International Journal of Social Trends
Volume 4, Issue 5 (Sept-Oct; 2026)

Review Article · Communication and Political Science

Machine Learning Against Misinformation: A Review of Detection, Diffusion and Intervention Research in Communication and Political Science

Freja Meier [1], Mihai Luca [2]

[1] Technical University of Denmark, DENMARK
[2] ETH Zurich, SWITZERLAND

ABSTRACT

Misinformation has become one of the most studied problems at the meeting point of computer science and the social sciences. This review synthesises research on how machine learning (ML) and related computational methods are used to detect false and misleading content, to describe how it spreads, and to design and evaluate interventions against it. Organised around a four-stage model of production, diffusion, detection and intervention, the review examines content-based, context-based and network-based detectors; automated fact-checking; bot detection; large-scale studies of exposure and sharing; and field and survey experiments on labels, accuracy prompts and prebunking. Three findings stand out. First, benchmark performance is a poor guide to real-world usefulness: on the widely used LIAR dataset, the best model in the original study reached only 27.4 percent accuracy on a six-way task, and gains on later benchmarks often rest on topical or source shortcuts. Second, the empirical literature on diffusion shows that exposure to and sharing of low-credibility sources is far more concentrated than public debate assumes, with about 1 percent of Twitter users accounting for 80 percent of exposures in the 2016 U.S. election. Third, the most robust behavioural evidence favours light-touch interventions that shift attention to accuracy or build resistance in advance, while the effects of labelling and removal depend on design choices that are rarely reported. The review closes with an agenda that calls for evaluation against downstream social outcomes, cross-platform and non-English data, and explicit attention to generative AI as a source of persuasive synthetic content.

Keywords: Misinformation; Fake News Detection; Machine Learning; Computational Social Science; Political Communication; Content Moderation; Generative AI

INTRODUCTION

Few topics have drawn computer scientists and social scientists into closer contact than misinformation. The term covers a wide range of phenomena, from fabricated news stories and manipulated images to misleading statistics and conspiracy narratives, and the concern it attracts reflects the central place of shared factual beliefs in democratic life. A landmark analysis of rumour cascades on Twitter found that false stories spread farther, faster, deeper and more broadly than true ones, and that the difference was driven by human users rather than automated accounts [1]. In the same issue of Science, a group of scholars from political science, psychology, communication and computer science called for a coordinated "science of fake news" that would combine large-scale data, causal designs and platform cooperation [2]. The years since have

produced a very large literature, much of it computational, spread across conferences and journals in disciplines that rarely read one another.

This review aims to bring that literature together for readers in communication and political science. Its premise is that computational methods have become indispensable for studying misinformation at the scale at which it now circulates, but that the questions those methods should answer are social-scientific: who produces false content and why, how it moves through networks of people and institutions, whom it reaches, what it changes, and which interventions protect the information environment without curtailing legitimate speech. Machine-learning research has often been driven by benchmark datasets and leaderboard metrics; social-science research has often been limited by small samples and survey self-reports. The most informative studies combine the two.

The review makes three contributions. First, it proposes an organising model of four linked stages (production, diffusion, detection and intervention) and maps computational research onto each stage (Figure 1). Second, it critically assesses the evidence on detection, diffusion and intervention, paying particular attention to how benchmark results translate, or fail to translate, into real-world performance. Third, it sets out a research agenda that addresses the new problems raised by generative artificial intelligence, which lowers the cost of producing fluent, persuasive and personalised false content [3, 4].

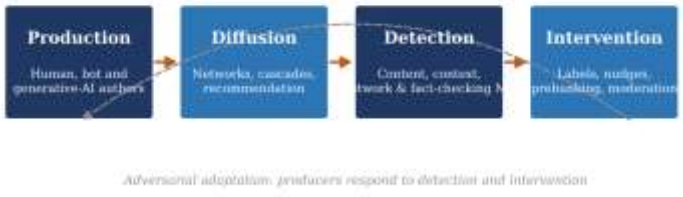

Figure 1: A four-stage model of the misinformation problem used to organise this review. Producers adapt to detection and intervention, so the stages form a cycle rather than a pipeline. Source: authors' design.

1.1 Scope and method of the review

The review is narrative and integrative rather than a formal systematic review. Sources were identified through the reference lists of major surveys [5-7], through citation tracing from influential empirical studies [1, 8-10], and through targeted searches of computer-science proceedings (ACL, EMNLP, KDD, WWW, ICWSM, CIKM) and social-science journals in communication, political science and psychology. Priority was given to peer-reviewed work, to studies with transparent data and designs, and to research that connects computational measurement to social outcomes. Because the field moves quickly, the review emphasises approaches and findings that have been replicated or widely adopted rather than every new model architecture.

1.2 Defining the object of study

Definitional precision matters because computational systems need operational labels. The most influential conceptual framework distinguishes misinformation (false content shared without intent to harm),

disinformation (false content shared knowingly to cause harm) and malinformation (genuine content shared to cause harm, such as leaked private information) [11]. Economists studying the 2016 U.S. election adopted a narrower definition of "fake news" as news articles that are intentionally and verifiably false and could mislead readers [12], while much computational work operationalises falsehood at the level of the source, labelling all content from domains that fact-checkers or journalists have flagged as low-credibility [8, 9]. Table 1 summarises these choices and their implications for measurement.

Table 1: Operational definitions of misinformation and their consequences for computational research.

Unit of analysis	Typical label source	Strengths	Limitations
Claim or statement	Professional fact-checkers (e.g., PolitiFact ratings)	Fine-grained; aligns with journalistic practice	Small, selective samples; label reflects checker priorities
News article	Article-level fact-checks; satire and hoax lists	Matches how users encounter content	Mixed-veracity articles are hard to label
Source or domain	Lists of low-credibility domains compiled by journalists and researchers	Scales to millions of URLs; replicable	Treats every item from a source as equally false
Rumour or cascade	Fact-check resolution of a circulating story	Captures diffusion dynamics	Depends on stories being checked at all
Account or actor	Bot or coordinated-behaviour classifiers	Targets producers rather than content	Classifier error; behaviour changes over time

Note. Compiled from the conceptual and empirical literature reviewed in Sections 1 to 4.

Each choice trades validity against scale. Claim-level labels are the most precise but cover a small and non-random fraction of what circulates; domain-level labels scale to whole platforms but misclassify true articles from unreliable outlets and false articles from reliable ones. Studies that report results under several definitions, as some diffusion studies do [8, 9], are therefore more credible than those that rely on a single list.

THEORETICAL FOUNDATIONS

2.1 Why people believe and share false content

Computational research is most useful when it is guided by theories of why misinformation succeeds. Psychological research on the continued influence effect shows that corrected falsehoods continue to shape reasoning, because people build mental models that are hard to revise once a causal explanation has been

accepted [13]. Political science long emphasised motivated reasoning, the tendency to accept information that supports one's political identity, and early experiments reported that corrections could even "backfire" and strengthen false beliefs among partisans [14]. Subsequent large-scale replications found such backfire effects to be rare, with most people moving toward the facts when corrected [15], and a meta-analysis of correction studies confirmed that corrections reliably reduce misperceptions on average, though their effectiveness varies with the source, the format and the prior strength of belief [16].

A complementary account emphasises inattention rather than motivation. In a series of studies, people who scored higher on a measure of reflective thinking were better at distinguishing true from false headlines regardless of whether the headlines favoured their own party, which suggests that susceptibility is better explained by "lazy" reasoning than by partisan bias [17]. This account has a direct design implication: if people share false content because they are not thinking about accuracy at the moment of sharing, then interventions that redirect attention to accuracy should improve the quality of what they share [10]. The two accounts are not mutually exclusive, and computational studies increasingly test them against each other using behavioural data at scale.

2.2 Why false content travels

Network and communication theories explain why some content spreads widely. The most prominent empirical finding is that false rumours on Twitter reached more people, and did so faster, than true ones; the authors attributed this partly to novelty, since false stories were judged more novel and elicited more surprise and disgust in replies [1]. Research on Facebook found that misinformation and conspiracy narratives circulate within homogeneous, polarised communities, or echo chambers, where they are reinforced by like-minded users [18]. Yet other work cautions against overstating isolation: a large study of Facebook users found that individual choices limited exposure to cross-cutting news more than the ranking algorithm did [19], and a study using browsing and television data concluded that fake news makes up a very small share of Americans' overall news diet [20]. These findings shape what detection systems should prioritise: if most exposure comes from a few highly active accounts and sources, then actor-level detection may matter more than classifying every item.

MACHINE-LEARNING APPROACHES TO DETECTION

Detection research seeks to label content, sources, cascades or accounts automatically. It has grown from feature-engineering studies of credibility on Twitter [21] into a large field built around benchmark datasets and deep neural networks [45, 62]. Table 2 lists the families of approaches discussed in this section and Table 3 lists widely used datasets.

Table 2: Families of computational detection approaches.

Approach	Signals used	Representative work	Main weakness
Content-based (linguistic)	Word use, style, readability, sentiment, deception cues	Pérez-Rosas *et al.*; Rashkin *et al.*	Learns topic and source shortcuts; brittle across domains

Approach	Signals used	Representative work	Main weakness
Knowledge-based (fact-checking)	Retrieval of evidence; claim-evidence entailment	FEVER; ClaimBuster	Needs up-to-date knowledge bases; struggles with new claims
Context-based (social)	User profiles, engagement, temporal patterns of responses	Castillo *et al.*; CSI	Signals arrive only after spread has begun
Propagation / network-based	Cascade structure, diffusion speed, network position	Ma *et al.*; Vosoughi *et al.*	Requires platform data rarely available to outsiders
Actor-based (bots, coordination)	Account metadata, posting rhythms, synchronised behaviour	Varol *et al.*; Shao *et al.*	Arms race with evolving bot designs

3.1 Content-based classifiers

Early content-based work treated fake news detection as text classification. Researchers built corpora of fake and legitimate news and trained classifiers on linguistic features such as n-grams, punctuation, psycholinguistic categories and readability, reporting accuracies well above chance within each corpus [22]. An analysis of articles from satire, hoax and propaganda sites compared with trusted news found systematic stylistic differences, such as greater use of subjective and dramatising language in unreliable sources [23]. These studies established that style carries signal, but they also revealed the central weakness of the approach: classifiers trained on one domain or time period transfer poorly to another, because much of what they learn is the topic, the outlet or the period rather than falsehood itself [6].

The release of the LIAR benchmark, consisting of 12,836 short statements labelled by PolitiFact into six categories from "pants on fire" to "true", marked a shift toward standardised evaluation [24]. Figure 2 shows the results reported in the original paper. A majority-class baseline achieved 20.8 percent accuracy on the test set; the best text-only model, a convolutional neural network, reached 27.0 percent; and a hybrid network that added speaker metadata such as party, job title and prior truthfulness record reached 27.4 percent. The task is hard because short political statements contain little textual evidence of their own truth, which is why metadata about the speaker helped. Later work using pretrained transformer models such as BERT [25] improved scores on many benchmarks, but the basic lesson of LIAR has held: without external evidence, text alone carries limited information about veracity.

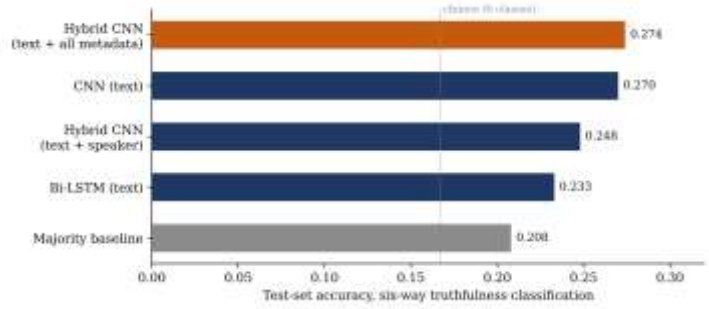

Figure 2: Test-set accuracy of models on the six-class LIAR benchmark, as reported by Wang (2017). Even the best hybrid model performs only modestly above the majority-class baseline. Data source: [24].

3.2 Knowledge-based detection and automated fact-checking

Because truth is a relation between a claim and the world, a second line of research builds systems that retrieve evidence and judge whether it supports or refutes a claim. The FEVER dataset framed this as a pipeline of document retrieval, sentence selection and natural-language inference, using 185,445 claims derived from Wikipedia and labelled as supported, refuted or lacking sufficient information [26]. Such systems are closer to how journalists verify claims, and they can produce explanations in the form of cited evidence. Related work prioritises which claims to check: ClaimBuster scores sentences in political speeches and debates by their "check-worthiness" so that fact-checkers can focus scarce attention where it matters most [27]. Knowledge-based approaches face their own limits. Evidence bases lag behind events, many viral claims concern recent or local developments not yet documented, and contested claims often turn on interpretation rather than verifiable fact.

3.3 Social-context and propagation-based detection

A third family uses the behaviour of people around a piece of content. Early work on Twitter credibility combined message features with the characteristics of users who posted and propagated a story and the structure of the propagation tree [21]. Recurrent neural networks have been used to model the time series of posts about an event, learning that false rumours generate distinctive patterns of questioning and denial over time [28]. The CSI model combined the text of articles, the responses they received and the characteristics of the users who engaged with them, outperforming text-only baselines [29]. Repositories such as FakeNewsNet were built specifically to support this kind of research by pairing news content with social context and spatiotemporal information [30]. A practical difficulty is timing: the social signals these models rely on accumulate only as content spreads, so a model that becomes accurate after a story has gone viral is of limited value for prevention.

Table 3: Widely used benchmark resources for misinformation research.

Resource	Content and labels	Typical use	Caveats
LIAR	12,836 short political statements; six PolitiFact truthfulness labels	Fine-grained claim classification	Short texts; speaker metadata can dominate
FEVER	185,445 claims generated from Wikipedia; supported / refuted / not enough information	Evidence retrieval and verification	Synthetic claims differ from real-world rumours
FakeNewsNet	News content plus social context from PolitiFact and GossipCop	Context and propagation models	Depends on platform data that can disappear
Rumour corpora (e.g., PHEME)	Twitter threads around breaking news, annotated for veracity and stance	Rumour detection and stance classification	Event-specific; English-dominated
Low-credibility domain lists	Domains flagged by journalists and researchers	Exposure and sharing studies at scale	Source-level, not item-level, labels

Note. Dataset descriptions follow the original publications cited in the text.

3.4 Detecting automated and coordinated accounts

Concern about automated accounts, or social bots, has driven a substantial body of detection research [31]. Supervised classifiers trained on account metadata, network position, content and temporal activity estimated that a meaningful minority of active Twitter accounts showed bot-like behaviour [32]. Analysis of millions of tweets linking to low-credibility sources found that bots were particularly active in the early moments of spread, amplifying content before it went viral, and that they targeted influential human users through mentions and replies [33]. At the same time, the most-cited diffusion study found that removing bot activity did not change the advantage of false over true news, because humans were responsible for most of the difference [1]. The policy-relevant target has therefore shifted from individual bots toward coordinated inauthentic behaviour, in which networks of human and automated accounts act in concert.

Box 1. Why high benchmark scores can mislead

Detection papers often report accuracy above 90 percent on balanced test sets. Three features of the evaluation explain much of the gap between such numbers and real-world value. First, test sets drawn from the same sources and periods as training data reward models for recognising outlets and topics rather

than falsehood. Second, balanced classes hide the base-rate problem: if misinformation is a small fraction of what users see [20], even a highly accurate classifier will produce many false positives in deployment. Third, most benchmarks are static, whereas producers adapt to detection. Evaluation designs that split data by time, by source and by topic, report precision at realistic base rates, and test against adversarially modified content give a more honest picture of what a detector will do in practice.

COMPUTATIONAL EVIDENCE ON DIFFUSION AND EXPOSURE

The second major contribution of computational methods has been descriptive: measuring how much misinformation circulates, who sees and shares it, and how it moves. This evidence has repeatedly corrected impressions formed from anecdotes and media coverage.

4.1 Cascades and speed

The study of 126,000 rumour cascades shared by about three million people on Twitter between 2006 and 2017 remains the reference point [1]. Falsehoods were 70 percent more likely to be retweeted than the truth, and true stories took about six times as long as false ones to reach 1,500 people. Because the stories were selected for having been fact-checked, the sample over-represents contested content, but the scale and the use of six independent fact-checking organisations make the core result unusually robust. Studies of Facebook pages found that conspiracy and science content spread in segregated communities with different cascade dynamics [18], and analyses of the early COVID-19 pandemic compared how questionable and reliable sources spread across five platforms, finding differences in amplification that depended on platform design and user base [34].

4.2 Concentration of exposure and sharing

A second generation of studies linked social-media accounts to survey respondents or voter records, which made it possible to measure exposure among identifiable populations rather than among all accounts. Two studies of the 2016 U.S. election are central (Figure 3). On Twitter, among a panel of more than 16,000 accounts matched to registered voters, only 1 percent of individuals accounted for 80 percent of exposures to fake news sources, and 0.1 percent accounted for nearly 80 percent of the sharing of such sources [8]. On Facebook, among survey respondents who allowed researchers to see their posts, only 8.5 percent shared at least one link from a fake news domain; users over 65 shared nearly seven times as many such articles as those aged 18 to 29, and age remained a strong predictor after controlling for ideology, education and overall posting volume [9].

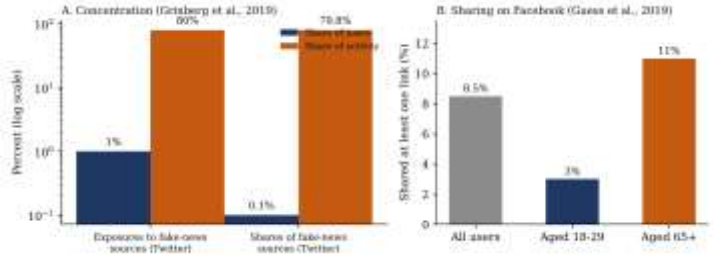

Figure 3: Concentration and demographics of fake-news engagement in the 2016 U.S. election. Panel A: share of panel members accounting for most exposures and shares on Twitter. Panel B: share of Facebook users who shared at least one fake-news link, overall and for the youngest and oldest age groups. Data sources: [8]; [9].

These findings have important consequences for the design of computational interventions. If engagement is heavily concentrated among a small number of "supersharers" and highly engaged older users, then systems that monitor sources and accounts can address much of the problem at far lower cost, and with fewer errors, than systems that classify every post. They also temper claims about the overall prevalence of misinformation. One study combining online and offline media consumption estimated that fake news accounted for about 0.15 percent of Americans' daily media diet [20], while earlier work using browsing data estimated that the average American adult saw and remembered only around one or a few fake news stories during the 2016 campaign [12]. Small averages can coexist with large effects among specific groups, so the relevant question is not only how much misinformation exists but where it is concentrated.

4.3 Algorithms, platforms and exposure

A central political-science question is whether recommendation algorithms cause exposure to misinformation and polarisation. Observational evidence is hard to interpret because algorithms respond to user behaviour. A large collaboration between academic researchers and Meta during the 2020 U.S. election provided experimental evidence: replacing the algorithmic feed with a reverse-chronological feed for consenting users changed what they saw, including increasing exposure to political content and to content from untrustworthy sources, but did not significantly change levels of affective polarisation or political knowledge over the three-month study [35]. The study shows both the value of platform-researcher collaboration and its limits, since such access depends on company decisions and the intervention period was short relative to the timescale of political attitudes.

INTERVENTIONS AND THEIR EVALUATION

The third stage concerns what to do once misinformation has been identified. Computational systems now power most large-scale interventions, but the evidence for their effectiveness comes largely from survey and field experiments in the social sciences. Figure 4 positions the main families of interventions by how far they depend on platform control and by their time horizon and cost.

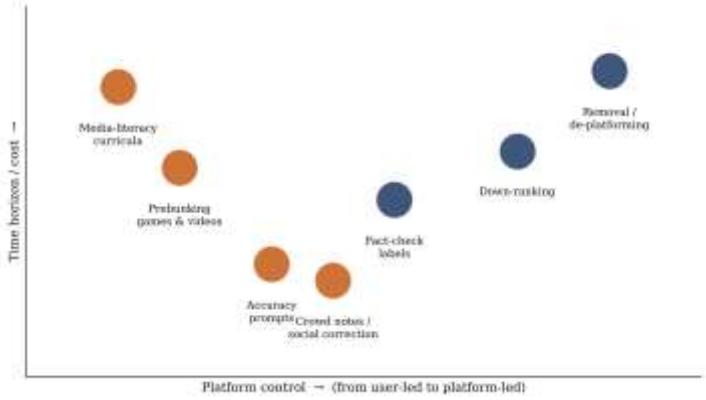

Figure 4: Conceptual map of misinformation interventions by degree of platform control and by time horizon or cost. Positions are illustrative rather than measured. Source: authors' design.

5.1 Labels and corrections

Correction is the oldest intervention, and the evidence is broadly encouraging: corrections reduce false beliefs on average [16], and "backfire" responses are uncommon [15]. Platform labels are a scaled form of correction. A survey experiment found that specific "disputed" or "rated false" tags reduced belief in false headlines more than a general warning about misinformation, although general warnings also reduced belief in true headlines [36]. Another study identified an "implied truth effect": when only some false headlines carry warnings, unlabelled false headlines are judged more accurate, because the absence of a label is read as verification [37]. Because automated systems can label only a fraction of content, this effect is directly relevant to the design of ML-assisted labelling. Partial coverage can backfire unless users understand what an unlabelled item means.

5.2 Accuracy prompts and attention-based nudges

Building on the inattention account, researchers tested whether reminding people about accuracy improves sharing. In survey experiments, the veracity of headlines had a large effect on accuracy judgements but little effect on sharing intentions, and asking participants to rate the accuracy of a single headline improved the quality of what they later chose to share. In a field experiment, private messages asking 5,379 Twitter users who had previously shared links to low-quality sites to rate the accuracy of a headline increased the quality of the news sources they subsequently shared [10]. The effects are modest for each individual but cheap to deliver at scale, which makes them attractive for platforms. Crowd-based judgements offer a related, scalable approach: aggregated ratings from small, politically balanced groups of lay people agreed with professional fact-checkers about as well as the fact-checkers agreed with each other [38]. That finding underpins community-based notes systems that let users add context to posts.

5.3 Prebunking and psychological inoculation

Inoculation theory proposes that exposure to a weakened form of a manipulation technique builds resistance to it. An online game in which players adopt the role of a fake news producer improved their ability to recognise manipulative headlines [39]. A later series of studies, including a field study on YouTube in which short videos explaining manipulation techniques were shown as advertisements to millions of users,

found improved recognition of techniques such as emotional language and false dichotomies [40]. Prebunking is attractive because it does not require classifying individual items as true or false, which sidesteps the hardest problem for detection systems. Open questions concern the durability of effects and whether improved recognition in tests translates into changed sharing behaviour.

Table 4: Summary of intervention evidence reviewed.

Intervention	Key evidence	Direction of effect	Design caveats
Corrections	Meta-analysis of correction studies; large replications of backfire	Reduces misperceptions on average	Weaker for strongly held beliefs; source matters
Fact-check labels	Survey experiments on tags and warnings	Specific tags reduce belief in false items	Implied-truth effect for unlabelled items
Accuracy prompts	Survey experiments and Twitter field experiment	Improves quality of shared news	Small per-person effects; possible habituation
Crowd judgements	Lay ratings compared with professional fact-checks	Comparable agreement at scale	Requires politically balanced crowds
Prebunking	Games and video experiments, including on YouTube	Improves recognition of techniques	Durability and behavioural transfer uncertain
Feed changes	Experimental replacement of algorithmic feeds	Changes exposure, not short-run attitudes	Short study periods; platform-dependent access

5.4 Moderation, removal and regulation

Removal and down-ranking are the most powerful interventions available to platforms and the least transparent to outside researchers. ML classifiers are essential for operating them at scale, but the trade-off between false positives, which suppress legitimate speech, and false negatives, which leave harmful content in place, is a political choice as much as a technical one. Regulation increasingly governs that choice. The European Union's Digital Services Act requires very large online platforms to assess and mitigate systemic risks, including risks from disinformation, to undergo independent audits and to provide vetted researchers with access to data [41]. For computational social scientists, data-access provisions of this kind may matter as much as any new algorithm, because the most informative studies of exposure and intervention have depended on data that platforms control.

GENERATIVE AI AND THE CHANGING THREAT

Large language models have changed the economics of producing misinformation. Earlier work already showed that neural models could generate news-like text that people found trustworthy, and that the best defence against such text was a model of the same kind trained to recognise machine-generated output [42]. More recent experiments found that participants could not reliably distinguish tweets written by GPT-3 from tweets written by humans, and that AI-generated disinformation was judged as credible as, or more credible than, human-written disinformation [4]. A widely cited report on influence operations argued that language models could lower costs, increase scale, enable personalised messaging and make content harder to attribute, while also identifying mitigations at the stages of model construction, model access, content dissemination and belief formation [3].

These developments challenge each stage of the model in Figure 1. Production becomes cheaper and more varied. Detection based on stylistic cues becomes less reliable as generated text grows more fluent, and detectors for machine-generated text are themselves brittle. Interventions that rely on users noticing low-quality content become less effective. At the same time, the same models offer new tools: they can assist fact-checkers in retrieving evidence and drafting explanations, generate counter-messages, and support content moderation in languages that previously lacked trained classifiers. The net effect is not yet known, and measuring it should be a priority.

METHODOLOGICAL CHALLENGES ACROSS THE FIELD

Beyond the stage-specific issues discussed so far, several methodological challenges cut across detection, diffusion and intervention research. They deserve separate treatment because they determine how far computational findings can be trusted as social-scientific evidence.

7.1 Ground truth and label quality

Every supervised detector and every prevalence estimate depends on labels, and labels for misinformation are produced by people working under constraints. Professional fact-checkers choose which claims to check, and their choices reflect newsworthiness, available resources and editorial priorities rather than random sampling [24]. Domain lists compiled by journalists and researchers differ from one another, which is why careful diffusion studies report results under several lists [8, 9]. Crowd judgements can approximate expert ratings when crowds are politically balanced [38], but they inherit the biases of the people who provide them. For social scientists, the implication is that "accuracy" in a detection paper is always accuracy relative to a particular labelling process, and that process should be described as carefully as any survey instrument.

7.2 Measurement validity of trace data

Digital trace data record behaviour rather than belief. A share, a like or a view is not the same as acceptance, and research on sharing intentions shows that people often share headlines they would not rate as accurate when asked [10]. Exposure measured from follow networks or browsing histories also differs from attention: a post that appears in a feed may never be read. Studies that link trace data to surveys, or that measure outcomes such as belief and knowledge directly [35], are therefore more informative about social

consequences than those that rely on engagement metrics alone. Conversely, survey self-reports of exposure are known to be unreliable, which is one reason why linked designs have become the standard for credible prevalence estimates [8, 9].

7.3 Causal inference with networked data

Most questions that matter for policy are causal: does exposure to misinformation change votes, vaccination or trust, and do interventions reduce harm? Observational trace data make causal inference difficult because exposure is driven by preferences, network position and algorithms that respond to behaviour. Randomised experiments embedded in platforms provide the strongest evidence [10, 35, 40], but they are rare, short and dependent on platform cooperation. Natural experiments created by policy changes, outages or staggered rollouts of platform features offer a complementary route, provided that researchers can show that the timing of these changes is unrelated to the outcomes being studied. Network interference, in which the treatment of one user affects others, complicates both designs and should be addressed explicitly in analysis plans.

7.4 Adversarial dynamics and concept drift

Misinformation is produced by actors who respond to detection. Bot designers change their behaviour as classifiers improve [31, 32], producers of false content shift topics and formats, and generative models reduce the stylistic signatures on which content-based detectors depend [3, 42]. Machine-learning research calls the resulting change in data distributions concept drift. A model that is not monitored and periodically re-evaluated on recent data will degrade silently. For research, this means that findings about the features of misinformation should be treated as time-stamped, and that replication across periods is as important as replication across datasets.

7.5 Ethics and the protection of research participants

Finally, misinformation research raises ethical questions that computational methods intensify. Field experiments that send messages to platform users, studies that link social-media accounts to voter records and analyses of private messaging groups all involve people who did not volunteer to be research participants in the ordinary sense. Labelling content or accounts as false or inauthentic carries reputational consequences if the label is wrong. Research teams should document consent procedures, minimise the personal data they retain, consider the risk that published detection methods help adversaries evade them, and pay particular attention to communities, including political minorities and non-English speakers, for whom classifier errors may be more frequent.

DISCUSSION: LESSONS FOR SOCIAL-SCIENCE RESEARCH

8.1 From benchmarks to social outcomes

The most important lesson of this review is that the value of computational misinformation research should be judged by its contribution to social outcomes, not by benchmark scores alone. A detector that achieves high accuracy on a static dataset may contribute little if most exposure is concentrated among a few sources and users, if its errors fall disproportionately on particular communities, or if the content it flags is a small fraction of what people actually consume [8, 20]. Conversely, a modest classifier can be valuable if it is

embedded in a well-designed intervention, such as prioritising claims for human fact-checkers [27]. Evaluation should therefore include measures that matter to social scientists: changes in exposure, belief, sharing, trust and participation.

8.2 Concentration, heterogeneity and targeting

A second lesson concerns heterogeneity. The concentration of exposure and sharing among a small group of highly engaged, often older and more partisan users [8, 9] suggests that average effects of interventions will understate their potential for those who matter most, and overstate it for the majority who rarely encounter misinformation. Interventions should be evaluated for these groups specifically. Targeting also raises normative questions. Directing interventions at identifiable demographic or political groups can be perceived as paternalistic or partisan, and the legitimacy of an intervention is a social-scientific question as much as a technical one.

8.3 Data access and reproducibility

A third lesson concerns data. Much of the most influential evidence comes from collaborations with platforms [19, 35] or from data that later became unavailable when platform policies changed. Research built on application programming interfaces is vulnerable to such changes, and results that cannot be replicated weaken the field's capacity to inform policy. Legislated access for researchers [41], privacy-preserving data donation by users and shared, versioned benchmark datasets with documented collection procedures [26, 30] are all part of the response.

8.4 Beyond English and beyond the United States

Finally, the literature is heavily concentrated on English-language content, on Twitter and Facebook, and on the United States. Misinformation in other languages, on encrypted messaging platforms and in societies with different media systems may follow different dynamics, and detection systems trained on English benchmarks are unlikely to transfer well. This geographic and linguistic imbalance is one of the clearest gaps in current knowledge.

A RESEARCH AGENDA

Drawing these lessons together, Table 5 proposes a research agenda organised by the four stages of the model in Figure 1.

Table 5: Research agenda organised by stage of the misinformation problem.

Stage	Priority question	Suggested design	Data requirement
Production	How does generative AI change the volume, variety and persuasiveness of false content?	Audits of influence operations; experiments comparing human and AI-written messages	Platform transparency reports; model-access logs

Stage	Priority question	Suggested design	Data requirement
Diffusion	Do concentration patterns found in 2016 hold across platforms, countries and languages?	Linked survey and trace-data panels in multiple countries	Data donation; legislated researcher access
Detection	Which detectors remain accurate under temporal, topical and adversarial shift?	Time-split, source-split and adversarial evaluations at realistic base rates	Versioned benchmarks with documented sampling
Intervention	Which interventions change behaviour durably, and for whom?	Long-run field experiments with pre-registered heterogeneous effects	Platform-researcher partnerships under independent oversight
Governance	How do moderation errors affect trust and political participation across groups?	Mixed-methods studies of affected communities	Moderation decision logs and appeals data

CONCLUSION

Computational methods have transformed the study of misinformation. Machine-learning classifiers, network analyses and large-scale behavioural experiments have produced evidence that no traditional social-science method could have generated, from the speed of false cascades [1] to the concentration of exposure among a small fraction of users [8, 9] and the effectiveness of light-touch accuracy prompts [10]. The same body of work has also shown the limits of purely technical solutions. Detection accuracy on benchmarks overstates real-world performance [6, 24], partial labelling can create new distortions [37], and the most important determinants of exposure lie in social networks and institutions rather than in the content alone. For communication and political science, the most productive path is a genuine integration of the two traditions: computational measurement at scale, guided by theories of belief, identity and institutions, and evaluated against the democratic outcomes those theories identify. The arrival of generative AI makes that integration more urgent. Cheaper production of persuasive falsehoods will not be met by better classifiers alone, but by a combination of detection, design, education, regulation and research access that this review has tried to map.

REFERENECS

[1] Potharaju, B., & Ravipudi, J. (2023). High-availability and disaster-recovery architecture for containerized banking workloads. *International Journal of Data Engineering and Intelligent Computing*, 6(1), 1–32.

[2] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151.

[3] Potharaju, A. (2021). Designing dynamic questionnaire engines for risk-based account review workflows. *International Journal of Applied Data Science & Modern Computing*, 4(1), 1–37.

[4] Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science, 359*(6380), 1094–1096.

[5] Karim, N. B., & Kovac, T. (2025). Cloud-Based Big-Data Architectures for AI-Driven Financial Planning and Analysis and Early Risk Signaling in Resource-Constrained Community Financial Institutions. *The Metascience, 3*(3), 29-40.

[6] Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). *Generative language models and automated influence operations: Emerging threats and potential mitigations* (arXiv:2301.04246). arXiv.

[7] Mallempati, A. (2024). Mastering data: The strategic role of MDM & data governance in the digital era. *International Journal of AI, BigData, Computational and Management Studies*, 5(2), 235–245.

[8] Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis)informs us better than humans. *Science Advances, 9*(26), eadh1850.

[9] Potharaju, B., & Ravipudi, J. (2025). Automated data validation and resilient infrastructure design for enterprise regulatory reporting platforms. *International Journal of Computer Science and Engineering Innovations*, 1(1), 46–64.

[10] Potharaju, A. (2026). Embedded compliance architecture for real-time financial systems modernization. *Journal of Information Systems Engineering and Management*, 11(2s), 1147–1158.

[11] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter, 19*(1), 22–36.

[12] Rahman, M. S., Kabir, A. A., Reon, M. A. H., & Rashid, S. U. (2025). From Financial Transactions to Clinical Records: A Unified Machine Learning Methodology for Real-Time Anomaly Detection and Institutional Decision Support In High-Stakes Domains. *International Journal of Information Technology and Management Information Systems (IJITMIS)*, *16*(6), 23-42.

[13] Movva, C. K., & Manohar, V. (2021). OCF protocol implementation for smart home IoT applications: An Android SDK development case study. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 2(1), 89–96.

[14] Karim, N. B., Kovac, T., & Song-Chun, Z. (2025). Model Risk Management for Adaptive Credit Early-Warning Systems in Supervised SmallBusiness Lending: Aligning Machine-Learning Recalibration with Federal Model-Risk Guidance and the NIST AI Risk Management Framework. *International Journal of Acta Informatica*, *4*(1), 80-94.

[15] Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys, 53*(5), Article 109.

[16] Potharaju, B. (2026). Architecting national-scale regulatory reporting platforms: A high-availability infrastructure framework for compliance-critical systems. *Journal of International Crisis & Risk Communication*.

[17] Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys, 51*(2), Article 32.

[18] Mallempati, A., Jaladi, D. S., & Jegajothi, B. (2026). HDLF-DGC: A hybrid deep learning framework for intelligent cybersecurity and data governance in distributed web systems. In *2026 Third International Conference on Innovations in Cybersecurity and Data Science (ICICDS)* (pp. 16–21). IEEE.

[19] Rahman, M. S., Bhowmik, P. K., Hossain, B., Tannier, N. R., Amjad, M. H. H., Chouksey, A., & Hossain, M. (2023). Enhancing fraud detection systems in the USA: A machine learning approach to identifying anomalous transactions. *Journal of Economics, Finance and Accounting Studies*, *5*(5), 145-160.

[20] Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science, 363*(6425), 374–378.

[21] Movva, C. K., & Mupparapu, H. K. (2026). MVVM architectural patterns for enterprise Android kiosk applications: State management and lifecycle handling. *International Journal of Emerging Trends in Computer Science and Information Technology*, 7(2), 320–330.

[22] Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances, 5*(1), eaau4586.

[23] Potharaju, A. (2026). Microservices-based modernization of financial regulatory systems. *International Journal of Computer Science and Engineering Innovations*, 2(1), 68–85.

[24] Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature, 592*(7855), 590–595.

[25] Rashid, S. U., Al Muzakkiruddin, S., & Møller, C. (2026). Detecting and Correcting Temporal Drift in Credit-Risk and Fraud-Detection Models at Community Banks, Credit Unions, and CDFIs: An Automated Model-Risk-Monitoring Framework Below the Supervisory Threshold. *The Computertech*, 26-37.

[26] Kumar, M. S., Yuvaraj, N., & Manoj, M. S. (2026). Intelligent orchestration of master data and privacy compliance via agentic AI for customer lifecycle optimization. In *2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 1715–1723). IEEE.

[27] Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making* (Report DGI(2017)09). Council of Europe.

[28] Yuvaraj, N., & Kumar, M. S. (2025). Agentic AI for zero-touch customer experience: A governance-constrained framework for autonomous service systems. *American International Journal of Computer Science and Technology*, 7(2), 100–111.

[29] Potharaju, A. (2026). Policy-driven authorization architectures for modern financial regulatory platforms. *Journal of International Crisis and Risk Communication Research*, 9(4), 36–44.

[30] Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211–236.

[31] Shammah, R. S., Sultana, S., Faruq, O., Kowser, S., Uddin, A., & Haque, S. A. (2026, July). Data-Driven Decision Support Framework for Engineering and IT Project Management Using Integrated Business Analytics and Management Information Systems. In *2026 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)* (pp. 1-6). IEEE.

[32] Aluri, G. N. C., & Manohar, V. (2025). Generative AI integration patterns for enterprise workflow automation: A practitioner framework. *International Journal of Emerging Research in Engineering and Technology*, 6(3), 165–172.

[33] Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13*(3), 106–131.

[34] Aluri, G. N. C., & Manohar, V. (2026). Model context protocol servers as orchestration layers for generative AI agents in enterprise software delivery. *International Journal of AI, BigData, Computational and Management Studies*, 7(2), 337–345.

[35] Karim, N. B., & Moreau, L. (2026). Quantifying the Prospective Impact of AI-Assisted Early Intervention in US Small-Business Lending: Default Avoidance, Credit-Access Preservation, and mall-Business Employment Effects Across Community Lending Portfolios. *The Metascience*, *4*(2), 33-44.

[36] Potharaju, B. (2026). Infrastructure-level intelligence: Embedding AI into data movement and validation layers. *International Journal of Computational and Experimental Science and Engineering*.

[37] Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior, 32*(2), 303–330.

[38] Manohar, V., & Mupparapu, H. K. (2025). Scalable data pipeline architecture for real-time supply chain analytics using PySpark and Snowflake. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 6(3), 203–212.

[39] Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior, 41*(1), 135–163.

[40] Kumar, M. S., & Yuvaraj, N. (2024). Predictive customer experience orchestration using governed data pipelines and intelligent service signals. *International Journal of Emerging Trends in Computer Science and Information Technology*, 5(1), 206–215.

[41] Rahman, M. M., Computer, M. M. R., Das, N., Uddin, M. S., Shammah, R. S., & Rahman, M. M. (2025, May). An Investigation of Face Detection and Feature Extraction Techniques in Deepfake

[42] Detection Using a Multi-Model Machine Learning Approach. In *2025 3rd International Conference on Business Analytics for Technology and Security (ICBATS)* (pp. 1-8). IEEE.

[43] Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs, 85*(3), 423–441.

[44] Aluri, G. N. C., & Mupparapu, H. K. (2021). Spring Batch framework patterns for large-scale data processing in cloud-native microservices: A GCP deployment study. *International Journal of AI, BigData, Computational and Management Studies*, 2(4), 130–142.

[45] Siddiqui, M. I. H., Rahman, M. S., Kabir, A. A., Mahmud, F. U., Rashid, S. U., & Shammah, R. S. (2023). Comparative analysis of explainable machine learning models for cancer classification using cytological features. *Journal of Medical and Health Studies, 4*(5), 110-150.

[46] Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition, 188*, 39–50.

[47] Zuberi, N., Rahman, M. M., Rashid, S. U., Ahsan, M. N., Rahman, M. D., & Mollah, M. H. O. R. (2026). A robust machine learning framework for detecting temporal drift in financial fraud prevention. *Scientific Reports*.

[48] Mupparapu, H. K., & Movva, C. K. (2020). Serverless migration patterns for compliance-critical financial applications on Microsoft Azure. *International Journal of Emerging Research in Engineering and Technology*, 1(2), 85–96.

[49] Potharaju, A. (2026). Human-guided artificial intelligence for adaptive compliance automation in financial institutions. *Journal of Computational Analysis and Applications*, 35(4), 57–68.

[50] Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences, 113*(3), 554–559.

[51] Karim, N. B., Tolosana, D., & Simko, R. (2025). Responsible and Legally Compliant Enterprise Artificial Intelligence in Small-Business Lending: Governance, Explainability, and Fair-Lending Safeguards for Early-Intervention Recommendation Engines. *International Journal of Modern Computing, 8*(1), 55-66.

[52] Jaladi, D. S., Mallempati, A., & Jegajothi, B. (2026). A hybrid federated privacy-aware deep learning framework for secure data governance in cloud systems. In *2026 4th International Conference on Sustainable Computing and Smart Systems (ICSCSS)* (pp. 1629–1636). IEEE.

[53] Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science, 348*(6239), 1130–1132.

[54] Mallempati, A., & Jaladi, D. S. (2023). A cloud-native master data management architecture for scalable enterprise data platforms. *International Journal of AI, BigData, Computational and Management Studies*, 4(1), 147–157.

[55] Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances, 6*(14), eaay3539.

[56] Shammah, R. S., Rahman, I., Guria, Z. M., Kowser, S., Faruq, O., & Morshed, N. (2025, December). Scalable Machine Learning Architecture for Securing IoT Networks with Digital Governance in the AI Era of Cyber-Resilient Smart Cities. In *2025 International Conference on AI-Driven STEM Education and Learning Technologies (AISTEMEDU)* (pp. 1-7). IEEE.

[57] Potharaju, B., & Sajja, R. R. (2024). Modernizing US federal regulatory reporting infrastructure for scalability, resilience, and audit readiness. *International Journal of Data Engineering and Intelligent Computing*, 7(1), 1–36.

[58] Movva, C. K., & Aluri, G. N. C. (2025). Enterprise mobile payment gateway architecture on IoT handheld devices: Security and reliability design. *International Journal of AI, BigData, Computational and Management Studies*, 6(1), 232–241.

[59] Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web* (pp. 675–684). ACM.

[60] Yuvaraj, N., Kumar, M. S., & Manoj, M. S. (2026). Unified agentic AI architecture for data lineage, governance, and predictive workflow automation in enterprise systems. In *2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 1817–1828). IEEE.

[61] Shahiduzzaman, M., Sourav, M. S. A., Kabtia, M., Shammah, R. S., Islam, M. A., Chakraborty, P., ... & Adnan, M. (2026). Attention-guided quantum-inspired extreme learning machine with grey wolf optimization for rare-attack detection in IIoT networks. *Discover Artificial Intelligence*.

[62] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3391–3401). Association for Computational Linguistics.

[63] Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2931–2937). Association for Computational Linguistics.

[64] Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 422–426). Association for Computational Linguistics.