Beyond Effectiveness: A Multi-Criteria Framework for Comparing Practical Socio-Technical Interventions

Catherine King
Carnegie Mellon University
Pittsburgh, USA
cking2@andrew.cmu.edu

Lynnette Hui Xian Ng
Carnegie Mellon University
Pittsburgh, USA
lynnetteng@cmu.edu

Kathleen M. Carley
Carnegie Mellon University
Pittsburgh, USA
kathleen.carley@cs.cmu.edu

Abstract

Designers and policymakers in sociotechnical domains like content moderation, privacy interfaces, recommender systems and beyond, must choose among a growing menu of proposed interventions, but typically lack a principled basis for comparing them. Prior work tends to evaluate interventions individually and mostly along the effectiveness criteria, while implementation constraints such as cost, effort and feasibility are often considered separately. We present a multi-criteria framework for evaluating sociotechnical interventions. This framework is instantiated through the case of misinformation, a domain of intense focus for proposed countermeasures. We survey $N = 39$ researchers on 40 operationalized interventions across five evaluative criteria: political feasibility, effectiveness, user acceptance, cost, and implementation effort. We find that the interventions that experts judge to be the most effective are not always the most acceptable to the public or the most feasible to implement. We also discuss how this tension has implications for the design of sociotechnical interventions beyond misinformation, and offer a decision framework for practitioners navigating the trade-offs of sociotechnical interventions.

CCS Concepts

• **Social and professional topics** → **Computing / technology policy**; • **Human-centered computing** → **Interaction design**.

Keywords

socio-technical interventions, effectiveness, misinformation, expert survey

ACM Reference Format:
Catherine King, Lynnette Hui Xian Ng, and Kathleen M. Carley. 2018. Beyond Effectiveness: A Multi-Criteria Framework for Comparing Practical Socio-Technical Interventions. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 21 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Social media platforms and the emerging generative-AI technology that shapes them are quickly transforming the online information environment, bringing with them various interventions used to address socio-technical harms [22, 62, 98]. Yet these interventions, such as those for countering misinformation or detecting fraud, are often developed and treated individually rather than analyzed as a whole using a standardized set of criteria [27, 41]. Socio-technical systems research has argued that such interventions cannot be meaningfully evaluated outside the social context in which they enter [99], and value-sensitive design calls for weighing a technology against multiple, sometimes competing values rather than a single metric [33]. Drawing on these two threads of research, we aim to establish a comprehensive framework for developing effective and practical countermeasures across the social media landscape. Researchers, companies, and policymakers can use this framework to evaluate the socio-technical interventions and consider their trade-offs.

We develop a multi-criteria framework for socio-technical intervention design through the case of misinformation. Throughout, we use "intervention" to mean a deliberate countermeasure (i.e., a policy, a platform design, an educational effort) that is deployed to reduce a socio-technical harm. Misinformation is a domain where the evaluation gap is especially visible. Countermeasures are proliferating faster than the field's capacity to compare them [93, 106], and platforms tend to implement interventions reactively and inconsistently [35]. Intervention design decisions typically require trading effectiveness against the practical and normative considerations that determine whether an intervention can or should be be deployed [42, 95]. However, to the best of our knowledge, no framework exists to support that reasoning. To develop the framework, we integrate prior research with an expert elicitation survey of $N = 39$ researchers who evaluated 40 interventions across five criteria: political feasibility, effectiveness, user acceptance, cost, and effort level. Our main **research question** is: How can we identify which interventions are both practical and effective, and under what circumstances? More specifically:

- RQ1: What trade-offs emerge when misinformation interventions are evaluated across political feasibility, effectiveness, acceptance, cost, and effort?
- RQ2: Which intervention categories best balance user-centered and implementation-centered criteria?
- RQ3: What design principles can guide practical misinformation intervention design?

The contributions of this work are threefold:

- First, we introduce a taxonomy of forty operationalized interventions based off a harmonization of literature review. These interventions are structured around eight categories,

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

consisting of account moderation, content moderation, content distribution, generative AI-specific measures, content labeling, user-based measures, media literacy and educational efforts, and other external and institutional measures.
- Second, we provide a multi-criteria expert evaluation of those interventions across five criteria: political feasibility, effectiveness, user acceptance, cost, and implementation effort. This expert elicitation survey shows that interventions frequently vary in effectiveness, acceptance or political feasibility, and the administrative effort or cost involved.
- Third, we translate these judgments into a decision framework that maps practical operating constraints (budget, political exposure, timeline, public trust) onto a starting cluster of interventions, and the design investments needed to unlock the next cluster.

Designers and policymakers often face similar tensions across many sociotechnical domains: how to moderate content [50], how much friction to add to privacy choices [2], or how to regulate dark patterns in interface design [72]. In each case, the question is not whether an intervention works, but which one to deploy given the contextual constraints. Here, we study misinformation as a case study to build a tool for this selection problem. This study provides insights that we discuss on how and which intervention types, scoped to the social media information environment, are most viable across different implementation constraints, and discusses how different stakeholders can use this framework.

2 Related Work

2.1 Misinformation interventions and their evaluation

Previous systematic review articles in the domain of misinformation interventions show that interventions are typically characterized by the part of the platform affected or the actor implementing them. Blair et al. [17] groups interventions into four general categories (informational, educational, sociopsychological, and institutional) and evaluates them primarily on efficacy, while suggesting that other dimensions like feasibility, scalability, and durability should also be considered when selecting and designing interventions. Courchesne et al. [27] find a mismatch between the countermeasures that are studied versus those that are implemented by platforms. Specifically, fact-checking interventions and interventions emphasizing source pre-bunking or user inoculation are widely studied, while account-level interventions such as removal and user notifications, and algorithmic interventions such as redirection and changes to monetization policy, remain understudied. Finally Kozyreva et al. [64] categorizes interventions into three general categories (nudges, boosts and educational interventions, and refutation strategies) and organizes implementation through the targeted outcome (i.e. behaviors, competence, and beliefs) and limitations.

While these three key review articles anchor the space of misinformation interventions, they stop short of the comparative evaluation that design practice needs. These review articles find that most prior research in this domain focuses narrowly on effectiveness or user acceptance of the interventions, but rarely both axes, and almost never alongside implementation constraints. Our framework closes this gap by comparing multiple dimensions of interventions at once to facilitate identifying the appropriate interventions for the context.

2.2 Intervention design features and trade-offs

To design and evaluate interventions based on their features using multiple criteria, we first need to define what those features are. A well-defined intervention has a set of descriptive, objective characteristics, as well as potentially changing user perceptions of those characteristics and the intervention as a whole. An intervention's objective characteristics refer to the relatively static features that define how it is implemented and by whom. These features include the implementer (which organization(s) are implementing the intervention), the targeted phase of the information pipeline (such as the creation, spread, or belief phases), the content being addressed, platform design changes, and the policy's description (how it is being implemented and communicated) [58]. User perception refers to how individuals evaluate the different objective characteristics of interventions and includes factors such as perceived effectiveness, fairness, intrusiveness, transparency, and problem awareness [42]. Interventions that are viewed as more effective (benefits outweigh costs) and are more supported by users (practical) are more likely to be implemented by platforms [39, 62, 68].

User perceptions are especially critical because they influence the overall level of policy support. These perceptions affect whether people believe a problem even needs intervention, whether they think the intervention will be effective and worth doing, and whether they trust the proposed implementer to enforce the new policies fairly. Public policy support often depends on problem awareness and the type of content addressed [42]. There is a higher level of support for anti-smoking initiatives compared to alcohol consumption or diet, which is partially explained by the public's awareness of the negative consequences associated with smoking [28]. Similarly, in a social media context, individuals have indicated they are more willing to engage in social corrections with close contacts or for seriously harmful content [59, 61, 102]. Additionally, some individuals may trust certain implementers more than others. For example, partisanship is associated with lower levels of support for government regulation [108].

Educational and messaging efforts could improve public perceptions and support for interventions. A meta-analysis of public policy interventions demonstrated that effectively communicating a potential policy's effectiveness increased overall support by approximately 4% [94]. Interventions that are more transparently implemented also tend to gather more public support [42, 50]. Clearly marked and communicated interventions can shape a user's perception of the effectiveness and intrusiveness of the intervention [116]. However, more opaque policies may in turn be less exploitable by malicious actors [50], illustrating the trade off with transparency.

2.3 Criteria selection for platform design

The field of Human-Computer Interaction has a substantial body of work on how platform affordances and intervention design shape user information behavior. Most of this work has independently arrived at the same trade-off-centered framing, where there are multiple criteria to balance simultaneously. Content moderation work treats moderation not as a single optimization target but

as a negotiation among competing values of efficiency, fairness, and user trust [35]. Ethnographic accounts of Reddit's moderation tool describe this negotiation as part of the platform's broader sociotechnical system, where automated and human judgment must be jointly accounted for [112]. Comparable trade-offs structure privacy interface design, where the amount of friction a nudge introduces is weighed against its protective benefit rather than maximized outright [2]. Within misinformation specifically, recent reviews converge on effectiveness, feasibility, scalability, and durability as the relevant criteria [17, 64]. Feasibility refers to how easily an intervention can be implemented and whether it aligns with user expectations and the intervention's intent [41]. Scalability and durability, both aspects of effectiveness, address how easily an intervention can be expanded to reach a larger audience and the longevity of its impact, respectively [47, 70, 86].

These criteria focus on the technical components or user interactions with platforms, without considering external influences outside the platforms. For that, we turn to public policy research. In public policy, when comparing strategies for regulating sociotechnical platforms, research often evaluates proposals using measures like effectiveness, administrative difficulty or burden, cost, and political acceptability [29, 65, 95]. The public policy perspective on administrative burden relates to the concept of implementation feasibility. Additionally, political acceptability is a similar concept to user acceptance of an intervention, which has been shown to influence support for that intervention, though it also incorporates external governmental constraints [39, 62, 95].

3 The PEACE Framework

In this work, we first define a categorization of intervention types and then develop a decision framework for evaluating intervention quality and balancing trade-offs. The **PEACE** framework assesses each intervention's **Political feasibility** in the current environment, its **Effectiveness** in achieving the intended outcome, the level of user **Acceptance** it is likely to receive, its **Cost** or cheapness, and the **Effort** required for implementation. These five evaluative criteria were drawn from literature, and insights from an expert elicitation survey was used to inform the design of a corresponding decision framework.

3.1 Intervention Categorization

To categorize intervention types, we selected and examined all the studies cited by four recent, comprehensive review articles from different fields: the Courchesne et al. [27] article published in an interdisciplinary journal focused on social sciences, the Blair et al. [17] article highlighting interventions from both the Global North and Global South in a psychology journal, the Aghajari et al. [3] article published in the proceedings of a prominent HCI conference, and the Kozyreva et al. [64] article in Nature Human Behavior. We augmented the literature review by pulling additional, relevant articles on socio-technical interventions identified through a Scopus search using the terms "misinformation" and "disinformation" combined with either "interventions" or "counter". We then categorized interventions on social media platforms by the stages of the information life cycle they target [60]. More details on our categorization and the specific literature referenced can be found in Appendix A.

The information lifecycle on social media generally includes three phases: the creation of information-disseminating accounts and the information itself, the dissemination of information through organic and viral amplification, and the engagement with the information by correcting, contesting, or contextualizing it [25, 62, 79, 110]. The life cycle lens provides a frame for thinking about where analysts could intervene. Because the volume and speed of information spread are too fast for journalists and users to respond promptly to incorrect or contested claims, such as with fact-checks, exploring ways to intervene earlier in the pipeline, like slowing the virality of misinformation, is an important consideration [25]. Table 1 provides a high-level overview of the intervention categories that will be discussed throughout the rest of the paper [62]. Refer to Appendix B for detailed information on the 40 specified intervention implementations. The appendix outlines how governments, platforms, and institutions can implement these measures in practice. The specific interventions were operationalized by reviewing the literature and platform policies to identify examples for each intervention type that have been implemented or proposed previously [1, 95].

The first six categories (account moderation, content moderation, content distribution, generative AI-specific, content labeling, user-based measures) are interventions that occur directly on the platforms. These interventions can be implemented by social media companies, users on those platforms, or imposed on the platforms through government regulation [95]. These platform-level interventions fall into two broad mechanisms: algorithmic interventions that govern what accounts and content can be created and distributed (account moderation, content moderation, content distribution, generative AI-specific) and front-end design choices that govern how content is displayed and the affordances in which it can be engaged with (content labeling, user-based measures). More specifically, account moderation strategies target the information creation phase by controlling account and content creation permissions; content moderation, distribution, and generative AI-assisted strategies target the information spread phase by reducing or removing algorithmic amplification and managing the information sharing process between users; and content labeling and user-driven strategies occur after the information has been shared to verify or contextualize claims.

The next two categories, media literacy and external institutional measures, are strategies that occur primarily beyond the platform, and can be implemented by platforms, governments or various civic organizations. Media literacy and educational initiatives focus on the prevention phase in the information pipeline, teaching students how to identify and discern truth from misinformation and scams. This includes learning how to recognize manipulation techniques, logical fallacies, and AI-generated or edited content. Structural strategies like regulation and data sharing may target any part of the pipeline, and are meant to support a healthier information environment.

In this study, we specifically examine the case of misinformation spread through social media platforms. The information lifecycle includes the creation of social media accounts and misinformation,

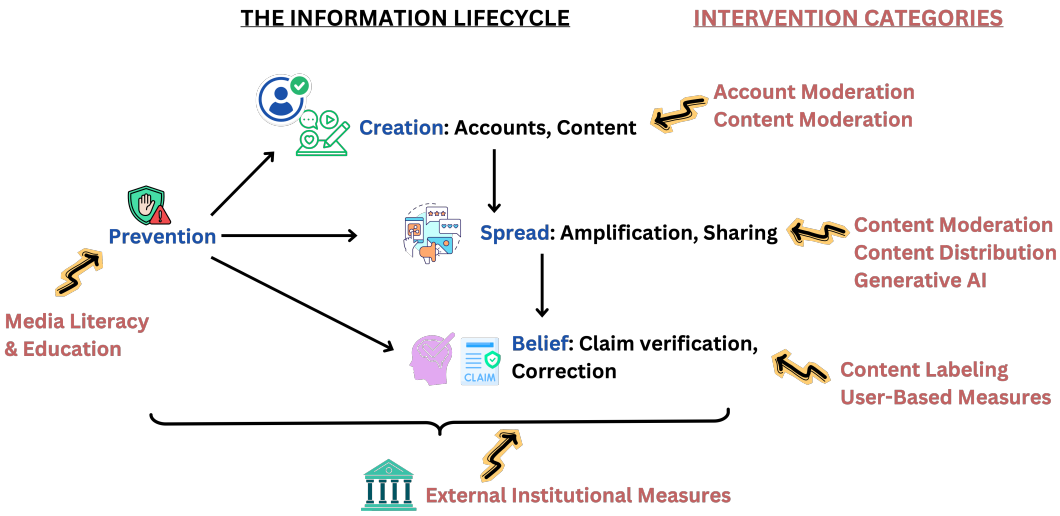

Figure 1: The information lifecycle [62] and corresponding intervention categories.

Table 1: Categories of Socio-Technical Interventions, based on the information lifecycle [62]

Information Lifecycle	Intervention Category	Definition
Creation: accounts, content	Account Moderation	The moderation of user accounts such as by suspending, banning, or limiting users
Spread: viral amplification, individual sharing	Content Moderation	The moderation of content such as by removing, downranking, or correcting it
	Content Distribution	Affecting the distribution/ sharing of content such as by using redirection, accuracy prompts, or friction
	Generative AI-specific	Employing genAI to address harmful or low-quality content (e.g., using chatbots to dialogue with users, using genAI to produce fact checks)
Belief: claim verification, correction	Content Labeling	The use of labeling or disclosure to notify users, provide additional context, or verify claims
	User-based Measures	Relating to user-driven responses to the information, such as social corrections or reporting
Prevention	Media Literacy and Education	Training efforts aimed at improving the public's media literacy and critical thinking skills
External	Institutional Measures	Measures taken outside the platform ecosystems such as regulation, data sharing, and investing in journalism

the dissemination of misinformation, and the correction or prevention of false beliefs and actions. The same lifecycle lens applies to adjacent harms – scams, fraud, coordinated inauthentic behavior, terrorist recruitment etc– that move through creation, spread, and belief on the same platforms.

3.2 PEACE Criteria

For this study, five evaluative criteria were selected after reviewing previous comprehensive review articles and policy analyses [17, 64, 95]. These criteria are: **political feasibility**, **effectiveness**, **acceptance**, **cost**, and **effort level**. The five main criteria are defined as follows:

- **Political feasibility:** The likelihood of implementing an intervention, incorporating the following factors as well as external, political factors, such as stakeholder perspectives, regulatory and legal constraints, and support from relevant officials or organizations.
- **Effectiveness:** The extent to which an intervention successfully targets the creation, spread, or belief phases in the information lifecycle under different circumstances, such as for some users, certain platforms, or particular types of content.
- **Acceptance:** The degree of user support for an intervention, which may be influenced by factors such as its perceived

effectiveness, fairness, intrusiveness, transparency, and the level of trust in the implementer.
- **Cost:** The financial burden on the users, platforms, or government entities involved in implementing and maintaining the intervention.
- **Effort:** The level of administrative effort required to implement and maintain the intervention for the users, platforms, or government entities involved.

The PEACE decision framework provides a balanced overview of each intervention's overall viability. The inherent characteristics of an intervention, along with how users perceive it, directly affect these five key evaluation criteria. More specifically, user perceptions can impact the acceptance of the intervention, thereby affecting public support for the policy, which in turn increases policy compliance and effectiveness [42]. For example, political feasibility likely affected Meta's US fact-checking program in 2025, where despite its effectiveness and broad user acceptance [91], external political pressures led to the program's termination [53, 115].

To provide a comprehensive set of recommendations across the intervention landscape that researchers, companies, and policymakers can use, we conducted an expert elicitation survey based on these five evaluative criteria. Design and decision research have long relied on structured expert judgment when the outcome data for a given intervention is costly, slow or impossible to obtain at the scale of comparison against alternatives [83]. Within misinformation research specifically, survey-based expert elicitation has previously been used to gauge which interventions experts believe should or would be effective, providing direct precedent for soliciting comparative judgments across our criteria, rather than waiting for outcome evidence for multiple interventions individually [7, 17].

4 Data and Methods

4.1 Survey Overview

We conducted an expert elicitation survey to obtain comparative judgments of across the five evaluative criteria of political feasibility, effectiveness, user acceptance, cost, and effort level. The survey was promoted to participants and invitees of the **SUMMIT 2025 [Full summit name, host location, and dates removed for peer review to preserve anonymity.]** Our use of expert elicitation for survey judgment captures how experts anticipate the impact of interventions, including their effectiveness in combatting misinformation, how the public and users are likely to respond to the interventions, and their expert evaluation of each intervention's feasibility, cost, and implementation effort.

The survey data was collected between January 23 and March 24, 2025 from thirty-nine researchers. The experts provided their professional assessments of the effectiveness and acceptability of the general intervention categories (Table 1) using a 1-7 Likert scale. They also each evaluated 12 out of the 40 operationalized interventions (Table 7) across the five evaluative metrics using a 1-5 Likert scale. The specific interventions assigned to each participant were randomly selected. Because of the survey's length, participants were not assigned all 40 interventions. Instead, 12 were assigned to each participant to ensure about 10 (or more) responses per intervention. Informed consent was obtained from all participants. See the supplemental file for a complete copy of the survey.

In the survey, a "1" on the Likert scale indicated low political feasibility, effectiveness, acceptance, cost, and effort; and "5" indicated high. During the analysis, cost and effort level scores were inverted (i.e., a Likert score of "5" indicated low cost and effort), ensuring all five metrics ran in the same direction, with higher scores reflecting better results. The metrics were also renamed to **cheapness** and **ease** of implementation for clarity in the Results section. This made it possible to analyze an overall average score for each intervention and conduct a cluster analysis.

4.2 Demographics

The demographics of the participants who completed the survey are shown in Table 2.

Table 2: Participant Demographics (N = 39)

Category	Response	N	%
Gender	Women	16	41.0
	Men	22	56.4
	Other / prefer not to say	1	2.6
Age	18–34	15	38.5
	35–44	12	30.8
	45+	10	25.6
	Prefer not to say	2	5.1
Academic Rank	Faculty	21	53.8
	PhD Student	12	30.8
	Industry	2	5.1
	Postdoc	1	2.6
	Other	3	7.7
Primary Discipline	Computational Social Sciences	17	43.6
	Computer Science	10	25.6
	Sociology	4	10.3
	Communications and Media Studies	4	10.3
	Political Science	2	5.1
	Other	2	5.1

5 Results and Analysis

5.1 Category-Level Tradeoffs

To better understand how experts perceive the misinformation interventions landscape, we first asked participants to rate the overall effectiveness and user acceptance of each of the 8 intervention categories among the American public. Figure 2 shows the fraction of participants who believed each category of countermeasures to be effective or acceptable to the public. Content moderation, account moderation, and media literacy / education are regarded as the most effective, with over 70% of respondents agreeing they would be effective. However, moderation, especially account moderation, is viewed by experts as one of the least acceptable to the public.

To further investigate the apparent tension between effectiveness and acceptance, Figure 3 illustrates the average Likert scores and 95% confidence intervals for the effectiveness and acceptance ratings. In this figure, we observe that media literacy / education and external institutional measures rise in the effectiveness rankings, bolstered by the proportion of participants who "strongly

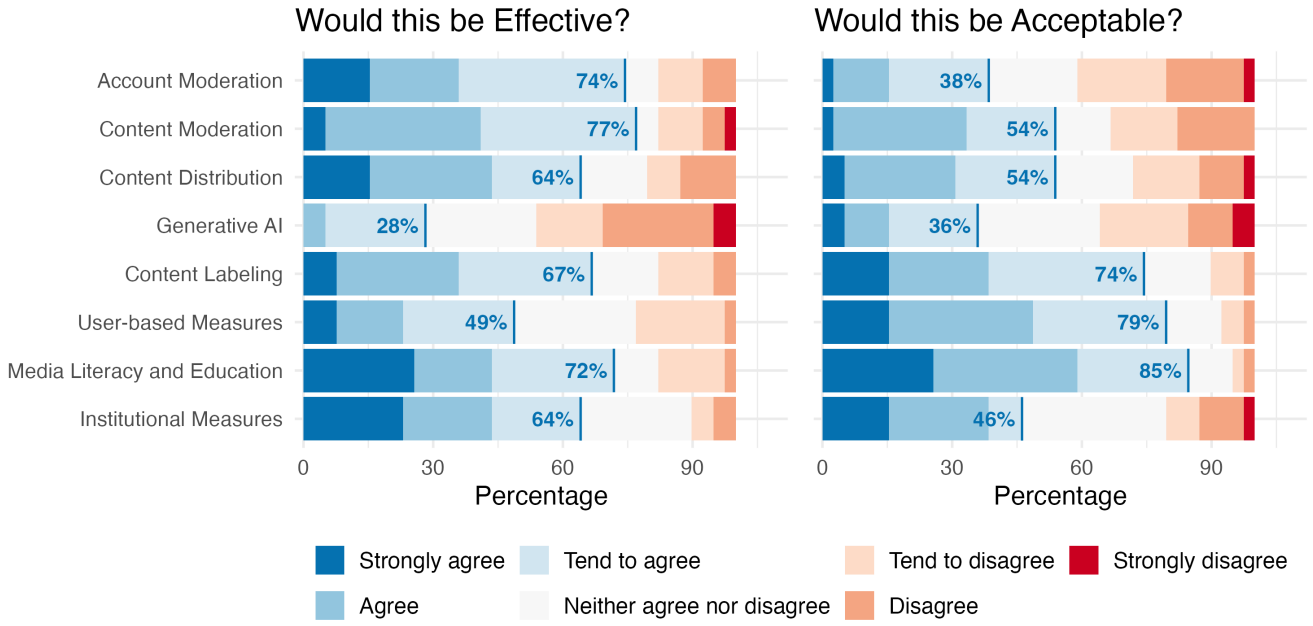

Figure 2: Stacked bar plot showing the percentage of responses on the effectiveness and acceptance of the general intervention categories.

agree" regarding their effectiveness rather than more weakly agreeing. Content and account moderation remain at high levels. While Figure 2 shows general agreement or disagreement among experts, this figure incorporates the strength of their opinions by using the Likert scores directly, making the differences in effectiveness and acceptance across some intervention categories more apparent. Account moderation and user-based measures exhibit the largest gap between effectiveness and acceptance, with effectiveness exceeding acceptance in account moderation, and acceptance exceeding effectiveness in user-based measures (both statistically significant at $p < 0.05$).

These results align with a similar survey of experts, conducted by Altay et al. [7], which found that experts believed that media literacy, labeling of false content, and fact-checking are the most effective interventions. Each of these interventions received around 70% approval. Meanwhile, accuracy prompts (a form of content distribution), source labeling, and inoculation ranked lower on the list. These were the only interventions included in that survey in a similarly structured question [7]. Altay et al. [7] also asked experts if certain types of actions *should* be taken, and concluded that platform design changes, algorithmic changes, and general content moderation as the most popular (over 75% agree), and penalizing misinformation sharing and shadow banning (a form of account moderation) as the least popular. In fact, shadow banning was the only suggested intervention where more experts disagreed with its implementation than agreed. Similarly, Blair et al. [17] found in a survey of experts that they recommended implementing educational and institutional interventions (such as media literacy, platform alterations, and journalist training) over other types of interventions. However, the authors note that these interventions tend to be among the least well-studied or have mixed evidence on actual effectiveness [17].

5.2 Metric Tradeoffs

Beyond effectiveness and acceptance of the high-level intervention categories, we also asked the experts to rate each individual intervention across five criteria: political feasibility, effectiveness, user acceptance, cheapness (inverted cost score), ease of implementation (inverted effort score). Figure 4 presents a heatmap that illustrates the overall average metric scores, averaged by category and ranked from highest to lowest. Content labeling, user-based measures, and content distribution topped the list among the five criteria. Content labeling and user-based measures scored well on effectiveness, acceptance, and feasibility. Content distribution was the most balanced category with the most similar average metric values and the only category where all five metrics averaged above a 3 on the Likert scale. Media literacy and external institutional measures were ranked last, primarily due to their high overall costs (low affordability) and required effort level. Taken together, these results suggest that while some countermeasures are seen as effective, and some are even considered both effective and acceptable by the public, they are not always regarded as practical by experts. Interventions that require more administrative effort or resources to implement tend to rank lower overall, despite their potential positive impact.

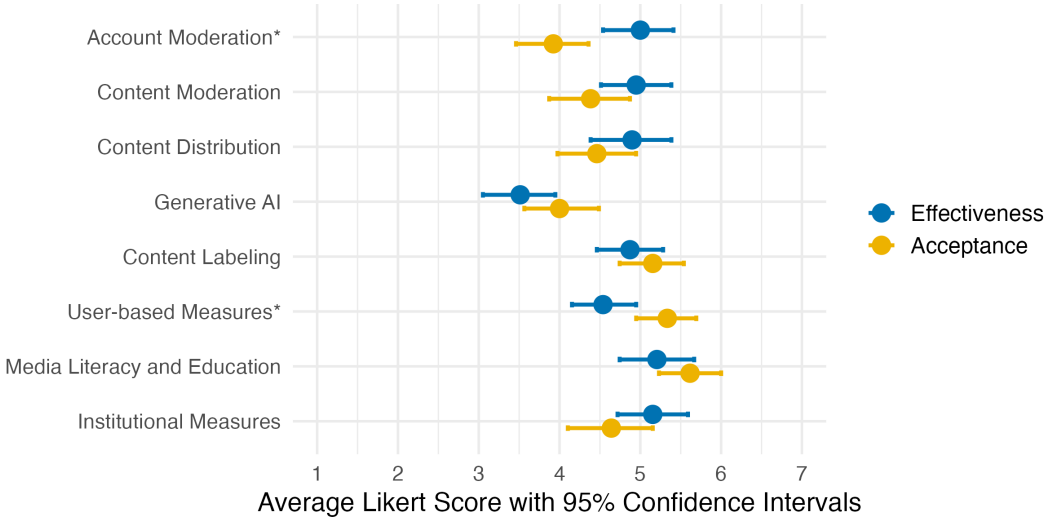

Figure 3: Average effectiveness and acceptance Likert scores per general intervention category.

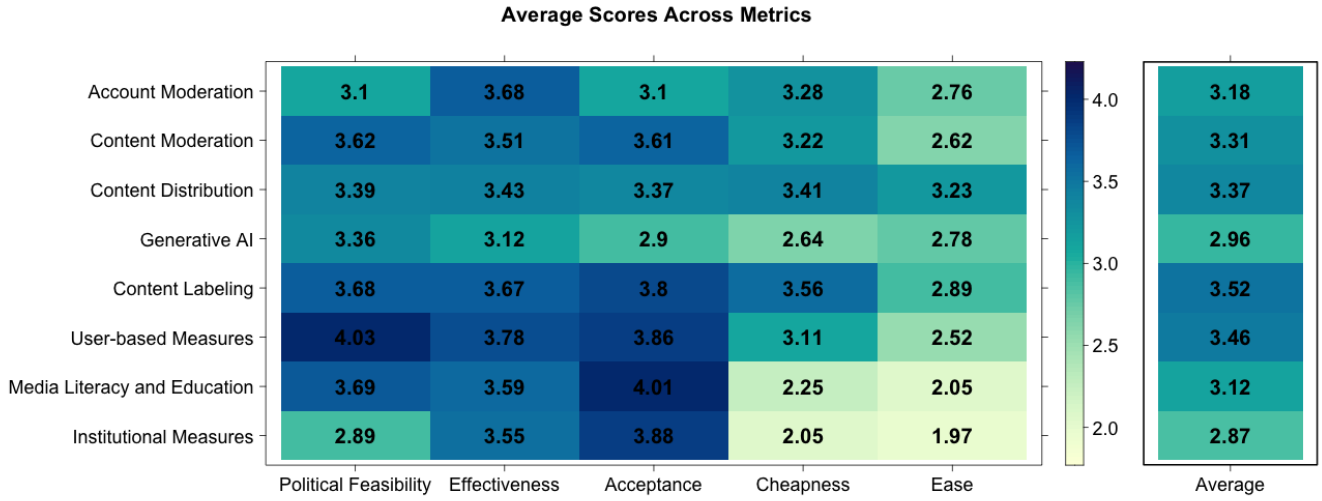

Figure 4: A heatmap showing the average score for each metric across general intervention categories.

5.3 Top Interventions

To identify which individual interventions drive these category-level tradeoffs, and what features they may have in common, we rank interventions by their overall scores (average of the five metrics) in Table 3. In general, interventions in the content labeling, user-based measures, and content distribution interventions dominate the top 10. Only two interventions from moderation categories (user control and demonetization) rank in the top 10 overall, and these moderation interventions are among the least restrictive or provide users with a high level of agency compared to other moderation interventions.

Analyzing only the highest overall scores among individual interventions provides a limited view and fails to highlight the trade-offs among the different metrics. When diving deeper into the top ten interventions, we find that only three are also in the top ten for effectiveness, five for acceptance, effort, and cost, and seven for political feasibility; the interventions rate differently for different metrics. Overall, most of the top interventions rank in the top ten for only two or three metrics. To better analyze the trade-offs, we used base R's implementation of the Ward's D2 clustering method to group interventions based on similar criterion scores [76, 90]. We identified four clusters that summarize recurring trade-off patterns. Table 4 presents the average criteria values for each intervention in each cluster:

(1) *Balanced High Performers (well-rounded, high performing interventions)*
The second largest cluster included 12 interventions. These were mostly platform-implemented interventions that scored

Table 3: Interventions ranked by adjusted average score. The top-ranked intervention in each category is bolded.

Intervention Rank	Category	Score	Intervention Rank	Category	Score
1. Government Labels	**Content Labeling**	**4.04**	21. Virality Circuit Breakers	Content Moderation	3.23
2. Limit Resharing	**Content Distribution**	**3.91**	22. Ban Deepfakes	Content Moderation	3.21
3. Limit Forwarding	Content Distribution	3.91	23. Platform Alterations	Content Distribution	3.20
4. Crowdsourcing	Content Labeling	3.76	**24. Data Sharing**	**Institutional Measures**	**3.20**
5. AI Disclosure	Content Labeling	3.62	25. Warning Labels	Content Labeling	3.18
6. Friction	Content Distribution	3.55	26. Redirection	Content Distribution	3.16
7. Reporting	**User-based Measures**	**3.52**	**27. Digital Media Literacy**	**Media Literacy**	**3.13**
8. User Control	**Content Moderation**	**3.45**	28. Inoculation	Media Literacy	3.10
9. Alter Platform Metrics	User-based Measures	3.45	29. Ban Political Ads	Content Distribution	3.06
10. Demonetization	**Account Moderation**	**3.43**	30. Account Suspensions	Account Moderation	3.01
11. Social Norms	User-based Measures	3.41	31. Deplatforming	Account Moderation	2.98
12. Targeted Ads	Institutional Measures	3.41	32. **Gen AI Chatbots**	**Generative AI**	**2.96**
13. Algorithmic Changes	Content Moderation	3.36	33. Gen AI Content	Generative AI	2.96
14. Accuracy Prompts	Content Distribution	3.35	34. Fact-Check Ads	Content Distribution	2.91
15. Content Removal	Content Moderation	3.31	35. Journalism Support	Institutional Measures	2.89
16. Downranking	Content Moderation	3.31	36. Media Support	Institutional Measures	2.86
17. Shadowbanning	Account Moderation	3.31	37. Taxes / Fines	Institutional Measures	2.85
18. Source Labeling	Content Labeling	3.27	38. Gov't Regulation	Institutional Measures	2.69
19. AI in Ads	Content Distribution	3.26	39. Anti-Trust Action	Institutional Measures	2.62
20. Fact-check Labels	Content Labeling	3.24	40. Privacy Legislation	Institutional Measures	2.45

relatively well across all five metrics. It included many of the less intrusive or higher agency affording interventions like demonetization of user accounts (rather than suspensions or shadowbanning), limits on forwarding or resharing content, and several labeling and user-based measures like fact-checking labels and improved user reporting.

(2) *Effective but Contested (effective and low-cost, but less popular interventions)*
The largest group of interventions (15) consisted of platform-implemented controls that were rated as being low cost to implement and effective to deploy. However, they have relatively low user acceptance. Generally, this category includes many of the more restrictive or intrusive account moderation, content moderation, and content distribution interventions, such as account suspensions, downranking, and banning political ads.

(3) *Effective but Costly (effective and acceptable, but resource-intensive interventions)*
This category included 9 interventions and mostly consisted of almost all of the external institutional measures (such as investing in journalism, pursuing privacy legislation or anti-trust action) and the media literacy/educational measures. These interventions had by far the highest costs and effort associated with implementing and maintaining these programs, and ranked on the lower end of the spectrum for political feasibility relative to other clusters.

(4) *Easy Starters (politically feasible and popular, but low-impact interventions)*
The smallest cluster included only 4 interventions, and it included the measures with by far the highest overall user acceptance scores and highest likelihood of implementation (such as crowdsourcing context labels, regulating deepfakes, and regulating AI in ads). However, in general, they were viewed as less effective than most other measures.

6 Discussion

In this study, 39 experts evaluated misinformation interventions based on five evaluative criteria: political feasibility, effectiveness, acceptance, cost, and effort. An intervention is never perfect; it is a trade-off between these evaluative criteria. The trade-offs between evaluative criteria for interventions are systematic and not random. Across both category-level and individual-level interventions, the data shows a recurring pattern, where interventions that are restrictive or opaque face acceptance and feasibility barriers regardless of their effectiveness. Informational and user-empowering interventions face the opposite problem of high acceptance but uncertain effectiveness. This finding should inform how platforms build their intervention portfolios, that they should diversify across multiple intervention types rather than maximizing on any single metric.

The main findings suggest that, like in many other public policy domains, people prefer informational interventions that afford them agency over more intrusive or restrictive ones [28, 44]. On average, content labeling, user-based measures, and content distribution interventions scored the highest among expert raters across the five metrics. Among these top three categories, content labeling interventions primarily target content after it has already spread, focusing on the belief phase of the misinformation pipeline. Content distribution interventions generally target the spread of misinformation, either addressing the direct sharing component (such as friction and accuracy prompts) or the amplification of the content (by limiting resharing or forwarding). The top-ranked user-based measures included improving user reporting, which also usually targets the spread phase. When determining which interventions to implement, these top categories unfortunately generally target

	Category	Intervention	N	Feasibility	Effectiveness	Acceptance	Cheapness	Ease	Overall
Balanced high performers	Account Moderation	Demonetization	12	3.58 (1.31)	4.00 (0.60)	3.75 (1.06)	2.92 (1.51)	2.92 (1.24)	3.43 (0.52)
	Content Moderation	Algorithmic Changes	11	3.64 (1.03)	4.09 (0.54)	3.55 (1.04)	3.00 (1.26)	2.55 (0.93)	3.36 (0.54)
	Content Distribution	Accuracy Prompts	8	3.38 (1.51)	3.62 (0.92)	3.50 (1.07)	2.88 (1.13)	3.38 (1.06)	3.35 (0.75)
	Content Distribution	Limit Forwarding	13	4.08 (0.86)	3.69 (0.63)	3.54 (1.05)	4.31 (0.85)	3.92 (1.04)	3.91 (0.49)
	Content Distribution	Limit Resharing	9	4.00 (1.22)	3.78 (0.44)	3.00 (1.22)	4.56 (0.53)	4.22 (0.67)	3.91 (0.45)
	Content Labeling	Fact-check Labels	15	3.53 (1.06)	3.87 (0.74)	3.80 (0.94)	2.79 (1.12)	2.21 (1.12)	3.26 (0.59)
	Content Labeling	Government Labels	5	4.00 (0.00)	3.80 (1.30)	4.00 (0.00)	4.40 (0.55)	4.00 (0.00)	4.04 (0.26)
	Content Labeling	AI Disclosure	14	3.71 (1.59)	3.71 (0.99)	4.07 (0.83)	3.31 (1.25)	3.31 (1.18)	3.64 (0.75)
	User-based Measures	Reporting	10	4.40 (0.52)	3.80 (0.79)	4.00 (0.94)	3.00 (1.49)	2.40 (1.07)	3.52 (0.45)
	User-based Measures	Social Norms	13	3.62 (0.77)	3.62 (0.96)	3.83 (0.94)	3.33 (1.23)	2.67 (1.37)	3.40 (0.60)
	User-based Measures	Alter Platform Metrics	12	4.08 (0.79)	3.92 (1.00)	3.75 (1.36)	3.00 (1.41)	2.50 (1.09)	3.45 (0.62)
	Institutional Measures	Targeted Ads	13	3.38 (1.19)	3.38 (0.65)	3.85 (0.80)	3.08 (1.38)	3.33 (1.30)	3.44 (0.64)
		Cluster 1 mean		**3.78**	**3.77**	**3.72**	**3.38**	**3.12**	**3.56**
Effective and low-cost, but less popular interventions	Account Moderation	Account Suspensions	15	2.80 (1.52)	3.53 (0.99)	2.86 (1.29)	3.47 (1.13)	2.40 (0.83)	3.02 (0.64)
	Account Moderation	Deplatforming	12	2.58 (1.31)	3.83 (0.83)	2.92 (1.00)	3.25 (1.36)	2.33 (1.07)	2.98 (0.38)
	Account Moderation	Shadow Banning	9	3.44 (1.24)	3.33 (1.32)	2.89 (1.05)	3.50 (1.69)	3.38 (1.51)	3.27 (0.90)
	Content Moderation	Content Removal	9	3.11 (1.17)	3.89 (0.93)	3.33 (0.87)	3.78 (0.97)	2.44 (1.24)	3.31 (0.66)
	Content Moderation	Downranking	13	3.15 (1.46)	3.92 (0.76)	3.23 (1.17)	3.15 (1.28)	3.08 (1.04)	3.31 (0.69)
	Content Moderation	Virality Circuit Breakers	12	3.58 (1.08)	3.58 (1.38)	3.08 (0.90)	3.00 (1.28)	2.92 (1.00)	3.23 (0.65)
	Content Distribution	Redirection	10	2.60 (1.35)	3.10 (0.99)	3.10 (1.10)	3.90 (0.88)	3.10 (1.10)	3.16 (0.45)
	Content Distribution	Friction	12	3.75 (0.97)	3.50 (1.00)	3.33 (1.07)	3.83 (1.47)	3.33 (1.15)	3.55 (0.67)
	Content Distribution	Platform Alterations	10	3.60 (1.07)	3.80 (1.32)	3.00 (0.94)	3.10 (1.37)	2.50 (0.97)	3.20 (0.65)
	Content Distribution	Ban Political Ads	15	2.20 (1.32)	3.00 (0.93)	3.21 (1.37)	3.27 (1.49)	3.60 (1.30)	3.05 (0.72)
	Generative AI	Gen AI Chatbots	18	3.39 (0.92)	3.24 (0.75)	2.71 (0.92)	2.65 (1.17)	2.82 (1.01)	2.96 (0.59)
	Content Labeling	Warning Labels	12	3.42 (1.08)	3.75 (0.97)	3.17 (1.03)	3.17 (1.40)	2.42 (1.08)	3.18 (0.67)
	Content Labeling	Source Labeling	9	3.11 (1.17)	3.67 (0.87)	3.67 (0.87)	3.56 (1.24)	2.33 (1.00)	3.27 (0.66)
	Generative AI	Gen AI Content	12	3.33 (0.65)	3.00 (0.85)	3.09 (0.83)	2.64 (0.81)	2.73 (0.79)	2.96 (0.50)
	Institutional Measures	Taxes / Fines	13	2.15 (0.99)	3.08 (1.26)	3.92 (1.32)	2.83 (1.64)	2.25 (1.48)	2.88 (0.77)
		Cluster 2 mean		**3.08**	**3.48**	**3.17**	**3.27**	**2.78**	**3.16**
Effective and acceptable, but resource-intensive	Content Distribution	Fact-Check Ads	9	3.33 (1.00)	3.44 (0.73)	3.75 (0.71)	2.00 (0.93)	2.00 (0.93)	2.95 (0.62)
	Media Literacy	Digital Media Literacy	9	3.78 (0.83)	3.78 (1.20)	4.11 (0.93)	2.00 (1.12)	2.00 (1.12)	3.13 (0.51)
	Media Literacy	Inoculation	10	3.60 (0.84)	3.40 (0.84)	3.90 (0.88)	2.50 (0.97)	2.10 (0.99)	3.10 (0.56)
	Institutional Measures	Media Support	10	3.30 (1.34)	3.40 (1.07)	3.78 (0.67)	1.80 (0.92)	2.00 (0.94)	2.83 (0.66)
	Institutional Measures	Journalism Support	13	3.46 (1.20)	4.00 (0.71)	3.77 (1.01)	1.38 (0.51)	1.85 (0.80)	2.89 (0.42)
	Institutional Measures	Data Sharing	11	3.00 (1.41)	4.09 (0.83)	4.18 (0.75)	2.73 (1.01)	2.00 (1.41)	3.20 (0.86)
	Institutional Measures	Government Regulation	11	2.91 (1.14)	3.64 (1.12)	3.55 (0.82)	1.82 (0.60)	1.55 (1.04)	2.69 (0.40)
	Institutional Measures	Privacy Legislation	12	2.75 (1.48)	3.25 (1.36)	3.80 (0.92)	1.36 (0.50)	1.09 (0.30)	2.51 (0.86)
	Institutional Measures	Anti-Trust Action	10	2.20 (1.40)	3.60 (1.07)	4.20 (0.63)	1.40 (0.70)	1.70 (1.06)	2.62 (0.55)
		Cluster 3 mean		**3.15**	**3.62**	**3.89**	**1.89**	**1.81**	**2.88**
Easy starters	Content Moderation	User Control	14	4.21 (1.12)	3.00 (1.18)	4.36 (0.84)	3.38 (1.39)	2.31 (0.95)	3.47 (0.70)
	Content Distribution	AI in Ads	18	3.61 (1.33)	2.94 (1.00)	3.89 (0.96)	2.83 (1.34)	3.00 (1.24)	3.26 (0.68)
	Content Moderation	Ban Deepfakes	11	4.00 (0.89)	2.55 (1.29)	4.10 (0.99)	3.00 (1.25)	2.40 (1.17)	3.26 (0.68)
	Content Labeling	Crowdsourcing	14	4.29 (0.61)	3.21 (0.80)	4.08 (0.64)	4.14 (1.17)	3.07 (1.27)	3.76 (0.60)
		Cluster 4 mean		**4.03**	**2.93**	**4.11**	**3.34**	**2.69**	**3.44**

Table 4: Interventions sorted by cluster and category with mean scores and standard deviations (in parentheses). Cluster mean rows show the average across interventions in that cluster.

only the **spread** and **belief** in misinformation rather than the **creation** of networks or content. An ideal enforcement strategy would be to determine and implement the best interventions across the misinformation pipeline.

However, two moderation strategies ranked among the top 10 highest-ranked interventions: altering platform metrics to reward accuracy rather than engagement and using demonetization as an account moderation strategy. These interventions target the incentives associated with creating misinformation content and can also affect the potential spread and amplification. Experts in the elicitation survey ranked them among the most effective interventions (5th and 3rd, respectively), although they are perceived to have middling user acceptance and require a decent amount of effort or cost to implement. The literature supports these results, suggesting that reducing incentives to create and share misinformation (or conversely, increasing either monetary or social incentives to share accurate information) can considerably improve the quality of content shared and improve user discernment [38, 54].

Further, comparing our results with a similar survey of public opinion [7], we note that expert judgment does not necessarily track public perception. For example, members of the public rate friction-based interventions, like delaying users from posting content they have not viewed, as very intrusive and among the least effective [62], but experts rate them as moderately effective. Notably, this expert-public mismatch is not uniform. For content moderation, experts and the public converge on the trade-off structure itself, independently judging that both account and content moderation are effective yet unpopular. This observation argues for participatory design for sociotechnical interventions, where expert elicitation are not treated as a direct stand in for public input, but user opinions

are also folded into the decision process. This would surface disagreements before deployment and create a more inclusive online space.

Finally, these sets of interventions go beyond the scope of misinformation. They target mechanisms of account and content visibility and distribution that recur across socio-technical systems more broadly. For example, account and content moderation are primary levers that platforms use to constrain the amplification of undesired narratives by automated and coordinated inauthentic accounts, regardless of whether the content is misinformation or not [81].

Consider the case of coordinated inauthentic behavior, where networks of automated and human-operated accounts amplify desired narratives [80]. The same five criteria apply, and they reproduce the same tension we find in misinformation. Account-level takedowns are effective, but opaque and contested [81], placing them in the low-acceptance, high-effectiveness cluster; transparency measures such as behavioral labels are acceptable but effects may vary, echoing the easy-starter cluster. The lifecycle gap recurs here too. Takedowns act only after a network has already amplified content, leaving account creation, the stage at which coordination is the easiest to disrupt, largely unaddressed, because it is difficult to predict if the account will later coordinate. Generally, the lifecycle lens and the trade-off structure carry over, though the interventions may need re-fitting to the domain.

6.1 Design Implications

Based on the main findings and the four calculated intervention clusters, we derive design implications and then propose a decision framework for practitioners.

6.1.1 Intervention Cluster Implications. The four calculated intervention clusters point to specific design implications. See Section 5.3 and Table 4 for details on each of the four clusters.

(1) Balanced High Performers - Design for user agency where the effectiveness case allows for it. Across the category-level comparison in Figure 4 and the intervention-level analysis (Table 4), user agency is a recurring feature of the most acceptable platform interventions in the cluster of well-balanced high performers. The interventions that users find most acceptable preserve their ability to interpret, contest, or act on information, rather than removing the choice for them. Among the top 10 overall interventions (Table 3), labeling government sources, crowdsourcing labels, AI disclosure, encouraging user reporting, and allowing users to control their news feeds directly all preserve a high level of information or agency for the user, rather than restrict it.

(2) Effective but Contested - Acceptance should be a necessary design target, not cost. Some of the most effective interventions, including many that are low-cost and have a low administrative burden, are unfortunately not always acceptable to users. As shown in Figure 2, large majorities of our expert participants rated the account and content moderation categories as effective (74% and 77%, respectively), but they also show the largest drop in percentage that believe these interventions would be acceptable to the general public (only 38% and 54%, respectively). This gap is also acknowledged in public policy literature, where acceptance is driven by the perceived fairness, transparency and trust in the implementation, or the implementer, rather than the intervention's actual effectiveness [42]. Therefore, implementers should focus on increasing user acceptance by improving trust in the systems and also improving the public messaging. Prior HCI work also shows that users tend to distrust what they cannot see or appeal [77, 81], so labeling systems should reduce opacity and include recourse and source transparency.

(3) Effective but Costly - Lower the cost of interventions that are already rated as highly effective and acceptable. The media literacy and external institutional measures in cluster 3 are rated as both highly effective and among the most acceptable, yet their overall scores on average are the lowest (see Table 4, Figure 4) because they are often too resource-intensive to be practiced. In this case, the trade-off is not values-based (as in moderation), but the burden of implementation. This points to a different design lever: reducing delivery cost rather than reducing scope. Meta-analytic work on media literacy interventions show that effectiveness is a measure that is sensitive to how and where an intervention is delivered [47], suggesting that institutions without dedicated funding should look to amortize these interventions into existing infrastructure. For example, literacy prompts could be embedded into product onboarding flows rather than a freestanding curricular.

(4) Easy Starters - Emerging interventions that are straightforward and low-cost are a good first step. Cluster 4 includes many newer interventions, such as regulating AI in social media ads, controlling the use of deepfakes, crowdsourcing fact-checking labels, and giving users control over their news feeds. These interventions have less evidence of effectiveness than the more established interventions in clusters 1-3, and their implementation would likely require creating and deploying new features. As Table 4 shows, these interventions score poorly on effectiveness and ease of implementation but rank highest among all clusters in political feasibility and user acceptance. This may partly be because these interventions are either high-agency-preserving or focus on regulating or controlling AI, a technology which is currently relatively unpopular across demographic groups in the US [73]. Since they are also low-cost and relatively simple to deploy, requiring only that platforms choose to do so, they present a low-risk option that platforms could adopt. Implementers can quickly adopt them, monitor user responses, and iterate to improve these interventions. These changes could enhance their future effectiveness and acceptability. We therefore recommend that practitioners view cluster 4 not as a final set of solutions but as a starting point: low-cost interventions that can be deployed quickly, are transparent, and empower users. This can help establish the legitimacy and trust needed between users and platforms to later implement more restrictive yet more effective interventions.

6.1.2 General Implications. Beyond the cluster-level patterns, a structural gap emerges when interventions are mapped onto the information lifecycle (Figure 1). The most viable interventions, i.e. those in Cluster 1, are concentrated at the spread and belief phases, leaving the creation phase largely unaddressed. Of the top 10 interventions overall (Table 3), only two (altering platform metrics, demonetization), directly target the incentives that drive

content creation. However, both have middling acceptance scores (both average a 3.75 on a 5-pt Likert scale, and rank 19th/ 20th out of 40 interventions on acceptance respectively). This aligns with research showing that misinformation persists because of platform incentive structures that favor engagement over accuracy at the point of content creation [38]. This is not merely an artifact of what has been studied, instead, it reflects a structural feature of the field: the most mature, acceptable and deployable tools act downstream of the incentive structures that generate the content in the first place, and are unable to address the root cause of content creation [38].

This pipeline coverage gap has direct implications for platform architecture. Closing it will require designing creation-phase interventions that can occupy the same viable space as the spread and belief phase interventions in Cluster 1 – transparency, agency-preserving and low-burden interventions. For the existing creation-phase interventions that come the closest (demonetization and altering platform metrics), they are already pointing in this direction: these interventions work by restructuring incentives rather than restricting behavior, which likely explains why they are better received than more overtly restrictive creation-phase measures like account suspension and deplatforming.

It is critical that all stages of the information lifecycle are addressed with interventions. Future intervention design should take this as a signal: creation-phase interventions framed as restructuring incentives, rather than restricting participation, may be more likely to achieve the acceptance needed for sustained deployment. More broadly, these findings argue against single-intervention strategies, and that combining interventions across different mechanisms reduces the spread of viral misinformation [13]. Because the most viable interventions cluster at the specific pipeline stages, and trade-off against each other on the five criteria, an effective implementation strategy requires a portfolio that deliberately spans the lifecycle. It should pair high-acceptance spread and belief-phase interventions with creation-phase measures whose acceptance can be incrementally built. In the next section, we propose a decision framework that is intended to support this kind of portfolio reasoning.

6.1.3 A Decision Framework. The four empirically derived clusters translate directly into a decision framework for practitioners. Organizations deploying socio-technical interventions typically operate under a specific set of constraints (i.e., budget, political exposure, timeline, or public trust), and the clusters align cleanly with the four constraint profiles. Table 5 maps each cluster to the conditions under which it is the appropriate starting point, with representative interventions and design considerations. This proposed decision framework should be used as a navigation tool. It begins with where the practitioner is constrained, which identifies the starting cluster of interventions, then, what design investments are needed to unlock the next cluster.

A key implication of this framework is that constraint profiles change over time, and the clusters are designed to be used sequentially as well as independently. Cluster 4 interventions (that are low-cost, high user acceptance and politically feasible), are most appropriate as a first step to build the legitimacy of the interventions, establishing the trust between users and platforms. This is a precondition for later deploying more effective but more contested interventions that are in Clusters 1 and 2. Cluster 3 interventions are both highly effective and well-received but are resource-intensive, and therefore should be pursued when coalition partners or dedicated funding can amortize the delivery cost. For example, embedding media literacy prompts into product onboarding flows rather than standalone curricula, is one way to reduce the structural barrier without reducing the intervention scope.

Critically, no single cluster is sufficient on its own. The pipeline coverage gap identified in Section 6.1.2 means that an intervention portfolio needs to balance across multiple clusters. For example, sole reliance on Cluster 1 interventions will systematically under-address account and content creation. Therefore, practitioners should pair any Cluster 1 deployment with at least one creation-phase measure, most likely from Cluster 2. They should also invest in the transparency and recourse mechanisms that can shift the measure's acceptance over time.

6.2 Limitations

While this study is among the first expert surveys on misinformation interventions and the only one so far to analyze multiple evaluative metrics at once, several limitations are noted. First, our sample was limited in both size (n = 39) and geographical reach. The researchers surveyed were based predominantly at a **Northern American academic institution, with researchers from [[Host university name withheld]]** in particular representing a high fraction of respondents. Second, while the intervention list had forty interventions, respondents only each rated twelve of the interventions to limit the survey length and improve response rates. This limited the sample size for each specific intervention, although all respondents rated the overall effectiveness and acceptance of the general countermeasures categories. Future work should consider a larger sample size or a more diverse set of respondents. We additionally weighted all five evaluative criteria equally, although some are plausibly correlated. An extension of this work could analyze the relationship between the metrics and how that may affect overall ratings. Further, there may be metrics not included in this work that may be relevant to practitioners or policymakers that could be valuable to investigate in future research. Finally, this survey captures expert judgment only, and not public perception of interventions. While our discussion compares our expert ratings against companion user findings from literature, that comparison spans separate data collection efforts. Future work should expand the survey to the general public for a paired measurement.

7 Conclusion

This work set out to give designers and policymakers a principled basis for comparing sociotechnical interventions, rather than evaluating them one at a time on effectiveness alone. We defined eight categories of interventions and forty operationalized interventions that can be deployed along the information lifecycle. Instantiating this through the case of misinformation, we surveyed $N = 39$ researchers on the forty interventions across five evaluative criteria. The five criteria are: political feasibility, effectiveness, acceptance, cost, and effort. Our results find that interventions that experts

Table 5: A practitioner decision framework for selecting misinformation interventions, mapped to the four empirically derived clusters in Table 4. † marks creation-phase interventions.

Cluster	When to choose it	Example interventions	Design considerations
1 Well-balanced, high-performing	Needs broadly viable wins without requiring unusual budget or political cover	Demonetization†; limits on resharing/forwarding; fact-check and government labels; reporting; AI disclosure labels	Preserves users' ability to interpret, contest, or act on information rather than removing the choice for them. Pair with at least one creation-phase measure from cluster 2 and invest in increasing that measure's acceptance
2 Effective and affordable, but low acceptance	Has an enforcement mandate or legal cover and can absorb public pushback	Account suspension†; deplatforming†; shadow banning†; downranking; content removal	Acceptance cost is addressable, not fixed. Invest in transparent criteria, reduced opacity, recourse options, and error minimization
3 Effective and acceptable, but resource-intensive	Has a long time horizon, coalition partners, or dedicated funding	Media literacy programs; journalism/media support; data sharing; regulation	Share cost across a coalition (platform + civil society + government); look for delivery channels that amortize costs
4 Emerging, straightforward, and low-cost	Needs to build public trust or political capital before bigger moves	Crowdsourcing; user control; deepfake/AI-in-ads regulation	Useful as a legitimacy-building first step or complement, designed with user agency in mind. Treat as an initial testbed to iterate and improve effectiveness, not as the primary countermeasure

judge to be the most effective are often not those judged to be most acceptable or feasible. User-based measures, content labeling and content distribution measures scored the highest in multiple evaluative metrics. Additionally, some provably effective interventions, such as content and account moderation techniques, lack user support and acceptance due to their perceived unfairness or intrusiveness.

An intervention that experts rate highly effective can still fail in deployment if it is unacceptable, too costly or politically infeasible. The framework we presented is a first step towards a comparative, multi-criteria evaluation of sociotechnical interventions. The framework is intended to support principled design decisions by policymakers and platform governance teams through identifying trade-offs. This framework also extends naturally beyond misinformation, including to interventions targeting the automated coordinated accounts that platforms moderate independently of content. The five criteria, the lifecycle lens and the cluster logic separates viable-but-limited interventions from effective-but-contested ones, providing a repeatable method for designing and evaluating such socio-technical interventions in different scenarios.

Acknowledgments

This work was supported by the Knight Foundation, the Office of Naval Research's MURI: Persuasion, Identity, & Morality in Social-Cyber Environments grant N00014-21-12749, Carnegie Mellon University's Graduate small Project Help (GuSH), the Center for Computational Analysis of Social and Organizational Systems (CASOS), and the Center for Informed Democracy and Social-cybersecurity (IDeaS). The views and conclusions contained in this document are those of the authors alone. The funders have no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] Accountable Tech. 2024. *Democracy by Design: Social Media's Policy Scores*. Technical Report. Accountable Tech. https://accountabletech.org/research/democracy-by-design-social-medias-policy-scores/

[2] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, et al. 2017. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys (CSUR)* 50, 3 (2017), 1–41.

[3] Zhila Aghajari, Eric P. S. Baumer, and Dominic DiFranzo. 2023. Reviewing Interventions to Address Misinformation: The Need to Expand Our Vision Beyond an Individualistic Focus. In *Proceedings of the ACM on Human-Computer Interaction (CSCW, Vol. 7)*. Association for Computing Machinery, 87:1–87:34. doi:10.1145/3579520

[4] Kathryn J. Aikin, Brian G. Southwell, Ryan S. Paquin, Douglas J. Rupert, Amie C. O'Donoghue, Kevin R. Betts, and Philip K. Lee. 2017. Correction of misleading information in prescription drug television advertising: The roles of advertisement similarity and time delay. *Research in Social and Administrative Pharmacy* 13, 2 (March 2017), 378–388. doi:10.1016/j.sapharm.2016.04.004

[5] Jennifer Allen, Antonio A. Arechar, Gordon Pennycook, and David G. Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science Advances* 7, 36 (Sept. 2021), eabf4393. doi:10.1126/sciadv.abf4393

[6] Jennifer Allen, Cameron Martel, and David G. Rand. 2022. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19. doi:10.1145/3491102.3502040

[7] Sacha Altay, Manon Berriche, Hendrik Heuer, Johan Farkas, and Steven Rathje. 2023. A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School Misinformation Review* 4 (July 2023). Issue 4. doi:10.37016/mr-2020-119

[8] Ahmer Arif, John J. Robinson, Stephanie A. Stanek, Elodie S. Fichet, Paul Townsend, Zena Worku, and Kate Starbird. 2017. A Closer Look at the Self-Correcting Crowd: Examining Corrections in Online Rumors. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*. Association for Computing Machinery, New York, NY, USA, 155–168. doi:10.1145/2998181.2998294

[9] Dennis Assenmacher, Derek Weber, Mike Preuss, André Calero Valdez, Alison Bradshaw, Björn Ross, Stefano Cresci, Heike Trautmann, Frank Neumann, and Christian Grimme. 2022. Benchmarking Crisis in Social Media Analytics: A Solution for the Data-Sharing Problem. *Social Science Computer Review* 40, 6 (Dec. 2022), 1496–1522. doi:10.1177/08944393211012268

[10] Mihai Avram, Nicholas Micallef, Sameer Patil, and Filippo Menczer. 2020. Exposure to social engagement metrics increases vulnerability to misinformation. *Harvard Kennedy School Misinformation Review* 1, 5 (July 2020). doi:10.37016/mr-2020-033

[11] Sumitra Badrinathan and Simon Chauchard. 2023. "I Don't Think That's True, Bro!" Social Corrections of Misinformation in India. *The International Journal of Press/Politics* 29 (Feb. 2023), 394–416. doi:10.1177/19401612231158770

[12] Bence Bago, David G. Rand, and Gordon Pennycook. 2020. Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General* 149, 8 (2020), 1608–1613. doi:10.1037/xge0000729

[13] Joseph B. Bak-Coleman, Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S. Schafer, Emma S. Spiro, Kate Starbird, and Jevin D. West. 2022. Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour* 6, 10 (June 2022), 1–9. doi:10.1038/s41562-022-01388-6

[14] Anton Barbashin. 2018. *Improving the Western Strategy to Combat Kremlin Propaganda and Disinformation.* Technical Report. Atlantic Council: Eurasia Center. https://www.atlanticcouncil.org/wp-content/uploads/2018/06/Improving_the_Western_Strategy.pdf

[15] Paul Barrett. 2019. *Tackling Domestic Disinformation: What the Social Media Companies Need to Do.* Technical Report. NYU Stern Center for Business and Human Rights. https://issuu.com/nyusterncenterforbusinessandhumanri/docs/nyu_domestic_disinformation_digital

[16] Libby Bishop and Daniel Gray. 2017. Ethical Challenges of Publishing and Sharing Social Media Research Data. In *The Ethics of Online Research*, Kandy Woodfield (Ed.). Advances in Research Ethics and Integrity, Vol. 2. Emerald Publishing Limited, 159–187. https://doi.org/10.1108/S2398-601820180000002007.

[17] Robert A. Blair, Jessica Gottlieb, Brendan Nyhan, Laura Paler, Pablo Argote, and Charlene J. Stainfield. 2024. Interventions to counter misinformation: Lessons from the Global North and applications to the Global South. *Current Opinion in Psychology* 55 (Feb. 2024), 101732. doi:10.1016/j.copsyc.2023.101732

[18] Leticia Bode and Emily K. Vraga. 2015. In Related News, That Was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media. *Journal of Communication* 65, 4 (Aug. 2015), 619–638. doi:10.1111/jcom.12166

[19] Leticia Bode and Emily K. Vraga. 2018. See Something, Say Something: Correction of Global Health Misinformation on Social Media. *Health Communication* 33, 9 (Sept. 2018), 1131–1140. doi:10.1080/10410236.2017.1331312

[20] Alexander Bor, Mathias Osmundsen, Stig Hebbelstrup Rye Rasmussen, Anja Bechmann, and Michael Bang Petersen. 2020. "Fact-checking" videos reduce belief in misinformation and improve the quality of news shared on Twitter. doi:10.31234/osf.io/a7huq

[21] Samantha Bradshaw and Lisa-Maria Neudert. 2021. *The Road Ahead: Mapping Civil Society Responses to Disinformation.* Technical Report. National Endowment for Democracy. https://www.ned.org/wp-content/uploads/2021/01/The-Road-Ahead-Mapping-Civil-Society-Responses-to-Disinformation-Bradshaw-Neudert-Jan-2021-2.pdf

[22] Kathleen M Carley. 2020. Social cybersecurity: an emerging science. *Computational and Mathematical Organization Theory* 26, 4 (2020), 365–381.

[23] Man-pui Sally Chan, Christopher R. Jones, Kathleen Hall Jamieson, and Dolores Albarracín. 2017. Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science* 28, 11 (Nov. 2017), 1531–1546. doi:10.1177/0956797617714579

[24] Lesley Chiou and Catherine Tucker. 2018. Fake News and Advertising on Social Media: A Study of the Anti-Vaccination Movement. doi:10.3386/w25223

[25] Giovanni Luca Ciampaglia. 2018. The Digital Misinformation Pipeline. In *Positive Learning in the Age of Information: A Blessing or a Curse?*, Olga Zlatkin-Troitschanskaia, Gabriel Wittum, and Andreas Dengel (Eds.). Springer Fachmedien, Wiesbaden, 413–421. doi:10.1007/978-3-658-19567-0_25

[26] Thomas H. Costello, Gordon Pennycook, and David G. Rand. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science* 385, 6714 (Sept. 2024), eadq1814. doi:10.1126/science.adq1814

[27] Laura Courchesne, Julia Ilhardt, and Jacob N. Shapiro. 2021. Review of social science research on the impact of countermeasures against influence operations. *Harvard Kennedy School Misinformation Review* 2 (Sept. 2021). doi:10.37016/mr-2020-79

[28] Stephanie Diepeveen, Tom Ling, Marc Suhrcke, Martin Roland, and Theresa M. Marteau. 2013. Public acceptability of government intervention to change health-related behaviours: a systematic review and narrative synthesis. *BMC Public Health* 13, 1 (Aug. 2013), 756. doi:10.1186/1471-2458-13-756

[29] Susan E. Dudley and Jerry Brito (Eds.). 2012. *Regulation: A Primer* (2nd ed ed.). Mercatus Center at George Mason University, Arlington, Va.

[30] Ullrich K. H. Ecker and Luke M. Antonio. 2021. Can you believe it? An investigation into the impact of retraction source credibility on the continued influence effect. *Memory & Cognition* 49, 4 (May 2021), 631–644. doi:10.3758/s13421-020-01129-y

[31] Ullrich K. H. Ecker, Jasmyne A. Sanderson, Paul McIlhiney, Jessica J. Rowsell, Hayley L. Quekett, Gordon DA Brown, and Stephan Lewandowsky. 2023. Combining refutations and social norms increases belief change. *Quarterly Journal of Experimental Psychology (2006)* 76, 6 (June 2023), 1275–1297. doi:10.1177/17470218221111750

[32] Ziv Epstein, Adam J. Berinsky, Rocky Cole, Andrew Gully, Gordon Pennycook, and David G. Rand. 2021. Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School Misinformation Review* 2, 3 (May 2021). doi:10.37016/mr-2020-71

[33] Batya Friedman and David G Hendry. 2019. *Value sensitive design: Shaping technology with moral imagination.* Mit Press.

[34] Mingkun Gao, Ziang Xiao, Karrie Karahalios, and Wai-Tat Fu. 2018. To Label or Not to Label: The Effect of Stance and Credibility Labels on Readers' Selection and Perception of News Articles. In *Proceedings of the ACM on Human-Computer Interaction (CSCW, Vol. 2).* Association for Computing Machinery, 55:1–55:16. doi:10.1145/3274324

[35] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

[36] Tarleton Gillespie. 2022. Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society* 8, 3 (July 2022), 20563051221117552. doi:10.1177/20563051221117552

[37] Henner Gimpel, Sebastian Heger, Christian Olenberger, and Lena Utz. 2021. The Effectiveness of Social Norms in Fighting Fake News on Social Media. *Journal of Management Information Systems* 38, 1 (Jan. 2021), 196–221. doi:10.1080/07421222.2021.1870389

[38] Laura K Globig, Nora Holtz, and Tali Sharot. 2023. Changing the incentive structure of social media platforms to halt the spread of misinformation. *eLife* 12 (June 2023), e85767. doi:10.7554/eLife.85767

[39] Robert Gorwa. 2019. What is platform governance? *Information, Communication & Society* 22, 6 (May 2019), 854–871. doi:10.1080/1369118X.2019.1573914

[40] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (Jan. 2020), 2053951719897945. doi:10.1177/2053951719897945

[41] Yasmin Green, Andrew Gully, Yoel Roth, Abhishek Roy, Joshua A. Tucker, and Alicia Wanless. 2022. *Evidence-Based Misinformation Interventions: Challenges and Opportunities for Measurement and Collaboration.* Technical Report. Carnegie Endowment for International Peace. https://carnegieendowment.org/2023/01/09/evidence-based-misinformation-interventions-challenges-and-opportunities-for-measurement-and-collaboration-pub-88661

[42] Sonja Grelle and Wilhelm Hofmann. 2024. When and Why Do People Accept Public-Policy Interventions? An Integrative Public-Policy-Acceptance Framework. *Perspectives on Psychological Science* 19, 1 (Jan. 2024), 258–279. doi:10.1177/17456916231180580

[43] Andrew M. Guess, Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences* 117, 27 (July 2020), 15536–15545. doi:10.1073/pnas.1920498117

[44] Désirée Hagmann, Michael Siegrist, and Christina Hartmann. 2018. Taxes, labels, or nudges? Public acceptance of various interventions designed to reduce sugar intake. *Food Policy* 79 (Aug. 2018), 156–165. doi:10.1016/j.foodpol.2018.06.008

[45] Todd C. Helmus and Bilva Chandra. 2024. *Generative Artificial Intelligence Threats to Information Integrity and Potential Policy Responses.* Technical Report. RAND Corporation. https://www.rand.org/pubs/perspectives/PEA3089-1.html

[46] Todd C. Helmus and Marta Kepe. 2021. *A Compendium of Recommendations for Countering Russian and Other State-Sponsored Propaganda.* Technical Report. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA894-1.html

[47] Guanxiong Huang, Wufan Jia, and Wenting Yu. 2024. Media Literacy Interventions Improve Resilience to Misinformation: A Meta-Analytic Investigation of Overall Effect and Moderating Factors. *Communication Research* (Oct. 2024), 00936502241288103. doi:10.1177/00936502241288103

[48] Se-Hoon Jeong, Hyunyi Cho, and Yoori Hwang. 2012. Media Literacy Interventions: A Meta-Analytic Review. *The Journal of Communication* 62, 3 (June 2012), 454–472. doi:10.1111/j.1460-2466.2012.01643.x

[49] Shagun Jhaver and Amy X. Zhang. 2023. Do users want platform moderation or individual control? Examining the role of third-person effects and free speech support in shaping moderation preferences. *New Media & Society* 27 (Dec. 2023). Issue 5. doi:10.1177/14614448231217993

[50] Jialun Aaron Jiang, Peipei Nie, Jed R. Brubaker, and Casey Fiesler. 2023. A Trade-off-centered Framework of Content Moderation. *ACM Transactions on Computer-Human Interaction* 30, Article 3 (March 2023), 34 pages. doi:10.1145/3534929

[51] Amelia Johns, Francesco Bailo, Emily Booth, and Marian-Andrei Rizoiu. 2024. Labelling, shadow bans and community resistance: did Meta's strategy to suppress rather than remove COVID misinformation and conspiracy theory on Facebook slow the spread? *Media International Australia* 197, 1 (March 2024), 222–241. doi:10.1177/1329878X241236984

[52] S. Mo Jones-Jang, Tara Mortensen, and Jingjing Liu. 2021. Does Media Literacy Help Identification of Fake News? Information Literacy Helps, but Other Literacies Don't. *American Behavioral Scientist* 65, 2 (Feb. 2021), 371–388. doi:10.1177/0002764219869406

[53] Joel Kaplan. 2025. More Speech and Fewer Mistakes. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

[54] Hansika Kapoor, Sarah Rezaei, Swanaya Gurjar, Anirudh Tagat, Denny George, Yash Budhwar, and Arathy Puthillam. 2023. Does incentivization promote sharing "true" content online? *Harvard Kennedy School Misinformation Review* 4, 4 (Aug. 2023). doi:10.37016/mr-2020-120

[55] Matthew Katsaros, Kathy Yang, and Lauren Fratamico. 2022. Reconsidering Tweets: Intervening during Tweet Creation Decreases Offensive Content. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM, Vol. 16)*. Association for the Advancement of Artificial Intelligence, 477–487. doi:10.1609/icwsm.v16i1.19308

[56] Tanveer Khan, Antonis Michalas, and Adnan Akhunzada. 2021. Fake news outbreak 2021: Can we stop the viral spread? *Journal of Network and Computer Applications* 190 (Sept. 2021), 103112. doi:10.1016/j.jnca.2021.103112

[57] Jisu Kim, Curtis McDonald, Paul Meosky, Matthew Katsaros, and Tom Tyler. 2022. Promoting Online Civility Through Platform Architecture. *Journal of Online Trust and Safety* 1, 4 (Sept. 2022). doi:10.54501/jots.v1i4.54

[58] Catherine King. 2025. *Effective and Practical Strategies for Combatting Misinformation*. PhD Dissertation. Carnegie Mellon University, Pittsburgh, PA, USA. doi:10.1184/R1/29316179

[59] Catherine King and Kathleen M Carley. 2025. Promoting Social Corrections: A Media Literacy Intervention for Misinformation on Social Media. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, Robert Thomson, Scott Renshaw, Samer Al-khateeb, Annetta Burger, Patrick Park, and Aryn A. Pyke (Eds.). Springer Nature, 223–232.

[60] Catherine King, Peter Carragher, and Kathleen M. Carley. 2025. Mapping the Scientific Literature on Misinformation Interventions: A Bibliometric Review. *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media* 2025 (June 2025), 10. doi:10.36190/2025.10

[61] Catherine King, Samantha C Phillips, and Kathleen M Carley. 2025. A path forward on online misinformation mitigation based on current user behavior. *Scientific Reports* 15, 1 (2025), 9475.

[62] Catherine King, Samantha C. Phillips, and Kathleen M. Carley. 2026. Public Support for Misinformation Interventions Depends On Perceived Fairness, Effectiveness, and Intrusiveness. *Journal of Online Trust and Safety* 3, 2 (March 2026). doi:10.54501/jots.v3i2.267

[63] Jan Kirchner and Christian Reuter. 2020. Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–27. doi:10.1145/3415211

[64] Anastasia Kozyreva, Philipp Lorenz-Spreen, Stefan M. Herzog, Ullrich K. H. Ecker, Stephan Lewandowsky, Ralph Hertwig, Ayesha Ali, Joe Bak-Coleman, Sarit Barzilai, Melisa Basol, Adam J. Berinsky, Cornelia Betsch, John Cook, Lisa K. Fazio, Michael Geers, Andrew M. Guess, Haifeng Huang, Horacio Larreguy, Rakoen Maertens, Folco Panizza, Gordon Pennycook, David G. Rand, Steve Rathje, Jason Reifler, Philipp Schmid, Mark D. Smith, Briony Swire-Thompson, Paula Szewach, Sander van der Linden, and Sam Wineburg. 2024. Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour* 8 (May 2024), 1044—-1052. doi:10.1038/s41562-024-01881-0

[65] Michael E. Kraft and Scott R. Furlong. 2017. *Public Policy: Politics, Analysis, and Alternatives* (6th ed.). CQ Press.

[66] Stephan Lewandowsky, John Cook, Ullrich Ecker, Dolores Albarracín, Michelle A. Amazeen, Panayiota Kendeou, Doug Lombardi, Eryn J. Newman, Gordon Pennycook, Ethan Porter, David G. Rand, David N. Rapp, Jason Reifler, Jon Roozenbeek, Philipp Schmid, Colleen M. Seifert, Gale M. Sinatra, Briony Swire-Thompson, Sander van der Linden, Emily K. Vraga, Thomas J. Wood, and Maria S. Zaragoza. 2020. Debunking Handbook 2020. doi:10.17910/b7.1182

[67] Stephan Lewandowsky and Sander van der Linden. 2021. Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology* 0, 0 (Feb. 2021), 1–38. doi:10.1080/10463283.2021.1876983

[68] Yi Liu, Pinar Yildirim, and Z. John Zhang. 2022. Implications of Revenue Models and Technology for Content Moderation Strategies. *Marketing Science* 41, 4 (July 2022), 831–847. doi:10.1287/mksc.2022.1361

[69] Zhuoran Lu, Patrick Li, Weilong Wang, and Ming Yin. 2022. The Effects of AI-based Credibility Indicators on the Detection and Spread of Misinformation under Social Influence. In *Proceedings of the ACM on Human-Computer Interaction (CSCW, Vol. 6)*. Association for Computing Machinery, 461:1–461:27. doi:10.1145/3555562

[70] Rakoen Maertens, Jon Roozenbeek, Melisa Basol, and Sander van der Linden. 2021. Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied* 27, 1 (2021), 1–16. doi:10.1037/xap0000315

[71] Francisco J. Martínez-López, Yangchun Li, and Susan M. Young. 2022. Social Media Monetization and Demonetization: Risks, Challenges, and Potential Solutions. In *Social Media Monetization: Platforms, Strategic Models and Critical Success Factors*, Francisco J. Martínez-López, Yangchun Li, and Susan M. Young (Eds.). Springer International Publishing, Cham, 185–214. doi:10.1007/978-3-031-14575-9_13

[72] Arunesh Mathur, Gunes Acar, Michael J Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark patterns at scale: Findings from a crawl of 11K shopping websites. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–32.

[73] Colleen McClain, Brian Kennedy, Jeffrey Gottfried, Monica Anderson, and Giancarlo Pasquini. 2025. *How the U.S. Public and AI Experts View Artificial Intelligence*. Technical Report. Pew Research Center. https://www.pewresearch.org/internet/2025/04/03/how-the-us-public-and-ai-experts-view-artificial-intelligence/

[74] Ariana Modirrousta-Galian and Philip A. Higham. 2023. Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. *Journal of Experimental Psychology: General* 152, 9 (2023), 2411–2437. doi:10.1037/xge0001395

[75] Garrett Morrow, Briony Swire-Thompson, Jessica Montgomery Polny, Matthew Kopec, and John P. Wihbey. 2022. The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology* 73, 10 (2022), 1365–1386. doi:10.1002/asi.24637

[76] Fionn Murtagh and Pierre Legendre. 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification* 31, 3 (2014), 274–295.

[77] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383.

[78] Jack Nassetta and Kimberly Gross. 2020. State media warning labels can counteract the effects of foreign disinformation. *Harvard Kennedy School Misinformation Review* 1, Special Issue on US Elections and Disinformation (Oct. 2020). doi:10.37016/mr-2020-45

[79] Lynnette HX Ng and Araz Taeihagh. 2021. How does fake news spread? Understanding pathways of disinformation spread through APIs. *Policy & Internet* 13, 4 (2021), 560–585.

[80] Lynnette Hui Xian Ng and Kathleen M Carley. 2022. Online coordination: methods and comparative case studies of coordinated groups across four events in the united states. In *Proceedings of the 14th ACM Web Science Conference 2022*. 12–21.

[81] Lynnette Hui Xian Ng, Ethan Pan, Michael Yoder, and Kathleen Carley. 2026. FATe of Bots: Ethical Considerations of Social Bot Detection. *ACM Journal on Responsible Computing* 3, 2 (2026), 1–35.

[82] Konrad Niklewicz. 2017. Weeding Out Fake News: An Approach to Social Media Regulation. *European View* 16, 2 (Dec. 2017), 335–335. doi:10.1007/s12290-017-0468-0

[83] Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. 2006. *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons.

[84] Orestis Papakyriakopoulos and Ellen Goodman. 2022. The Impact of Twitter Labels on Misinformation Spread and User Engagement: Lessons from Trump's Election Tweets. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*. Association for Computing Machinery, New York, NY, USA, 2541–2551. doi:10.1145/3485447.3512126

[85] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592 (March 2021), 590—-595. doi:10.1038/s41586-021-03344-2

[86] Ethan Porter and Thomas J. Wood. 2021. The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences* 118, 37 (Sept. 2021), e2104235118. doi:10.1073/pnas.2104235118

[87] Jon Porter. 2020. WhatsApp says its forwarding limits have cut the spread of viral messages by 70 percent. https://www.theverge.com/2020/4/27/21238082/whatsapp-forward-message-limits-viral-misinformation-decline

[88] Carlie Porterfield. 2020. Twitter Begins Asking Users To Actually Read Articles Before Sharing Them. https://www.forbes.com/sites/carlieporterfield/2020/06/10/twitter-begins-asking-users-to-actually-read-articles-before-sharing-them/

[89] Elaine S. Povich. 2021. Internet Ads Are a Popular Tax Target for Both Parties. https://pew.org/3psT1Rq

[90] R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-

project.org/
[91] David Rand and Cameron Martel. 2025. Americans actually do want expert content moderation. https://thehill.com/opinion/technology/5109667-sorry-zuckerberg-americans-actually-do-want-expert-content-moderation/
[92] Adrian Rauchfleisch and Jonas Kaiser. 2024. The impact of deplatforming the far right: an analysis of YouTube and BitChute. *Information, Communication & Society* 27, 7 (May 2024), 1478–1496. doi:10.1080/1369118X.2024.2346524
[93] Christian Reuter, Amanda Lee Hughes, and Cody Buntain. 2025. Combating information warfare: state and trends in user-centred countermeasures against fake news and misinformation. *Behaviour & Information Technology* 44, 13 (2025), 3348–3361.
[94] J. P. Reynolds, K. Stautz, M. Pilling, S. van der Linden, and T. M. Marteau. 2020. Communicating the effectiveness and ineffectiveness of government policies and their impact on public support: a systematic review with meta-analysis. *Royal Society Open Science* 7, 1 (Jan. 2020), 190522. doi:10.1098/rsos.190522
[95] Alex Rochefort. 2020. Regulating Social Media Platforms: A Comparative Policy Analysis. *Communication Law and Policy* 25, 2 (April 2020), 225–260. doi:10.1080/10811680.2020.1735194
[96] Jon Roozenbeek, Sander van der Linden, and Thomas Nygren. 2020. Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School Misinformation Review* 1, 2 (Feb. 2020). doi:10.37016//mr-2020-008
[97] Brennan Schaffner, Arjun Nitin Bhagoji, Siyuan Cheng, Jacqueline Mei, Jay L. Shen, Grace Wang, Marshini Chetty, Nick Feamster, Genevieve Lakier, and Chenhao Tan. 2024. "Community Guidelines Make this the Best Party on the Internet": An In-Depth Study of Online Platforms' Content Moderation Policies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, Article 486, 16 pages. doi:10.1145/3613904.3642333
[98] Sarita Schoenebeck, Amna Batool, Giang Do, Sylvia Darling, Gabriel Grill, Daricia Wilkinson, Mehtab Khan, Kentaro Toyama, and Louise Ashwell. 2023. Online harassment in majority contexts: Examining harms and remedies across countries. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–16.
[99] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.
[100] Filipo Sharevski, Raniem Alsaadi, Peter Jachim, and Emma Pieroni. 2022. Misinformation warnings: Twitter's soft moderation effects on COVID-19 vaccine belief echoes. *Computers & Security* 114 (March 2022), 102577. doi:10.1016/j.cose.2021.102577
[101] Ciarra N. Smith and Holli H. Seitz. 2019. Correcting Misinformation About Neuroscience via Social Media. *Science Communication* 41, 6 (Dec. 2019), 790–819. doi:10.1177/1075547019890073
[102] Edson C. Tandoc, Darren Lim, and Rich Ling. 2020. Diffusion of disinformation: How social media users respond to fake news and why. *Journalism* 21, 3 (March 2020), 381–398. doi:10.1177/1464884919868325
[103] Daniel Robert Thomas and Laila A. Wahedi. 2023. Disrupting hate: The effect of deplatforming hate organizations on their online audience. *Proceedings of the National Academy of Sciences* 120, 24 (June 2023), e2214080120. doi:10.1073/pnas.2214080120
[104] Benjamin Toff and Nick Mathews. 2021. Is Social Media Killing Local News? An Examination of Engagement and Ownership Patterns in U.S. Community News on Facebook. *Digital Journalism* 0, 0 (Oct. 2021), 1–20. doi:10.1080/21670811.2021.1977668
[105] Melissa Tully, Emily K. Vraga, and Leticia Bode. 2020. Designing and Testing News Literacy Messages for Social Media. *Mass Communication and Society* 23, 1 (Jan. 2020), 22–46. doi:10.1080/15205436.2019.1604970
[106] João Varela da Costa and Miguel Mira da Silva. 2025. Countermeasures against fake news: a Delphi study. *Transforming Government: People, Process and Policy* 19, 2 (2025), 392–413.
[107] Tiago Ventura, Rajeshwari Majumdar, Jonathan Nagler, and Joshua A. Tucker. 2023. Misinformation Exposure Beyond Traditional Feeds: Evidence from a WhatsApp Deactivation Experiment in Brazil. doi:10.2139/ssrn.4457400
[108] Emily A. Vogels. 2022. Support for more regulation of tech companies has declined in U.S., especially among Republicans. https://www.pewresearch.org/short-reads/2022/05/13/support-for-more-regulation-of-tech-companies-has-declined-in-u-s-especially-among-republicans/
[109] Nathan Walter and Sheila T. Murphy. 2018. How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs* 85, 3 (July 2018), 423–441. doi:10.1080/03637751.2018.1467564
[110] Claire Wardle and Hossein Derakhshan. 2017. *Information Disorder: Toward an interdisciplinary framework for research and policy making*. Technical Report. Council of Europe. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c
[111] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (Nov. 2018), 4366–4383. doi:10.1177/1461444818773059
[112] Lucas Wright. 2022. Automated platform governance through visibility and scale: On the transformational power of automoderator. *Social Media+ Society* 8, 1 (2022), 20563051221077020.
[113] Kamya Yadav. 2020. *Countering Influence Operations: A Review of Policy Proposals Since 2016*. Technical Report. Carnegie Endowment for International Peace. https://carnegieendowment.org/2020/11/30/countering-influence-operations-review-of-policy-proposals-since-2016-pub-83333
[114] Kamya Yadav. 2021. *Platform Interventions: How Social Media Counters Influence Operations*. Technical Report. Carnegie Endowment for International Peace. https://carnegieendowment.org/2021/01/25/platform-interventions-how-social-media-counters-influence-operations-pub-83698
[115] Max Zhan. 2025. Here's why Meta ended fact-checking, according to experts. https://abcnews.go.com/US/why-did-meta-remove-fact-checkers-experts-explain/story?id=117417445
[116] Hong Zhou, Yaobin Lu, Ling Zhao, Bin Wang, and Ting Li. 2024. Effective reporting system to encourage users' reporting behavior in social media platforms: an empirical study based on structural empowerment theory. *Behaviour & Information Technology* 43, 14 (Oct. 2024), 3490–3509. doi:10.1080/0144929X.2023.2281491

A Intervention Categorization

This appendix describes the literature of interventions, in which we used to develop our categorization taxonomy. We selected four prominent review articles about misinformation interventions, all published in different fields and journals (Aghajari et al. [3] in an HCI conference, Blair et al. [17] in a psychology venue, Courchesne et al. [27] in an interdisciplinary social science journal, and Kozyreva et al. [64] in Nature Human Behavior). Table 6 summarizes the intervention categories analyzed by each paper in alphabetical order.

The Courchesne et al. [27] article classifies types of interventions based on those publicly announced as having been implemented by various social media platforms. These 10 categories are advertisement policy, content labeling, content / account moderation, content reporting, content distribution / sharing, disinformation disclosure, disinformation literacy, redirection, security / verification, and other. The article notes that fact-checking was included in the disinformation disclosure category, so this is categorized as fact-checking in Table 6.

The Aghajari et al. [3] article categorizes interventions primarily based on the driver of the misinformation that each intervention targets: Content, Source, Individual Users, or Community. For content-based interventions, they discuss fact-checking, warning labels, and platform alterations (reducing the size of misinformation). For source-based interventions, they discuss source credibility labels and the crowdsourcing of those labels. For user-based strategies, they consider media literacy, accuracy prompts, and account removal. Lastly, for community-based interventions, they address social norms.

The Blair et al. [17] and Kozyreva et al. [64] articles categorize interventions not just in terms of which parts of the platforms are affected, but also what part of the social media misinformation pipeline is targeted (such as the creation, spread, belief phases as well as prevention - see Figure 1). More specifically, the Blair et al. [17] article categorizes 11 types of interventions into four main groups: Informational (targets belief), Educational (targets prevention), Sociopsychological (targets spread), and Institutional (external). Information interventions include inoculation / prebunking, debunking, credibility labels / tags, and contextual labels / tags. These interventions aim to correct misinformation. Under educational interventions, they define media literacy. This category of

Table 6: Categorizations of misinformation interventions in the literature.

Type	Courchesne et al. (2021) [27]	Aghajari et al. (2023) [3]	Blair et al. (2024) [17]	Kozyreva et al. (2024) [64]
Account Moderation	×		×	×
Account Removal		×		
Accuracy Prompts		×		×
Advertising Policy	×			
Content Distribution	×			
Content Labeling	×			
Content Moderation	×			
Context Labels			×	
Crowdsourcing		×		
Debunking			×	×
Fact-Checking	×	×		
Friction			×	×
Inoculation			×	×
Journalist Training			×	
Lateral Reading			×	
Media Literacy	×	×		
Platform Alterations		×		
Politician Messaging			×	
Redirection	×			
Reporting	×			
Security/Verification	×			
Social Norms		×	×	×
Source Credibility Labels		×	×	
Warning Labels		×	×	
Other	×			

interventions seeks to prevent belief in misinformation. Sociopsychological interventions include accuracy prompts, friction, and social norms. These interventions aim to discourage users from spreading misinformation. Lastly, they define platform alterations, politician messaging, and journalist training under institutional interventions.

The Kozyreva et al. [64]. article classifies nine types of interventions into three main categories [64]: Nudges (targets spread), Boosts and Educational Interventions (targets prevention), and Refutation Strategies (targets belief). The nudges category includes accuracy prompts, friction, and social norms. These interventions address sharing behavior by discouraging users from further distributing misinformation. The boosts/educational interventions category includes inoculation, lateral reading and verification strategies, and media literacy tips. These interventions aim to increase user competence in evaluating content online and discerning misinformation or untrustworthy content from reputable information. Lastly, the refutation strategies category includes debunking and rebuttals, warning and fact-checking labels, and source credibility labels.

Many of the review articles used a similar categorization of countermeasures; however, there is no common typology [3, 17, 27, 64]. Many of these defined categories overlap (e.g., lateral reading skills vs. media literacy) or are sub-categories of other categories. For example, redirection is a form of content distribution. Similarly, context labels and source credibility labels are types of content labeling. Additionally, specific categories are absent, particularly user-led or institutional interventions such as government regulation. User-based measures are an often overlooked aspect in the fight against misinformation. Individual-level debunking, especially from trusted messengers, is effective in various contexts [11, 19, 66]. In addition to user reporting and social norms, users can block others or engage in social corrections. When misinformation is successfully posted on social media, other users serve as the first line of defense since they can flag or debunk it. While social corrections can be considered a type of debunking or fact-checking, the social context may be important to distinguish.

By synthesizing the categorizations used by four review articles, we developed eight general categories of countermeasures, as shown in Table 1. These categories are primarily classified by the part of the information lifecycle targeted and the part of the platform affected. The following subsections describe these eight general categories in more detail, as well as the subcategories of intervention types often described in the literature for each general category.

A.1 Account Moderation

Account moderation refers to platform interventions directed at user accounts. These interventions include actions such as account suspensions or account removals, limiting account visibility through shadowbanning, or demonetizing accounts, which restricts their ability to monetize content[27, 111]. Specific interventions include:

- **Account Suspension** - Account suspension involves temporarily suspending or permanently banning accounts that repeatedly violate platform policies, like repeatedly disseminating misinformation. Deplatforming describes a coordinated effort to remove a high-risk account across multiple platforms, rather than just on a single platform [92, 103].
- **Shadow Banning** - Shadow banning suppresses an account's reach by limiting its visibility to others, while leaving the account itself still active. Users are often not notified that their content has been throttled [51, 111].
- **Demonetization** - Demonetization refers to platforms removing or limiting an account's monetization privileges [71]. Platforms typically apply such restrictions after repeated policy violations.

A.2 Content Moderation

Content moderation refers to a broad range of interventions regarding the way content is presented, distributed, or displayed on social media. These interventions range from evaluating the accuracy of information to modifying its visibility or circulation, and may be implemented through human moderators, users or automated systems [50, 97]. Specific interventions include:

- **Misinformation Detection** - Misinformation detection refers to the use of computational methods to identify content that is potentially false, misleading, or otherwise unreliable. These systems typically serve as an upstream component of content moderation, identifying content that may

be subsequently reviewed, labeled, downranked or removed [56, 69].
- **Fact-checking** - Fact-checking assesses the accuracy of a claim by comparing it against available evidence. It can be carried out by experts, journalists, and platforms, and includes both text and video-based formats [20, 109].
- **Debunking** - Debunking goes beyond determining whether a claim is accurate by providing additional evidence, explanations or contextual information to correct misleading claims. In this sense, debunking can be understood as a form of "narrative intervention" that identifies falsehoods and provides a more coherent and accurate account of the issue [23, 66].
- **Algorithmic Moderation** - Algorithmic content moderation refers to the use of automated systems to identify, evaluate and manage content at scale. Such systems would automatically fact-check, label, remove, downrank or otherwise reduce the visibility of content that violate policies, like misinformation [15, 19, 36, 40]. Algorithmic moderation can also function as a virality circuit breaker, temporarily limiting the amplification of rapidly spreading, unverified content, until its accuracy can be assessed [1].
- **User Control** - User control shifts some moderation responsibilities from the platforms to the users, allowing users to take a more active role in deciding what appears in their feeds, rather than than leaving that decision entirely to the platform's algorithms [49].

A.3 Content Distribution

Content distribution refers to a broad category of interventions that shape how content circulates, reaches audiences and is encountered on social media platforms. Rather than directly removing or modifying content, these interventions influence the pathways through which information is spread on social media. Specific interventions include:

- **Account Limits** - Another type of content distribution is to impose account limits on users to restrict their forwarding capabilities, which caps the number of recipients the message can be forwarded to [87], resharing limits that constrain repeated dissemination, or disable sharing functionality after content has had several rounds of redistribution [1].
- **Friction** - Friction introduces additional steps or delays, such as an extra click or a confirmation window, into the process of sharing content so that users can pause and possibly reconsider before sharing [12, 55]. One example of such friction would be to prevent users from re-sharing content that they had not opened and prompting them to consider reading the article first instead [88].
- **Accuracy Prompts** - Accuracy prompts, sometimes referred to as "nudges," are brief questions or reminders about the importance of accurate information sent to users and intended to bring accuracy to the front of their minds when a user is deciding whether or not to post a piece of content [85]. One implementation involves asking users to evaluate the accuracy of one or more headlines to reflect the importance of sharing accurate news before they can continue using the platform [32].
- **Redirection** - Redirection intercepts a search query on a potentially harmful or dangerous topic and directs the user toward vetted and official information. In some cases, it may suppress the search results entirely [18, 101]. Sometimes, platforms may limit access to the requested content altogether. A common example is surfacing guidelines from a public health organization, such as the CDC or the WHO, above or next to general search results when users search for a topic like COVID-19 vaccines. Similarly, in the immediate lead-up to an election in the United States, many platforms highlight how to find one's polling place using official government sources. [27, 114].
- **Advertising Policy** - Advertising policies govern how advertisements can be created, distributed, targeted or displayed on social media platforms [15, 27, 46]. These policies govern the rules that platforms or regulators can impose on advertising content, and they may include restrictions on the type of content that can run and how it must be labeled or displayed [15, 27, 46]. Examples include banning political ads, prohibiting the use of AI or manipulated content in political or targeted ads, requiring ads to undergo a fact-checking process before posting, and labeling ads as "paid for" [4, 24, 27].
- **Platform Alterations** - Platform alterations refer to interface- or architecture-level changes on social media platforms that shape how content is surfaced, sized, or displayed to users, which in turn affects how users interact with it [63]. Such interventions can reduce the size or visibility of a post [36, 63, 100] or alter the platform interface to restructure how users can share content, such as directing posts toward more specific rather than general groups [57].

A.4 Generative AI-specific

This category is for interventions that explicitly use generative AI tools like chatbots or LLM-written content to counter misinformation. Specific interventions include:

- **Generative AI Content** - Using AI-generated content to create rebuttals against misinformation or to develop educational initiatives.
- **Generative AI Chatbots** - Deploying conversational AI systems to dialogue with users as a way to reduce their belief in conspiracy theories or false claims [26].

There are other generative AI-adjacent interventions, such as prohibiting the use of AI or manipulated content to produce deepfakes (Content Moderation), banning the use of AI in political or targeted advertising (Content Distribution), or requiring clear disclosure on any ads that incorporate AI-generated images, videos, or audio (Content Labeling) [45]. However, these types of interventions already fit into pre-existing categories.

A.5 Content Labeling

Content labeling refers to interventions that attach disclosures, warnings or supplementary information to content in order to communicate its accuracy, credibility, source, or broader context. These interventions provide users with information through labels that help them evaluate the content they encounter on social media.

These labels include fact-checks (see Content Moderation section) or source information or credibility [75, 114]. Specific interventions include:

- **Crowdsourcing** - Crowdsourcing involves drawing on evaluations from ordinary users to access information and provide labels or contextual information. It shifts the labeling task away from professionals and onto the user base, then aggregates user judgments into a label or note [5]. One prominent implementation of this approach is X's Community Notes program, in which users collaboratively contribute and evaluate contextual notes that attached to the tweets [6].
- **Context Labels** - Context labels attach additional contextual information to posts so that users can better interpret or evaluate the content. These labels provide relevant background, clarification or evidence to the post. X's Community Notes are context labels, and the program uses crowdsourcing to create and maintain them [6].
- **Source Labels** - Source credibility labels provide users with information about the reliability, identity, or institutional affiliation of the source behind a post. These interventions may, for example, communicate assessments of a news source's reliability [34] or identify accounts associated with government officials or state-run media sources [78].
- **Warning Labels** - Warning labels flag posts based on the post's content or sources that may be false, misleading, or otherwise warrant additional scrutiny [84]. Some implementations require an active click to bypass an interstitial screen before the user can view or share the underlying content [100].
- **Notification** - Notifications directly inform users when content they have been posted, shared or otherwise interacted with has subsequently been flagged to contain harmful or false information, or originating from state-run media sources [27].

A.6 User-based Measures

User-based measures refer to interventions that empower individuals and communities to respond directly to misinformation and shape their own information environments [102]. Specific interventions include:

- **Social Norms** - Social norm interventions leverage community expectations and social signals to encourage more responsible information-sharing behaviors. Platforms can support these norms by designing features or metrics that reward accuracy rather than engagement [10, 31, 37].
- **Social Corrections** - Social corrections are peer-to-peer corrections in which one user directly fact-checks another's post, sometimes occurring in the open (via a reply) and other times occurring privately (via a direct message) [11, 19].
- **Retractions** - Retractions occur when users or organizations delete or correct a false post that was previously published. It may also involve efforts to notify users who previously viewed the post. [8, 30].
- **User Reporting** - Reporting enables users to flag potentially harmful or misleading posts or accounts for platform review [82, 116].
- **Blocking** - Blocking gives users a way to reduce their exposure to unwanted content, such as content from specific users or on specific topics [102].
- **Other/User Empowerment** - Individualized measures provide users with tools or interventions that encourage self-moderation and behavioral change, such as nudges which will encourage users to reduce their social media use [107].

A.7 Media Literacy and Education

Media literacy and education encompasses interventions designed to strengthen individuals' ability to critically evaluate and engage with information. These approaches seek to improve the public's ability to navigate digital sources and to think critically about the media they encounter [43, 48]. Specific interventions include:

- **Media Literacy** - Media literacy and related educational interventions include a range of initiatives intended to develop or assess media literacy skills. These efforts may provide practical guidance for identifying misinformation [43], evaluate individuals' information, digital, or news literacy [52], or develop and test communication strategies for improving the effectiveness of media literacy education [105].
- **Inoculation** - Inoculation, sometimes called "pre-bunking" in the literature, aims to build resistance to false information before exposure occurs. These interventions typically warn individuals about common misleading arguments or manipulation techniques, and explain how they operate, making users better prepared to recognize and resist subsequent misinformation [67].
- **Games** - Games like fake news games are a specific type of educational tool that use interactive, gamified experiences to teach users how to recognize misinformation and strengthen their critical evaluation skills. These games often expose players to common misinformation strategies in a controlled environment, helping them identify similar techniques in future encounters [70, 74, 96].

A.8 Institutional Measures

Institutional measures refer to interventions undertaken by governments, civil society organizations, news media, and other institutions [21]. Specific interventions include:

- **Government Regulation** - Government regulation encompasses laws and regulations designed to govern platforms and increase their accountability. These measures can be implemented at the local, state, or federal level [82, 95, 113]. Examples include modifying Section 230, developing comprehensive privacy laws similar to Europe's GDPR, taking anti-trust action to break up monopolistic technology companies [95], or restriction of micro-targeted advertising [89].
- **Media Support** - Media support involves efforts to strengthen the broader news ecosystem, including funding local media outlets, giving credible local reporting greater visibility on social media platforms [104], and providing training and resources to support high-quality reporting [14, 21].

- **Data Sharing** - Data sharing involves social media platforms sharing raw data and internal findings with researchers, thereby supporting greater transparency and independent investigation [9, 15, 16].
- **Other Institutional Measures** - Other institutional measures include developing resources and tools to support civil society and strengthening coordination and collaboration among governments, researchers, media organizations, platforms and other relevant institutions [21, 110].

B Operationalized Interventions

This section describes the 40 operationalized interventions participants were asked about in the expert survey. Each intervention is a specific implementation of one or more interventions described in the literature and in the previous appendix. On the survey, participants were given the definitions of the specific intervention implementations, as shown in the first column of the following table.

Specified Intervention Implementations	Intervention Type(s)
Account Moderation	
Account Suspensions - Permanently or temporarily ban certain users.	Account Suspension
Deplatforming - Coordinated efforts to ban prominent misinformation spreaders.	Account Suspension
Shadow Banning - Limit the visibility and the spread of posts from certain policy-violating accounts.	Shadow Banning
Demonetization - Remove or restrict monetization features for a user account.	Demonetization
Content Moderation	
Content Removal - Remove posts verified to contain misinformation.	Misinformation Detection
Ban Deepfakes - Prohibit the use of AI to misrepresent the speech or actions of public figures.	Misinformation Detection
Downranking - De-emphasize posts in news feeds that contain misinformation.	Algorithmic Moderation
Upranking - Uprank high-quality news sources and downrank low-quality ones.	Algorithmic Moderation
Virality Circuit Breakers - Briefly halt algorithmic amplification of unverified content.	Algorithmic Moderation
User Control - Give users more control over the algorithms powering their news feeds.	User Control
Content Distribution	
Limit Forwarding - Cap the number of users one can forward a given message to.	Account Limits
Limit Resharing - Remove share buttons on posts after several levels of sharing.	Account Limits
Friction - Temporarily delay users from posting content they did not open via a pop-up.	Friction
Accuracy Prompts - Nudge or remind people to consider accuracy before posting or sharing content.	Accuracy Prompts
Redirection - Redirect users to other content, such as official content or no content.	Redirection
Ban Political Ads - Ban political ads on social media platforms.	Advertising Policy
Fact-Check Ads - Fact-check ads on social media platforms.	Advertising Policy
AI in Ads - Prohibit the use of AI or manipulated content in political or targeted ads.	Advertising Policy
Platform Alterations - Reduce the size or visibility of a post containing misinformation.	Platform Alterations
Generative AI	
Gen AI Content - Generate rebuttals to misinformation or create educational initiatives.	Generative AI
Gen AI Chatbots - Use chatbots to reduce belief in conspiracy theories or misinformation.	Generative AI
Content Labeling	
Crowdsourcing - Labels generated by the public rather than professionals (e.g., Community Notes).	Crowdsourcing, Context Labels
Fact Check Labels - Link information from third-party fact-checkers on misinformation posts.	Fact-Checking, Context Labels
Source Labels - Labels indicating the reliability of news sources.	Source Labels
Government Labels - Label the accounts of government officials or state-run media.	Source Labels
AI Disclosure - Require disclosures on advertisements that use AI-generated content.	Source Labels, Context Labels
Click-through Warning Labels - Place misinformation posts behind click-through labels containing facts and context.	Warning Labels
User-Based Measures	
Social Norms - Use social or community norms to encourage social corrections and self-corrections.	Social Corrections, Retractions
Alter Platform Metrics - Reward accuracy rather than engagement to discourage misinformation sharing.	Social Norms
Reporting - Improved reporting functionality and transparency.	User Reporting
Media Literacy / Education	
Digital Media Literacy - Invest in and promote content on how to critically evaluate online information.	Media Literacy
Inoculation - Inoculate people against misinformation with games or videos.	Inoculation, Games
Institutional Measures	
Government Regulation - Hold companies accountable for content on their platforms.	Government Regulation
Privacy Legislation - Develop comprehensive privacy legislation, similar to Europe's GDPR.	Government Regulation
Anti-Trust Action - Break up monopolistic big tech companies.	Government Regulation
Taxes / Fines - Tax or fine social media companies for their use of personal user data.	Government Regulation
Targeted Advertising - Limit or ban micro-targeted advertising.	Government Regulation
Media Support - Promote and invest in local media.	Media Support
Journalism Support - Support and train the next generation of journalists.	Media Support
Data Sharing - Have companies regularly release data and internal research reports to researchers.	Data Sharing

Table 7: Operationalized misinformation interventions across the eight general categories.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009