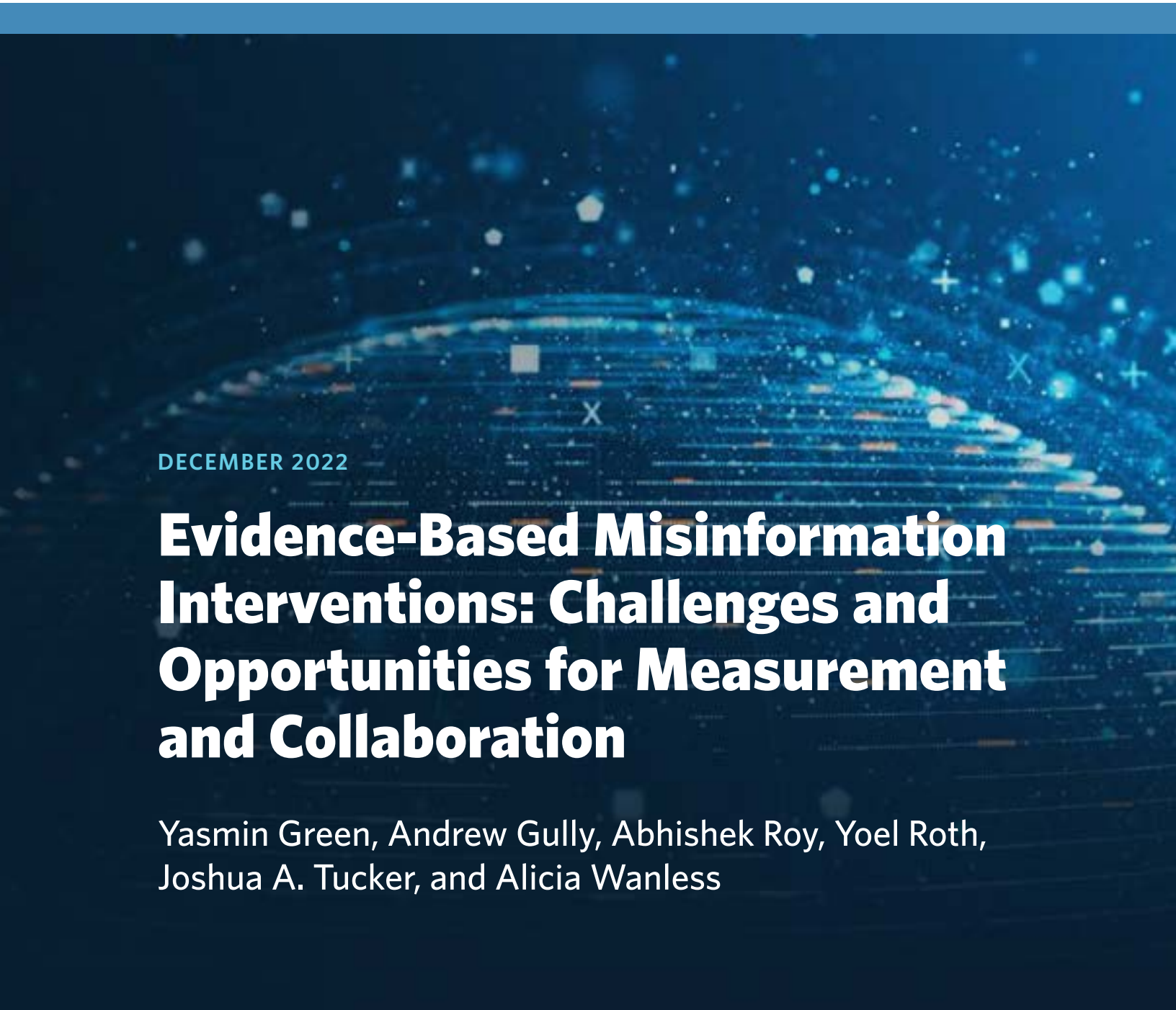
CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

ESOC
Empirical Studies of Conflict



DECEMBER 2022

Evidence-Based Misinformation Interventions: Challenges and Opportunities for Measurement and Collaboration

Yasmin Green, Andrew Gully, Abhishek Roy, Yoel Roth, Joshua A. Tucker, and Alicia Wanless

Evidence-Based Misinformation Interventions: Challenges and Opportunities for Measurement and Collaboration

Yasmin Green, Andrew Gully, Abhishek Roy, Yoel Roth, Joshua A. Tucker, and Alicia Wanless

This paper was funded by generous support from Microsoft, the John S. and James L. Knight Foundation, Craig Newmark Philanthropies, and the William and Flora Hewlett Foundation.

© 2022 Carnegie Endowment for International Peace and Princeton University. All rights reserved.

Carnegie does not take institutional positions on public policy issues; the views represented herein are those of the author(s) and do not necessarily reflect the views of Carnegie, its staff, or its trustees.

No part of this publication may be reproduced or transmitted in any form or by any means without permission in writing from the Carnegie Endowment for International Peace. Please direct inquiries to:

Carnegie Endowment for International Peace
Publications Department
1779 Massachusetts Avenue NW
Washington, DC 20036
P: + 1 202 483 7600
F: + 1 202 483 1840
CarnegieEndowment.org

This publication can be downloaded at no cost at CarnegieEndowment.org.

Contents

Contents Continued

Summary

The lingering coronavirus pandemic has only underscored the need to find effective interventions to help internet users evaluate the credibility of the information before them. Yet a divide remains between researchers within digital platforms and those in academia and other research professions who are analyzing interventions. Beyond issues related to data access, a challenge deserving papers of its own, opportunities exist to clarify the core competencies of each research community and to build bridges between them in pursuit of the shared goal of improving user-facing interventions that address misinformation online. This paper attempts to contribute to such bridge-building by posing questions for discussion: How do different incentive structures determine the selection of outcome metrics and the design of research studies by academics and platform researchers, given the values and objectives of their respective institutions? What factors affect the evaluation of intervention feasibility for platforms that are not present for academics (for example, platform users' perceptions, measurability at scale, interaction, and longitudinal effects on metrics that are introduced in real-world deployments)? What are the mutually beneficial opportunities for collaboration (such as increased insight-sharing from platforms to researchers about user feedback regarding a diversity of intervention designs). Finally, we introduce a measurement attributes framework to aid development of feasible, meaningful, and replicable metrics for researchers and platform practitioners to consider when developing, testing, and deploying misinformation interventions.

Introduction

Users of social media continue to be confronted with misinformation despite the progress made by major social media companies in scaling enforcement and ranking approaches to addressing the harms caused by misleading or inaccurate information online.[1] Research suggests both that informed users can slow the spread of misinformation[2] and that users want the tools to make these judgment calls for themselves, but we lack a robust foundational understanding of how to achieve this goal.[3] In this paper, we explore how greater collaboration between research communities—in particular, those inside technology platforms, the academy, and civil society—can accelerate progress toward empowering users to be safe and informed online.[4]

Bridge-building is necessary to unlock what types of interventions are best suited to address threats within the information environment, particularly in the context of democracies. One way forward to fostering evidence-based decisionmaking to tackle misinformation online is to establish a shared understanding of the aims of interventions and the metrics for assessing them. With advances in platform data access to external parties emerging through the European Union's Digital Service Act and the European Digital Media Observatory's outline guiding such a regime, this paper invites both academics and platform researchers to engage in a dialogue about how measurement research can evolve through collaboration.[5]

Research communities have different perspectives, priorities, and practices. In the context of studying misinformation interventions, taking stock of what generally sets research constituencies apart can illuminate the untapped opportunities for collaboration between them, or at the very least should suggest ways for overcoming existing barriers to collaboration. To

generalize, academics tend to be motivated by a basic scientific understanding of a phenomenon—in this case the consuming, engaging with, and sharing of misinformation, and secondarily interested in the design of platform features to affect those phenomena. Platform researchers, on the other hand, are primarily motivated by addressing problems on their own product surfaces and secondarily interested in advancing science. Similarly, research communities prioritize different contexts for impact: academics are inclined to optimize for advancing scientific understanding with high-quality and thoroughly cited publications; platform researchers aim to improve or inform new feature design.

Unsurprisingly, research communities lack a common framework for conceptualizing human behavior and for evaluating the efficacy of the online tools and strategies that aim to affect that behavior; we also miss opportunities for collaboration to develop these frameworks. Compounding this challenge, most research conducted internally by the platforms will only make its way into the public domain if the platforms choose to release the research publicly. This is similar to the "file drawer" problem in academia, where less interesting (and often null) results often fail to be published, causing bias in the overall accumulation of knowledge.[6]

External access to user data is front and center in the debate about the role of tech platforms in supporting public interest research, but is not the primary focus of this paper. There is important work happening on this front and the ultimate scope and governance around platform data sharing will be driven at least in part by regulation.[7] Instead, this paper attempts to advance the dialogue between researchers across sectors, recognizing that data sharing is uneven across platforms (and likely to remain so absent regulatory changes). We begin with the hope that the goals and motivations of researchers may productively align despite this unevenness.

Focusing on the development and evaluation of user-facing interventions,[8] we explore the following key questions that would benefit from collective problem-solving across research communities:

- How do the goals of developing meaningful and feasible measures factor into the selection of outcome metrics by researchers within and outside the major online platforms?

- What factors affect the impact of on-platform interventions that are not at play in lab settings, and how should measurement approaches account for them?

- Given the current data-sharing environment, what opportunities exist to improve the transferability of analyses across contexts and researchers?

- Given the highly sensitive nature of research and experimentation concerning informational dynamics and misinformation, how can we improve upon the current consent practices across sectors to protect and empower users as potential research subjects?

Outcome Measures

Research on misinformation intervention today occurs within the silos of tech firms and large platforms and within academic, nonprofit, and independent research circles. While the ultimate goals of most (if not all) of those working in the field include reducing the spread and consumption of misinformation online, the specific questions asked, methodologies, and motivations may differ considerably. Because of these inherent differences in motivations, platforms and external researchers may prioritize differently the types of studies they conduct and, indeed, the very outcomes that are measured.

Research in this space generally falls into one of three types of experiment: small-n lab studies, which may examine the same subjects over time; live field experiments to simulate, with constraints, the act of changing features of the information environment; and live on-platform experiments. While platforms may conduct all three types of experiment at different stages of research or product design, the third category (on-platform experiments) almost always falls exclusively within the domain of experimental research conducted inside tech companies using privileged data access, much of which goes unpublished.[9]

Challenges in Defining Misinformation

Designing and measuring the impact of interventions to counter misinformation reveals a fundamental challenge within the field: determining which content qualifies as misinformation and which does not.[10] Without some ability to identify and categorize the content in question (for example, misinformation), such that interactions with or perceptions of the content may be observed, researchers will have no ability to properly observe treatment effects.

Determining if a message or behavior is false and misleading is often a nuanced challenge whose outcome depends on context-specific examples. Consensus around underlying facts evolves over time, meaning that any decisions to adjudicate misinformation today could be revisited when new information arises. Facebook faced this challenge when it introduced a policy classifying claims that the coronavirus was human-made, labeling such content as violative, only to reverse that decision later.[11] It can be a challenge to apply such definitions in practice, where at-scale solutions require clear criteria against which an intervention can be designed.

Within academic scholarship, researchers have taken numerous approaches to manage this problem, such as relying on fact-checkers to make true/false decisions using their own judgment or leveraging compiled lists of problematic domains when attempting to measure the effects of interventions on misleading content.[12] Bounded definitions of misinformation assist academic researchers in producing replicable research, and third-party definitions help mitigate potential biases that could be introduced from simultaneously defining and analyzing problematic content. Platforms, in comparison, are necessarily operating outside the constraints of any one individual study and are necessarily charged with creating and enforcing policies that are both rigorous toward known harms and capable of covering future ones.

The central element of platform definitions of misinformation is the criterion that content be untrue, misleading, or deceptive—although the precise mechanics of this definition, and who is empowered to make that determination, varies significantly from service to service. For example, TikTok defines misinformation as "content that is inaccurate or false."[13] Twitter similarly targets statements that "advance a claim of fact, expressed in definitive terms" and are "demonstrably false or misleading, based on widely available, authoritative sources."[14] YouTube, in its misinformation policy, illustrates where misleading or deceptive content could pose a serious risk of egregious harm—especially when such content is aimed at promoting dangerous remedies or cures, suppressing census participation, distributing hacked material, or interfering with participation in democratic processes.

Most large online platforms today also incorporate in their continually evolving misinformation policies an emphasis on harm. Platforms may be clear that content is false and therefore misinformation, yet unclear on the potential to cause harm. They may be further unclear internally or collectively on what is the appropriate response (particularly given differences in the product capabilities associated with misinformation interventions at each platform). This lack of a commonly accepted and operationalized definition of false and misleading information makes it difficult, if not impossible, to accurately compare the efficacy of different interventions across studies or platforms. These definitional problems also complicate translation of academic research to on-platform tests if researchers and platforms have differing definitions of problematic content or are using content examples drawn from outside those platforms entirely.[15]

In the absence of a single common definition of misinformation that satisfies researchers, platforms, and the public, approaches that emphasize transparent and transferable content classification and measurement strategies may provide effective means for platforms and researchers to work toward shared outcomes.

Platforms, for their part, could consider publishing canonical lists or examples of content that meet their own definitions of problematic content and behavior for post-hoc research. This would incentivize the study of interventions that are optimized directly for platform

response and would enable a more informed, mutual conversation with external researchers. Facebook and Twitter have done this as part of their efforts to publish information about coordinated activity from inauthentic accounts.[16]

When human raters are involved in the evaluation of content, platforms also should strive to publish rater guidelines and data sets that make these judgments replicable. Google Search, for example, publishes the guidelines that humans use to evaluate pages for search quality.[17] Platforms will, however, need to weigh the potential risks of releasing data that could provide adversaries with insights to work around misinformation filters.

Academic research institutions are investing in similar areas, such as demonstrating the potential application (or limitations) of crowd-sourced judgments of information quality.[18] If proved successful, these would provide additional avenues for platforms and researchers to collaborate on research involving information quality decisions in an open and transparent manner.

Attributes of Ideal Metrics: Feasible, Meaningful, and Replicable

For platforms or researchers that intend to evaluate the impact of interventions on people's assessments of information credibility, there are several attributes of ideal metrics to keep in mind. Depending on the goals of the intervention in question, measures should be feasible, meaningful, and replicable.[19]

The feasibility of a measure is determined by whether it is technically and legally possible to compute and is aligned with user expectations and the design of the product. Meaningful measures are those that quantifiably demonstrate that a change in platform performance produced changes in the attitudes or behaviors of users. Replicable measures are those that achieve consistent results among a target user population. Researchers inside and outside of the major platforms have different perspectives on these three attributes, given their vantage points and capabilities.

The choice of any particular metric must begin with the objectives of the intervention in mind.[20] Researchers may be interested in motivating people to more correctly identify the veracity of content, or more specifically, the ability to discern true from false content. Other researchers may be interested in correcting false beliefs. Still others may be interested in impacting future browsing behavior.[21] And some will want to focus not on beliefs, but on

sharing. The goal of platform researchers and product designers is to reduce people's exposure to misinformation and, when eliminating exposure is impossible, to give them the tools necessary to correctly identify misinformation as such.

Large platforms, because of their bias toward studying user outcomes that are observable at scale, and their privileged access to large data sets that allow for such analysis, tend to favor outcomes that fall along the causal chain of events between a user being exposed to a specific piece of content and a feasibly observable intent toward or engagement with that content. Such outcomes could be defined with measures that observe rates of sharing, clicks on pages, dwell time on content, or subsequent browsing behavior. Additionally, measures that are observable at scale also are replicable such that they enable key decisionmakers or executives to track the impact of interventions over time in order to evaluate overall product heath and performance of the intended feature. These measures, once defined, also have the added benefit of integration into automated A/B testing systems that directly impact product decisionmaking (such as through exposure in executive dashboards or briefings), providing a feedback loop of incentives for individual researchers to have measured impact on users at scale.

External researchers, on the other hand, do not have access to large-scale, high-resolution data that might enable them to study at scale the holistic relationship between user exposure and engagement. Academics cannot simply run large-scale observational studies using platform data whenever they choose, but rather are limited to data made available by platforms. Without the cooperation of the platforms, academics are unable to run causal studies in the way that platform researchers do because they lack the ability to manipulate platform users' experiences and because platforms don't release data on users' exposure to different interventions. While replicability is important to the broader scientific community to be able to independently and reliably say that the results are indeed true and replicable, these standards may or may not be met among platform researchers. For the latter, replication of a phenomenon outside their platform generally is a secondary concern or even a nongoal. For instance, an intervention tested on Twitter may aim to be internally reliable, such that repeated experiments will yield the same results among Twitter users; however, such a design is not necessarily intended or designed to work on any other platform or user population.

Behavioral scientists, in particular, have utilized small-n lab studies that often focus on how belief formation mediates content engagement following exposure, either immediately or after several days or weeks. Grounded in theory, these studies allow for advances in basic research that inform online intervention design. Additionally, external researchers are more likely to be incentivized to make contributions to foundational theory and advance scientific knowledge; this incentivization may or may not have direct and immediate applied impact on internet products and services.

Examples of Ideal Metrics in Practice

Due to the general lack of access to internal platforms and the need to closely collaborate with platforms to conduct platform manipulations, academic researchers often lean toward controlled lab studies with specific outcome measures that are clear, easy to understand, and relatively simple to calculate. Examples of such metrics that have been proposed in the literature are discernment,[22] the ability of users to identify false from true content;[23] rates of sharing content (either observed or self-reported);[24] or the identification of a given piece of content's underlying rhetorical strategy.[25] It can be said that these metrics are meaningful, feasible, and replicable since they measure an outcome of interest to the researcher, can be readily measured in a controlled observational setting, and are transparently published.

With a mandate to evaluate the efficacy of new online strategies and access to full-scale platform data, researchers inside the platforms can study behavioral outcomes and attribute them to changes in features (for example, Gina Hernandez, "New Prompts").[26] They can also study indicators that are not observable to external clients (for example, YouTube Blog, "The Four Rs of Responsibility").[27]

Metrics may be chosen by platform researchers for combinations of reasons: because they fit within product-specific design considerations and user expectations; because key business leaders recognize and understand their importance; or because engineering decisions make it practically more feasible to measure one metric over another. While platform metrics may occasionally mirror or look similar to those used in smaller-scale lab studies, a key difference and consideration for platforms are the previously mentioned definitional attributes of what constitutes misinformation. Making decisions about perhaps several dozen content examples to conduct a lab study requires significantly different considerations than identifying and classifying that content at scale.

Measures of online user behavior are distinct from measures of offline real-world outcomes, which are less frequently studied in academic or platform contexts. Even if online interventions do shift user behavior, such as through changing the spread of online misinformation, there may be no real-world effect of such interventions. For example, rates of vaccine adoption may remain unchanged.

While there is a critical need to assess the causal relationships between online interventions and real-world behavior, it remains difficult and rare. A recent review of 223 studies examining misinformation countermeasures found only one study that attempted to link the impact of countermeasures to subsequent real-world behavior.[28] This likely is because studies that measure real-world outcomes are more complex to execute. It is difficult to establish causal inference, which requires the identification, follow-up, and potentially observation of the treated population in the real world. For their part, platforms may shy away from studying off-platform behavior for fear of pushing the boundaries of expected conduct by platforms (such as the privacy concerns inherent in evaluating offline behavior).

Improved collaboration with platform research partners may increase the likelihood that misinformation intervention research will apply to platforms and influence platform decisionmaking. Closer collaboration may also reveal to nonplatform researchers why, for instance, individual products within the technology industry (for example, Facebook Feed, Google Search) may differ in choosing one measurement strategy over another. These products must consider each feature's purpose, its users' expectations, and even their core values when attempting to evaluate the efficacy of misinformation interventions. This rationale often is opaque to both end users and researchers.

One interesting new step in this regard is the U.S. 2020 Facebook & Instagram Election Study,[29] a first-of-its-kind collaborative research effort between a team of over fifteen external academics with internal Meta researchers, engineers, and project managers to conduct over a dozen pre-registered observational and experimental studies aimed at understanding the impact of Facebook and Instagram on the 2020 U.S. elections.[30] The researchers identified four key areas of interest: "Political participation, political polarization, knowledge and misperceptions, and trust in US democratic institutions." In December 2021, Twitter announced a new consortium of external researchers to which it plans to disclose data about content moderation and platform governance issues, with the stated goals of providing "data-driven transparency" and encouraging public-interest research.[31] More collaborations like these are needed to meaningfully bridge the gap between the impactful work being done outside platforms and the ways in which these and other ideas may be implemented to help users.

Estimation of Treatment Effects

In order to estimate the impact of a counter-misinformation intervention on an outcome measure, we must establish a causal relationship between the two. This causal effect of an intervention or treatment on the outcome measure(s) is called the "treatment effect." Approaches to designing experiments can broadly be classified into two categories: experimental approach (where the experimenter controls randomization of assignment of users to treatment and control) and quasi-experimental studies (where the experimenter has little or no control over randomization of assignment of users to treatment and control).[32] In this section, we discuss some considerations and challenges involved in the design of experiments, how they differ between practitioners working at digital platforms and those working in academia or the public sector, and how these challenges might be addressed.

Considerations for Experimental Approaches

Assessing the efficacy of an intervention typically involves measuring the average causal effect of a treatment on units of population. The most robust form of causal inference can be drawn when the mechanism for assignment of the participants to the treatment groups is not dependent on factors endogenous to the system being analyzed. A common approach, both to industry and academia, to such measurements is the randomized controlled trial (RCT) or "A/B test." RCTs typically involve measuring the difference between the potential outcome with and without the treatment by randomizing the allocation of treatment in the experiment by a chance mechanism.[33] In theory, if designed and executed properly, since this difference in treatment is induced by factors exogenous to the system being analyzed (random allocation), the difference in the outcome measure can be attributed to the intervention. For example, in Pennycook et al. (2020),[34] participants in the experimental trial were randomly allocated to a treatment and a control group, where the presentation of accuracy reminders (that is, judging the accuracy of non-coronavirus-related headlines) nearly tripled the level of truth discernment for subsequent headlines in the treatment group compared to those in the control group who didn't receive an accuracy prompt. In this section, we discuss some salient challenges and considerations for researchers when trying to estimate these "treatment effects" and compare them between research conducted in a lab setting and research conducted on a live product.

One worry in any RCT is the possibility of a spillover effect, whereby an untreated unit in the experiment is impacted by the treatment of another unit. Lab experiments are generally the most secure from such threats to validity of inference, due to the fact that researchers can closely monitor the environment of the experiments. On the other hand, field experiments are the most at risk to spillover effects, as by definition they take place "in the wild." Most experiments run by platforms generally take place in the field, because platforms often run live A/B tests of potential new product features with a small representative section of their traffic to determine whether those features should be deployed.[35] In the lab environment, unmodeled spillover effects can be curtailed, but this is difficult on large social network platforms as a change in the behavior of user A could have spillover effects on the potential outcomes for user B if they are connected in the same network.

For example, in a paper exploring potential causal links between virality of misinformation and echo chambers—isolated networks of users with similar dispositions about a topic—Petter Törnberg found that a polarized network (or presence of echo chambers) increases

the likelihood of spread of misinformation (because misinformation seems more authoritative and potentially also because users feel the need to conform to their group's point of view) suggesting presence of "network effects."[36] This makes the task of causal inference additionally difficult as any intervention that affects the outcome measures of a treatment group user within an echo chamber network (like reduction in number of false claims seen by the user on the feed) would affect the potential outcomes of all users connected to them (some of whom might be in the control group). Moreover, the strength and directionality of connections between users might also have implications for the experiment design, as not all users are equally influential in a social network and not all users are influential on all their network peers.[37]

Finally, it is hard to replicate the full effects of a social media feed—or a content recommendation feed—on the attention of the user in a lab setting (for example, by using a simulated static feed to assess the potential effect of the intervention on the subject). By contrast, in live A/B tests, since a small fraction of daily active users view the intervention in their organic feed, such attention effects are also accounted for. For example, users may exhibit lower recall for a media literacy banner intervention on a live feed as compared to a simulated feed in a lab since they have more familiar and interesting content competing for their attention.[38] An increase in perceived workload for the user can thus lead to what is often called banner blindness, which it has been suggested may affect the recall of the salient stimuli for the user over time.[39] As such, there is both scope and need for collaboration between industry and academia to study how lab experiments may inform complex mechanisms in the wild, such as those involved in interventions designed to affect how users process salient stimuli.

Considerations for Quasi-experimental Approaches

In some cases, randomized allocation of the treatment and control conditions may not be feasible, equitable, or otherwise desirable, such as where geographic location determines exposure to a treatment (for example, a public information campaign run by a local government agency) or individuals self-select into the treatment group (and therefore make random allocations infeasible). In such cases, investigators might consider a type of observational study, also known as a quasi-experimental approach, to measure average treatment effects. While quasi-experimental approaches have their advantages (like lower costs), they also have some major disadvantages as compared to experimental approaches in that they are

vulnerable to selection biases and therefore more challenging for making causal inferences. We briefly discuss some of the challenges practitioners face when deploying such approaches to estimate treatment effect.

Quasi-experimental designs try to address the concern of selection bias by attempting to mimic randomization by constructing a control group that is as similar as possible to the treatment group in terms of observable baseline pre-intervention characteristics. This acts as a counterfactual and aims to capture the outcome that would have resulted if the intervention had not been implemented. Hence, the treatment effect is calculated by measuring the difference in outcomes between the treatment group and the control group.[40]

As an illustrative example, to measure the effectiveness of a public advisory and counter-misinformation labeling campaign during a disease outbreak scenario in a country,[41] a social network might construct a quasi-experimental study by comparing the change in some predefined outcome measures (like recall of the counter-misinformation label, click-through rates or shares for labeled articles, and more) for users with prior exposure to nonlabeled posts who were then exposed to the labeling campaign compared to a control group of similar users who were exposed to the labeling campaign without any prior exposure to nonlabeled posts (identified based on a score-matching technique like propensity score–matching[42] or Mahalanobis distance matching).[43] The primary challenge in conducting such studies, faced by both researchers at digital platforms and those working in academia and public sector, is that they may lack sufficient evidence to establish causality at the end of the study or may not account for reverse causality or two-way causal relationships. For example, users who were not exposed to an online travel advisory ad campaign related to a hypothetical disease outbreak might receive such information from other sources like news on television or friends and family.

Some recent innovations have been made in studying online user behavior, particularly with respect to tackling misinformation, by conducting randomized experiments using a hybrid lab-field approach. In such studies, online treatment allocation is randomized within survey experiments or directly on social media platforms (administered by direct messages, public posts, or social-tie invitations), and then the subsequent social media behavior of the participants is observed. Since this study setup would involve not making the users explicitly aware that they are participating in an experiment, there are several ethical considerations with this approach beyond the obvious need to protect the user's privacy and avoiding user exposure to harmful content. These considerations have been discussed in detail by Mohsen Mosleh, Gordon Pennycook, and David G. Rand in their 2022 work on "Field Experiments on Social Media."[44] An alternative approach—employed in the U.S. 2020 Facebook & Instagram Election Study—is to enroll subjects with their consent into studies with experiments carried out on-platform and then evaluated with both survey and on-platform measures.[45]

User Experience Research

In the nascent field of countering misinformation, where new interventions are often without precedent, it is imperative for platform researchers to check foundational assumptions about their ease of use, or their "usability." User experience (UX) research aims to help uncover issues that interventions may cause for some users relating to their understanding and perceptions of the design, and how interventions may impact their views on associated products or features.[46]

Applying UX thinking and evaluation to proposed interventions can help inform the iterative process of designing interventions to work for on-platform environments. While academic research on misinformation interventions closely evaluates the effects of various approaches on attitudes or behaviors about the misinformation itself, there tends to be less scholarly research about the users' perceptions of the interventions.[47] Platforms always and continuously conduct UX research to gather novel insights into the preferences, pain points, and utility of new features, including misinformation interventions. This research is largely uncontroversial. Greater sharing of it may help the academic community tailor interventions research or help answer basic questions around the user-perceived utility of such designs.

Product teams must consider user perceptions and expectations from specific features. For example, a misinformation intervention that was highly effective by some measure (for example, reducing the spread of misinformation) yet greatly disliked by users would be unlikely to be prioritized for development compared to features that could strike a balance between the two. In the context of developing consumer products, developers and policymakers must consider not only what is effective at mitigating a problem, but also what will be received positively and enthusiastically by a user base. Academic research that measures user likability in addition to other key metrics of interest would increase the probability that these ideas would be further tested, and potentially adopted, by industry.

There are many options for UX research. For example, platforms often conduct qualitative user research studies involving a small group of users who are shown a new feature or application while observers watch, listen, and take notes. Alternatively, survey studies may be conducted to measure the likability or perceived utility of a feature, or sentiment toward it. These questions can be added directly to intervention lab studies, providing an opportunity for researchers to gather data about how users may react to an intervention. Platforms, with their vast UX teams and deep knowledge on user expectations, could openly publish design guidelines and product-specific user expectation principles so academics who are researching misinformation interventions could, if they so choose, conduct research informed by the feasibility of product implementation.

Finally, user research will allow researchers to learn whether users understand a new policy and resulting enforcement action and whether they recognize and perceive a new contextual feature in the way it was intended. Such research can offer opportunities for researchers

to optimize interventions that balance advancement of scientific knowledge, impact, and user perceptions. Evaluating users' understanding of interventions is also an essential part of researching efficacy—preliminary small-n interviews with mockups can reveal design shortcomings, a potentially key consideration in misinformation interventions such as labeling that may be broadly divisive or poorly understood.[48] User-centered research thus helps avoid developing a flawed theory of change from shortcomings in intervention design and can ultimately improve the efficacy of translating promising lab studies to effective online interventions.

Unintended Consequences

An awareness of potential biases and blind spots can help improve experimental design and is imperative when moving from the lab to the applied context. While both lab and on-platform research allow for the study of treatment effects, the opportunities for discovery of unintended consequences are far greater with on-platform experiments.

Research proposals to study intervention efficacy typically include a theoretical framework that explains the hypothesized impact on affected individuals. Because the lab environment is by definition a simplification of the live environment, there will be less opportunity for researchers to observe interaction effects between variables of interest, which may affect how well findings in the lab translate to platform deployments. The large differences between the user populations of each platform (for example, the younger demographic of TikTok users compared to Facebook users) are also likely to cause outcomes to differ depending on the product context in which the intervention is deployed.

For example, academic studies suggest that fact checking features—which aim to provide users with authoritative context alongside search results or online posts about dubious claims—can reduce false belief.[49] But in certain circumstances they can lead to an "implied truth effect," whereby labeling a subset of false information as "false" promotes overconfidence in the accuracy of unlabeled information.[50] The implied truth effect exemplifies the potential for unanticipated, undesired dynamics associated with deploying an intervention that is typically studied in a lab environment at scale. With more dialogue between the academic and platform research communities, we can develop shared knowledge of how interaction effects play out online that can inform both academic interventions research and appreciation of platforms' decisionmaking variables.

Individuals have different needs, preferences, and risk profiles with respect to information online, and studies examining aggregate effects may miss patterns that exist among subgroups. Research indicates, for example, that information literacy differs for groups with low versus high cognitive reflectiveness,[51] that elderly people are more likely to share

misinformation,[52] and that some communities (for example, service members[53] and those with racial grievances[54]) are at heightened risk of being targeted by misinformation campaigns. Any average treatment effects across the general population will likely mask important nuances among communities, including potentially adverse effects.

Academic researchers with the expertise and mandate to explore the likelihood of differential outcomes can use their vantage point to anticipate where population-level effects may vary within subpopulations, such as across demographic factors like age, ethnicity, gender, or political affiliations. Platforms, though they may have enormous volumes of user data, are often constrained by data collection practices and internal policies from collecting very specific demographic information that is commonly gathered via the battery of questions completed by research subjects in lab studies.

Platforms therefore may not have the data available from live studies to analyze differences in outcomes at a demographic level, while researchers working in the lab environment can more carefully control for demographics. While the latter context may allow for a better control of confounding variables and thereby a clearer understanding of the treatment effect in the lab environment, the prohibitive marginal cost of recruiting a large pool of participants across demographic groups can often hamper the generalizability of such findings.

As we have seen around the globe, online interventions to tackle misinformation can be sensationalized and manipulated by political actors. In India, members of the ruling party accused Twitter of selective targeting when their tweets were labeled as "manipulated media."[55] Russia threatened YouTube with "retaliatory measures" following the removal of its German-language state media RT channels for containing COVID-19 misinformation,[56] and YouTube faced twin lawsuits in the United States in 2019 from groups on both the right and the left of the political spectrum for labeling content "restricted."[57] Similarly, authoritarian actors seeking to control the information environment have abused interventions for their own benefit, as was the case during Syria's civil war when Bashar al-Assad's government attempted to censor opposition voices on Facebook by falsely reporting their content for policy violations and copyright infringement.[58] Regardless of its lack of empirical basis,[59] such politicization risks undermining the intended impacts of interventions on the users they were designed to help.

Though the problem of unintended consequences is unbounded, platforms might anticipate and build in mitigations through "red teaming," a process that's commonly deployed in technology companies, in which teams are challenged to take an adversarial approach to identifying weaknesses in a system. Indeed, platforms have embraced red teaming to help identify hacking vulnerabilities,[60] and even deepfakes that might propagate misinformation.[61] Scenarios concerning proposed broader misinformation interventions might also benefit from this type of exercise.

Platforms can also address the limitations of their own expertise and potential bias by expanding fellowships, research grants, and engagements with a broad and diverse representation of users and experts from across geographic, ideological, and political perspectives.

Consent

Approaches to obtaining consent from subjects studied in research differs considerably between projects conducted by industry and academia. Often the Terms of Service (TOS) to which end users agree as part of using a digital platform is the basis for gaining consent of participants in on-platform live experiments, with some platforms seeking further consent from users to be part of deeper studies.[62]

In Facebook's current TOS, for example, the platform says user data is used "to develop, test and improve our Products," as well as "to conduct and support research and innovation on topics of general social welfare, technological advancement, public interest, health and well-being."[63] Such clauses are vague and tell users little about what data they generate and how it is used by companies for research purposes. Moreover, most people do not read the TOS; some surveys have suggested that upwards of 90 percent of users click yes without reading the terms,[64] although the way the agreement is presented can increase consumption of it.[65] According to some surveys, social media users are not aware that their public posts are used in research and feel that they should be asked explicitly before such data can be studied.[66] That being said, most citizens are also not aware that their votes or employment statuses are used in research all the time, and few would suggest that we need consent from voters or workers to study election results or unemployment trends over time.

In contrast to the platforms' research consent norms, academics often rely on institutional review boards (IRBs), many of which consider publicly available social media data not to be related to human subjects. This may exempt researchers from seeking informed consent from participants, raising ethical concerns around the use of such data, particularly in studies of interventions that do not involve direct engagement with end users.[67] This exemption has led individual researchers or research teams to navigate their own ethics for using social media data.[68]

In a highly networked environment, users might be agreeing to give up data related to other users, without their knowledge. For example, should the consent of a single user in a private WhatsApp group to share data about the communications of the entire group be sufficient grounds for such a data transfer? Similarly, what considerations might be given to the impact of other users exposed to a participant who agreed to take part in experimental research around the labeling of misinformation should they decide to share content related to the study? These questions, however, also point to the importance of taking a nuanced approach to what it means to share data for research: surely there is a difference between reporting that user A shared misinformation and including user A's behavior in a study that examines the prevalence of misinformation on a platform or the impact of a particular intervention on the overall prevalence of misinformation among the treatment as opposed to the control group.[69] The issue of consent from users to be studied as part of research is fraught with ethical challenges.

Issues around consent can become murky when industry collaborates with academics. In one well-known example, Facebook garnered significant backlash for its collaborative 2014 study with researchers at Cornell University that deliberately manipulated the emotional responses of unsuspecting users.[70] The study was designed in consultation with the academics and relied solely on Facebook's TOS. The researchers only sought IRB approval from Cornell after the experiment had been conducted. Cornell would later claim that because the professor in question "did not participate in data collection and did not have access to user data" and Facebook conducted the analysis, "he was not directly engaged in human research and that no review by the Cornell Human Research Protection Program was required."[71] However, the journal where the academics published their paper based on the experiment did require one for research on human subjects.

Alternative approaches have been proposed, drawing from models in the field of health, seeking "waivers of normative expectations" to gain consent, whereby participants must fully understand what the specific act is that they are allowing to happen to them as participants in a study within the context it occurs.[72] That level of understanding is not likely achieved through vague language in a TOS outlining that personal data might be used for a variety of research purposes. Other approaches could include offering features or browser extensions that require a user to not just add the extension but also clearly opt in to be a subject in a research project.

The field of intervention research would benefit from norms and standards around user awareness on how their data can be used for research purposes, mechanisms for gaining consent (particularly at scale for a large group of users), and existing academic approaches such as IRBs. Future measurement research would benefit from dialogue and alignment between the platforms, external researchers, academic administration, human rights advocates, and ethicists on questions such as: What types of studies are ethically conducted using a TOS as the basis for consent? How can IRBs be updated to address social media data? What other ways can informed consent be acquired? What sort of biases might be introduced by opt-in-only consent approaches to participating as a research subject? What is the right balance between acquiring sufficient consent from participants in research and conducting necessary studies involving nontrivial portions of a population on interventions to reduce serious risks and harms?

Conclusion

Major social media and technology companies continue to make algorithmic, user interface, and policy changes to their products to address information integrity challenges on their platforms. Concurrently, researchers in academia and the public sector continue to study and advance the science of misinformation discernment among the general public, its impact

on sociopolitical outcomes, and the efficacy of methods to mitigate the problem of spread of misinformation online. Major social media companies and academic/public interest researchers agree on the desirability of providing the public with more skills, context, and tools to evaluate the information they encounter online.[73] Yet, collaboration between the two sectors has been limited by, among other things, various definitional, methodological, and data stewardship challenges.

In order to define common measures of efficacy for interventions to help users confront misinformation, there should be a shared approach toward categorizing the problematic content (such as "misinformation," "disinformation," "influence operation," and so on). One step in helping overcome this challenge would be for platforms to publish canonical lists outlining the definitions they use for problematic content and how they categorize it (providing examples), thus allowing academic researchers to use such information in their work.

Since there also exists, between industry and academia/the public sector, a difference in preference and ability to measure certain outcomes of intervention efficacy (owing to design constraints or access to data), there is an increasing need for research partnerships that align experts across disciplines to conduct novel experiments. Such partnerships should also help overcome methodological challenges in measuring the efficacy of intervention measures that exist for researchers working in academia or the public sector, as they are able to design live experiments without having to simulate or approximate the product experience, and industry experts gain credibility for their impact reports.

These definitional challenges and a lack of consilience in approaches to measurements research between fields of study and sectors could also be addressed by creating a multinational research center to study the information environment and threats within it, such as misinformation. Indeed, an independent multinational research facility could bring different types of researchers together to develop a shared understanding and related terminology, while protecting the independence of those involved. In building shared engineering infrastructure, such an institution could also speed up measurements research.

The only path to knowing, and agreeing, that we are deploying effective interventions is to establish a shared understanding of the goals of each intervention and the metrics that tell us how well they are working. Accessible, responsible external research with platform data will not only unlock greater insights but also deepen cross-sector collaboration. As the data sharing environment evolves, we can expedite progress by exploring and alleviating the methodological and applied complexities of collaboration between industry and academia. The authors hope to invite both social media platforms and academic researchers to participate more actively in dialogue and contribute toward this evolving space of measurement research.

About the Authors

Yasmin Green is the CEO of Jigsaw, a unit within Google focused on solving global security challenges through technology. She previously pioneered approaches to counter violent extremism and state-sponsored disinformation, including seeding the first online network of former violent extremists and survivors of terrorism, launching the Redirect Method advertising-based program to confront online radicalization, and informing cross-platform responses to coordinated disinformation campaigns.

Andrew Gully leads the product research team at Jigsaw. He and his team conduct user and general research that combine mixed-methods approaches to uncover unique insights about users, the issues they face, and ultimately inform technology solutions to overcome those challenges.

Yoel Roth is a former head of Safety & Integrity at Twitter. He led Twitter's policy and threat investigation teams responsible for a wide range of security, authenticity, and content issues, including platform manipulation, misinformation, election security, data privacy, and user identity.

Abhishek Roy leads user research efforts on extremism, misinformation, and user harms in Google's Trust & Safety team. In this role, he leads quantitative and qualitative research studies focused on gaining insights into user behavior, perceptions, and preferences in order to drive actionable changes that promote user protection and advocate for changes with product and policy teams.

Joshua A. Tucker is professor of politics, affiliated professor of Russian and Slavic studies, and affiliated professor of data science at New York University. He is the director of NYU's Jordan Center for Advanced Study of Russia, co-director of the NYU Center for Social Media and Politics, and served for over a decade as a writer and editor of the award-winning politics and policy blog *The Monkey Cage* at the *Washington Post*.

Alicia Wanless is the director of the Partnership for Countering Influence Operations at the Carnegie Endowment for International Peace, which aims to foster evidence-based policymaking to counter threats within the information environment. Alicia also leads a multistakeholder network in partnership with the G7 Rapid Response Network to support efforts in Ukraine.

Acknowledgments

We greatly appreciate the contribution of the following peers who helped us with invaluable feedback that made this work possible: Jacob Shapiro, Jason Lipshin, Paree Zarolia, Jozef Janovský, Erin Saltman, Clement Wolf, Beth Goldberg, Rocky Cole, Alek Chakroff.

Notes

1 Harms stemming from exposure to misinformation have been studied and documented in various forms, for example in relation to humanitarian crises, addictions, vaccine misinformation during the coronavirus pandemic, climate change, and democracies. Terrence Neumann, Maria De-Arteaga, and Sina Fazelpour, "Justice in Misinformation Detection Systems: An Analysis of Algorithms, Stakeholders, and Potential Harms," arXiv [cs.CY], last revised April 29, 2022, http://arxiv.org/abs/2204.13568.; Paarth Neekhara, Brian Dolhansky, Joanna Bitton, and Cristian Canton Ferrer, "Adversarial Threats to DeepFake Detection: A Practical Perspective," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Nashville: IEEE, September 1, 2021): 923–932, https://doi.org/10.1109/CVPRW53098.2021.00103; Caroline Wright, Philippa Williams, Olga Elizarova, Jennifer Dahne, Jiang Bian, Yunpeng Zhao, and Andy S. L. Tan, "Effects of Brief Exposure to Misinformation about E-Cigarette Harms on Twitter: A Randomised Controlled Experiment," *BMJ Open* 11, no. 9: e045445, https://doi.org/10.1136/bmjopen-2020-045445; Claire Wardle and Eric Singerman, "Too Little, Too Late: Social Media Companies' Failure to Tackle Vaccine Misinformation Poses a Real Threat," *BMJ* 372 (January 21, 2021): n26, https://doi.org/10.1136/bmj.n26; Kathie M. D'I. Treen, Hywel T. P. Williams, and Saffron J. O'Neill, "Online Misinformation about Climate Change," *Wiley Interdisciplinary Reviews: Climate Change* 11, no. 5 (June 18, 2020), https://doi.org/10.1002/wcc.665; Spencer McKay and Chris Tenove, "Disinformation as a Threat to Deliberative Democracy," *Political Research Quarterly* 74, no. 3 (July 4, 2020): 703–17, https://doi.org/10.1177/1065912920938143). See also Cristos Goodrow, "On YouTube's Recommendation System," YouTube Official Blog, September 15, 2021, https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/; Adam Mosseri, "Working to Stop Misinformation and False News," Meta, April 7, 2017, https://www.facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news.

2 Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand, "Shifting Attention to Accuracy Can Reduce Misinformation Online," *Nature* 592, no. 7855 (March 17, 2021): 590–595, https://doi.org/10.1038/s41586-021-03344-2.

3 Twitter's 2020 large-scale survey and interview research before launching their first misinformation labeling policy found that "nearly 9 out of 10 individuals said placing warning labels next to significantly altered content would be acceptable [...] respondents were somewhat less supportive of removing or hiding Tweets that contained misleading altered media. For example, 55 percent of those surveyed in the US said it would be acceptable to remove all of such media." Yoel Roth and Ashita Achuthan, "Building Rules in Public: Our Approach to Synthetic and Manipulated Media," Twitter Blog, February 4, 2020, https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.

4 We intend this paper to be useful to a wide range of research communities, including those inside and outside technology platform companies, such as researchers in academic institutions, the public sector, and civil society. Our aim is to be inclusive with our language, but for the sake of simplicity we use the term "academia" or "academic" to represent all researchers working outside of technology platform companies to advance the understanding of misinformation interventions.

5 "Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access," European Digital Media Observatory, May 31, 2022, https://edmoprod.wpengine.com/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf.

6 Annie Franco, Neil Malhotra, and Gabor Simonovits, "Publication Bias in the Social Sciences: Unlocking the File Drawer," *Science* 345, no.6203 (August 28, 2014): 1502–5, https://doi.org/10.1126/science.1255484.

7 For an overview of the importance of the topic, see Joshua A. Tucker and Nathaniel Persily, "Conclusion: The Challenges and Opportunities for Social Media Research," in *Social Media and Democracy: The State of the Field, Prospects for Reform*, ed. Joshua A. Tucker and Nathaniel Persily (Cambridge: Cambridge University Press, 2020), 313–331; Nathaniel Persily and Joshua A. Tucker, "How to Fix Social Media? Start With Independent Research," Brookings, December 1, 2021, https://www.brookings.edu/research/how-to-fix-social-media-start-with-independent-research/. For examples of recent regulatory efforts see Senator Chris Coons, "Coons, Portman, Klobuchar Announce Legislation to Ensure Transparency at Social Media Platforms," press release, December 9, 2021, https://www.coons.senate.gov/news/press-releases/coons-portman-klobuchar-announce-legislation-to-ensure-transparency-at-social-media-platforms; Senator Michael Bennet, "Bennet Introduces Landmark Legislation to Establish Federal Commission to Oversee Digital Platforms," press release, May 12, 2022, https://www.bennet.senate.gov/public/index.cfm/2022/5/bennet-introduces-landmark-legislation-to-establish-federal-commission-to-oversee-digital-platforms; European Digital Media Observatory, "Platform-to-Researcher Data Access."

8 Emily Saltz and Claire Leibowicz describe three broad intervention types that platforms use to approaches misinformation: labels ("any kind of partial or full overlay on a piece of content that is applied by platforms to communicate information credibility to users"), ranking (use of "various signals to rank what and how content appears to users") and removal ("the temporary or permanent removal of any type of content on a platform"). Our focus in this paper is on the first category, which we refer to as user-facing interventions. The authors of this paper have experience from working in the tech sector and collaborating with external researchers to advance evidence-based innovation in these user-facing interventions. We have directly benefited from the creativity, rigor, and theoretical underpinnings of our research collaborators and are motivated to bridge the knowledge and communication gap between those working towards the same goal from inside and outside of the major platforms. Emily Saltz and Claire Leibowicz, "Shadow Bans, Fact-Checks, Info Hubs: The Big Guide to How Platforms Are Handling Misinformation in 2021," Neiman Lab, June 15, 2021, https://www.niemanlab.org/2021/06/shadow-bans-fact-checks-info-hubs-the-big-guide-to-how-platforms-are-handling-misinformation-in-2021/.

9 Moreover, the decision of what gets published is not random. As one of us has written previously (Tucker, with Nathaniel Persily), "most research conducted internally by the platforms will only make its way into the public domain if the platforms choose to release the research publicly. In academia, this is known as the 'file drawer' problem, where less interesting (and often null) results fail to be published, and as a consequence the overall accumulation of knowledge is biased" (see Franco, Malhotra, and Simonovits, "Unlocking the File Drawer"). "When we consider this from the perspective of for-profit corporations, the net result can be even more pernicious, which is that the overall accumulation of knowledge would likely be biased in the direction of research that puts the platforms in a better light. However, knowing the potential for such biases to exist should lead outside observers to discount such research accordingly, making knowledge accumulation that much more difficult" (Persily and Tucker, "How to Fix Social Media").

10 It is worth noting that in addition to the myriad issues around determining which specific narratives are not appropriately labeled "misinformation," there also are practical considerations around the breadth and specificity of content that platforms contend with that off-platform researchers can sidestep. For instance, video sharing platforms may debate the utility of labeling an entire video "misinformation" if there is one claim made within hours of content.

11 Cristiano Lima, "Facebook No Longer Treating 'Man-Made' Covid as a Crackpot Idea," *Politico*, May 26, 2021, https://www.politico.com/news/2021/05/26/facebook-ban-covid-man-made-491053.

12 Jennifer Allen, Antonio A. Arechar, Gordon Pennycook, and David G. Rand, "Scaling Up Fact-Checking Using the Wisdom of Crowds," *Science Advances* 7, no. 36 (September 1, 2021): eabf4393, https://doi.org/10.1126/sciadv.abf4393; William Godel, Zeve Sanderson, Kevin Aslett, Jonathan Nagler, Richard Bonneau, Nathaniel Persily, and Joshua A. Tucker, "Moderating With the Mob: Evaluating the Efficacy of Real-Time Crowdsourced Fact-Checking," *Journal of Online Trust and Safety* 1, no. 1 (October 28, 2021), https://doi.org/10.54501/jots.v1i1.15; Jon Roozenbeek, Claudia R. Schneider, Sarah Dryhurst, John Kerr, Alexandra L. J. Freeman, Gabriel Recchia, Anne Marthe van der Bles, and Sander van der Linden, "Susceptibility to Misinformation About COVID-19 Around the World," *Royal Society Open Science* 7, no. 10 (October 14, 2020): 201199, https://doi.org/10.1098/rsos.201199; Alexandre Bovet and Hernán A. Makse, "Influence of Fake News in Twitter During the 2016 US Presidential Election," *Nature Communications* 10, no. 1 (January 2, 2019): 7, https://doi.org/10.1038/s41467-018-07761-2; Andrew M. Guess, Brendan Nyhan, and Jason Reifler, "Exposure to Untrustworthy Websites in the 2016 US Election," *Nature Human Behaviour* 4, no. 5 (March 2, 2020): 472–80, https://doi.org/10.1038/s41562-020-0833-x; Lisa Singh, Leticia Bode, Ceren Budak, Kornraphop Kawintiranon, Colton Padden, and Emily Vraga, "Understanding High- and Low-Quality URL Sharing on COVID-19 Twitter Streams," *SIAM Journal on Scientific Computing: A Publication of the Society for Industrial and Applied Mathematics* 3, no. 2 (November 27, 2020): 343–66, https://doi.org/10.1007/s42001-020-00093-6.

13 "TikTok Community Guidelines," TikTok, accessed August 23, 2022, https://www.tiktok.com/community-guidelines?lang=en#37.

14 "COVID-19 Misleading Information Policy," Twitter Help Center, December 2021, https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy.

15 Alicia Wanless and James Pamment, "How Do You Define a Problem like Influence?" *Journal of Information Warfare* 18, no. 3 (Winter 2019): 1–14.

16 Vijaya Gadde and Yoel Roth, "Enabling Further Research of Information Operations on Twitter," Twitter Blog, October 17, 2018, https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.

17 Danny Sullivan, "An Overview of Our Rater Guidelines for Search," Google, October 19, 2021, https://blog.google/products/search/overview-our-rater-guidelines-search/.

18 Allen, Arechar, Pennycook, and Rand, "Scaling Up Fact-Checking"; Godel, Sanderson, Aslett, Nagler, Bonneau, Persily, and Tucker, "Moderating With the Mob."

19 Researchers may consider focusing on identifying robust outcome metrics that allow for greater study comparability and meta-analysis, both on platforms and in lab environments. While any single outcome measure may not be desirable to platforms, which each have their own distinct constraints, design, and applications to users, the development of a core set of indices will aid the research community, and industry, to use a common set of language and measures to enhance a variety of collaborations.

20 We use the term metric to broadly mean dependent variable, outcome variable, objective function, or measure of business interest.

21 See for example Kevin Aslett, Andrew M. Guess, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker, "News Credibility Labels Have Limited Average Effects on News Diet Quality and Fail to Reduce Misperceptions," *Science Advances* 8, no. 18 (May 6, 2022): eabl3844, https://doi.org/10.1126/sciadv.abl3844.

22 Pennycook, Epstein, Mosleh, Arechar, Eckles, and Rand, "Shifting Attention to Accuracy."

23 Godel, Sanderson, Aslett, Nagler, Bonneau, Persily, and Tucker, "Moderating With the Mob."

24 Mohsen Mosleh, Gordon Pennycook, and David G. Rand, "Field Experiments on Social Media," *Current Directions in Psychological Science* 31, no. 1 (December 1, 2021): 69–75, https://doi.org/10.1177/09637214211054761; Andrew Guess, Jonathan Nagler, and Joshua Tucker, "Less Than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook," *Science Advances* 5, no. 1 (January 9, 2019): eaau4586, https://doi.org/10.1126/sciadv.aau4586.

25 Stephan Lewandowsky and Sander van der Linden, "Countering Misinformation and Fake News Through Inoculation and Prebunking," *European Review of Social Psychology* 32, no. 2 (February 22, 2021): 348–384, https://doi.org/10.1080/10463283.2021.1876983.

26 Gina Hernandez, "New Prompts to Help People Consider Before They Share," TikTok Newsroom, February 3, 2021, https://newsroom.tiktok.com/en-gb/new-prompts-to-help-people-consider-before-they-share-uk.

27 "The Four Rs of Responsibility, Part 2: Raising Authoritative Content and Reducing Borderline Content and Harmful Misinformation," YouTube Blog, December 3, 2019, https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/.

28 Laura Courchesne, Julia Ilhardt, and Jacob N. Shapiro, "Review of Social Science Research on the Impact of Countermeasures Against Influence Operations," *Harvard Kennedy School Misinformation Review*, September 3, 2021, https://doi.org/10.37016/mr-2020-79.

29 Talia Stroud, Joshua A. Tucker, Annie Franco, and Chad P. Kiewiet de Jonge, "A Proposal for Understanding Social Media's Impact on Elections: Peer-Reviewed Scientific Research," *2020 Election Research Project* (blog), Medium, August 31, 2020, https://medium.com/@2020_election_research_project/a-proposal-for-understanding-social-medias-impact-on-elections-4ca5b7aae10.

30 Stroud, Tucker, Franco, and Kiewiet de Jonge, "Social Media's Impact on Elections."

31 Yoel Roth and Vijaya Gadde, "Expanding Access Beyond Information Operations," Twitter Blog, December 2, 2021, https://blog.twitter.com/en_us/topics/company/2021/-expanding-access-beyond-information-operations-.

32 Donald T. Campbell and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research*, (Ravenio Books, 2015).

33 The simplest RCT designs typically have a treatment group and a control group; however, variations on the design can have either multiple treatment groups (treatment A versus treatment B versus control) or a factorial design (treatment A versus treatment B versus treatment A and B versus control). For example, Toni G.L.A. Meer and Michael Hameleers measured whether news media literacy interventions on users could promote a more cross-cutting media consumption by randomly assigning participants to three groups where they were exposed to news media literacy messages with injunctive norms, descriptive norms, or none at all (Toni G.L.A. Meer and Michael Hameleers, "Fighting Biased News Diets: Using News Media Literacy Interventions to Stimulate Online Cross-cutting Media Exposure Patterns," *New Media & Society* 23, no. 11 (July 30, 2020): 3156–3178, https://doi.org/10.1177/1461444820946455).

34 Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G. Lu, and David G. Rand, "Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention," *Psychological Science* 31, no. 7 (June 30, 2020): 770–80, https://doi.org/10.1177/0956797620939054.

35 James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, et al., "The YouTube Video Recommendation System," in *Proceedings of the Fourth ACM Conference on Recommender Systems* (New York: Association for Computing Machinery, 2010), 293–96.

36 Petter Törnberg, "Echo Chambers and Viral Misinformation: Modeling Fake News as Complex Contagion," *PloS One* 13, no. 9 (September 20, 2018): e0203958,https://doi.org/10.1371/journal.pone.0203958.

37 For example, in a recent study by the Center for Countering Digital Hate that examined a sample of shared anti-vaccine content on Facebook and Twitter between February 2021 and March 2021, about 65 percent of anti-vaccine content was attributable to a small group of twelve influencers producing such content at a rapid pace See "The Disinformation Dozen," Center for Countering Digital Hate, March 21, 2021, https://www.counterhate.com/disinformationdozen. Experimenters may solve the issue of network effects by adopting a cluster-randomized sampling scheme that first partitions users into clusters (based on certain network attributes) and then randomly samples these clusters into the treatment or the control group, computing average treatment effects at a cluster level instead of user level. See Huan Gui, Ya Xu, Anmol Bhasin, and Jiawei Han, "Network A/b Testing: From Sampling to Estimation," in *Proceedings of the 24th International Conference on World Wide Web* (New York: Association for Computing Machinery, 2015), 399–409; Brian Karrer, Liang Shi, Monica Bhole, Matt Goldman, Tyrone Palmer, Charlie Gelman, Mikael Konutgan, and Feng Sun, "Network Experimentation at Scale," in *Proceedings of the 27th ACM*

SIGKDD Conference on Knowledge Discovery & Data Mining (New York: Association for Computing Machinery, 2021), 3106–3116, https://doi.org/10.1145/3447548.3467091. It is also worth noting that there are approaches that allow for direct estimation of peer influence in a social network using procedures like randomized edge allocation or Bayesian estimation of network uncertainty. See Sean J. Taylor and Dean Eckles, "Randomized Experiments to Detect and Estimate Social Influence in Networks," in *Complex Spreading Phenomena in Social Systems: Influence and Contagion in Real-World Social Networks*, ed. Sune Lehmann and Yong-Yeol Ahn (Cham: Springer International Publishing, 2018), 289–322, https://doi.org/10.1007/978-3-319-77332-2_16; Panos Toulis and Edward Kao, "Estimation of Causal Peer Influence Effects," in *Proceedings of the 30th International Conference on Machine Learning*, vol. 28 (Atlanta: jmlr.org, 2013), III-1489–III-1497, https://dl.acm.org/doi/10.5555/3042817.3043103.

38 "We analyze the user-generated content in Sina Weibo, and find evidence that the spread of popular messages often follow a mechanism that differs from the spread of disease, in contrast to common belief. In this mechanism, an individual with more friends needs more repeated exposures to spread further the information. Moreover, our data suggest that for certain messages the chance of an individual to share the message is proportional to the fraction of its neighbors who shared it with him/her, which is a result of competition for attention." Ling Feng, Yanqing Hu, Baowen Li, H. Eugene Stanley, Shlomo Havlin, and Lidia A. Braunstein, "Competing for Attention in Social Media Under Information Overload Conditions." *PloS One* 10, no. 7 (July 10, 2015): e0126090, https://doi.org/10.1371/journal.pone.0126090.

39 Moira Burke, Anthony Hornof, Erik Nilsen, and Nicholas Gorman, "High-Cost Banner Blindness: Ads Increase Perceived Workload, Hinder Visual Search, and Are Forgotten," *ACM Transactions on Computer-Human Interaction: A Publication of the Association for Computing Machinery* 12, no. 4 (December 2005): 423–45, https://doi.org/10.1145/1121112.1121116.

40 Howard White and Shagun Sabarwal, "Quasi-Experimental Design and Methods," *Methodological Briefs: Impact Evaluation* 8 (2014): 1–16. Some commonly used quasi-experimental design approaches are regression discontinuity designs and propensity score–matching designs. Another alternative causal inference approach using observational data involves constructing a counterfactual using a statistical model like a regression analysis to estimate what would have happened sans an intervention. Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott, "Inferring Causal Impact Using Bayesian Structural Time-Series Models," *The Annals of Applied Statistics* 9: 247–74.

41 Elizabeth Culliford, "Facebook to Label All Posts About COVID-19 Vaccines," Reuters, March 15, 2021, https://www.reuters.com/article/us-health-coronavirus-facebook/facebook-to-label-all-posts-about-covid-19-vaccines-idUSKBN2B70NJ; Ian Carlos Campbell, "Twitter Will Label COVID-19 Vaccine Misinformation and Enforce a Strike System," The Verge, March 1, 2021, www.theverge.com/2021/3/1/22307919/twitter-covid-19-vaccine-labels-five-strike-system.

42 It is worth noting that Gary King and Richard Nielsen showed that propensity score–matching techniques often fail to use all the information available and thereby end up unnecessarily increasing imbalance, inefficiency, model dependence, and bias. Gary King and Richard Nielsen, "Why Propensity Scores Should Not Be Used for Matching," *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association* 27, no. 4 (May 7, 2019): 435–54, https://doi.org/10.1017/pan.2019.11.

43 Marco Caliendo and Sabine Kopeinig, "Some Practical Guidance for the Implementation of Propensity Score Matching," *SSRN Electronic Journal* (May 11, 2005), https://doi.org/10.2139/ssrn.721907.

44 Mosleh, Pennycook, and Rand, "Field Experiments on Social Media."

45 "Research Partnership to Understand Facebook and Instagram's Role in the U.S. 2020 Election," *Meta Research* (blog), December 10, 2021, https://research.facebook.com/2020-election-research/.

46 Oliver Lindberg, "User Research: Best Practices and Methodologies," Adobe, January 29, 2020, https://xd.adobe.com/ideas/process/user-research/user-research-best-practices-methodologies/.

47 Emily Saltz, Soubhik Barari, Claire Leibowicz, Claire Wardle, "Misinformation Interventions Are Common, Divisive, and Poorly Understood," *Harvard Kennedy School Misinformation Review*, October 27, 2021, misinforeview.hks.harvard.edu/article/misinformation-interventions-are-common-divisive-and-poorly-understood.

48 Saltz, Barari, Leibowicz, Wardle, "Misinformation Interventions."

49 Nathan Walter, Jonathan Cohen, R. Lance Holbert, and Yasmin Morag, "Fact-Checking: A Meta-Analysis of What Works and for Whom," *Political Communication* 37, no. 3 (October 24, 2020): 350–75.

50 Gordon Pennycook, Adam Bear, Evan T. Collins, and David G. Rand, "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines without Warnings," *Management Science* 66, no. 11 (February 21, 2020): 4944–57, https://doi.org/10.1287/mnsc.2019.3478.

51 Laura D. Scherer and Gordon Pennycook, "Who Is Susceptible to Online Health Misinformation?" *American Journal of Public Health* 110, no. S3 (October 1, 2020): S276–77, https://doi.org/10.2105/AJPH.2020.305908.

52 Guess, Nagler, and Tucker, "Less Than You Think."

53 Robert C. Francis, Jr., "Protecting Our Heroes From Disinformation on Social Media," Council on Foreign Relations, January 27, 2021, https://www.cfr.org/blog/protecting-our-heroes-disinformation-social-media.

54 "(U)Report of the Select Committee on Intelligence, United States Senate, on Russian Active Measures Campaigns and Interference in the 2016 Election, Volume 2: Russia's Use of Social Media with Additional Views," Select Committee on Intelligence, United States Senate (116th Congress, first session), https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf.

55 Akhilesh Singh, "BJP Accuses Twitter of Bias, Wants Centre to Take Action," *Times of India*, May 23, 2021, https://timesofindia.indiatimes.com/india/bjp-accuses-twitter-of-bias-wants-centre-to-take-action/articleshow/82873041.cms.

56 Lexi Lonas, "Russia Threatens 'Retaliatory Measures' after YouTube's Removal of RT Channels," *Hill*, September 29, 2021, thehill.com/policy/international/russia/574497-russia-threatens-to-block-youtube-after-removal-of-rt-channels.

57 Sara Harrison, "No One's Happy With YouTube›s Content Moderation Policies," *Wired*, August 28, 2019, https://www.wired.com/story/no-ones-happy-youtubes-content-moderation/.

58 Michael Pizzi, "The Syrian Opposition Is Disappearing From Facebook," *Atlantic*, February 4, 2014, https://www.theatlantic.com/international/archive/2014/02/the-syrian-opposition-is-disappearing-from-facebook/283562/.

59 Paul M. Barrett and J. Grant Sims, "False Accusation: The Unfounded Claim That Social Media Companies Censor Conservatives," NYU Stern Center for Business and Human Rights, https://www.stern.nyu.edu/experience-stern/faculty-research/false-accusation-unfounded-claim-social-media-companies-censor-conservatives.

60 Lily Hay Newman, "Facebook's 'Red Team X' Hunts Bugs Beyond the Social Network's Walls," *Wired*, March 18, 2021, https://www.wired.com/story/facebook-red-team-x-vulnerabilities/.

61 Neekhara, Dolhansky, Bitton, and Ferrer, "Adversarial Threats to DeepFake Detection."

62 For example, in Facebook and Instagram's 2020 Election Research Project described earlier in this paper, all individual-level participants in experimental treatments consented to participation in the study See Stroud, Tucker, Franco, and Kiewiet de Jonge, "Social Media's Impact on Elections".

63 "Meta Privacy Policy - How Meta Collects and Uses User Data," Facebook, July 26, 2022, https://www.facebook.com/privacy/policy/?entry_point=data_policy_redirect&entry=0.

64 Jessica Guynn, "What You Need to Know before Clicking 'I Agree' on That Terms of Service Agreement or Privacy Policy," *USA Today*, January 28, 2020, https://www.usatoday.com/story/tech/2020/01/28/not-reading-the-small-print-is-privacy-policy-fail/4565274002/.

65 Nili Steinfeld, "'I Agree to the Terms and Conditions': (How) Do Users Read Privacy Policies Online? An Eye-Tracking Experiment," *Computers in Human Behavior* 55, part B (February 2016): 992–1000, https://doi.org/10.1016/j.chb.2015.09.038.

66 Casey Fiesler and Nicholas Proferes, "'Participant' Perceptions of Twitter Research Ethics," *Social Media + Society* 4, no. 1 (March 10, 2018): 2056305118763366, https://doi.org/10.1177/2056305118763366.

67 Health Communication and Informatics Research Branch, "Human Subjects Considerations for Social Media Research," National Cancer Institute, August 26, 2019, https://cancercontrol.cancer.gov/sites/default/files/2020-06/human-subjects-considerations-for-social-media-research.pdf; Jessica Pater, Casey Fiesler, and Michael Zimmer, "No Humans Here: Ethical Speculation on Public Data, Unintended Consequences, and the Limits of Institutional Review," *Proceedings of the ACM on Human-Computer Interaction* 6, no. GROUP (January 2022): 1–13, https://doi.org/10.1145/3492857.

68 Graciela Gonzalez-Hernandez, "On the Ethics of Using Social Media Data for Health Research," *NLM Musings from the Mezzanine* (blog), National Institutes of Health National Library of Medicine, June 25, 2019, https://nlmdirector.nlm.nih.gov/2019/06/25/on-the-ethics-of-using-social-media-data-for-health-research/.

69 This is analogous to not reporting in a medical study that person A has disease X, but using data about person A's medical history in order to form population-level estimates of the prevalence of disease X in particular communities.

70 Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock, "Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks," *Proceedings of the National Academy of Sciences of the United States of America* 111, no. 24 (June 2, 2014): 8788–90, https://doi.org/10.1073/pnas.1320040111.

71 Gail Sullivan, "Cornell Ethics Board Did Not Pre-Approve Facebook Mood Manipulation Study," July 1, 2014, https://www.washingtonpost.com/news/morning-mix/wp/2014/07/01/facebooks-emotional-manipulation-study-was-even-worse-than-you-thought/.

72 As Catherine Flick explains, "for any procedure or situation that may violate a person's normative expectation for behaviour, such as using a knife to cut into a person's body, the person needs to waive that particular behaviour (for surgery, for example). If they do not waive the expected norm, then consent has not been given (as would be expected in cases such as a stabbing)." Catherine Flick, "Informed Consent and the Facebook Emotional Manipulation Study," *Research Ethics* 12, no. 1 (August 11, 2015): 14–28, https://doi.org/10.1177/1747016115599568.

73 David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, et al., "The Science of Fake News," *Science* 359, no. 6380 (March 9, 2018): 1094–96, https://doi.org/10.1126/science.aao2998.

Carnegie Endowment for International Peace

The Carnegie Endowment for International Peace is a unique global network of policy research centers in Russia, China, Europe, the Middle East, India, and the United States. Our mission, dating back more than a century, is to advance peace through analysis and development of fresh policy ideas and direct engagement and collaboration with decisionmakers in government, business, and civil society. Working together, our centers bring the inestimable benefit of multiple national viewpoints to bilateral, regional, and global issues.

Partnership for Countering Influence Operations

The Partnership for Countering Influence Operations (PCIO) is a multistakeholder initiative focused on fostering evidence-based policymaking to address issues within the information environment at the Carnegie Endowment for International Peace. Since launching in January 2020, PCIO has built an active community across governments, industry, academia, and civil society.

Princeton University

Princeton University is a vibrant community of scholarship and learning that stands in the nation's service and in the service of humanity. Chartered in 1746, and known as the College of New Jersey until 1896, it was British North America's fourth college. Princeton is an independent, coeducational, nondenominational institution that provides undergraduate and graduate instruction in the humanities, social sciences, natural sciences, and engineering.

Empirical Studies of Conflict Project

The Empirical Studies of Conflict Project (ESOC) is a multi-university consortium launched in 2009 to support research on insurgency, civil war, and other politically motivated violence, worldwide. Based at Princeton University, ESOC empowers scholarship and to help address pressing security threats ranging from civil war to misinformation campaigns.

CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

CarnegieEndowment.org