and potentially other interested communities).

---

[5] **Research:Detox** - Meta. Retrieved August 22, 2016, from meta.wikimedia.org/wiki/Research:Detox Research showcase video at www.youtube.com/watch?v=eZgqzVuRDRs

The team is in the process of releasing their initial results, datasets, and machine-trained algorithms. This workshop provided the team with feedback on their research approach, next steps, and possible collaborations for future research. The workshop provided insights on new questions this system could help answer, as well as guidance on concrete experiments to explore the demographics and potential biases hidden in crowd-sourced annotations. In sessions, the team was also able to link their work to related research from psychology.

Infrastructures for Experiments in Moderation and Platform Design

(J. Nathan Matias, Cliff Lampe, Eric Gilbert)

Very few attempts to reduce or respond to harassment have been evaluated with causal methods. With harassment styles and their corresponding responses changing frequently across many communities, we need capacities for knowledge that work at the same rate. Infrastructures of experimentation can offer the potential to grow our knowledge on harassment responses at the pace and scale of the issue.

Although we did not have time to workshop this topic, several participants have active projects to build infrastructures of experimentation:

- Nathan is building CivilServant, software for online platform users and volunteer moderators to run experiments to test the outcomes of their work to support community. He is also writing about the implications of citizen-led field experiments for policies in digital communications and beyond. The alpha version of CivilServant is complete, and Nathan is recruiting communities to do experiments together.[6]
- Cliff is working with Nicholas Diakopoulos on a grant proposal to develop partnerships with news organizations to test the effects of platform design changes on moderation work and harassment.

High Impact Research Projects

One main goal of our workshop was to support participants to make progress on high impact research projects related to online harassment. Participants brought their questions and ideas, and we worked to offer support, help them refine the ideas, and discuss next steps. Of the 18 projects that attendees brought, here are four that we are able to share publicly.

[6] Sample CivilServant experiment at osf.io/gn6kx/

Estimating the Chilling Effects from Online Harassment

(J. Nathan Matias, Sarah Sobieraj, Justin Cheng, Ellery Wulczyn, and others)

Does receiving hateful or antagonistic comments have a chilling effect on people's future speech? What demographic and psychological factors predict someone's risk of withdrawing their participation after receiving harassment? Insight into chilling effects could change how we interpret the first amendment, transform how we think of the harms and costs of harassment, and offer insights on who might be most at risk of these harms.

Answering this question is hard to do without forcing harassment upon people to observe the effects. However, matching-based causal inference methods[7] allow us to use historical data to estimate the effect of harassment on people's future speech and participation. Using historical data, surveys, and interviews, this project tests a fundamental assumption of online harassment discussions: that harassment has the effect of pushing women and people of color out of the public sphere, or at least reducing the rate at which they contribute.

"this project tests a fundamental assumption of online harassment discussions: that harassment has the effect of pushing women and people of color out of the public sphere"

At the workshop, several of us realized that we were trying to answer the same question using similar matching-based causal inference, and that we can support each other on study design:

- Nathan and Justin have developed an observational study for asking this question among journalists and opinion contributors, using article, comment, moderation data from newspapers, and surveys of those contributors
- For Sarah's book project, she is also asking this question by interviewing people who have faced harassment
- Ellery is designing an observational study to ask this question on Wikipedia, building on the classifier of personal attacks that Jigsaw and Wikimedia recently developed together
- One workshop participant has already begun to survey journalists about harassment, and we will be discussing ways to collaborate
- We worked to refine exactly what we mean by "future speech and participation," developing three specific, measurable outcomes.

Key next steps include:

[7] Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical science: a review journal of the Institute of Mathematical Statistics, 25(1), 1.

- Nathan, Justin, and Sarah need access to data from a content publishing platform (whether text, or video) that includes information about the comments or harassment received. They are reaching out to news organizations and other platforms.
- Sarah is looking for a way to contact people who have experienced various levels of reported or moderated harassment, so she can interview the people who tend to be silent about their experiences for her book. A platform with internal records on who has reported or received harassment may be able to help her contact people to interview without disclosing private data.

Testing The Outcomes of Peer Interventions Against Harassment

(Kevin Munger)

What kind of peer (nonadmin) sanctions are the most effective in decreasing the incidence of harassment? Admin responses are often heavyhanded and perceived as capricious or self interested, and can sometimes backfire. If we knew how individuals in a community could better encourage good behavior, we could empower them to clean up their online communities more effectively. This problem seems especially relevant to Twitter, with its mix of high profile individuals and anonymous accounts. Kevin has been been conducting randomized field experiments using bots to sanction bad actors and see what kinds of sanctions work.

“if we knew how individuals in a community could better encourage good behavior, we could empower them to clean up their online communities more effectively”

At the workshop, Kevin shared his experiment designs and worked together on new studies focused on reducing personal attacks across partisan lines during the US presidential election. The experiment will consider whether discouragements from personal attack are more effective if they come from an account that appears to have the same political affiliation.

Progress during the workshop included:

- Sharing resources on classifying Twitter accounts by political partisanship. For example, Derek Ruths has published research on political affiliation classification.[8]
- Sharing resources on classifying personal attacks by Jigsaw and Wikimedia
- Discussing other contexts and platforms where experiments might be meaningful

Key next steps include:

- Replicating Kevin's finding on other platforms and communities
- Generating new ideas for interventions to test

[8] Cohen, R., & Ruths, D. (2013, June). Classifying Political Orientation on Twitter: It's Not Easy!. In *ICWSM*.

What Motivates Online Harassing Behavior?

(Alice Marwick and Lindsay Blackwell)

What motivates people to harass others online? While there are many presuppositions and presumptions based on anecdotal evidence, there is a lack of empirical work in this area. While there is a body of research on specific forms of online misbehavior, such as trolling, flaming, and "griefing," harassment is a broader spectrum of behaviors, enabled by different technical affordances and motivated by diverse factors.

There is considerable data on the victims of online harassment, but without knowing why people harass others online, it is difficult to mitigate or prevent harassing behaviors. There is concern that this lack of data will result in brute force solutions that do not recognize or solve the underlying social factors that fuel online harassment.

Access to people who have been accused of harassment is a chief difficulty for Alice, since platform records of alleged harassers are private and because such research represents substantial risks for both researchers and harassment victims. While she looks for access to a more representative sample, Alice is conducting focused studies of specific communities associated with harassment. Lindsay is conducting experiments to investigate the underlying social and psychological factors that may motivate people to harass others online, as well as the factors that contribute to an individual user's understanding of what "counts" as online harassment.

"Qualitative studies offer tremendous value by revealing the structural and cultural factors in play, offering deep insights for the design of harassment prevention and response."

The workshop included two participants who have been interviewing Twitter users added to public block lists, and another who has done ethnography with street gangs. Participants discussed the challenges and importance of building out general knowledge from many in-depth, context-specific studies. Qualitative studies offer tremendous value by revealing the structural and cultural factors in play, offering deep insights for the design of harassment prevention and response. Many such studies are needed to better understand the full spectrum of motivations behind harassing behaviors online.

Key needs include:

- Help identifying appropriate research contexts and cases
- Access to those contexts
- Support for developing ethical recruitment procedures that also protect the researcher

Predicting Who Is Most Likely to Non-Consensually Share Photos Online

(Reynol Junco and Amy Hasinoff)

Since the moral panic about sexting began in 2008, researchers, educators, and legislators have been scrambling to find ways to prevent sexting. However, like abstinence-only sex education, these approaches have largely failed—around one-third of high school students are still engaging in sexting despite being aware of the risks. Much of the empirical research on sexting is based on the assumption that there is something unusual and troubling about sexters. A number of studies examine correlations between sexting and other perceived risky or unsafe behaviors, such as alcohol consumption and unprotected sex, and the links to undesirable personality traits, such as impulsivity. One review article found that 79 percent of the studies published on sexting emphasize the risks and link sexting to negative outcomes. The implicit question in a lot of the existing research on sexting is: What kind of person would sext? And what exactly is wrong with this person?

This project is one of the first to focus on the prevalence, correlates, and prevention of sexual privacy violations. This unique form of digital abuse, which is the key problem with sexting, has to date been directly examined in only a handful of studies and is not yet well understood. The dozens of studies on sexting typically do not distinguish consensual sexting from violations of privacy and instead view all forms of sexting as a deviant and risky behavior. This project remedies this oversight by focusing our attention on understanding sexual privacy violations in order to ultimately help prevent this form of digital abuse. The project collects trace data from cell phones as well as survey data on empathy, maladaptive attitudes about women's willingness to engage in sexual activity, and sexual satisfaction to develop models to predict who will engage in a sexual privacy violation. Furthermore, these models will use trace data as proxies for the measurement of more complex constructs like empathy in order to provide information about how we might use easy-to-obtain data to identify and educate those likely to engage in such sexual privacy violations.

At the workshop, participants worked to refine the definitions and measures used in the predictive model. While this project has a dataset and is underway, this dataset is small, and substantial advances could be made with further resources.

Key needs include:

- Partnerships with platforms that might test these predictive models at scale
- Discussions with platforms and other key stakeholders on the practical application and oversight of predictive models for these behaviors in an evolving legal landscape

Acknowledgements

The authors of this report are deeply grateful to Jigsaw, who funded the workshop, to Ellie Klerlein and Dennis Polin of Spitfire Strategies, for offering strategy advice and facilitation, and to Emilie Reiser, Nicole Freedman, and Lorrie LeJeune of the MIT Media Lab, for organizing and managing many of the workshop logistics. Thanks, everyone!