مجتمع التحقق العربي
Arabi Facts Hub

# Countering Misinformation in the Digital Age

## A Comprehensive Handbook for Detecting and Combatting Inauthentic Online Content

Mahmoud Hadhoud

IDRC · CRDI
International Development Research Centre
Centre de recherches pour le développement international

Canada

**This work would not have been possible without the generous support of the International Development Research Centre (IDRC)**


**This research was developed based on trainings supported by the United Nations Democracy Fund**



Open Transformation Lab Inc.
110 Duane St, Ste 1C
New York, NY 10007
(202) 733-8338

This publication can be downloaded at no cost at Arabifactshub.com.

# Table of Contents

# Introduction

In the age of digital media, misinformation has become one of the most pressing challenges facing societies around the world. The rapid spread of false or misleading information, whether driven by political agendas, economic interests, or societal polarization, undermines trust in institutions, fuels hatred, and distorts public understanding of critical issues. The problem is particularly acute on social media platforms, where unverified content can reach millions within seconds, often before fact-checking or reliable sources can intervene. Combatting misinformation requires more than just technological tools; it demands a deep understanding of the sources, tactics, and impact of such content.

This handbook is designed as a comprehensive guide for journalists, fact-checkers, researchers, and anyone concerned with addressing the spread of misinformation, especially in the context of the Arab world. It offers practical methodologies, ethical guidelines, and tools to help detect, verify, and counter false information. The handbook is organized into seven detailed chapters, each focused on a specific aspect of misinformation detection and intervention. We also include two appendices that address essential considerations for digital rights and the use of valuable data resources.

Chapter 1, **Methodology and Tools for Detecting Inauthentic Online Campaigns**, provides a clear framework for identifying inauthentic campaigns that manipulate public opinion through fake accounts, bots, and coordinated disinformation efforts. We explore the technical tools and methodologies used to track these campaigns, focusing on both manual and automated techniques for uncovering digital manipulation.

Chapter 2, **Guidelines for Verifying Gender-based Information**, explains that gender-based misinformation can have far-reaching consequences, especially in regions where gender inequality persists. This chapter offers guidelines for evaluating content that perpetuates stereotypes or promotes gender-based violence, focusing on how to identify and challenge harmful narratives in digital media.

Chapter 3, **Guidelines for Verifying Content That Incites Against Refugees**, focuses on identifying and countering misinformation that fuels hate and discrimination against refugees. We outline methods for detecting misleading narratives and provide strategies for fact-checking content related to refugees' rights and conditions, with a focus on fostering empathy and accurate representation.

Chapter 4, **Ethical Guidelines for Using Facial Recognition Tools in Fact-Checking**, addresses facial recognition technology, which has become a valuable tool in fact-checking but raises significant ethical concerns related to privacy and bias. This chapter discusses the ethical considerations surrounding the use of facial recognition tools, offering guidelines for responsible implementation in digital investigations.

Chapter 5, **Guide to Auditing Online Advertisements During Election**s, focuses on election periods, which are often marked by the manipulation of information through targeted ads and misinformation campaigns. This chapter explains how to audit online advertisements, track misleading political ads, and evaluate the role of big data in influencing voter behavior, with a focus on maintaining transparency in digital electioneering.

Chapter 6, **Guidelines for Fact-Checking Disinformation Campaigns During Climate Conferences**, examines the growing issue of climate change misinformation, particularly during international summits like COP meetings. We offer practical tools for fact-checking claims related to climate science and policy, while also exploring strategies to debunk greenwashing and other misleading tactics used by various stakeholders.

Chapter 7, **Guidelines and Warnings for Using Artificial Intelligence in Data Journalism**, delves into the world of AI tools, which have transformed data journalism, but they come with limitations and ethical challenges. This chapter provides essential guidelines for journalists who use AI in their work, offering warnings about algorithmic bias, data privacy issues, and the importance of maintaining journalistic integrity in the use of AI-driven data tools.

Appendix 1, **Digital Rights: Protecting Privacy, Freedom, and Access in the Digital Age**, explores the intersection of digital rights and misinformation, offering a comprehensive overview of how to protect privacy, ensure freedom of expression, and guarantee access to accurate information in the online world. It covers the importance of maintaining digital rights while combating disinformation and how policies and regulations can support these goals.

Appendix 2, **Arabi Facts Hub Database: How to Access and Make Use of It?**, introduces the Arabi Facts Hub database, a powerful resource for fact-checkers and researchers seeking to verify information related to the Arab world. We provide instructions on how to access and navigate the database, along with tips on how to effectively utilize it for comprehensive digital investigations.

Two case studies was meticulously developed by the **Arabi Facts Hub** team under the expert supervision of AFH senior editor Osama Elsayyad, in order to provide a practical model for fact-checkers and enhance best practices.

This book was authored in close collaboration with the **United Nations Democracy Fund** (UNDEF) as part of the **Identifying Data Patterns to Understand and Counter Misinformation on Arabic Social Media** project. This project brought together the strengths of both organizations to tackle the pressing issue of misinformation in the digital age. The partnership between Arabi Facts Hub and UNDEF was driven by a shared commitment to enhancing media literacy, safeguarding democratic processes, and upholding the integrity of information across Arabic social media platforms. We are deeply grateful to the UNDEF for their generous support, which enabled the creation of this vital resource. Their ongoing dedication to fostering transparent, reliable information helps empower individuals and organizations working to identify, counter, and prevent the spread of misinformation in the Arab world and beyond.

Chapter I

# Methodology and Tools for Detecting Inauthentic Online Campaign

## Introduction

With the rapid expansion of digital communication, social media has evolved into one of the most influential tools for shaping public discourse, influencing opinions, and framing narratives. However, this digital space is increasingly being manipulated through inauthentic online campaigns designed to mislead the public, amplify certain perspectives while suppressing others, and create artificial trends that distort reality. These campaigns are often orchestrated through networks of bot accounts, coordinated digital activities, and paid influencers who propagate misleading or politically motivated content. Their objectives vary, ranging from influencing electoral outcomes and swaying public sentiment on geopolitical conflicts to targeting activists, journalists, and opposition figures with disinformation and harassment.

In the Arab world, the use of inauthentic online campaigns has gained significant prominence, particularly during diplomatic crises, political upheavals, and times of social unrest. These campaigns have been strategically deployed to manufacture consent, discredit opponents, and shift public focus away from critical issues by flooding digital spaces with misleading narratives. Governments, intelligence agencies, and interest groups have increasingly relied on these tactics, leveraging social media platforms to push state-controlled narratives or discredit independent media. In some cases, these disinformation campaigns are transnational, coordinated across different regions to reinforce certain ideologies or political stances on a global scale.

Recognizing the indicators of inauthentic activity and equipping journalists, fact-checkers, and media professionals with the necessary tools to investigate and expose these operations is crucial to safeguarding information integrity. Effective detection methods

include analyzing engagement patterns, identifying bot-like behavior, tracing the origins of viral content, and cross-referencing sources for authenticity. Understanding the mechanisms behind fake online campaigns—such as the characteristics of fake accounts, their coordination strategies, and their amplification techniques—allows journalists and researchers to dismantle these operations more effectively.

This chapter provides a comprehensive analysis of the structure and impact of fake online campaigns, detailing the technical and strategic mechanisms behind them. It also explores the latest investigative tools and methodologies that can be employed to track, analyze, and counteract these manipulative digital efforts, ultimately reinforcing the credibility of online discourse and protecting the public from deception.

## What Are Fake Online Campaigns?

Fake online campaigns involve the coordinated use of social media accounts to artificially amplify a narrative, suppress opposing views, or generate misleading content, often serving political, economic, or ideological interests. These campaigns rely on a combination of fake bot networks, troll armies, and manipulated accounts designed to create an illusion of widespread support for certain ideas while drowning out dissenting voices. By leveraging automation, artificial engagement, and deceptive messaging, these campaigns aim to manipulate public perception, distort facts, and influence political or social debates.

One of the most common tactics in these operations is the deployment of bot accounts—automated profiles programmed to interact with trending topics, retweet specific messages, and artificially boost engagement. These bots are often controlled by centralized command systems that allow a single operator to manage thousands of accounts, enabling rapid dissemination of propaganda or disinformation. Additionally, troll armies—human operators tasked with spreading misleading narratives—work alongside bots, engaging in direct harassment of critics, journalists, and activists who challenge the promoted discourse. Trolls often coordinate their efforts through private messaging groups, receiving instructions on what content to push and whom to target.

Another key element of fake online campaigns is hashtag manipulation. Coordinated networks of bots and trolls work together to flood social media platforms with pre-approved hashtags, making them trend artificially. This tactic creates the illusion that a particular viewpoint has organic support, pressuring users and media outlets to take the fabricated narrative seriously. Additionally, fake engagement—such as likes, shares, and

comments—further boosts visibility, ensuring that manipulated content reaches wider audiences.

These campaigns are not only used to promote specific narratives but also to suppress or discredit opposing views. By overwhelming discussions with misleading content, fake accounts dilute factual information, making it harder for genuine voices to be heard. In some cases, coordinated smear campaigns target individuals, spreading false accusations, doctored images, or fabricated quotes to undermine credibility. The goal is often to create confusion, erode trust in independent journalism, and polarize public opinion.

Understanding the mechanisms behind these manipulative efforts is essential for journalists, fact-checkers, and researchers working to expose digital deception. Recognizing patterns of inauthentic activity, identifying coordinated engagement, and using investigative tools to trace the origins of viral content are crucial steps in dismantling these campaigns. By shedding light on these operations, media professionals can help safeguard public discourse from misinformation and ensure that online spaces remain platforms for authentic and informed debate.

## Types of Fake Online Campaigns

**1. Propaganda Campaigns**

Propaganda campaigns are orchestrated efforts to spread political, ideological, or social narratives that serve the interests of a particular entity, whether a government, political party, or corporate organization. These campaigns leverage social media, news outlets, and digital influencers to push curated messaging, often exaggerating positive aspects of an entity while demonizing its opponents. By saturating digital platforms with carefully crafted narratives, propaganda campaigns aim to shape public perception, reinforce loyalty among supporters, and suppress dissent by drowning out alternative viewpoints.

**2. Astroturfing**

Astroturfing is a deceptive tactic in which fake grassroots movements are artificially created to simulate widespread public support for a cause, policy, or product. This strategy involves deploying bot accounts, paid influencers, and coordinated social media activity to fabricate the illusion of organic public backing. Governments, corporations, and lobbyists often use astroturfing to manipulate public opinion, sway elections, or promote policies under the guise of genuine popular demand. By making it appear as though a large number

of people support a specific narrative, astroturfing can pressure policymakers, mislead audiences, and distort democratic discourse.

**3. Character Assassination**

Character assassination campaigns are designed to target individuals—such as activists, journalists, or political opponents—by discrediting their reputation through smear tactics. These campaigns often involve spreading false accusations, manipulated images, deepfake videos, or fabricated stories to damage the credibility of the target. Coordinated online harassment, defamatory content, and doxxing (revealing personal information) are frequently employed to intimidate and silence critics. Character assassination can be particularly harmful in repressive environments, where such attacks may be used as a pretext for legal action, imprisonment, or public ostracization.

**4. Agenda Manipulation**

Agenda manipulation involves controlling public discourse by amplifying specific topics while suppressing or distracting from others. This tactic is often used to divert attention from critical issues, steer political debates in a particular direction, or frame narratives that benefit a certain group. Social media trends, algorithmic manipulation, and coordinated influencer campaigns are tools frequently employed to execute this strategy. By strategically directing the public's focus, agenda manipulation can influence elections, shape policy discussions, and control the media landscape to favor certain interests.

**5. Misinformation Campaigns**

Misinformation campaigns are organized efforts to disseminate false or misleading information with the intent to deceive and manipulate audiences. These campaigns can take various forms, including fake news articles, doctored images, misleading statistics, and fabricated quotes. They are commonly used to polarize societies, undermine trust in legitimate institutions, or sway public opinion on contentious issues. Misinformation campaigns are particularly dangerous during election seasons, public health crises, and geopolitical conflicts, where they can have real-world consequences, such as influencing voter behavior or inciting violence.

**6. Crisis Exploitation**

Crisis exploitation refers to the strategic use of ongoing political, economic, or social crises to spread misinformation and advance specific agendas. In times of instability—such as wars, protests, pandemics, or natural disasters—bad actors exploit public fear and confusion to push propaganda, create division, or justify authoritarian measures. False reports, conspiracy theories, and fabricated narratives are often spread to manipulate

emotions, incite panic, or legitimize aggressive policies. Crisis exploitation is particularly effective because people are more susceptible to misinformation when they are anxious, overwhelmed, or desperate for information.

## Fake Bots: What Are They and How Do They Work?

Fake bots are automated or semi-automated accounts programmed to execute repetitive tasks on social media platforms, such as liking, sharing, commenting, or posting content, often at a rapid pace and in large volumes. These accounts can be controlled individually or as part of a coordinated network, commonly referred to as a botnet, to manipulate social media trends, amplify specific narratives, and influence public perception. By flooding online discussions with predetermined content, fake bots create the illusion of widespread support or opposition, shaping debates and misleading audiences into believing that certain viewpoints are more popular than they actually are.

These bots come in different forms, ranging from simple script-based accounts that follow and like posts to more sophisticated AI-driven profiles that can generate human-like interactions, respond to comments, and engage in conversations. Some bots are designed to operate passively, boosting engagement numbers for targeted content without direct interaction, while others actively participate in discussions by posting pre-scripted replies or responding to real users to create the appearance of genuine discourse. By leveraging algorithms and automation, these fake accounts can artificially boost the visibility of hashtags, news articles, and social media posts, ensuring that their promoted messages reach a broad audience.

Beyond mere engagement manipulation, fake bots play a crucial role in spreading misinformation and propaganda. They can be used to share misleading or completely fabricated stories, amplifying narratives that benefit a specific political agenda, corporate interest, or ideological movement. During political campaigns, fake bots can be deployed to attack opponents, distort public debates, or create confusion by sharing conflicting or misleading information. In crisis situations, such as protests, conflicts, or health emergencies, bot networks often exploit public anxiety by pushing conspiracy theories, promoting divisive rhetoric, or drowning out fact-based reporting with fabricated content.

One of the most concerning aspects of fake bots is their ability to evolve and adapt to countermeasures deployed by social media platforms. While companies like Twitter (now X), Facebook, and Instagram implement policies to detect and remove automated accounts, bot developers continually refine their techniques to bypass these restrictions.

Some bots use machine learning algorithms to mimic human behavior, such as varying their posting patterns, using conversational language, and even engaging in debates to avoid detection. Others operate in hybrid networks, where human operators oversee and direct automated activity to maintain a more authentic appearance.

The widespread use of fake bots poses significant risks to the integrity of online discussions, the credibility of news dissemination, and the democratic process. By artificially shaping public perception, they can distort election outcomes, fuel societal divisions, and erode trust in legitimate media sources. Combating this phenomenon requires a combination of advanced detection tools, regulatory oversight, and public awareness initiatives to educate users on identifying and countering bot-driven manipulation. Understanding how these automated entities function is crucial for journalists, fact-checkers, and policymakers seeking to safeguard digital spaces against manipulation and ensure that social media remains a platform for authentic and informed discourse.

## Characteristics of Fake Bots

**- Low-Quality Profile Pictures**

Many fake bot accounts use easily identifiable, low-quality profile pictures to appear legitimate at first glance. These images often include AI-generated faces, stock photos, or stolen images from real people, sometimes altered slightly to avoid detection. AI-generated images may contain subtle distortions, such as asymmetrical facial features or unnatural eye alignment, which can serve as red flags. Some bots also use pictures of celebrities or public figures, banking on the assumption that users will not scrutinize the profile closely. By analyzing reverse image searches and checking for inconsistencies, researchers and journalists can spot these deceptive tactics.

**- Unrealistic Activity Levels**

One of the most telling signs of a fake bot account is its abnormally high level of activity. While an average human user may post several times a day, bots often exceed human capabilities by tweeting, commenting, or sharing content dozens or even hundreds of times within a short period. This hyperactive engagement is designed to manipulate social media algorithms by increasing the visibility of specific posts or hashtags. Additionally, these bots rarely take breaks, often operating 24/7, which is another indicator of automated behavior. Tracking engagement patterns over time can help reveal unnatural posting frequencies.

**- Recent Account Creation**

Many fake bot accounts are created shortly before a disinformation or propaganda campaign begins. Since these operations are often short-term efforts aimed at influencing a particular event—such as an election, geopolitical crisis, or social movement—bot networks are frequently set up just days or weeks in advance. These newly created accounts typically have minimal or no prior posting history, few personal details, and a sudden surge in activity. Journalists and fact-checkers can use account age as a key factor when determining whether an account is part of a coordinated inauthentic campaign.

**- Irregular Engagement Patterns**

Fake bots often display highly irregular and synchronized engagement behavior, making them distinguishable from real users. They tend to amplify specific hashtags, posts, or political narratives in coordinated bursts, creating the illusion of organic support. These spikes in engagement, often occurring at the same time across multiple accounts, suggest that the activity is centrally controlled. Additionally, bots may exhibit engagement patterns that lack natural variation, such as liking and retweeting in rapid succession or following hundreds of accounts within minutes. Identifying these engagement anomalies can help expose coordinated disinformation efforts.

**- Repetitive Content**

Another hallmark of bot activity is the repetitive nature of the content they share. Fake bots frequently copy and paste the same messages across multiple accounts, sometimes with slight variations to evade detection. This tactic is commonly used to spread misinformation, reinforce propaganda narratives, or boost engagement metrics artificially. Even when variations exist, the structure, wording, and themes of the messages often remain strikingly similar. Researchers can detect this behavior by analyzing patterns in wording and cross-referencing posts from suspected accounts.

**- Follow Patterns**

Fake bot accounts typically exhibit abnormal follow behaviors, often clustering together in an artificial network. They frequently follow each other or a central coordination account, which serves as the command hub for a broader disinformation campaign. This closed-loop following pattern ensures that content spreads rapidly within the network, boosting visibility before it reaches external audiences. Additionally, bot accounts tend to follow high-profile influencers, political figures, or media outlets that align with the intended narrative, further amplifying their impact. Analyzing follower relationships can provide crucial insights into bot networks.

**- Anonymity**

Many fake bot accounts lack basic personal information, making them easily identifiable upon closer inspection. They often use generic names, random strings of letters and numbers in their usernames, or usernames that resemble one another in a structured way. These accounts also tend to avoid providing details such as location, biography, or a history of personal interactions, which are typical of genuine users. A lack of personal engagement—such as conversations with friends or posts about daily life—further indicates inauthentic behavior. Journalists and researchers can use metadata analysis to detect anonymity patterns and assess the credibility of an account.

## Detecting and Investigating Fake Online Campaigns

**Step 1: Identifying Suspicious Hashtags and Trends**

One of the first indicators of a coordinated fake campaign is an unnatural surge in hashtag activity. Journalists and fact-checkers should closely monitor social media for hashtags or topics that experience sudden, unexplained spikes in engagement. While organic trends typically build momentum over time through diverse user interactions, manipulated trends often emerge abruptly, with large numbers of accounts posting identical or similar content within a short period. Analyzing when and how a hashtag gains traction, including the timestamps of the first few tweets, can reveal whether it is being artificially amplified. Comparing this activity to previous organic trends can also help differentiate between authentic public discourse and inauthentic manipulation.

**Step 2: Examining Participating Accounts**

Once a suspicious trend is detected, the next step is to analyze the accounts that are driving the conversation. Fake accounts often share distinct characteristics, such as being recently created, having limited followers, and posting excessive retweets with little original content. By examining these factors, fact-checkers can determine whether the engagement is authentic or part of a coordinated campaign. Additionally, inconsistencies in profile images and bios—such as generic or AI-generated photos, mismatched usernames, or a lack of personal details—can further indicate inauthentic behavior. Investigating the linguistic patterns and timestamps of posts can also provide insight into whether the activity originates from a centralized operation rather than independent users.

**Step 3: Network Analysis of Fake Accounts**

Beyond individual account analysis, understanding the connections between participating accounts is crucial in uncovering larger bot networks. Fake campaigns often rely on clusters of accounts that follow each other, engage in synchronized activity, and amplify the same content in a structured manner. By mapping out these relationships, researchers can identify command hubs, automated patterns, and coordination strategies. Advanced network analysis techniques—such as visualizing retweet relationships, shared content, and follower overlap—can help expose the backbone of a fake campaign. In some cases, identifying a few key accounts can lead to the discovery of an entire network designed to manipulate public discourse.

**Step 4: Using Investigative Tools**

A variety of investigative tools can assist in detecting fake campaigns and bot activity. Social media analytics platforms like Hoaxy, Botometer, and TruthNest help researchers assess the likelihood of an account being a bot based on its activity patterns, posting frequency, and network behavior. Reverse image search tools, such as Google Images or TinEye, can verify whether a profile picture has been stolen or AI-generated. Metadata analysis tools enable investigators to track anomalies in post timing and geographical origins. By leveraging these tools alongside manual analysis, journalists and fact-checkers can systematically uncover inauthentic campaigns and mitigate the spread of misinformation.

## Investigative Tools for Detecting Fake Accounts and Online Campaigns

**1. Tools for Measuring Influence and Reach**

- Twitonomy

Twitonomy is an analytics tool designed to offer in-depth insights into Twitter activity. It provides valuable data on user engagement, including retweets, mentions, and hashtags used. Through visual reports and interactive charts, Twitonomy helps users track how often certain hashtags are used and identify trends within a particular timeline. The tool also allows for the monitoring of follower growth patterns and the analysis of which users are generating the most influence within a specific topic. Twitonomy's ability to uncover high-level engagement trends makes it useful for journalists and fact-checkers when identifying potential manipulation or inauthentic campaigns. It helps assess whether an account or hashtag is organically gaining traction or being artificially amplified through bot activity.

- Meltwater

Meltwater is a paid social media monitoring and analytics tool that provides in-depth analysis of social media influence and engagement. It allows users to track the performance of specific hashtags, measure the reach of particular content, and analyze tweet geolocation to detect patterns of manipulation. Meltwater can identify suspicious spikes in hashtag usage that may indicate artificial amplification or astroturfing campaigns. Its geolocation feature also enables the detection of coordinated activity across different regions or countries, which is particularly useful in uncovering international or politically motivated disinformation efforts. Meltwater's comprehensive suite of tools helps to assess both the impact and authenticity of social media interactions, allowing for a clearer understanding of whether an online movement or campaign is being manipulated.

**2. Tools for Detecting Account Networks**

- Gephi

Gephi is a robust open-source network analysis tool used to visualize and analyze relationships between entities, such as social media accounts. By mapping out connections between users, Gephi allows researchers to uncover coordinated networks of activity. This is especially valuable when investigating potential bot networks or coordinated disinformation campaigns. The tool offers powerful visualization options that enable analysts to view clusters of interconnected accounts, detect patterns of behavior, and identify central nodes in the network, such as bot controllers or orchestrating accounts. Gephi is often used in combination with social media data collected through APIs to uncover hidden connections and analyze the dynamics of fake account networks.

- Fedica Bot Checker

Fedica Bot Checker is an advanced tool that specializes in detecting fake accounts and identifying clusters of coordinated activity. This tool uses sophisticated algorithms to analyze account behavior, looking for signs of automation, such as high posting frequency, low engagement diversity, and unnatural patterns of activity. It can also identify accounts that are part of larger, interconnected networks by analyzing follower relationships and engagement patterns. Fedica's ability to detect clusters of activity makes it an essential tool for uncovering fake campaigns that rely on multiple accounts to artificially inflate support for certain narratives or manipulate public opinion on social media platforms.

- Pipl

Pipl is a powerful tool for uncovering an individual's digital footprint across various online platforms. It collects publicly available data from a wide range of sources, such as social

media, blogs, and websites, to help verify the authenticity of an online identity. Pipl's ability to search through extensive databases makes it ideal for investigating whether an account is legitimate or part of a broader network of fake profiles. It can also identify inconsistencies in the information provided by individuals across different platforms, allowing journalists and fact-checkers to cross-check identities and verify whether the accounts involved in a campaign are connected or fraudulent.

- Webmii

Webmii is a search tool that helps uncover an individual's digital traces across the web. It provides insights into how and where a person or account appears online by aggregating public information from a variety of online sources. Webmii can be particularly useful for identifying individuals involved in suspicious online activity and verifying the authenticity of their digital presence. By searching through social media profiles, news outlets, and public databases, Webmii can provide an overview of an individual's online identity, helping to verify if they are real users or fabricated personas created to manipulate online discussions.

**3. Tools for Analyzing Social Media Data**

- Python Programming & Twitter API

Python programming, when combined with the Twitter API, offers a powerful method for extracting and analyzing social media data. Researchers and journalists can use Python scripts to collect large datasets from Twitter, including tweets, user information, hashtags, and trends. These scripts can be customized to monitor specific topics or keywords over time, enabling the detection of anomalous activity patterns that suggest manipulation. Python can also be used to clean, analyze, and visualize this data, providing insights into how content is spreading, which accounts are driving engagement, and whether a campaign is authentic. The flexibility and scalability of Python, coupled with the Twitter API, make it a go-to tool for in-depth social media analysis.

- Social Bearing

Social Bearing is a tool that provides detailed insights into Twitter activity, including sentiment analysis, engagement metrics, and network connections. It is especially useful for tracking user behavior and understanding the impact of specific posts or hashtags. Social Bearing can reveal which accounts are driving conversations, the level of engagement a particular tweet receives, and the sentiment surrounding a topic. It also offers network analysis features, helping to uncover relationships between users and detect coordinated activity. Fact-checkers and journalists can use Social Bearing to

analyze patterns in online discourse, identify fake trends, and assess whether certain narratives are being artificially promoted.

**4. Tools for Detecting AI-Generated Content**

- AI Content Detector & Hugging Face AI Content Detector

AI Content Detectors, such as those provided by platforms like Hugging Face, are tools designed to identify whether a piece of content—whether text, article, or social media post—has been generated by artificial intelligence. These tools use machine learning algorithms to analyze language patterns, structure, and coherence to distinguish human-written content from AI-generated text. The AI Content Detector can be especially useful for fact-checkers who suspect that automated systems may be responsible for creating misleading narratives or spreading disinformation. It helps identify AI-generated content that could be part of a larger misinformation campaign aimed at manipulating public opinion or shaping political discourse.

- Is It AI?

Is It AI? is a tool designed to assess the likelihood that a piece of content is AI-generated. By analyzing textual features such as sentence structure, coherence, and word choice, it determines whether the content was written by a human or a machine. Is It AI? uses deep learning models trained on large datasets of human and AI-generated text to predict the origin of the content. This tool is particularly valuable when investigating suspicious posts or articles that may be part of an artificial disinformation campaign. It helps journalists and researchers quickly assess whether a text is authentic or potentially part of an automated effort to manipulate the online discourse.

**5. Tools for Detecting AI-Generated Images**

- AI or Not

AI or Not is a tool specifically designed to identify AI-generated images and deepfakes. It analyzes the image for inconsistencies that might indicate that it was created using artificial intelligence, such as irregularities in pixel patterns or visual distortions that are not typically present in natural photos. This tool is useful for detecting images that may have been fabricated or altered to spread misleading information or create false narratives. In the context of disinformation campaigns, identifying AI-generated images is crucial in ensuring that the visual content being shared online is authentic and not part of a coordinated effort to manipulate public perception.

- JPEGSnoop

JPEGSnoop is a detailed image analysis tool that examines the metadata and compression artifacts of JPEG images to detect signs of modification. It provides forensic analysis of an image's file structure, allowing users to determine whether the image has been edited or artificially generated. JPEGSnoop is particularly useful in verifying the authenticity of images that are being used to manipulate public opinion, as it can detect alterations in both the image content and its digital footprint. Journalists and fact-checkers can use this tool to investigate suspicious images that may be part of a fake news campaign, ensuring the integrity of visual content shared online.

- Hugging Face AI Image Detector

Hugging Face offers an AI Image Detector tool designed to analyze images and determine if they were created by AI applications. This tool leverages advanced machine learning models to detect visual patterns that are characteristic of AI-generated images, such as unnatural lighting, reflections, or backgrounds. It is especially useful for identifying deepfakes and other forms of fabricated media that can be used to deceive the public. By using the Hugging Face AI Image Detector, fact-checkers can quickly determine whether images circulating on social media or news websites have been artificially generated, helping to combat visual disinformation in real-time.

## Best Practices for Investigating Inauthentic Campaigns

**1. Cross-check Information**

Cross-checking information is a vital step in verifying the credibility of suspicious posts or claims. Fact-checkers and journalists should compare the content with reliable and trustworthy news sources to verify the accuracy of the information being shared. This process involves checking whether the claims are reported by reputable outlets, cross-referencing dates, details, and context, and ensuring that the sources align with the facts. By cross-referencing with multiple established news sources, it's possible to uncover discrepancies and identify whether a story is based on misinformation or disinformation. This step is essential to prevent the spread of false narratives and ensure that accurate information is shared with the public.

**2. Monitor Geolocation Inconsistencies**

Geolocation inconsistencies are another red flag when investigating suspicious online activity. By analyzing the geographic location from which posts or hashtags are originating, fact-checkers can identify if a hashtag is being artificially amplified in areas where it is

unlikely to gain organic traction. For example, if a hashtag that is relevant to a particular country or region is gaining significant engagement from accounts in another part of the world with no apparent connection to the topic, this could indicate manipulation through bot networks or coordinated activity. Monitoring these inconsistencies helps to detect inauthentic activity and prevents the spread of fabricated narratives that may be influenced by external sources with a vested interest.

**3. Analyze Metadata**

Metadata analysis is a powerful method for verifying the authenticity of digital content, such as images and videos. By extracting metadata from digital files, researchers can uncover valuable information about the creation process, such as the time, date, and location where a file was created or modified. This information can help verify whether an image or video is original or has been tampered with. Metadata can also reveal the tools and software used to edit or manipulate the content, allowing journalists and fact-checkers to determine whether the content has been altered to support a specific narrative. Analyzing metadata is an essential tool for uncovering digital forgeries and maintaining the integrity of online content.

**4. Leverage Open-Source Intelligence (OSINT)**

Open-source intelligence (OSINT) involves gathering publicly available information from a variety of online platforms to build a comprehensive picture of a situation or event. By combining data from social media, news outlets, official reports, and other accessible sources, fact-checkers can cross-reference details and uncover inconsistencies. OSINT is particularly useful in verifying suspicious online content, as it allows researchers to gather a wide range of perspectives and corroborate the information. This process involves piecing together various digital clues to build a holistic investigation, which can be especially valuable in uncovering disinformation campaigns or identifying the origins of misleading content. OSINT is a cost-effective and efficient method for countering misinformation and ensuring information accuracy.

**5. Engage in Crowd-sourced Verification**

Crowd-sourced verification involves leveraging the collective knowledge and efforts of online communities and fact-checking organizations to validate or debunk content. By collaborating with other researchers, journalists, and experts, fact-checkers can access a wider range of tools, perspectives, and resources to examine suspicious claims more thoroughly. This approach taps into the power of crowdsourcing, where multiple individuals or organizations contribute to the verification process, often leading to faster and more accurate conclusions. Engaging in crowd-sourced verification allows for a more

comprehensive investigation, as it brings together diverse expertise and insights that can uncover hidden manipulation tactics or false narratives. This method fosters transparency and collaboration, ultimately helping to ensure the accuracy and integrity of the information shared online.

## Conclusion

The rise of inauthentic online campaigns presents a profound and escalating challenge to the integrity of information in the digital age. These deceptive tactics, often employed by malicious actors or entities with political or ideological agendas, not only distort public opinion but also undermine trust in credible sources of news and information. With the rapid proliferation of social media platforms, which are central to modern communication, it has become increasingly difficult to discern between genuine interactions and orchestrated disinformation efforts. The pervasive use of fake accounts, coordinated content manipulation, and viral misinformation can significantly alter the course of political, social, and economic discourse, creating confusion and mistrust within the public sphere. As such, the ability to identify, expose, and counteract these campaigns is crucial for maintaining the transparency and authenticity of digital communication.

However, through the strategic and rigorous use of investigative methodologies, digital tools, and an understanding of the evolving tactics used by those who manipulate online narratives, journalists, researchers, and fact-checkers can effectively mitigate these threats. Employing techniques such as network analysis, metadata verification, and cross-referencing with credible sources enables media professionals to detect and dismantle fake accounts and orchestrated campaigns. By staying vigilant and continuously adapting to the ever-changing landscape of online manipulation, individuals and organizations involved in safeguarding information integrity can play an active role in curbing the negative effects of disinformation. Moreover, knowledge-sharing among journalists, fact-checkers, and the wider public is essential in staying ahead of the sophisticated methods employed by those who seek to deceive and manipulate.

This chapter serves as a comprehensive guide for identifying fake accounts, detecting misinformation campaigns, and utilizing cutting-edge tools to expose inauthentic online activity. By detailing best practices and providing a thorough understanding of how digital disinformation works, this resource empowers media professionals to strengthen their investigative capabilities. It equips them with the knowledge necessary to navigate the increasingly complex world of online manipulation. By implementing these strategies and

adopting a proactive, collaborative approach to information verification, journalists and researchers can enhance digital resilience, ensuring the ongoing integrity of public discourse. As the fight against disinformation intensifies, it is imperative that the media and fact-checking communities work together to counter the growing threat of fake online campaigns and preserve the truth in the digital sphere.

## Further readings


- Keller, T., Graham, T., Angus, D., Bruns, A., Nijmeijer, R., Nielbo, K. L., … Veiga de Oliveira, V. (2020). 'COORDINATED INAUTHENTIC BEHAVIOUR' AND OTHER ONLINE INFLUENCE OPERATIONS IN SOCIAL MEDIA SPACES. AoIR Selected Papers of Internet Research, 2020. https://doi.org/10.5210/spir.v2020i0.11132
- Michele Mazza, Guglielmo Cola, Maurizio Tesconi. "Ready-to-(ab)use: From fake account trafficking to coordinated inauthentic behavior on Twitter?" Online Social Networks and Media (Volume 31, September 2022, 100224). https://doi.org/10.1016/j.osnem.2022.100224

Chapter II

# Guidelines for Verifying Gender-based Information

## Introduction

A common misconception about gendered disinformation is the belief that it is solely directed at women, aiming to undermine them or reinforce restrictive and misogynistic stereotypes about their role in society. While women are frequently targeted by such campaigns, gendered disinformation is not exclusive to them. Men, too, have often been victims of gender-based falsehoods, with narratives that seek to manipulate perceptions of masculinity, distort their roles in public life, or use gendered narratives to serve broader political, social, or economic objectives.

The Internet Governance Forum (IGF) defines gendered disinformation as false or misleading information created to "attack or undermine people based on their gender or weaponize gendered narratives for political, social, or economic objectives." This broad definition highlights the role of disinformation in shaping societal perceptions and reinforces the idea that gendered narratives are often instrumentalized for strategic gains. Such disinformation may take the form of fabricated stories, manipulated images, or coordinated campaigns aimed at discrediting individuals based on gendered stereotypes.

Gendered disinformation does not exist in isolation. It often intersects with technology-facilitated gender-based violence (TFGBV)—a form of harm that occurs through digital and technological platforms, targeting individuals specifically based on their gender. TFGBV can include a wide range of online abuses, such as harassment, stalking, doxxing (the unauthorized release of private information), deepfake technology used for defamation, and the spread of false narratives aimed at silencing individuals.

The consequences of such digital violence are profound, leading to reputational harm, psychological distress, and, in extreme cases, threats to personal safety. Women in politics, journalism, and activism often face orchestrated disinformation campaigns designed to push them out of public discourse. Meanwhile, men may be targeted through

attacks that question their masculinity or falsely link them to scandals to diminish their credibility.

Recognizing the full scope of gendered disinformation is critical to developing effective responses. It is not simply an issue of protecting women from misogynistic attacks but addressing the systematic use of gendered narratives to manipulate public opinion, discredit individuals, and advance particular agendas. By increasing awareness, implementing stronger digital protections, and holding perpetrators accountable, societies can work toward mitigating the harmful effects of gendered disinformation on all individuals, regardless of their gender.

## Forms of Disinformation Associated with Gender-Based Violence

Gendered disinformation plays a significant role in reinforcing gender-based violence by spreading misleading or harmful narratives that specifically target individuals based on their gender. These tactics are not random; they are often strategically coordinated efforts aimed at silencing, discrediting, or marginalizing individuals, particularly women and gender minorities, in public discourse. The following are key forms of gendered disinformation, accompanied by real-world examples that illustrate their impact.

**1. The Creation or Distribution of Misleading or Harmful Information or Images Based on Gender**

One of the most prevalent forms of gendered disinformation involves the fabrication or manipulation of information—whether through text, images, or video content—to attack individuals based on their gender. These campaigns often seek to tarnish reputations, erode public sympathy, or justify discriminatory treatment.

A notable example is the report by Arabi Facts Hub titled "Smear Campaign against the Arish Student Led by Accounts Affiliated with the Egyptian Regime." The report investigated a coordinated campaign designed to defame a young Egyptian girl who tragically passed away earlier this year. Instead of allowing space for public mourning, disinformation networks worked systematically to distort the narrative surrounding her death, framing her in ways that would diminish public sympathy and shift the focus away from systemic issues related to her case. This kind of targeted disinformation serves as a tool for reputational damage, often leading to secondary victimization of individuals who are already vulnerable.

**2. The Use of Stereotypes and Gender-Based Profiling**

Another widespread tactic in gendered disinformation is the reinforcement of harmful stereotypes. By framing individuals in ways that align with deeply ingrained societal biases, these campaigns manipulate public perception and limit the roles that certain groups—especially women—are seen as capable of fulfilling.

A clear example of this is found in the report by the Yemeni platform "Saddaq" (Believe) titled "A Doctored Image of Judge Sabah Al Alwani Returning from Travel." The report examined a case in which photoshopped images were used to falsely accuse a Yemeni judge of undergoing cosmetic surgeries at the state's expense. The campaign leveraged misogynistic narratives that suggest women in professional roles are more interested in appearance and luxury than in their work. By focusing on superficial and gendered tropes, the orchestrators of such disinformation sought to undermine the credibility of a highly qualified female judge—a clear attempt to dissuade women from seeking leadership positions.

**3. The Use of Hate Speech and Its Evolution into Incitement and Hostility Through Disinformation**

Gendered disinformation is closely linked to hate speech, as false narratives often serve as a catalyst for targeted harassment, threats, and even physical violence. Hate-driven disinformation campaigns dehumanize individuals, particularly women and gender minorities, making them more vulnerable to real-world harm.

A compelling case study comes from Heya Tatahaqaq (She Checks), in their report titled "Disinformation Fuels Hate Speech Against Algerian Athlete Imane Khelif." The report analyzed how misinformation campaigns deliberately spread false claims about Khelif's biological sex. The disinformation suggested that she was not a cisgender woman but a transgender athlete, a false claim intended to undermine her achievements and delegitimize her participation in professional sports. Such campaigns not only target women in male-dominated spaces but also weaponize public sentiment against gender minorities, fostering hostility and discrimination within broader society.

**4. Targeting Women with Intersecting Identities (Race, Religion, or Other Social Categories)**

Women who belong to marginalized groups—including racial, ethnic, and religious minorities—often face compounded discrimination when gendered disinformation is used against them. These campaigns strategically exploit existing prejudices, making individuals more susceptible to public vilification.

One of the most egregious examples of this phenomenon can be seen in the treatment of Algerian Olympic boxer, Imane Khelif. The coordinated misinformation campaign against her not only targeted her gender but also played into stereotypes associated with her identity as an Arab and a Muslim woman. By questioning her biological sex, the campaign attempted to portray her as an "outsider" or someone who did not belong in competitive sports, thereby reinforcing multiple layers of discrimination. This underscores how gendered disinformation is often intersectional, affecting individuals in ways that extend beyond gender alone.

**5. Undermining Capabilities Solely Based on Gender**

A core strategy of gendered disinformation is challenging the competence and credibility of individuals solely based on their gender. This tactic is particularly evident in politics and governance, where women in leadership positions are systematically subjected to disinformation campaigns designed to dissuade them from seeking or maintaining power.

A striking example is highlighted in Arabi Facts Hub's report titled "Not Cut for It: Female Presidential Candidates in Algeria Face an Electronic Discrimination Campaign." The report exposed a large-scale digital smear campaign targeting female candidates in the 2024 Algerian presidential elections. The campaign deliberately discredited the candidates, not by critiquing their policies or leadership skills, but by promoting the false narrative that women are inherently unfit to lead. Misinformation networks circulated claims that leadership positions are naturally reserved for men, implying that women, regardless of their qualifications, are less capable or undeserving. This systematic use of disinformation perpetuates gender-based discrimination and discourages women from entering or staying in political spaces.

## Methodology for detecting gender-based disinformation

Ruwayda Al Arabi, founder of the "Heya Tatahaqaq" platform (She Checks), which specializes in fact-checking, emphasizes the importance of thoroughly analyzing the content of a message to determine whether it specifically targets women. This involves assessing whether the message is directed at female figures or addresses issues directly related to gender, such as gender rights. Disinformation targeting women often takes the form of misleading narratives, exaggerated claims, or outright falsehoods designed to undermine their credibility, reinforce harmful stereotypes, or discourage their participation in public life. By carefully dissecting the language, tone, and framing of such messages,

fact-checkers can identify underlying biases and hidden agendas that seek to discredit women based on gender rather than factual merit.

Al Arabi also stresses the need to go beyond surface-level content analysis and examine the broader context in which the message is being spread. Fact-checkers should investigate the source of the information, the possible motivations behind it, and whether similar patterns of gendered disinformation have emerged from the same actors in the past. Understanding whether a message is part of a coordinated effort to smear or intimidate women can provide crucial insights into the mechanisms of gender-based disinformation. Additionally, recognizing how such narratives align with existing gender biases can help expose systematic attempts to manipulate public perception. By adopting a methodical approach that includes verifying sources, tracking dissemination patterns, and debunking false claims, fact-checkers can play a vital role in countering gendered disinformation and ensuring a more accurate representation of women in the media and public discourse.

## What skills are needed to uncover gendered disinformation campaigns?

**First**, mastering general fact-checking skills is the foundation for uncovering gendered disinformation campaigns. This involves developing the ability to track coordinated campaigns that rely on inauthentic activity, such as bot networks, fake accounts, and manipulated engagement metrics. Fact-checkers must be proficient in identifying the origins of misleading narratives, tracing the amplification of false information, and recognizing patterns that suggest a deliberate attempt to deceive. Additionally, they should be well-versed in using digital verification tools to assess the authenticity of various types of content, including written posts, images, and videos. The rise of AI-generated content and deepfake technology has made verification even more challenging, requiring fact-checkers to stay updated on the latest advancements in synthetic media detection. By combining investigative techniques with technological tools, they can effectively analyze and dismantle gendered disinformation campaigns that attempt to distort public perception and manipulate online discourse.

**Second**, understanding the concept of hate speech in relation to disinformation is another crucial skill for uncovering gendered disinformation campaigns. Hate speech, when intertwined with disinformation, is often used to degrade, intimidate, or silence women and gender minorities by spreading false narratives designed to incite hostility. Fact-checkers

need to develop the ability to distinguish between genuine criticism and orchestrated disinformation that exploits hate speech to achieve political or social objectives. This requires familiarity with the language patterns and framing techniques commonly used in gendered attacks, as well as knowledge of the legal and ethical implications of online harassment. Additionally, analyzing the impact of hate speech-driven disinformation involves monitoring how it spreads, who is targeted, and the potential consequences for the individuals involved. By systematically identifying and debunking gender-based hate speech within misinformation campaigns, fact-checkers can help prevent the further marginalization of vulnerable groups and ensure that online spaces remain accountable and inclusive.

**Third**, utilizing open-source databases related to women and gender minorities is essential for exposing the realities behind gendered disinformation and its broader social implications. These databases provide valuable insights into the political representation of women, their economic conditions, and the systemic biases that contribute to misinformation. Fact-checkers can use this data to cross-reference claims made in disinformation campaigns, revealing discrepancies between false narratives and actual statistics. Additionally, understanding the policies of social media platforms regarding gender-based misinformation is critical, as it allows fact-checkers to assess whether tech companies are effectively addressing gendered disinformation or inadvertently enabling its spread. Furthermore, familiarity with local and international laws and agreements related to gender-based violence helps fact-checkers place gendered disinformation within a legal and human rights framework. By leveraging open-source data and legal knowledge, fact-checkers can provide a more comprehensive and factual perspective on the challenges faced by women and gender minorities in the digital space.

## Data sources to reference when preparing gender-related reports

**1. Academic research:**

- Google Scholar is an invaluable tool for conducting academic research across a wide range of disciplines. It offers users the ability to perform advanced searches by specifying geographic and temporal ranges, as well as including or excluding specific keywords to refine the results. This advanced searching capability makes it particularly useful for narrowing down results and finding research relevant to specific regions or time periods. Google Scholar also provides citation methods and

documentation for sources, which is essential for researchers looking to properly attribute and track academic contributions. The platform indexes a vast array of academic articles, journals, theses, books, conference papers, and patents, allowing researchers to access scholarly material from across the globe. Furthermore, Google Scholar's citation tracking feature allows researchers to gauge the impact of a paper by showing how many times it has been cited by others, which can help in identifying influential research in a particular field. With its user-friendly interface and free access, Google Scholar is a go-to resource for both novice and experienced researchers.

- ResearchGate is another excellent platform for academic research, particularly for accessing high-quality studies based on sound methodology. ResearchGate is a social networking site designed for researchers to share papers, ask and answer questions, and find collaborators. It is free to use, which makes it an accessible option for academics from various fields. One of the key features of ResearchGate is that it allows researchers to upload their papers and make them available for others to read, often including access to full-text versions, as opposed to just abstracts. While some studies may require payment for full access, many researchers upload their own publications or preprints, making it easier to obtain important research without paywalls. ResearchGate also facilitates direct communication with authors, enabling users to request copies of papers or ask for clarifications about their work. This platform is particularly useful for staying up-to-date on recent developments in your area of interest, as researchers frequently share their latest publications, presentations, and findings.

- JSTOR is another widely recognized and valuable database for academic research. Specializing in humanities, social sciences, and arts, JSTOR offers access to a vast collection of scholarly articles, books, and primary sources. The database is subscription-based, but many institutions, such as universities and libraries, provide access to their students and patrons. JSTOR's advanced search functionality allows users to narrow their results by disciplines, publication types, and date ranges, making it highly useful for finding authoritative and peer-reviewed academic work in the relevant field. It also provides tools for citation management and organizing references, which streamlines the research process. JSTOR is widely respected for its robust collection of back issues of academic journals, making it an essential tool for anyone conducting research in the humanities and social sciences, especially for exploring historical trends and perspectives.

- Scopus is another essential academic database known for its comprehensive coverage of scientific, technical, and medical research. Unlike other databases, Scopus not only provides access to journals, conference papers, and patents, but also offers citation and abstract data that helps researchers assess the impact and relevance of particular articles. One of Scopus' standout features is its citation analysis tools, which allow researchers to track how frequently a particular article or author has been cited. This can be especially useful for identifying key research trends, influential authors, and emerging fields of study. Scopus also offers tools for analyzing publication patterns, such as journal rankings and metrics for individual articles. It is a subscription-based service, but access is often available through academic institutions. Scopus is particularly valuable for researchers in the fields of science, technology, and medicine due to its in-depth coverage of these disciplines and its integration of citation data.

**2. Specialized Databases:**

- The interactive map published by the Arab Women Organization serves as a valuable tool for analyzing the role of Arab women in political life. This map provides users with the ability to search by country to determine the percentage of women holding ministerial positions and their representation in parliament. By offering an easily accessible visual representation of women's political participation, the map helps researchers, policymakers, and advocates track progress, identify disparities, and push for greater gender inclusivity in governance. The platform is particularly useful for comparative analysis across the region, allowing users to examine the political status of women in different Arab countries and assess the impact of policies designed to enhance female political participation.

- The Arab Women Organization's database, "Bright Marks in Women's History," is an extensive archive dedicated to documenting the contributions of pioneering women in the Arab world. It offers users the ability to conduct searches by year, keywords, or specific names of influential female figures, making it a crucial resource for scholars, journalists, and researchers interested in women's history. The database sheds light on the achievements of Arab women in various fields, from politics and activism to science and the arts, ensuring that their legacies are preserved and widely recognized. By compiling this wealth of historical information, the database not only celebrates the accomplishments of these women but also provides inspiration for future generations striving for gender equality.

- United Nations databases are instrumental in facilitating research on cross-sectional issues involving women, enabling users to conduct keyword searches related to development, security, peace, economic status, conflicts, wars, and more. These databases provide researchers with access to extensive global data, allowing for in-depth analysis of how gender dynamics intersect with broader socioeconomic and political trends. By integrating gender-specific statistics into discussions on peace-building, economic growth, and crisis response, these databases help policymakers and advocates develop more inclusive strategies that address the unique challenges faced by women worldwide. The ability to cross-reference different UN datasets also strengthens efforts to track progress on international gender-related commitments, such as the Sustainable Development Goals.

- The World Bank's Gender Data Portal offers a comprehensive collection of gender-related statistics that shed light on critical issues such as workforce participation, political representation, and economic indicators affecting women in various countries. This portal serves as an essential tool for policymakers, researchers, and gender advocates seeking data-driven insights into the economic conditions of women. By providing access to reliable and up-to-date statistics, the portal helps identify trends, measure progress in gender equality, and assess the effectiveness of policies aimed at closing gender gaps in employment, wages, and leadership roles. Additionally, the data allows for comparisons across different regions, contributing to a more nuanced understanding of the structural barriers women face in economic and political spheres.

- Arab Barometer is a research network that has been conducting public opinion surveys in the Arab world since 2006, providing crucial insights into social, political, and economic issues, including gender-related topics. The platform offers survey data and reports on key gender issues, such as women’s political participation and societal perceptions of gender equality. By systematically collecting public opinion data, Arab Barometer helps track shifts in attitudes toward gender roles, discrimination, and women’s rights across different Arab societies. This data is particularly valuable for understanding how cultural and political changes influence gender norms and for designing interventions that promote gender inclusivity in public life.

- The UN Women - Arab States website is a critical resource offering a wealth of reports and data on gender equality, women's empowerment, and violence against

women in various Arab countries. It provides insights into regional efforts to enhance women's rights, highlighting progress, challenges, and policy recommendations. The platform also serves as a hub for best practices in promoting gender equality, featuring case studies, policy briefs, and research findings that inform advocacy efforts and government strategies. By offering a regional perspective on gender-related issues, the UN Women - Arab States website supports policymakers, activists, and organizations working toward meaningful change in gender policies and women's rights across the Arab world.

- The World Economic Forum's Gender Gap Report presents detailed data and rankings on gender equality across various indicators, including political empowerment, economic participation, and education, with a particular focus on the Middle East and North Africa. This report provides a comparative analysis of gender disparities globally, allowing stakeholders to identify areas where progress has been made and where challenges persist. By offering a structured framework for assessing gender gaps, the report plays a crucial role in shaping policy discussions and guiding efforts to address structural inequalities. The inclusion of regional breakdowns helps shed light on the specific barriers women face in different cultural and economic contexts, making it a valuable tool for governments, organizations, and advocates committed to gender equity.

- The Women, Peace, and Security Index evaluates the status of women on an annual basis, measuring three key dimensions: inclusion, justice, and security. It further breaks down these dimensions into subcategories such as education, labor market participation, gender-based violence, and empowerment. This index provides a comprehensive overview of the challenges and progress in achieving gender equality, particularly in fragile and conflict-affected regions. By tracking these indicators over time, the index helps identify trends, highlight persistent gaps, and inform policy recommendations aimed at improving women's safety and opportunities. The data-driven insights from this index serve as a critical tool for organizations and policymakers working to strengthen gender-sensitive policies and address the intersection of gender and security in global development initiatives.

# Tips for writing reports on fact-checking gendered disinformation campaigns

- Understanding the social and cultural context is critical when preparing reports on gender-based misinformation. According to the She-Checks platform, it is essential to first gain a deep understanding of the cultural and social dynamics in the region or community where the misinformation is circulating. This knowledge can provide important insights into why certain misinformation may spread more readily in one country or culture than in another. It can also help to identify the factors that contribute to the dissemination of gender-based misinformation, whether they are related to political, social, or cultural issues. By understanding the local context, the reporter can clarify and explain the nuances to the reader, offering more comprehensive, accurate, and sensitive reporting. Contextual understanding not only helps in verifying misinformation but also ensures that the report considers the broader societal and cultural implications, making the work more informative and relevant to the audience.

- Using gender-sensitive language is crucial for ensuring that reports on gender-based misinformation are respectful and accurate. Language plays a significant role in shaping perceptions, and using terms preferred by the communities or groups being discussed is essential for building empathy and avoiding stigmatization. For example, referring to “transgender individuals” instead of “transsexuals” aligns with the language used by the community and avoids outdated or offensive terminology. Similarly, framing issues like “child or minor marriage” emphasizes the coercive nature of the violation, rather than normalizing harmful practices. Avoiding victim-blaming language is also vital—terms such as “familial rape” should be used instead of “familial fornication” to avoid stigmatizing the survivor. In addition to this, it's important to refrain from mentioning details that might cause social stigma or reduce empathy, such as references to women's clothing in harassment or assault cases. Gender-sensitive language helps build a respectful and responsible narrative that acknowledges the dignity of those involved and promotes a more just understanding of their experiences.

- Prioritizing the safety of participants and sources is a fundamental principle in ethical journalism, especially when reporting on sensitive topics like gender-based violence or misinformation. This involves taking every measure possible to protect the identities and well-being of those featured in a report. For example, anonymizing

names, concealing images, and altering voices are necessary precautions to prevent exposing participants to potential harm or danger. In some cases, participants may be vulnerable to retaliation, violence, or legal consequences, so it is essential to avoid disclosing any information that could compromise their safety. The Dart Center's guidelines emphasize respecting the rights of victims and survivors to refuse documentation or recordings. Journalists should never pressure individuals to share their experiences or go against their wishes. Additionally, asking questions that implicitly blame survivors or place undue responsibility on them—such as questioning why they posted a controversial statement or why they didn't speak out earlier—should be avoided. The primary concern must always be the safety and well-being of those involved in the reporting process.

- Gender representation balance is essential in creating accurate and fair reports on gender-based misinformation. It's not enough to simply report on the experiences of women and gender minorities as victims of disinformation; it is equally important to highlight them as active participants and experts in the field. Ensuring that women and gender minorities are adequately represented in all aspects of the story—from sharing their personal experiences to offering their insights as professionals—is vital for a balanced portrayal. This approach challenges the traditional narrative that often depicts women only as victims or passive subjects, and instead acknowledges their agency and expertise. Moreover, giving women and gender minorities a platform to speak directly in the report, with their voices and perspectives reflected, is crucial for combating the marginalization of these groups in media. This balanced representation also serves to normalize their roles in various spheres of influence, such as politics, media, and academia, and ensures that their voices are not only heard but valued.

- Highlighting the actual impact of online misinformation, hate speech, and incitement is a vital aspect of reporting on gender-based issues. Rather than simply describing the spread of misinformation, it is important to investigate and illustrate the real-world consequences these actions have on individuals, especially those targeted based on their gender. This could include examining how misinformation results in harm to individuals or specific gender groups, whether through online harassment, physical violence, or reputational damage. Moreover, the report should explore the broader societal impacts of misinformation, such as how it might influence public attitudes or policies. For example, disinformation campaigns targeting women might reinforce harmful stereotypes, hinder gender equality, or lead to legal and political actions that negatively affect women's rights.

Understanding the tangible effects of misinformation helps readers grasp the full scope of its consequences and encourages accountability for those responsible for spreading false narratives.

- Legal and legislative knowledge is crucial when addressing gender-based misinformation in a comprehensive and informed manner. Familiarity with both international charters and local laws relating to gender equality and the protection of gender minorities is necessary for assessing the legality of actions and behaviors involved in misinformation campaigns. Journalists and researchers must understand the legal frameworks that address issues such as gender-based violence, defamation, and the rights of individuals to live free from discrimination. International agreements like the Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW) and local constitutional provisions often provide guidelines that can be used to evaluate whether certain actions or policies are in violation of women's rights. By grounding reports in legal and legislative knowledge, reporters can provide a more accurate analysis of the legitimacy of the actions under investigation and hold perpetrators accountable for any violations of rights, whether local or global.

- The responsibility of social media and digital platforms in combating gender-based incitement and misinformation is an increasingly important topic in the digital age. Social media platforms have immense influence over the spread of information and misinformation, and their policies on hate speech, misinformation, and gender-based incitement need to be critically evaluated. It is essential to analyze whether these platforms are effectively implementing their policies and adhering to their stated guidelines on issues like harassment and harmful content. Journalists should also assess any gaps or challenges in the platforms' policies, such as the lack of enforcement or the failure to remove harmful content in a timely manner. This can involve examining case studies where platforms have failed to take appropriate action or where their policies have been insufficient in curbing the spread of gender-based disinformation. Evaluating the effectiveness of platform policies and practices not only informs the public but also holds digital platforms accountable for their role in exacerbating or mitigating gendered misinformation.

## Further Readings

- Dean Spade, Documenting Gender, 59 *HASTINGS L.J.* 731 (2008). https://digitalcommons.law.seattleu.edu/faculty/349
- Stacy Larson, *Intersexuality and Gender Verification Tests: The Need to Assure Human Rights and Privacy*, 23 Pace Int'l L. Rev. 215 (2011) DOI: https://doi.org/10.58948/2331-3536.1316
- Ashley, Florence, Recommendations for Institutional and Governmental Management of Gender Information (January 18, 2019). New York University Review of Law & Social Change, Vol. 44, No. 4, 2021. Available at SSRN: https://ssrn.com/abstract=3398394 or http://dx.doi.org/10.2139/ssrn.3398394

**Chapter III**

# Guidelines for Verifying Content That Incites Against Refugees

## Introduction

There are over 43 million refugees worldwide, many of whom are classified as highly vulnerable groups due to their exposure to incitement campaigns and hate speech. These campaigns, which often contain misleading or entirely false information, are particularly dangerous for refugees, as they amplify negative stereotypes, fuel discrimination, and contribute to societal division. The situation has become even more pressing with the rise of extremist movements and nationalist trends, which frequently target refugees as scapegoats for broader societal issues. These campaigns exploit existing fears and anxieties in the population, manipulating public perception and increasing the risk of further marginalization and harm to refugees and migrants. The spread of misinformation about refugees can have far-reaching consequences, such as hindering their access to essential services, increasing social isolation, and deepening their vulnerability to violence.

Digital tools have played a significant role in the ease with which misleading content can be created and distributed. In particular, the use of social media platforms has made it easier than ever to target refugees and migrants with harmful narratives. The ability to quickly generate and disseminate false information has made it more difficult to counter misinformation and protect these vulnerable populations. Social media provides platforms for users to share content rapidly, often without any effective oversight or moderation. These platforms also create digital spaces that can be exploited for covert and coordinated campaigns aimed at discrediting refugees. The anonymity and reach offered by digital spaces allow bad actors to spread harmful content with relative impunity, making it a challenge for fact-checkers and journalists to track and debunk these disinformation campaigns in real time.

Minorities, including refugees, account for a disproportionate share of victims of online hate speech. According to a report from the 13th Forum on Minority Issues, three-quarters of online hate speech victims globally belong to minority groups, with women facing even

higher levels of targeting. Refugees, who are often already marginalized and face systemic discrimination, are particularly vulnerable to online hate speech and digital harassment. The intersectionality of gender, race, and refugee status can exacerbate the impact of these harmful campaigns. Women refugees are particularly at risk of being targeted with gendered disinformation, which further isolates them and perpetuates harmful stereotypes. As online platforms become more central to public discourse, understanding the scale of this issue and the mechanisms through which refugees are targeted is essential for mitigating the harms caused by digital hate speech.

This article outlines effective methods for detecting misinformation aimed at refugees, emphasizing the need for a careful and thorough approach to verifying the accuracy of information circulating online. The task of identifying false or misleading content targeting refugees requires a combination of fact-checking skills, digital literacy, and contextual awareness. Fact-checkers must utilize a range of tools designed to verify the authenticity of content, such as image verification tools, social media monitoring platforms, and databases that track disinformation campaigns. These tools help identify when misinformation is being deliberately spread and allow fact-checkers to trace the origins of false narratives. Additionally, the article offers key verification tools that can be used to confirm the authenticity of images, videos, and written content. For example, reverse image searches and metadata analysis can uncover whether a photo or video has been manipulated, while cross-referencing claims with reputable sources helps determine if the information is accurate. The combination of these techniques is crucial for building a robust defense against disinformation.

The article also provides valuable tips for fact-checkers based on extensive experience analyzing incitement campaigns. One of the key tips is to always consider the context in which the misinformation is being spread. Fact-checkers should investigate not only the claims being made but also the broader social, political, and historical factors that may be fueling the spread of false information. Understanding the motives behind the disinformation, such as political agendas or efforts to stoke fear, can help fact-checkers identify patterns in the spread of false claims and develop more effective strategies to counter them. Another tip is to be vigilant in tracking coordinated campaigns, which often involve the use of fake accounts, bots, and other tactics to amplify misleading messages. By identifying these patterns early on, fact-checkers can prevent the spread of harmful content and provide timely corrections. Lastly, building partnerships with local organizations and experts who have in-depth knowledge of refugee issues is crucial for understanding the nuances of misinformation campaigns and providing more accurate and relevant fact-checking.

In conclusion, the issue of misinformation targeting refugees is complex and requires a multifaceted approach to effectively combat it. Fact-checkers must be equipped with the right tools, knowledge, and strategies to navigate the challenges posed by digital disinformation campaigns. By understanding the social and political context, utilizing advanced verification methods, and collaborating with local experts, fact-checkers can play a key role in safeguarding refugees from the harm caused by misleading and harmful online content. As the digital landscape continues to evolve, it is essential to stay ahead of emerging disinformation trends and work collectively to protect vulnerable groups from the dangers of online hate speech and misinformation.

## Hate Speech Against Refugees and Migrants

First, fact-checkers must be well-versed in understanding the precise terminology used in hate speech. Not all insults or derogatory terms toward an individual or a group qualify as hate speech. For language to be categorized as hate speech, it must directly target an individual's or a group's identity based on inherent characteristics, such as their religion, ethnicity, nationality, race, color, descent, gender, sexual orientation, or any other identity factor. The United Nations defines hate speech as any form of communication—whether in speech, writing, or behavior—that attacks or uses discriminatory language with the intent to degrade or dehumanize a person or group based on these attributes. Hate speech is often dangerous because it can incite violence, hatred, or discrimination, and in some cases, may threaten the social peace of communities. It is crucial for fact-checkers to recognize that while offensive or harmful language can be prevalent online, not all such language rises to the level of hate speech unless it directly targets identity. Understanding this distinction ensures that fact-checkers do not mislabel harmless criticism or general negative speech as hate speech but also helps them identify and address the more dangerous, discriminatory rhetoric that can contribute to societal harm.

Furthermore, it is important to differentiate between the terms "migrant" and "refugee," as confusion between the two can fuel misinformation and increase tensions. The term "migrant" refers to individuals who voluntarily move from one country to another, often in search of better economic opportunities, safety, or a different lifestyle. In contrast, "refugees" are forced to flee their home countries due to violence, persecution, or the threat of harm, and they are protected under international law. This distinction is critical when addressing issues related to migration and refugee crises because conflating these two terms can distort the reality of the situation and contribute to misunderstanding and the spread of hate speech. The exaggeration of refugee numbers by including migrants in

statistics can amplify fears and resentment in host countries, particularly during times of economic hardship. For example, the Matsad'sh platform in Egypt debunked a widely circulated claim that Egypt was hosting nine million refugees, clarifying that this figure included both refugees and migrants. The actual number of refugees in Egypt is approximately 575,000. By making these distinctions clear, fact-checkers can reduce the misinformation that exacerbates tensions between refugees and local populations, ensuring more accurate and informed public discourse on these issues.

Second, the concept of "size of the monster" plays a critical role in assessing the impact and prevalence of misinformation campaigns. In journalism, this term refers to understanding the scope of a problem—how widespread it is and how many people are affected by it. Fact-checkers must evaluate whether a particular incident of disinformation or incitement is an isolated case or part of a larger, coordinated effort. In cases where misinformation is part of a broader campaign, fact-checkers must investigate the scale of the operation, including the use of digital tools to amplify the message. Coordinated disinformation campaigns are often executed by organized groups, sometimes referred to as "electronic flies" (trolls) or "bots." These automated accounts are programmed to spread misinformation at an accelerated pace, creating the illusion of widespread sentiment or consensus. These campaigns can be especially damaging when they target vulnerable groups like refugees, spreading hate speech and fueling social division. Fact-checkers must be able to discern whether a single, isolated post is part of a larger pattern of disinformation or whether it is merely an isolated incident with minimal impact. For example, Arabi Facts Hub, in one of its reports, tracked a coordinated hate campaign against refugees in Egypt, in which dehumanizing language and expressions were used to incite hostility toward these displaced populations. This campaign was not just a single instance but part of a broader trend of anti-refugee sentiment that was being systematically amplified online through coordinated efforts. By identifying such patterns early, fact-checkers can help prevent the spread of hateful rhetoric and provide timely corrections that limit its impact.

Third, fact-checkers must focus on uncovering misinformation that is embedded within hate campaigns. Hate speech often goes beyond harmful language, leveraging digital tools to spread disinformation and manipulate public opinion. One of the most powerful tools in modern disinformation campaigns is the use of image and video editing software to create fake or misleading content. These digital tools can be used to fabricate evidence, alter the context of real-world events, or create entirely false narratives that are difficult for the average person to identify as fraudulent. For instance, the platform Misbar recently uncovered a series of misleading photos shared on social media, which were purported to show the Algerian government expelling large numbers of migrants during the 2022 Arab

League Summit. The images, which were doctored to exaggerate the scale of the expulsion, were used to create an atmosphere of fear and mistrust regarding migrants. By analyzing the metadata and conducting reverse image searches, Misbar was able to reveal that the photos had been altered and were not an accurate representation of the event. This is an example of how hate campaigns targeting vulnerable populations like refugees often use fake visuals to amplify their message and create a sense of urgency or danger that is not supported by the facts. Fact-checkers must be vigilant in verifying the authenticity of such images, videos, and other forms of media, utilizing a range of digital tools to ensure that false narratives do not go unchecked. By exposing the manipulation behind these materials, fact-checkers play a crucial role in dismantling hate campaigns and ensuring that the public receives accurate and truthful information.

In addition to the above, fact-checkers should be aware of the role that echo chambers and algorithmic amplification play in the spread of hate speech and misinformation. Social media platforms, by design, amplify content that generates engagement, and this often includes sensationalist or divisive content. When hate speech is amplified by algorithms, it can lead to the creation of online echo chambers where individuals are repeatedly exposed to the same false narratives, reinforcing their beliefs and further entrenching their biases. This creates a feedback loop that can make it more difficult to break through with accurate information. Fact-checkers must not only verify the facts but also understand the broader dynamics of social media platforms and how their algorithms contribute to the spread of misinformation. By raising awareness of these issues and working to expose the mechanisms behind the amplification of false content, fact-checkers can help mitigate the damage caused by hate speech and contribute to a healthier, more informed digital environment.

## Tools for Detecting Misleading Campaigns Against Refugees

**Hashtag Analysis Tools on Social Media Platforms**

Hashtag analysis tools provide critical insights into the online discourse surrounding specific topics or movements by tracking engagement with hashtags over time. These tools help identify not only the volume of posts associated with a particular hashtag but also geographic locations with the highest levels of activity and the most frequently used words within the posts—forming a "word cloud." This feature is especially valuable for analyzing how particular terms or hashtags evolve, and whether they are being used to spread hate

speech or incite hostility. In the context of refugees and migration, hashtag analysis can help fact-checkers identify the emergence of potentially harmful narratives or misinformation related to these vulnerable groups. For instance, identifying patterns of harmful language around a hashtag can reveal coordinated hate speech campaigns targeting refugees, allowing fact-checkers to intervene early.

Social Media Activity Analysis Tools includes:

**1. Meltwater**

Meltwater is a comprehensive media monitoring and social media analytics platform that helps track and analyze content across a wide range of sources. It allows users to monitor trends, analyze engagement, and track hashtags in real-time. It offers powerful tools to perform sentiment analysis, track mentions of specific keywords or brands, and generate detailed reports on social media conversations. In the context of refugee-related misinformation, Meltwater can be used to monitor the evolution of hashtags associated with refugee issues and track the sentiment around these conversations to identify if hate speech or incitement is occurring.

**2. CrowdTangle**

CrowdTangle is a social media analytics and monitoring tool that helps track how content spreads across social media platforms, particularly Facebook, Instagram, and Reddit. By providing insights into the performance of specific content—such as engagement rates and interactions—CrowdTangle can be used to analyze the virality of posts and track the public sentiment surrounding particular topics. For fact-checkers, CrowdTangle is a valuable tool for monitoring the spread of refugee-related misinformation and hate speech, helping them identify popular posts that may need further verification or correction.

**3. Social Bee**

Social Bee is a social media management tool that includes features for monitoring social media content, scheduling posts, and analyzing engagement across various platforms. It provides users with insights into audience interactions and helps track keywords, hashtags, and trends. Social Bee can be used to track hashtag performance in real time, especially those related to sensitive issues like refugee rights and migration. By using this tool, fact-checkers can observe shifts in discourse and identify emerging hate speech campaigns, enabling timely interventions to prevent the spread of harmful content.

**Reverse Search for the Origin of Viral Clips and Images**

Reverse image and video search techniques are essential for verifying the authenticity of viral media, especially when it is used to manipulate narratives or spread misinformation.

These tools allow users to trace the origin of media, detect manipulations, and determine whether content has been taken out of context to support harmful or false claims. This process is critical for fact-checkers dealing with hate speech, as viral videos or images can often be fabricated or edited to incite hatred or perpetuate harmful stereotypes about refugees.

**1. Reverse Search through Search Engines**

Reverse image search through search engines like Google can help identify the first appearance of an image or video online. By uploading the media to search engines, fact-checkers can trace its origins, identify similar images, or locate instances where it has been previously debunked. Google's reverse search tool is one of the most widely used resources for media verification due to its broad database and easy-to-use interface.

**2. Duplichecker**

Duplichecker is an online tool that offers a reverse image search feature. It allows users to upload an image or input a URL to find similar images across the web. It can be used to verify whether an image has been manipulated or if it has been taken from a different context. This tool is particularly useful when verifying images shared in social media posts that claim to depict specific events, such as violence against refugees.

**3. TinEye**

TinEye is a reverse image search engine that specializes in identifying the origin of images on the internet. It provides advanced search algorithms that can detect where images have been used and whether they have been altered. TinEye's database is frequently updated, and it supports reverse searches using both uploaded images and URLs. For fact-checkers, TinEye is valuable for verifying whether an image that is being used to spread misinformation about refugees has been manipulated or misattributed.

**4. PimEyes**

PimEyes is a facial recognition tool that goes beyond traditional reverse image search by offering a more advanced method to identify people in images using facial recognition technology. It is particularly useful for verifying viral videos or images that feature individuals, allowing fact-checkers to trace whether the person depicted is related to the claim being made in the post. While PimEyes provides high accuracy in its facial recognition, it is important for fact-checkers to use it alongside other tools to cross-verify the origin and context of the media.

**Verification of Visual Media Metadata**

Verifying the metadata of images and videos is a critical part of the fact-checking process, as this data can provide crucial insights into the origin, time, and context of the media. Visual media metadata includes details such as the location where an image or video was captured, the device used for recording, and the date and time of the recording. This information can be invaluable for fact-checkers in determining whether a piece of media is being manipulated or misused in a disinformation campaign targeting refugees.

**1. Jimpl**

Jimpl is a tool that helps verify the metadata of images by analyzing the EXIF (Exchangeable Image File Format) data embedded in the files. EXIF data contains information such as the date, time, and location where an image was captured, as well as the type of camera or device used. Jimpl makes it easy for fact-checkers to access this data and verify whether it aligns with the claims made in a social media post. For instance, if an image purports to show an incident involving refugees but the EXIF data indicates it was taken in a different location or on a different date, it could be flagged as misleading or false.

**2. EXIFTool**

EXIFTool is a powerful software tool that reads, writes, and edits the metadata of a wide variety of file types, including images and videos. It provides detailed information about the media file, such as the location, device used, and other technical data that can help verify its authenticity. For fact-checkers, EXIFTool is an indispensable resource for determining whether visual media has been manipulated or taken out of context, as changes to the metadata can often reveal attempts to deceive or mislead audiences.

By utilizing these digital verification tools, fact-checkers can more effectively combat misinformation and hate speech targeting refugees. These tools allow for the thorough verification of media, tracking of online campaigns, and analysis of language and rhetoric used to incite hostility, ultimately contributing to more accurate and responsible reporting on the issue of refugees in the digital age.

## Who is Responsible for Protecting Refugees?

Accountability and oversight are fundamental elements of journalism, ensuring that those who spread misinformation and engage in harmful activities are held responsible. However, these crucial aspects are often not fully applied in fact-checking reports or are reserved for investigative journalism. When addressing the issue of hate speech and misinformation targeting refugees, it is important to recognize the wide-ranging

responsibility for their protection—both from governments and private sector actors like social media platforms.

The primary responsibility for the protection of refugees and asylum seekers lies with host countries, which are bound by international treaties and agreements. Under the 1951 Refugee Convention and various human rights treaties, countries have an obligation to provide refugees with safety and security. This includes not only ensuring physical safety from violence but also offering legal protection against discrimination, abuse, and exploitation. These treaties compel states to create a safe environment where refugees can live free from harassment or harm. However, in many cases, the implementation of these responsibilities is not always upheld. Some governments may fail to provide adequate protection, either through neglect, insufficient enforcement of laws, or even direct hostility toward refugees, which can exacerbate the dangers they face. Moreover, when hate speech against refugees becomes widespread, it can legitimize or encourage discriminatory policies and actions, further marginalizing already vulnerable populations.

In addition to governments, social media platforms also share a critical responsibility in combating hate speech. Digital platforms, by providing users with a space to express themselves and form communities, have an outsized influence on the way information spreads. They enable individuals and groups to reach vast audiences with minimal oversight, often without regard for the content they are sharing. The impact of this unchecked speech can be significant: online hate speech can quickly lead to real-world consequences, including physical violence, social exclusion, and discrimination against the targeted group—in this case, refugees. Sudan's Beam Reports highlighted that hate speech on platforms like Facebook and Twitter can foster radical virtual communities that exchange extremist ideas, which sometimes manifest in offline acts of violence or discrimination. These platforms, which are integral to modern communication, have a duty to actively monitor and manage the content that circulates on their sites, ensuring that it does not escalate into harmful actions.

For instance, Facebook's hate speech policies are designed to maintain a safe environment for users by removing content that violates their guidelines. This includes posts that promote violence or discrimination based on race, ethnicity, nationality, or other protected characteristics. If users post inciting content, Facebook typically issues warnings and may suspend or delete accounts in cases of repeated violations. Similarly, X (formerly known as Twitter) prohibits hate speech, stating that individuals may not directly attack others based on a range of protected categories, including race, ethnicity, and religion. These policies, on paper, are essential in preventing the spread of harmful rhetoric online, which could contribute to an atmosphere of hostility and violence in the real world.

However, the enforcement of these policies is inconsistent. A report by Arabi Facts Hub revealed that social media platforms, including Facebook and X, have not consistently removed posts inciting violence or discrimination against refugees, particularly in Egypt. Despite the platforms' established policies, these posts often remain visible for extended periods, allowing hate speech to spread and proliferate. This failure to uphold guidelines effectively results in an environment where harmful content continues to thrive. In some cases, this can even lead to a rise in hate speech over time, as the lack of accountability emboldens individuals to share more extreme views. These platforms are meant to act as gatekeepers to prevent harmful content from reaching a wide audience, but their failure to remove harmful material contributes to the normalization of discriminatory ideas and fuels the perpetuation of misinformation.

These coordinated hate campaigns that spread misinformation about refugees do not remain confined to the digital space. The effects can spill over into the physical world, where refugees may face increased hostility, discrimination, and even violence. The pervasive nature of online hate speech can create a climate in which refugees are seen as undeserving of protection, resources, or empathy, leading to more exclusionary and discriminatory policies. For example, in Egypt, inflammatory posts about refugees have contributed to rising anti-refugee sentiment, which has translated into public resentment and, in some cases, physical violence against refugees. As these campaigns gain traction on social media, the gap between digital incitement and real-world consequences grows, leading to further societal risks.

In light of these risks, journalists and fact-checkers have a crucial role in monitoring and addressing these harmful narratives. By focusing on both the digital and real-world implications of online hate speech, journalists can help bring awareness to the issues facing refugees. Fact-checkers play an essential part in debunking misinformation that fuels hostility toward these vulnerable groups, ensuring that accurate information is made available to the public. In addition, their work highlights the role that social media platforms play in either curbing or facilitating the spread of hateful content, holding these platforms accountable for their responsibility in protecting marginalized communities.

Moreover, the role of international actors and advocacy groups cannot be overstated. Global organizations, such as the UNHCR (United Nations High Commissioner for Refugees) and other humanitarian bodies, have an important responsibility to monitor trends in hate speech and advocate for stronger protections for refugees. Governments must also be held accountable for upholding international legal obligations and addressing the needs of refugees within their borders.

Protecting refugees from hate speech and misinformation requires a coordinated effort across multiple sectors. Host countries, social media platforms, fact-checkers, and the international community must all take responsibility for creating a safer environment for refugees—both online and offline. Journalists and media organizations are essential players in holding these actors accountable, ensuring that the spread of hate speech is combated and that refugees' rights and dignity are respected and protected.

## Further readings


- Kasapoğlu, Tayfun et al. “Unpacking algorithms as technologies of power: Syrian refugees and data experts on algorithmic governance.” Digital Geography and Society (2021): 2(6373):100016.
- Sefa Secen. (2024) Between Worlds: Ontological Security among Syrian Refugees in Germany and Turkey. Journal of Global Security Studies 9:4.

**Chapter IV**

# Ethical Guidelines for Using Facial Recognition Tools in Fact-Checking

## Introduction

The rapid technological advancements in artificial intelligence (AI) models have significantly transformed many fields, including journalism. AI tools now offer the capability to generate highly realistic images and videos, which, while remarkable in their potential for creative and educational uses, also present significant challenges for journalists and fact-checkers. One of the most prominent issues is the ease with which images and videos can be manipulated or fabricated, making it increasingly difficult to discern between authentic and altered content. The rise of deepfake technology, for example, allows individuals to create highly convincing videos that depict people saying or doing things they never actually did. These manipulated images and videos have the potential to mislead audiences, distort public opinion, and contribute to the spread of misinformation. Consequently, fact-checkers and journalists face a growing challenge of verifying the authenticity of digital media before it can be used in reports or published for wider consumption.

However, despite these challenges, AI itself offers a powerful tool for detecting media manipulation. Through advanced techniques such as facial recognition and image forensics, AI can help journalists identify and verify manipulated content. Facial recognition technology, in particular, has become an essential tool in investigative journalism, enabling reporters to trace the identities of individuals in images or videos, confirm their participation in certain events, and detect inconsistencies that may suggest manipulation. By analyzing unique facial features, such as the geometry of the eyes, nose, and mouth, these AI models can cross-reference images with databases of known individuals to determine whether the content has been altered or misrepresented. This capability is particularly useful in tracking individuals who may be involved in high-profile events, criminal investigations, or political reporting.

In fact-checking, facial recognition tools can be utilized to confirm the authenticity of images or videos circulating on social media platforms or news outlets. For example, if a

video is widely shared claiming to show a political leader making controversial statements, facial recognition software can help fact-checkers determine if the person in the video is indeed the stated individual. Additionally, these tools can be used to detect instances where a video or image has been fabricated by swapping faces or digitally altering features. With the increasing prevalence of user-generated content online, this kind of verification is crucial for maintaining the integrity of news reporting.

Beyond facial recognition, AI-driven image analysis tools offer other techniques for detecting manipulation. These tools can identify inconsistencies within images, such as unnatural lighting, pixelation around edited areas, or mismatched image quality that could suggest digital alteration. Image forensics tools like FotoForensics, InVID, and others can also provide valuable insights by analyzing image metadata, examining the pixel-level structure of images, and identifying digital fingerprints that reveal tampering. These tools are invaluable for journalists seeking to verify the authenticity of content that may be intentionally misleading or fabricated.

Despite the potential benefits, the use of AI in journalism, especially facial recognition, raises several important ethical and privacy concerns. One of the most prominent concerns is the potential misuse of facial recognition technology. While it can be a powerful tool for verification, it can also be used to violate individuals' privacy or invade personal spaces. For example, using facial recognition technology to track individuals without their consent or to monitor specific groups, such as political dissidents, can lead to surveillance and an erosion of personal freedoms. In some countries, governments have used facial recognition for purposes of mass surveillance, which has sparked debates about its impact on civil liberties.

Another issue is the potential for bias in AI models, which could lead to inaccuracies or unfair targeting. Research has shown that facial recognition systems can be less accurate when applied to individuals of certain ethnicities or genders, leading to false identifications or missed verifications. This bias can skew investigative reporting and lead to unreliable conclusions. As AI technology continues to evolve, developers must be vigilant in ensuring that these systems are trained on diverse datasets and that their applications adhere to ethical guidelines that prioritize fairness and accuracy.

Moreover, the increasing reliance on AI for content verification also presents a challenge in terms of transparency. AI models, particularly deep learning algorithms, often operate as “black boxes,” meaning that their decision-making processes are not always fully understood or explainable. This lack of transparency can be problematic, especially in high-stakes situations where accuracy is critical. Journalists and fact-checkers must be

cautious about over-relying on AI tools without understanding their limitations, and they should always apply human judgment to verify AI-generated conclusions.

Ethical challenges related to AI and image manipulation also include questions about consent and the potential harm caused by publishing manipulated content, even if it is exposed as fake. For instance, in some cases, revealing the manipulation of a video or image can cause harm to individuals who were not involved in the incident in question or were victims of digital manipulation. These individuals may face reputational damage or undue public attention that could have lasting consequences. Journalists must therefore be responsible in how they handle and report on manipulated content, ensuring that they protect the privacy and dignity of those affected.

AI-powered tools, especially facial recognition, offer significant potential for improving the accuracy and reliability of journalism, particularly in the verification of images and videos. They provide a crucial means of detecting manipulated content and ensuring that fact-checkers can accurately assess the authenticity of digital media. However, the use of these tools also comes with important ethical considerations, including issues related to privacy, bias, consent, and transparency. Journalists and fact-checkers must balance the benefits of AI with a thoughtful and responsible approach to its application, ensuring that it serves the public interest without compromising individual rights or ethical standards. As AI technology continues to advance, it will be important to keep refining the tools used for verification, while simultaneously addressing the ethical challenges they present.

## How to Use AI for Facial Recognition?

Facial recognition technology relies on advanced AI algorithms, machine learning, and artificial neural networks to identify and verify human faces within images or videos. These systems are designed to differentiate faces from other visual elements in an image, such as backgrounds, objects, or landscapes, making them particularly useful for verification purposes in journalism and fact-checking.

The process begins with gathering images or videos circulating on social media or other platforms that may contain potentially manipulated or misleading content. Journalists and fact-checkers may encounter viral images or videos that appear to show specific individuals involved in a particular event, and their primary goal is to confirm the authenticity of the content and determine whether the individuals in the media are who they are purported to be.

To begin the analysis, the first step is to select an image or video that needs to be verified. In many cases, these images are shared widely online, and identifying the individuals within them is crucial for understanding the context. A fact-checker or journalist isolates the relevant section of the image or video that contains the face of interest. This may involve cropping the image or selecting a video frame that features a clear view of the person's face. Once the image is prepared, it is uploaded to a facial recognition tool.

At this point, the algorithm processes the image to enhance its quality, which may involve improving resolution, adjusting lighting, or removing noise to ensure that the facial features are as clear as possible for analysis. This step is essential, as it helps overcome issues such as low-quality images or distorted video frames that might otherwise hinder the accuracy of facial recognition tools. AI algorithms focus on detecting key facial landmarks, such as the eyes, nose, and mouth, to create a distinct and unique "digital fingerprint" for the face.

This fingerprint is a set of data points that represent the unique characteristics of a person’s facial features, including the distance between their eyes, the shape of their nose, the curve of their jawline, and the contours of their cheekbones. These features are captured in a mathematical format and serve as a template that can be compared against databases of known individuals or images to identify potential matches.

The next step involves searching facial recognition databases for matches. These databases can include publicly available images from social media, professional databases, or other sources, depending on the tools being used. Some facial recognition platforms also allow users to search the broader internet for similar images, which can help uncover whether the individual in the image has appeared elsewhere online, possibly in different contexts. The search results provide a set of images that feature individuals who resemble the target face or match it in certain ways. This step can be particularly useful for identifying public figures or well-known personalities, as these individuals’ images are more likely to be cataloged in facial recognition databases.

Some facial recognition tools also provide links to the source of the images, which may lead to social media profiles, news articles, or other online platforms where the image or similar images have been posted. This feature streamlines the process of gathering information about the individual and helps journalists quickly access context around the images or videos in question. In some cases, the tool may provide metadata, such as the date and location of the images, further assisting with verification.

However, while facial recognition technology offers powerful tools for verifying the authenticity of images and videos, it is important to note that it has limitations. The

accuracy of facial recognition systems can be influenced by several factors, including the quality of the image, the angle from which the photo is taken, lighting conditions, and the presence of accessories like glasses, hats, or masks that may obscure key facial features. Additionally, facial recognition models can sometimes struggle with recognizing faces in low-resolution images or those that have been heavily manipulated or altered.

Moreover, facial recognition systems have faced criticism for bias and inaccuracy, particularly when applied to certain demographic groups. Studies have shown that these systems can exhibit higher error rates for people of color, women, and younger or older individuals. This is largely due to biases in the training datasets used to develop these models, which may not include a diverse range of facial images. Such biases can lead to misidentifications, further complicating the use of facial recognition technology for fact-checking purposes.

In addition to these challenges, the ethical implications of using facial recognition technology must also be considered. Privacy concerns are at the forefront of discussions about its use, particularly when it comes to tracking individuals without their consent. While facial recognition can be useful for verifying identities in the context of journalism, it must be applied in a responsible manner that respects individuals' rights and protects against potential misuse. Journalists and fact-checkers must ensure that they are not infringing on privacy or contributing to surveillance practices by using this technology.

Despite these limitations, when used properly and responsibly, facial recognition technology can be a powerful asset for journalists and fact-checkers. By enabling them to track down the origins of images or videos, verify the identities of individuals in question, and uncover inconsistencies or manipulations, these tools help ensure the integrity of journalistic work in the digital age.

As AI and facial recognition technology continue to evolve, it is likely that the accuracy and capabilities of these tools will improve, making them even more effective for use in fact-checking and investigative journalism. However, it is essential that these advances are accompanied by ethical considerations and a commitment to transparency and fairness, ensuring that the technology is used to enhance the accuracy of reporting without compromising individuals' privacy or rights.

## Uses of Facial Recognition Technology

Facial recognition technology has rapidly expanded beyond its initial use in security and surveillance, becoming an integral tool in a variety of industries and applications. Its ability

to analyze and identify individuals based on their unique facial features makes it highly valuable in enhancing security, streamlining operations, and improving user experiences. The following are some key uses of facial recognition technology:

**1. Digital Security and Surveillance Systems**

Facial recognition is widely used in digital security systems, particularly in preventing unauthorized access to sensitive areas or securing property. In corporate or government buildings, the technology helps ensure that only authorized personnel can enter restricted zones by scanning employees' faces at entry points, providing a seamless and contactless alternative to traditional security measures such as ID cards or passwords. Surveillance systems in public spaces, airports, and other high-security locations also use facial recognition to monitor individuals and track potential threats, helping law enforcement identify and track suspects, locate missing persons, or prevent criminal activity.

**2. Finding Missing Persons or Tracking Suspects**

Facial recognition technology has shown promise in locating missing persons or tracking down suspects in criminal investigations. By scanning public footage or images from surveillance cameras, authorities can match faces with databases of missing persons or criminal suspects. This technology has been particularly useful in locating children or vulnerable adults who may be at risk. In some cases, it is even used to identify perpetrators in crimes where suspects may attempt to conceal their identities.

**3. Access Control, Attendance, and Time Management**

Facial recognition has become an efficient tool in managing access control in various environments, from office buildings to schools. By using facial recognition, organizations can monitor employees' attendance and departure without the need for physical timecards or biometric fingerprint systems. In addition, this technology ensures that only authorized individuals can access certain areas, such as server rooms or confidential meeting spaces. This provides a higher level of security and helps prevent time theft or unauthorized access to sensitive information.

**4. Identity Verification for Equipment or Service Access**

Many industries are adopting facial recognition as an alternative method of identity verification, particularly in environments where security is paramount. For example, in healthcare, doctors and medical professionals may use facial recognition to verify their identity before accessing patient records or entering restricted areas. Similarly, in financial sectors, customers can use facial recognition to access their accounts or execute transactions, enhancing security and reducing the risk of identity theft.

**5. Financial Applications and Electronic Know Your Customer (eKYC)**

Facial recognition has become a key tool in financial applications, particularly in the realm of digital banking and financial services. By integrating facial recognition into the eKYC process, financial institutions can verify a customer's identity remotely without requiring physical documentation. This process is particularly useful for online banking, mobile payments, and cryptocurrency exchanges, where customers can securely register and authenticate their identities. As a result, eKYC has greatly enhanced the convenience and security of digital transactions while minimizing the risk of fraud.

**6. Two-Factor Authentication (2FA) and Protection Against Impersonation**

Facial recognition technology plays an essential role in two-factor authentication (2FA), an increasingly common method of securing online accounts and devices. Rather than relying solely on passwords, facial recognition adds an extra layer of security, ensuring that only the authorized user can access their accounts or devices. This is especially useful in preventing impersonation or unauthorized access to sensitive information, as it’s much harder to fake someone’s facial features compared to traditional authentication methods like PIN codes or passwords. In many mobile devices and laptops, facial recognition has become the default authentication method, replacing fingerprint or password-based logins.

**7. Identifying Travelers at Airports and Border Checkpoints**

Facial recognition has revolutionized the way travelers pass through airports and border checkpoints. Many airports now use facial recognition to match passengers' faces with their travel documents, streamlining the check-in process and reducing wait times. This technology can also be used to track passengers through various stages of travel, including security screening, boarding, and customs, improving efficiency and security. At border crossings, facial recognition allows authorities to quickly verify the identities of travelers, ensuring that individuals don’t try to enter a country using fraudulent or expired documents.

**8. Personalized Advertising**

Facial recognition is also making its way into the world of advertising. With the help of AI and machine learning algorithms, advertisers can analyze users' facial expressions, age, gender, and even emotional reactions to tailor advertisements based on a viewer's preferences and mood. For example, digital signage in malls or public transport stations can use facial recognition to display ads relevant to the person standing in front of it. This targeted approach aims to increase engagement and make advertisements more personalized and effective. However, this application has raised privacy concerns, as it

involves collecting personal data without explicit consent, prompting discussions on the ethics and regulations surrounding the use of facial recognition in advertising.

**9. Retail and Consumer Experience Enhancement**

In retail, facial recognition technology is used to personalize customer experiences and enhance security. By recognizing frequent shoppers, retailers can tailor offers and promotions to individual customers based on their purchasing history and preferences. Some stores have also implemented facial recognition at checkout counters, allowing customers to pay for their purchases through facial scans, bypassing the need for traditional payment methods. Moreover, facial recognition is used in retail security systems to identify shoplifters or prevent fraudulent returns by comparing customers' faces to previously stored data of known offenders.

**10. Entertainment and Media**

In the entertainment industry, facial recognition technology is being utilized for various purposes, such as in live events and sports. For example, stadiums or concert venues may use facial recognition to grant entry to ticket holders, improving crowd management and reducing the chances of counterfeit tickets. Additionally, facial recognition can be used in personalized media services, such as streaming platforms that tailor content recommendations based on viewers' facial expressions, reactions, or preferences. This technology could also be leveraged for virtual reality (VR) or augmented reality (AR) applications, allowing users to interact with digital content in a more immersive and personalized manner.

**11. Law Enforcement and Security in Public Spaces**

Facial recognition is increasingly being used by law enforcement agencies in public spaces to identify criminals or persons of interest. Surveillance cameras equipped with facial recognition can scan crowds at public events, protests, or large gatherings to monitor for suspicious activity or identify individuals with arrest warrants. While this use has raised concerns about surveillance and privacy, it has proven effective in certain scenarios, such as tracking down fugitives, solving cold cases, or preventing acts of terrorism. However, there are ongoing debates about the potential for misuse and the need for strict regulations on government use of facial recognition technology.

Facial recognition technology continues to evolve, and its applications are rapidly expanding across various industries. From improving security and access control to enhancing user experiences and personalizing services, this technology holds immense potential. However, its implementation must be approached with caution, as it raises concerns related to privacy, consent, and potential biases in identification. As its use

becomes more widespread, it will be important for governments, businesses, and individuals to carefully consider the ethical implications and establish appropriate safeguards to ensure responsible usage of facial recognition technology.

## Facial Recognition and Its Importance in Journalism and Fact-Checking

Facial recognition technology has increasingly become an essential tool in the field of journalism, particularly in the context of the rise of AI-generated content, such as fabricated images of fictional characters, and the widespread sharing of videos depicting violence and vandalism on social media platforms. These videos often circulate without proper context, and their responsible parties or instigators are frequently unidentifiable. As a result, journalists are under greater pressure to verify the authenticity of such content, ensuring it does not contribute to misinformation or distort public understanding.

In such an environment, facial recognition technology offers a vital solution for fact-checking, providing journalists with the means to track the identities of individuals in images and videos, which can be crucial for verifying claims and investigating the roots of misinformation. This technology is particularly useful during high-stakes events, such as conflicts, elections, political disputes, or wars, where disinformation is prevalent and can have significant impacts on public perception and political outcomes.

One of the key advantages of facial recognition is its ability to help journalists discern the true identity of individuals depicted in viral content, enabling the tracing of key figures involved in crucial events. In investigative journalism, especially when tracking networks of individuals in positions of power or influence, this technology can play a crucial role in verifying claims and uncovering connections. For example, during the investigation into the poisoning of Russian opposition leader Alexei Navalny, a team of journalists from Bellingcat and Insider utilized facial recognition technology alongside open-source intelligence (OSINT) to identify individuals involved in the plot. By cross-referencing facial data from publicly available sources and analyzing videos from various angles, the journalists were able to track down the operatives responsible for the attack, shedding light on a covert operation that would have otherwise gone unnoticed.

The application of facial recognition in such investigative efforts goes beyond simply identifying individuals in images. It allows journalists to create connections between

people and events, building a clearer picture of networks of influence and uncovering hidden relationships, especially in politically sensitive or secretive situations. This can be vital for exposing corruption, political manipulation, and state-sponsored actions, providing a more accurate and comprehensive narrative to the public.

Furthermore, the use of facial recognition in journalism is not limited to investigative pieces. It has increasingly been applied to verify the authenticity of user-generated content during breaking news situations, such as protests or natural disasters, where images and videos are shared rapidly on social media platforms. By comparing faces in these pieces of content with databases or other publicly available information, journalists can quickly confirm whether the images are authentic or manipulated.

The use of facial recognition in journalism is not without its ethical challenges. Issues of privacy, consent, and bias in facial recognition algorithms are ongoing concerns that require careful consideration. Facial recognition tools may not always be accurate, especially when applied to certain demographic groups, which could lead to the misidentification of individuals or the reinforcement of discriminatory practices. As a result, it is crucial for journalists and media outlets to maintain transparency about how they use these tools, ensure that they are applying them responsibly, and be aware of the potential for harm.

The increasing reliance on AI-generated content and facial recognition highlights the importance of media literacy and digital responsibility. Journalists must be vigilant in not only using these technologies but also educating the public about their limitations and potential for misuse. This will help to build trust in media outlets while ensuring that audiences are equipped with the knowledge needed to critically assess the content they encounter online.

## How to Select and Choose Images for Facial Recognition Analysis

The "infodemic," characterized by the rapid spread of false and misleading information, poses significant challenges for journalists and fact-checkers. With the overwhelming volume of visual media circulating online, the task of prioritizing which images or videos need scrutiny becomes daunting. Facial recognition tools are vital in such cases to verify the identity of individuals depicted in these visual materials. However, prioritizing what to examine can be challenging given the sheer number of images or videos that need

verification. To effectively manage this process, fact-checkers are advised to operate based on four main criteria: time, location, reach, and relevance.

1. **Time**: This criterion focuses on the age of the visual media, with a particular emphasis on recently published images and videos. Given the fast-paced nature of news and social media, older media may not be as relevant to ongoing narratives or misinformation campaigns. Recent images or videos, however, can be indicative of current events, and their verification can have a more immediate impact on public understanding. Fact-checkers should prioritize new content that is spreading quickly or appears to be gaining traction, as this is more likely to be part of an ongoing misinformation campaign.

2. **Location**: The geographical area covered by the media organization or the target audience of a specific campaign is also crucial when deciding which images or videos to focus on. For example, an image that circulates in a specific country or region may have a different impact depending on the context of that place. A location-based approach helps ensure that fact-checking efforts are relevant to the audience most at risk of being misled. Additionally, this can prevent the misallocation of resources to content that is not as significant in the areas of greatest concern.

3. **Reach**: This refers to the number of people exposed to a particular piece of visual media. Images or videos with significant reach (through shares, retweets, or reposts) are more likely to shape public opinion and spread disinformation. Monitoring the viral nature of content allows fact-checkers to prioritize media that is most likely to cause harm, especially if it is inciting violence or spreading harmful narratives about specific groups such as refugees, minorities, or political opponents.

4. **Relevance**: Even if an image or video has wide reach, its relevance to ongoing public discourse is equally important. Fact-checkers need to consider whether the content is related to critical events, such as political elections, protests, conflicts, or crises. Visual media tied to high-profile issues is more likely to influence public opinion or fuel political polarization, thus requiring urgent verification. Public interest is a vital component here—content related to sensitive topics, such as refugee rights or election integrity, demands more attention than other, less significant images.

Once content is prioritized based on these criteria, the verification process itself must involve several critical skills and tools for detecting altered or AI-generated images. As digital manipulation becomes more sophisticated, it's essential for journalists to have a set of tools and techniques at their disposal to identify and confirm the authenticity of visual media.

**1. Familiarity with Editing Techniques**

Fact-checkers must have a basic understanding of image editing programs like Adobe Photoshop, GIMP, or other similar software. Knowing common editing techniques allows them to recognize potential signs of manipulation. Some of the most recognizable alterations include:

- Inconsistent lighting: Edits may result in unnatural lighting or shadows that don't match the rest of the image.

- Unnatural reflections: Objects in photos might have reflections that don't align with the rest of the environment.

- Edge distortion: Edits may leave slight distortions around the edges of objects that are manipulated.

Having this foundational knowledge helps fact-checkers quickly identify red flags in images and decide whether facial recognition technology is required for further verification.

**2. Using Image Analysis Tools**

Image analysis software can also play an essential role in detecting manipulation. Adobe Photoshop, for instance, offers tools that can highlight areas of an image that have been altered. Some software can analyze pixel-level inconsistencies or differences in compression, helping to identify tampered or artificially created content.

Additionally, reverse image search engines like Google Images or TinEye can trace the source of an image or video. These tools are helpful for uncovering the original context of media, allowing fact-checkers to see whether an image has been taken out of context or misattributed.

Tools like the AI OR NOT tool can estimate the likelihood of an image being AI-generated, providing valuable insight into whether an image is likely to be a deepfake or otherwise artificially created. This can help distinguish between real events and fabricated content.

Another advanced tool, Sensity, specializes in detecting deepfake manipulation in media. This tool can identify alterations in facial features, audio, and even video manipulations, making it useful for verifying not only still images but also video content, such as manipulated speeches or fake video clips shared on social media.

**3. Paying Attention to Subtle Image Distortions**

Fact-checkers should be mindful of subtle signs of manipulation that may not be immediately obvious but can reveal inconsistencies in digital images. Examples of subtle distortions include:

- Irregular proportions: Faces or objects in images may have unusual proportions due to digital manipulation.

- Unnatural reflections: Reflections in glass or water might appear distorted or inconsistent with the rest of the image.

- Abnormal sharpness: Areas of an image may appear unnaturally sharp or pixelated due to editing.

Such details may not always be immediately noticeable but could suggest that an image has been altered. A trained eye can identify these anomalies and prompt further investigation.

**4. Consulting Experts and Collaborating with Specialists**

In cases of particularly challenging or suspicious images, consulting with digital media analysis experts or collaborating with specialists who have deep knowledge of image forensics can provide a more comprehensive perspective. Specialists in the field of digital forensics can assist in validating the authenticity of images or identifying advanced manipulation techniques, such as deepfakes.

Journalists can also benefit from collaborating with other fact-checkers and organizations specializing in digital content verification. Partnerships with global fact-checking networks or institutions like Bellingcat, which has extensive experience in open-source investigation, can help enhance the credibility and effectiveness of the verification process.

## Tools and Techniques for Facial Recognition in Images and Detection

**1. Amazon Rekognition**

Amazon Rekognition is a powerful tool offered by Amazon Web Services (AWS) that leverages machine learning algorithms to analyze images and videos for face recognition. This tool can detect and compare faces, and even analyze attributes such as age, gender, and facial expressions, making it particularly useful for identifying individuals in images and video content. By using facial features, Amazon Rekognition can verify whether a face in a new image matches any faces in a pre-existing database, providing a reliable method for verifying identities. Its high level of accuracy and scalability makes it popular among organizations needing large-scale face recognition solutions.

**2. Azure AI Video Indexer**

Azure AI Video Indexer, provided by Microsoft, is an advanced tool designed to extract metadata from videos and images, with a focus on identifying faces and other visual elements. It utilizes artificial intelligence to analyze video content in detail, detecting and indexing key features such as faces, speech, objects, and even emotions. In terms of facial recognition, Azure AI can cross-reference faces with known individuals, providing verification or tracking information. This tool is ideal for organizations and journalists who need to analyze large volumes of video content quickly, making it an essential tool in investigative journalism and media verification.

**3. FaceCheck.id**

FaceCheck.id is a user-friendly tool that allows users to upload images and search for matches based on facial features. This tool uses facial recognition technology to compare the faces in the image against a large database to identify individuals. It is particularly useful for journalists and fact-checkers who need to verify the identity of people in visual media, especially when trying to trace the origins of a photo or locate a person associated with specific content. FaceCheck.id is an accessible tool for those in need of quick and efficient facial verification in images.

**4. Betaface**

Betaface is a comprehensive facial recognition tool that not only helps identify individuals in images but also provides additional analytical features. In addition to recognizing faces, Betaface can estimate attributes such as age, gender, and ethnicity, offering deeper insight into the characteristics of the people in the images. This can be particularly useful in contexts where the goal is to categorize individuals or identify specific traits in a group of people. The tool also supports various image formats and integrates easily into web applications, making it a versatile solution for different image analysis needs.

**5. Pimeyes**

Pimeyes is a powerful reverse image search engine that specializes in facial recognition. This tool allows users to upload an image and find where that image or similar images have been previously published online. It scans a vast array of websites and platforms, helping to identify the context of the image and determine if it has been misattributed or manipulated. Pimeyes can be especially helpful for verifying images used in misinformation campaigns or identifying the original sources of viral content. By tracking where an image has appeared across the web, Pimeyes provides journalists and fact-checkers with valuable tools for investigating visual content.

# Limitations and Ethical Challenges of Facial Recognition Technology

It is important to approach the results generated by facial recognition tools with caution and not treat them as definitive conclusions. While these tools offer valuable insights, their accuracy can vary significantly depending on the quality of the training data used in their machine learning processes. If the dataset used to train the algorithms is incomplete, biased, or insufficiently diverse, the resulting facial recognition analysis can be flawed, leading to misidentifications or missing important matches. For instance, some tools may struggle with accurately identifying faces under certain lighting conditions, in crowded settings, or when faces are partially obscured or altered. These factors can reduce the reliability of the results and make it necessary to corroborate findings with additional evidence.

An illustrative example of this limitation can be seen in an investigation conducted by Belgian journalists who used Pimeyes to identify a man wearing military clothing in a photo. Despite the potential of the tool, Pimeyes failed to correctly identify the individual, instead providing results for other people shown in the same image. This case underscores the need for caution when relying solely on facial recognition technology for investigative purposes. It serves as a reminder that no tool is infallible, and that cross-checking with other sources of evidence—such as metadata, historical context, or other visual verification methods—is crucial to ensure the accuracy of the findings.

Additionally, the widespread use and development of facial recognition technology have sparked growing concerns over individual privacy and the potential misuse of biometric data. Since these tools rely on analyzing unique biological traits, such as facial features, they raise significant ethical questions about consent and data protection. Many facial recognition systems gather and store biometric data without individuals' knowledge or consent, which could lead to the unauthorized collection of personal information. This lack of consent is particularly concerning in the context of surveillance and mass data collection, where individuals may have their identities tracked and stored without their approval.

As facial recognition technology becomes more ubiquitous, its use must be governed by strict ethical guidelines and regulations to protect individual rights. Journalists, fact-checkers, and media organizations must ensure they apply these technologies responsibly and transparently, taking care to minimize privacy violations and to balance the benefits of

accuracy with the need for privacy protection. The ethical use of facial recognition requires respecting the rights of individuals, safeguarding sensitive data, and being mindful of how such information is used to avoid potential harms, such as unjust surveillance or discrimination.

## Ethical Use of Open Sources in Investigations

**1. Ensuring the Proper Data Collection in Tools Used**

Journalists must take the time to thoroughly understand the tools they use for facial recognition, including the nature of the data these tools rely on, their sources, and any ethical concerns surrounding data collection. Some tools may use publicly available images, while others may operate on private or proprietary databases, raising concerns about consent and data security. Additionally, no decision or conclusion should be drawn based solely on the results of a facial recognition tool, particularly if the results seem questionable or inconclusive. Instead, journalists should either refrain from using the tool altogether or incorporate its findings as just one element in a broader verification process, cross-referencing with multiple sources, including metadata analysis, witness testimonies, and other digital forensic techniques.

**2. Compliance with Values of Transparency and Disclosure**

Maintaining transparency in investigative work is critical, especially when using AI-powered tools for fact-checking and verification. Journalists should disclose the specific tools they use, explaining their functionalities and any inherent limitations that could affect the accuracy or reliability of results. This also involves detailing the steps taken to validate findings, whether through additional open-source intelligence (OSINT) techniques, human verification, or collaboration with experts. Being open about the methodologies used in an investigation not only strengthens public trust in journalism but also allows for accountability and independent verification of findings by other researchers or fact-checkers.

**3. Serving the Public Interest**

The use of facial recognition in journalism should always be guided by the principle of serving the public interest. This means that revealing the identity of individuals in images or videos should be justified by the need to uncover misinformation, expose wrongdoing, or protect vulnerable communities. Journalists must weigh the potential consequences of their work, ensuring that the publication of such information does not put individuals at

unnecessary risk or contribute to harmful narratives. In cases where revealing identities is not essential to the integrity of the story, anonymization or redaction should be considered as ethical alternatives.

**4. Respecting the Right to Privacy**

Journalists have an obligation to respect the privacy of individuals and ensure that their data is handled responsibly. This includes considering whether the individuals in an image or video have consented to their likeness being analyzed and publicly disclosed. In cases where facial recognition tools identify individuals in crowded or public settings, the faces of bystanders or unrelated persons should be blurred or covered to protect their anonymity. This practice is especially important when reporting on sensitive topics, such as protests, conflicts, or refugee communities, where unintended exposure could lead to serious personal or legal consequences for those involved.

**5. Availing the Right of Reply**

Accountability is a core journalistic principle, and offering the right of reply is essential in any investigation, particularly those involving allegations or accusations. Individuals who are identified using facial recognition tools should be given the opportunity to respond to the findings before publication. This ensures fairness, prevents potential misrepresentation, and allows for corrections if needed. Providing individuals with a platform to clarify or contest findings upholds journalistic integrity and ensures that reporting remains balanced and just.

## Further readings


- Almeida D, Shmarko K, Lomas E. The ethics of facial recognition technologies, surveillance, and accountability in an age of artificial intelligence: a comparative analysis of US, EU, and UK regulatory frameworks. AI Ethics. 2022;2(3):377-387. doi: 10.1007/s43681-021-00077-w. Epub 2021 Jul 29. PMID: 34790955; PMCID: PMC8320316.
- Smith, M., Miller, S. The ethical application of biometric facial recognition technology. AI & Soc 37, 167–175 (2022). https://doi.org/10.1007/s00146-021-01199-9
- Chan, Jason. (2022). Facial Recognition Technology and Ethical Issues. Proceedings of the Wellington Faculty of Engineering Ethics and Sustainability Symposium. 10.26686/wfeess.vi.7647.

Chapter V

# Guide to Auditing Online Advertisements During Elections

## Introduction

Electoral advertising plays a crucial role in shaping public opinion and informing voters about candidates, their policies, and their proposed programs. Traditionally, this form of advertising has relied on methods such as banners on streets, printed flyers distributed in areas frequented by potential voters, and collaborations with marketing agencies specializing in political communication. Additionally, television, radio, and newspapers have long been used to reach broad audiences, providing candidates with platforms to communicate their messages effectively.

However, with the increasing dominance of digital media, electoral advertising has expanded beyond traditional formats to include social media campaigns, paid digital advertisements, and influencer partnerships. Platforms such as Facebook, Instagram, X (formerly Twitter), YouTube, and TikTok have become critical tools for political outreach, allowing candidates and political entities to reach millions of users in targeted ways. While this shift has enhanced campaign outreach and engagement, it has also introduced several challenges, including the spread of misleading information, hate speech, and violations of electoral transparency regulations.

One of the most significant issues with digital political advertising is the lack of full transparency regarding the funding, targeting, and messaging strategies behind these ads. While platforms have introduced ad libraries and transparency tools—such as Meta's Ad Library and Google's Political Ads Transparency Report—many political advertisers attempt to circumvent these measures. Some deploy vague or deceptive messaging, while others use networks of anonymous pages and accounts to promote political content without clear attribution.

In regions such as the Middle East and North Africa (MENA), where political landscapes are often polarized and heavily influenced by misinformation campaigns, monitoring and fact-checking these ads becomes even more essential. Disinformation campaigns can exploit

social media ads to spread false claims about electoral candidates, undermine political opponents, or manipulate public perception in ways that can impact electoral outcomes. Moreover, some campaigns employ divisive narratives, amplifying sectarian, ethnic, or ideological tensions to influence voter behavior.

## Fact-checking Sponsored Electronic Advertisements

Sponsored advertisements serve as a legitimate and lawful means for candidates and political parties to reach audiences, provided they maintain transparency and comply with the legal framework of the country where the elections take place. These advertisements play a crucial role in modern electoral campaigns, allowing candidates to communicate their policies, engage with voters, and counter misinformation. However, transparency remains a key issue, as not all political advertisers disclose their funding sources or the extent of their spending. Fact-checkers and media watchdogs can play an essential role in monitoring political ad campaigns to ensure they align with electoral regulations, particularly regarding sponsorship disclosure, content accuracy, and respect for democratic principles.

One important aspect of monitoring political advertisements is ensuring adherence to electoral silence periods—legally mandated timeframes when campaigning must cease to allow voters to make decisions without undue influence. Fact-checkers can also assess whether candidates and political parties comply with spending limits imposed on campaign advertising. In many cases, the absence of a cap on digital advertising expenditure during election periods disproportionately benefits financially dominant parties, enabling them to reach a wider audience and exert greater influence over public discourse. This imbalance undermines the democratic principle of equal opportunity among candidates, as wealthier individuals or parties can flood digital spaces with their messaging while less-funded competitors struggle to gain visibility.

Beyond financial concerns, the content of sponsored advertisements must also be closely scrutinized to prevent the spread of harmful rhetoric. Fact-checkers should identify political ads that contain hate speech, incitement against competitors or marginalized groups, or defamatory content designed to manipulate public opinion. Historically, political campaigns have leveraged sponsored ads to spread misleading narratives and discredit opponents. For example, during the 2016 U.S. presidential election, Donald Trump’s campaign strategically used digital ads to attack Hillary Clinton, influencing voter perceptions and gaining a competitive edge. Such tactics highlight the need for rigorous

oversight to ensure that electoral advertising serves the public interest rather than distorting democratic processes through targeted misinformation and divisive messaging.

## Auditing Paid Digital Advertisements in the Arab World

The greatest risk associated with paid digital advertisements in political campaigns lies in their potential to spread misleading or manipulative information. Unlike traditional political advertising, which is subject to stricter regulations and public scrutiny, digital ads can be highly targeted and difficult to track, making them a powerful tool for shaping public opinion in ways that may not always align with democratic values. According to the annual report by Oxford University on social media manipulation by political actors, approximately $10 million was spent on political advertisements across various digital platforms. This significant investment highlights the growing reliance on digital advertising as a means of influencing voters, sometimes through deceptive tactics.

The same report also documented the removal of over 317,000 accounts and pages by social media platforms due to policy violations or the dissemination of misleading political content. This demonstrates that while platforms have taken steps to curb misinformation, bad actors continue to exploit digital spaces to manipulate narratives and sway elections. The challenge lies in the ability of these actors to rapidly create and disseminate new content, often circumventing platform policies by using fake accounts, misleading domain names, and paid influencers. This underscores the urgent need for rigorous fact-checking and monitoring of political advertisements to ensure that voters receive accurate and transparent information.

In 2024, several countries are holding elections at different levels, making the issue of digital political advertising even more pressing. Jordan, for instance, is set to conduct parliamentary elections, while Tunisia, Algeria, Mauritania, and the Comoros are preparing for presidential elections. At the local level, Libya and Somalia are organizing municipal and council elections. Given the significance of these elections in shaping political landscapes, there is a heightened risk of political actors using digital advertising to spread propaganda, suppress dissent, or discredit opposition candidates. The lack of oversight in many of these regions makes it even more critical to analyze and fact-check online political ads.

The urgency of scrutinizing paid political advertisements is especially pronounced in the Middle East and North Africa (MENA) region, where transparency in electoral processes remains a significant challenge. Many countries in the region have limited access to

publicly available campaign finance data, making it difficult to track who is funding political advertisements and how much is being spent. This opacity allows well-funded political entities to dominate digital spaces, shaping public discourse in ways that may not be visible to fact-checkers or voters. Without clear regulations on digital advertising expenditures, elections in the MENA region risk being influenced by financial power rather than democratic debate.

Beyond financial concerns, the political environments in many MENA countries further complicate efforts to ensure fair and transparent elections. In nations with authoritarian tendencies or weak democratic institutions, state-affiliated actors and ruling parties often exploit digital advertising to reinforce their dominance while suppressing opposition voices. In conflict-ridden contexts, such as Libya and Somalia, political advertising can become a tool for exacerbating tensions, fueling divisions, and even inciting violence. The ability to fund and distribute online advertisements with little accountability makes it easier for actors to spread disinformation, target specific groups, and manipulate public perceptions.

Given these challenges, fact-checkers, journalists, and election watchdogs must prioritize the monitoring of paid political advertisements. This includes tracing the sources of funding, analyzing the content of these ads, and assessing whether they adhere to electoral laws and platform policies. Additionally, digital literacy campaigns can help voters critically engage with online political messaging, making them less susceptible to manipulation. In an era where digital advertising is a key battleground for political influence, ensuring transparency and accountability in this space is essential for protecting democratic processes and fostering informed electoral participation.

## Platform Policies

Online platforms that host and regulate political advertisements have established varying policies to either limit or control their use, ensuring they adhere to platform rules and legal frameworks. While some platforms impose strict restrictions, others allow political advertising under specified conditions, including transparency requirements and adherence to local laws. These policies aim to balance the need for political expression with the risk of misinformation, manipulation, and undue influence in electoral processes. However, challenges remain, as some platforms struggle to enforce these regulations effectively, allowing loopholes that political actors and interest groups exploit.

Meta, the parent company of Facebook and Instagram, permits political and election-related advertisements as long as they comply with its advertising policies and do not violate platform guidelines. Advertisers must follow specific requirements, including providing clear disclosure about the entity funding the ad. Additionally, Meta requires that political ads include information such as the duration of publication, the amount spent, and a disclaimer that absolves the platform of liability. These measures aim to enhance transparency and allow users to access key details about the advertisements they encounter.

Furthermore, Meta enforces spending limits to prevent excessive financial influence by any single political entity. The platform also maintains an Ad Library, where users and researchers can review a database of all political ads, their sponsors, and spending patterns. Despite these efforts, concerns persist regarding the platform's ability to adequately fact-check political advertisements, as misinformation can still circulate widely before being flagged or removed. Moreover, enforcement of these policies varies across different regions, with some countries experiencing weaker oversight due to a lack of local monitoring capacity.

X (formerly known as Twitter) has taken a more permissive stance on political advertising, allowing funded political ads as long as they adhere to national election laws and respect election silence periods. This means that advertisers must comply with country-specific regulations, including any restrictions on campaign spending and timing. However, X has been criticized for its inconsistent enforcement of these policies, particularly in regions where misinformation is rampant.

Following Elon Musk's acquisition of the platform, X reinstated political advertisements in the U.S. after previously banning them under its previous management. This decision sparked debates over the potential for misuse, especially given the platform's role in past controversies surrounding election-related misinformation. Unlike Meta, X does not provide as detailed a database for tracking political ads, making it harder for researchers and journalists to monitor spending and identify potential disinformation campaigns.

Unlike Meta and X, TikTok has adopted a strict prohibition against paid political advertisements and politically themed sponsored content. This policy aims to prevent the platform from being used as a tool for election interference, given its large, young user base and the virality of its content. However, while TikTok's official policy bans political ads, there have been multiple instances where political content has been promoted in less transparent ways, raising concerns about covert influence campaigns.

One of the biggest criticisms of TikTok's approach is the platform's lack of transparency in enforcing its own policies. Investigations have revealed cases where influencers were paid to promote political content without disclosing their financial ties. This practice raises ethical concerns, as it blurs the line between organic user engagement and covert political advertising. Such incidents highlight the challenges of regulating political messaging in the digital space, where influencers can act as intermediaries for undisclosed campaigns.

A notable example of TikTok's enforcement shortcomings was exposed in a BBC investigation, which revealed that certain influencers had been compensated by a marketing firm to post anti-Trump content during the U.S. presidential election. These influencers failed to disclose that they were receiving payments, violating ethical standards for transparency. The revelation sparked concerns about the platform's ability to prevent hidden political advertising and its susceptibility to being used for influence operations.

While TikTok maintains that it strictly prohibits political advertising, this investigation demonstrated how political actors can circumvent the ban by leveraging third-party marketing agencies. Unlike traditional advertisements, where platforms require disclaimers about sponsorship, influencer-driven campaigns can appear organic, making them harder to detect and regulate. This raises the need for stricter policies to ensure that all political messaging, whether direct or indirect, is transparent to the public.

The differences in policies among Meta, X, and TikTok illustrate the complexities of regulating political advertisements across digital platforms. While some platforms prioritize transparency by requiring financial disclosures and ad libraries, others impose outright bans, which can sometimes be circumvented. This fragmentation creates an uneven playing field, where political actors may shift their strategies across platforms depending on where they can gain the most influence with the least oversight.

Moreover, global variations in election laws further complicate the enforcement of platform policies. Some countries have strict regulations on political advertising, while others lack clear legal frameworks, leaving digital platforms to set their own rules. This inconsistency makes it difficult to implement universal standards, allowing political advertisers to exploit regulatory gaps. As digital advertising becomes an increasingly powerful tool in elections, greater coordination between governments, platforms, and civil society is needed to ensure fair and transparent electoral processes.

# Sponsored Advertisement Libraries on Various Platforms

The process of researching sponsored electronic advertisement libraries serves as the starting point for preparing reports on election campaigns conducted on social media platforms. These libraries enable the collection of data that can later be analyzed and audited to derive insights.

**1. Meta Ad Library**

The Meta Ad Library serves as a comprehensive repository for sponsored advertisements with political or electoral relevance across Facebook and Instagram. One of its most significant features is its commitment to transparency, offering users access to periodic reports that detail the nature of political advertisements, their sponsors, and their reach. Users can search through the library using various filters, including keywords, political parties, candidates, and trending topics, making it an essential tool for researchers, journalists, and fact-checkers. By providing detailed information on ad spending and targeting strategies, the library helps expose potential manipulation, foreign influence, or misleading campaigns that might affect public opinion. Additionally, the transparency reports published by Meta offer insights into advertising trends, the demographics of targeted audiences, and the financial power behind different political campaigns. Despite these advantages, concerns remain regarding Meta's enforcement of its policies, as certain misleading ads or undisclosed sponsorships have occasionally bypassed its regulations. Nevertheless, the library remains a valuable resource for those investigating the impact of digital political advertising.

**2. X Ad Repository**

The X (formerly Twitter) Ad Repository provides transparency regarding advertisements running on the platform, offering insights into the accounts responsible for paid promotions, the popularity of the ads, and their audience reach. This database allows users to analyze the effectiveness and engagement levels of political advertisements while also shedding light on the strategies employed by various political actors. The repository is particularly useful for tracking how political campaigns or interest groups use X's advertising services to influence discourse, especially during election periods. By displaying ad metadata, including impressions and interactions, the repository aids in monitoring potential trends in misinformation, propaganda, or voter manipulation. However, X has faced criticism for its inconsistent transparency measures, particularly after shifting its policies under Elon Musk's ownership. Researchers and watchdog organizations have pointed out that the platform provides less detailed information compared to its competitors, making it more difficult to track the influence of political ads.

While the X Ad Repository remains a helpful tool, its effectiveness depends on the platform's willingness to maintain and improve its transparency efforts.

**3. Google Ad Transparency Center**

Google's Ad Transparency Center is a centralized platform that allows users to access information about advertisements running across various Google services, including Google Search, YouTube, and partner websites. One of its key features is the Google Political Ads section, which provides a dedicated space for political advertising, allowing users to see details about political campaigns, their sponsors, and the applications through which the ads are published. This database enables users to track how political messaging is distributed across Google's vast advertising network, offering insights into funding sources and audience targeting strategies. Google's approach to transparency is notable for its broad scope, covering multiple platforms and services where political advertisements are displayed. Additionally, Google provides historical data on political ad campaigns, making it easier to analyze long-term trends in digital political advertising. However, while Google's transparency efforts are commendable, the complexity of its advertising ecosystem means that some ads may still evade scrutiny, particularly when campaigns use indirect marketing strategies or third-party intermediaries.

**4. TikTok Ads Transparency**

TikTok's Ads Transparency Library is designed to offer insight into sponsored advertisements running on the platform, particularly in relation to political and electoral content. The library includes transparency reports that detail ad spending and audience targeting, helping to monitor how political actors use TikTok's advertising system. Additionally, TikTok states that it complies with government requests to remove ads that violate local election regulations, ensuring that its platform does not become a vehicle for political manipulation. However, TikTok has been widely criticized for a lack of enforcement and inconsistencies in its ad policies. Investigations have revealed cases where politically motivated content has been promoted through influencers and third-party marketing agencies, circumventing the platform's formal ad system. Unlike other major platforms, TikTok has a strict ban on direct political advertising, but this policy has not entirely prevented covert political campaigns from taking place. The transparency library is a step in the right direction, but experts argue that more stringent measures are needed to prevent hidden political influence on the platform.

**5. Snap Political Ad Library**

Snapchat's Political Ad Library is a dedicated repository for tracking paid political advertisements on the platform, with a strong emphasis on transparency. One of its

distinguishing features is that it provides an archive of political ads dating back to 2018, allowing researchers and journalists to analyze historical trends in political advertising on Snapchat. This archive can be downloaded and reviewed for further analysis, making it an invaluable tool for those studying digital campaigning strategies. Snapchat's transparency efforts extend beyond merely displaying ad content; the platform enforces strict guidelines for political advertisers, ensuring that all political ads disclose their sponsors and comply with electoral regulations. Given Snapchat's younger audience demographic, the political ad library also helps monitor how campaigns are targeting younger voters, which has become an increasing concern in modern elections. Despite these advantages, Snapchat remains a smaller player in the digital advertising space compared to Meta and Google, meaning its impact on large-scale election campaigns is relatively limited. Nevertheless, the availability of a transparent political ad archive strengthens accountability in digital political advertising.

**6. Reddit Political Ads Library**

Reddit's Political Ads Library offers a detailed look into political advertisements running on the platform, including information about the advertisers, the amounts spent, and the target audience. What sets Reddit apart is its proactive approach to ad moderation—platform staff review all political ads before they are published to ensure compliance with community guidelines and transparency policies. This pre-publication review process is designed to minimize the risk of misinformation and ensure that political advertising adheres to ethical standards. Additionally, Reddit's transparency measures help users track how political campaigns engage with niche online communities, where grassroots digital activism and political discourse often flourish. Given that Reddit operates through user-driven discussions and forums, political advertising on the platform tends to be more subtle and integrated into specific community conversations. The political ads library allows researchers and users to scrutinize these interactions, identifying potential attempts at influence operations or targeted propaganda. However, since Reddit's advertising reach is smaller than that of larger platforms like Meta or Google, its political ad library plays a more specialized role in digital election monitoring.

Before analyzing the content of paid electronic ads and their compliance with the laws, it is essential to document and archive the results of various searches for electronic ads. This helps prevent them from being lost or deleted by the publishers or by the platforms themselves in case of policy violations.

This information can be organized into a spreadsheet for archiving evidence, which should include the following:

- The platform displaying the advertisement
- The nature of the advertisement
- The advertiser
- The amounts paid
- The target audience
- The timing of the advertisement
- The advertisement link after it has been archived using various archiving tools
- Screenshots to confirm documentation

Notes on the content of the advertisement; for example: it contains AI-generated images, targets a specific group, or promotes a party other than the advertiser.

## Analysis of Funded Political Advertisements

Tracking the data collection and archiving phase, analyzing the content of the advertisement to identify the funder, their objectives, and the methods used to influence.

**1- Analyzing the Messages According to the Target Audience**

Targeting specific groups with tailored messages has always been a fundamental aspect of election campaigns. However, funded advertisements on digital platforms take this practice to a new level by allowing campaigns to micro-target audiences based on detailed data, such as demographics, interests, and online behavior. This enables candidates to craft messages designed to appeal to specific groups while shielding these messages from the broader public. Such a strategy allows campaigns to present different, and sometimes contradictory, narratives to different audiences. For example, a candidate may run ads that promote policies favoring business owners, such as tax cuts and deregulation, while simultaneously targeting workers with promises of labor protections and wage increases. This selective exposure raises concerns about transparency and voter manipulation, as the general public may not be aware of the multiple and potentially conflicting messages a candidate is disseminating.

Moreover, targeted advertising can exacerbate societal divisions by reinforcing pre-existing beliefs and preventing voters from being exposed to alternative viewpoints. If political actors use data-driven targeting to deliver hyper-personalized messages that play to individuals' biases or fears, they can effectively deepen political polarization. This practice,

known as "dark ads," has been widely criticized for its role in influencing elections without public scrutiny. Fact-checkers and journalists must scrutinize political advertisements to identify inconsistencies in messaging and assess whether campaigns are deliberately misleading different segments of the population to secure votes. By comparing targeted ads across different demographics, they can uncover discrepancies and bring them to public attention, ensuring greater transparency in digital political advertising.

**2- Advertisement Dates**

The timing of election advertisements is strictly regulated in many countries, with authorities imposing specific periods during which campaigns are allowed to run paid political ads. Any advertisement that appears outside these designated timeframes is a violation of electoral laws and can have serious consequences for the fairness of the election process. Fact-checkers and journalists must pay close attention to the dates when political ads are published and compare them with official campaign schedules. Early advertisements can give certain candidates an unfair advantage by allowing them to shape public perception before their opponents have the opportunity to respond. This can be especially concerning in countries where electoral institutions lack the capacity to enforce regulations effectively.

Additionally, some political actors may attempt to circumvent restrictions on advertisement timing by using proxy organizations or unofficial supporters to publish ads on their behalf. These indirect campaigns can be harder to track, as they may not explicitly disclose their affiliation with a candidate or political party. Platforms like Meta and Google provide tools that allow users to search for political ads based on publication date, but these tools are not foolproof. Investigative journalists and election monitors must cross-check advertisements with financial disclosures, media reports, and other sources to determine whether a candidate or political party is violating electoral rules. Holding campaigns accountable for respecting official advertising periods is crucial in ensuring fair competition and preventing undue influence on voters.

**3- Including Emotional Appeals or Stirring Emotions**

Political advertisements often rely on emotional appeals to connect with voters, but funded digital ads take this to an extreme by using data-driven techniques to trigger specific emotional responses. Campaigns may craft advertisements that provoke fear, anger, or sympathy to influence voter behavior. For example, fear-based messaging is frequently used to warn voters about the alleged dangers of electing an opponent, while sympathy-driven content highlights personal stories to generate emotional support for a candidate. One of the most notable examples of this strategy was Kamala Harris’s use of

funded advertisements to criticize Donald Trump's stance on women's rights. Her campaign produced videos featuring personal testimonies from women who feared for their health and reproductive rights under Trump's policies, aiming to mobilize female voters against him.

While emotional appeals are a natural part of political persuasion, their use in digital advertisements can be problematic when they exploit false or exaggerated claims to manipulate public sentiment. Misleading emotional narratives can shape voter perceptions in ways that are difficult to counter, particularly when these messages are tailored to specific individuals. Unlike traditional political ads aired on television or radio, digital ads often escape public scrutiny because they are only visible to the targeted audience. Fact-checkers should closely analyze the emotional framing of political advertisements, assessing whether they are based on verifiable facts or designed to manipulate public perception unfairly. They must also educate voters on how emotional influence works in digital campaigning, helping them critically evaluate the political messages they encounter online.

**4- Budget**

Election laws in many countries impose spending limits on political advertisements to ensure fair competition between candidates. However, digital advertising has made it easier for well-funded campaigns to bypass these restrictions. By spreading expenditures across multiple online platforms, using third-party organizations, or leveraging influencer marketing, campaigns can conceal the true scale of their spending. Investigative journalists and fact-checkers play a vital role in tracking these expenditures by analyzing financial disclosures, cross-referencing spending reports from advertising platforms, and identifying any undeclared political ad campaigns. Comparing reported budgets with actual advertisement presence on social media can reveal discrepancies that indicate potential violations.

Moreover, the absence of strict regulations on digital ad spending in some countries allows wealthy candidates or political parties to dominate online spaces, undermining the principle of equal opportunity in elections. Fact-checkers must scrutinize ad libraries provided by platforms like Meta, Google, and Snapchat to assess how much candidates are investing in digital campaigns. They should also pay attention to indirect spending, such as corporate-sponsored ads that promote a political agenda without directly affiliating with a candidate. By exposing excessive or unregulated spending, fact-checkers contribute to a more transparent electoral process and help prevent financial advantages from distorting democratic competition.

**5- Misinformation or Inauthentic Behavior**

One of the most concerning aspects of funded digital advertising is its potential to spread misinformation. A study on criminal protection from media misinformation during election campaigns found that deceptive advertisements can significantly impact voter perceptions, particularly when they involve smear campaigns, false claims, or manipulated narratives. Political actors may use misleading ads to tarnish the reputation of opponents, create confusion about election procedures, or push fabricated statistics to sway public opinion. These tactics undermine the integrity of elections by distorting facts and misleading voters.

Inauthentic behavior, such as coordinated disinformation campaigns, further compounds the problem. Some campaigns employ fake accounts, automated bots, or networks of influencers to amplify misleading political ads, making them appear more credible and widespread. Digital platforms have made efforts to combat such tactics, but enforcement remains inconsistent. Fact-checkers must be vigilant in identifying signs of misinformation in political ads, cross-referencing claims with reliable sources, and exposing disinformation networks. By debunking false narratives and highlighting manipulation tactics, they can help voters make more informed decisions.

**6- Using the Ad Observatory Tool**

The Ad Observatory Tool, developed by New York University's Information Security in Democracy initiative, provides a valuable resource for tracking political advertisements on Meta's platforms, Facebook and Instagram. This tool allows users to search for ads based on keywords, topics, political actors, or geographical regions, offering insights into spending patterns and campaign strategies. By analyzing this data, fact-checkers can uncover trends in digital political advertising, such as which candidates are investing the most in online outreach and how different demographics are being targeted.

One of the most important applications of the Ad Observatory Tool is detecting potential violations of electoral advertising laws. Fact-checkers can use it to monitor whether campaigns are exceeding spending limits, running ads outside the designated timeframes, or engaging in misleading messaging. Additionally, the tool helps researchers identify foreign influence in domestic elections by tracing advertisements linked to international actors. While Meta has faced criticism for limiting researchers' access to ad data, tools like the Ad Observatory remain essential in holding political campaigns accountable for their digital advertising practices. Investigative efforts using such tools contribute to greater transparency and safeguard democratic processes from covert manipulation.

## Further readings


- Moore, M., & Ramsay, G. N. (2024). Local News in National Elections: An “Audit” Approach to Assessing Local News Performance During a National Election Campaign. Digital Journalism, 1–20. https://doi.org/10.1080/21670811.2024.2333827
- Ranjan, A., & Upadhyay, A. K. (2024). Exploring the continuity and change in political advertising research: a systematic literature review. Cogent Social Sciences, 10(1). https://doi.org/10.1080/23311886.2024.2376853
- Calvo D, Cano-Orón L, Baviera T. Global Spaces for Local Politics: An Exploratory Analysis of Facebook Ads in Spanish Election Campaigns. *Social Sciences*. 2021; 10(7):271. https://doi.org/10.3390/socsci10070271
- Rose, K., & Rohlinger, D. A. (2024). Political Influencers and Their Social Media Audiences during the 2021 Arizona Audit. Socius, 10. https://doi.org/10.1177/23780231241259680

Chapter VI

# Guidelines for Fact-Checking Disinformation Campaigns During Climate Conferences

## Introduction

Climate misinformation has emerged as a significant barrier to global efforts to address climate change. Disinformation campaigns, often fueled by interest groups opposed to climate action, aim to cast doubt on scientific consensus and weaken public trust in environmental policies. These campaigns exploit social and traditional media to spread misleading narratives, discouraging individuals from adopting sustainable behaviors and pressuring governments to delay or abandon climate-related regulations. The impact of such misinformation is especially pronounced during high-profile climate events like the Conference of the Parties (COP), where global leaders gather to negotiate climate commitments.

The spread of climate misinformation often intensifies around COP sessions, as opponents of climate policies seek to discredit scientific findings and undermine policy discussions. False claims range from denying the existence of climate change to exaggerating the economic costs of climate action. Some misinformation also takes the form of greenwashing—where corporations or governments falsely present themselves as environmentally responsible to deflect criticism. Journalists and fact-checkers play a crucial role in identifying and debunking such claims, ensuring that public discourse is informed by accurate and science-based information.

To effectively counter climate misinformation, journalists and fact-checkers must adopt a structured approach. This includes monitoring key narratives before, during, and after COP sessions, identifying sources of misinformation, and verifying claims using reliable scientific data. Fact-checkers should cross-reference statements with reports from authoritative institutions such as the Intergovernmental Panel on Climate Change (IPCC) and leading climate research organizations. Additionally, they must consider the context in

which misinformation is spread—whether it is driven by political, economic, or ideological motives—and expose the underlying interests behind misleading claims.

Several digital tools can aid journalists in their fact-checking efforts. Google's Fact Check Explorer allows users to verify climate-related statements against existing fact-checks, while Climate Feedback provides expert-reviewed analyses of media coverage on climate topics. Satellite imagery tools, such as NASA's Earth Observatory and the European Space Agency's Sentinel Hub, offer real-time environmental data to counter claims that deny observable climate changes. Social media monitoring tools like CrowdTangle can help track the spread of misinformation and identify influential sources amplifying false narratives. By leveraging these tools, fact-checkers can quickly assess the accuracy of climate-related claims and provide timely corrections.

Despite the availability of verification tools, fact-checkers face significant challenges in countering climate misinformation. One of the biggest obstacles is the rapid spread of false claims on social media, where algorithms often prioritize engagement over accuracy. Additionally, some climate misinformation is presented in sophisticated ways, making it difficult for the general public to distinguish between genuine scientific debate and intentional disinformation. Coordinated efforts between journalists, fact-checking organizations, and social media platforms are essential to curb the influence of climate misinformation. Enhancing media literacy among audiences can also empower individuals to critically assess climate-related information and recognize misleading narratives.

The fight against climate misinformation is integral to advancing meaningful climate action. As misinformation continues to evolve, journalists and fact-checkers must remain vigilant in monitoring emerging false narratives, strengthening verification methods, and collaborating with scientific institutions. Providing accessible and clear fact-checks can help counter misleading claims before they gain widespread traction. Ultimately, ensuring that the public receives accurate climate information is essential not only for fostering environmental responsibility but also for holding governments and corporations accountable for their climate commitments.

## What is Climate-Related Disinformation?

Climate-related disinformation is a strategic effort to mislead the public about climate change, often driven by political, economic, or ideological motives. The European Union defines it as the deliberate spread of false or misleading narratives designed to sow doubt about scientific consensus and delay climate action. These narratives vary widely, from

outright denying the existence of climate change to distorting data by downplaying its severity or questioning its human-induced causes. By fostering confusion and skepticism, such disinformation weakens public support for climate policies and allows polluters and governments to evade responsibility for addressing the crisis.

A key tactic of climate disinformation involves promoting conspiracy theories that frame climate change as a hoax or a tool for global control. Some narratives suggest that climate policies are merely a pretext for imposing restrictions on economic growth or personal freedoms, while others falsely claim that climate activists and scientists manipulate data to advance hidden agendas. This type of disinformation often spreads rapidly through social media, where algorithms amplify sensationalist content, making it difficult for accurate scientific information to reach the public. The deliberate distortion of climate science not only misguides individuals but also hinders collective action, as it creates divisions and undermines the legitimacy of environmental initiatives.

Another widespread form of climate disinformation is "greenwashing," a deceptive practice highlighted by the United Nations. This occurs when corporations or governments present themselves as environmentally responsible while continuing to engage in unsustainable activities. Companies may use misleading marketing campaigns to portray their products as eco-friendly, even when they contribute significantly to pollution and carbon emissions. Similarly, some governments announce ambitious climate pledges but fail to implement meaningful policies, using symbolic gestures to gain public approval while avoiding systemic reforms. By diverting attention from genuine solutions, greenwashing not only misleads consumers but also delays critical actions needed to combat climate change effectively.

## Why Combating Climate Disinformation During Climate Summits is a Priority?

The proliferation of climate disinformation during international climate conferences has transformed from sporadic occurrences into well-coordinated campaigns aimed at shaping policies, influencing decisions, and even altering the agendas of these crucial global gatherings. These efforts present a formidable challenge to climate action, as they exploit the high-profile nature of these events to spread false narratives that weaken public trust in climate science and delay urgent interventions. By infiltrating discussions and manipulating media narratives, these campaigns serve the interests of industries and governments seeking to avoid stricter climate regulations.

A striking example of this phenomenon was observed during COP28, held in the UAE in 2023. A report by *The New York Times* exposed how internet influencers, fossil fuel companies, and even certain participating nations actively contributed to the spread of misleading narratives. These actors strategically used digital platforms to cast doubt on the scientific consensus regarding climate change, promoting rhetoric that downplayed its severity or framed climate policies as threats to economic stability. By amplifying such disinformation, they sought to erode public pressure for decisive action, ultimately hindering the adoption of meaningful climate commitments.

Beyond undermining climate policy discussions, these disinformation campaigns have also fueled hostility toward environmental activists. *The New York Times* report highlighted a surge in online hate speech targeting climate advocates, particularly those pushing for stronger climate justice measures. By portraying activists as extremists or painting their demands as unrealistic and harmful to economic interests, these campaigns aim to delegitimize grassroots climate movements and weaken their influence in policy negotiations. This tactic not only stifles public discourse but also deters broader societal engagement with climate issues.

The Climate Action Against Disinformation (CAAD) coalition, an alliance of organizations combating climate misinformation, released a report during COP28 detailing the efforts of fossil fuel lobbyists and state-backed actors to obstruct climate policies. The report presented case studies on how these entities leverage online networks to manipulate public opinion, using misleading statistics, cherry-picked data, and deceptive arguments to justify continued fossil fuel dependency. These campaigns often rely on sophisticated digital advertising strategies and coordinated bot activity to amplify false narratives and distort the climate conversation.

A particularly alarming aspect of these campaigns is their ability to influence policymakers directly. In many cases, fossil fuel interests fund think tanks and research institutions that produce reports downplaying the urgency of climate action. These reports are then cited by politicians and industry leaders during climate summits to argue against ambitious climate policies. By creating an illusion of scientific debate, these tactics enable governments to justify inaction or weak commitments, effectively stalling global efforts to curb emissions and transition to renewable energy sources.

Rahma Diaa, founder of the Climate School initiative, has noted that the rise in climate disinformation during COP summits is a direct response to growing pressure from climate advocates demanding radical actions, such as phasing out fossil fuels and ensuring climate justice funding. As calls for systemic change intensify, resistance from economic and commercial circles also escalates, leading to aggressive counter-campaigns aimed at

discrediting climate activists and shifting public perception in favor of maintaining the status quo. This dynamic underscores the ongoing battle between scientific truth and vested economic interests.

The economic and political stakes surrounding climate action have created an environment where misinformation thrives. Fossil fuel companies, concerned about the financial implications of climate regulations, invest heavily in disinformation to protect their interests. Similarly, some governments—especially those heavily reliant on fossil fuel exports—actively participate in these efforts to delay global climate commitments. By shaping public opinion through targeted disinformation, these actors create obstacles to policy changes that could significantly reduce emissions and accelerate the transition to sustainable energy sources.

Countering this growing wave of climate disinformation requires a proactive approach from journalists, fact-checkers, and climate organizations. Strengthening digital literacy among the public, exposing deceptive tactics, and holding disinformation networks accountable are crucial steps in safeguarding climate discourse. As climate summits continue to be targeted by misinformation campaigns, ensuring transparency and amplifying credible voices will be essential in maintaining momentum toward meaningful global climate action.

## Notable Reports Exposing Climate Disinformation at Summits

The CAAD coalition's report "Robo-COP29: Bots Boosted Propaganda Promoting Petrostate Host" revealed a large-scale disinformation campaign orchestrated through thousands of fake social media accounts that sought to portray Azerbaijan, the host of COP29, as a leader in climate action. These bots systematically amplified narratives positioning Azerbaijan as a climate champion, despite its deep reliance on fossil fuel exports. By artificially inflating positive sentiment around the country's climate policies and suppressing criticism, the campaign attempted to shape public perception and minimize scrutiny of Azerbaijan's environmental record. The report highlighted how digital manipulation tactics, including coordinated bot activity and algorithmic boosting, are increasingly being used to distort climate discourse and manufacture legitimacy for fossil fuel-dependent states hosting global climate summits.

The report “Deny, Deceive, Delay: New Trends in Climate Disinformation at COP27” documented how fossil fuel companies aggressively pushed misleading narratives through digital advertising, spending over $4 million on Meta’s platforms during the 2022 climate conference in Egypt. These ads aimed to downplay the role of fossil fuels in the climate crisis, promote false solutions such as carbon capture, and shift blame away from major polluters onto individual consumers. By leveraging targeted advertising, these companies exploited Meta’s vast reach to spread disinformation at a critical moment when global leaders were negotiating climate policies. The report underscored how the fossil fuel industry continues to use digital platforms to undermine climate action, delaying the transition to renewable energy by misleading the public and policymakers about the feasibility of continued fossil fuel dependence.

## Methodology for Fact-Checking Coordinated Climate Disinformation Campaigns

First, we need to monitor content shared on social media platforms closely and analyze user activity to detect any misleading information. This requires observing patterns of engagement, the types of content shared, and identifying unusual spikes in activity, which can suggest artificially driven campaigns. By identifying whether user behavior is organic or driven by coordinated bots, fact-checkers can better understand the dynamics of content dissemination and evaluate its authenticity. This analysis is essential to uncover disinformation campaigns that target vulnerable audiences, exploiting social media algorithms to spread false narratives and create confusion. Tracking these patterns allows us to trace the origins of disinformation, flagging suspicious accounts, and taking appropriate steps to stop the spread of false or harmful information before it reaches a wider audience.

Additionally, it is crucial to track and verify official statements made by public figures, especially those in positions of power, such as politicians, ministers, presidents, and other influential figures. Fact-checking the claims made by these individuals is key to ensuring the integrity of the climate discourse. For example, platforms like the “Ozone” initiative have played an essential role in fact-checking several misleading statements made by former U.S. President Donald Trump on climate change. These fact-checking efforts exposed logical fallacies, misleading terminology, and delays in policy initiatives that hinder meaningful action on climate change. The ability to scrutinize such statements with a rigorous, evidence-based approach ensures that the public is not misled by high-profile

figures who downplay the urgency of addressing the climate crisis, particularly when these individuals hold significant sway over public opinion and policy.

News websites can also serve as a source of climate-related misinformation, though often unintentionally. Many errors originate from a lack of deep understanding of climate science or the use of misleading terminology by journalists or bloggers who do not have the necessary expertise in the subject. These errors can range from misinterpreting scientific data to oversimplifying complex climate issues. To address this, it is essential for fact-checkers to work closely with journalists and editors to ensure that the information they share is accurate and based on credible sources. Additionally, providing reporters with accessible, clear, and scientifically sound information about climate issues can help curb the spread of misinformation in the media. Journalists have a responsibility to ensure that they report on climate-related matters with clarity and accuracy, avoiding sensationalism or oversights that could mislead their audiences.

Sponsored advertisements play an increasingly significant role in the spread of climate disinformation, as various interest groups fund ads that promote misleading content about climate science or policy. These ads often aim to shift public opinion away from supporting necessary climate actions, such as transitioning to renewable energy or implementing stronger environmental regulations. Therefore, it is necessary to monitor and examine paid advertisements on platforms like Meta, X (formerly Twitter), and TikTok that contain climate-related content. By scrutinizing these ads, fact-checkers can detect misinformation campaigns funded by fossil fuel companies, political groups, or other vested interests, which often use digital advertising to spread false claims and shape public discourse on climate change. The ability to track sponsored content and investigate its funding source is a powerful tool for combating misinformation and promoting transparency in how climate-related narratives are shaped online.

The sources mentioned above can be instrumental in uncovering misinformation that falls under **the Five Techniques of Science Denial**.

The first of these techniques is the use of **Fake Experts**, where individuals or organizations with no legitimate expertise in climate science are presented as credible authorities. These fake experts are often amplified by media outlets or interest groups looking to sow doubt about the scientific consensus on climate change.

The second technique, **Logical Fallacies**, involves offering arguments based on flawed reasoning or erroneous conclusions. For example, claims that climate change is not real because the weather was colder than usual in one particular region during a brief period

are examples of logical fallacies that mislead the public. These arguments often ignore the broader patterns of climate data that point to long-term global warming trends.

**Impossible Expectations** is the third technique, where climate action is delayed by demanding unrealistic or impractical standards before any steps can be taken. This includes arguments that environmental policies cannot be implemented unless all countries simultaneously agree to take drastic action, which serves as a convenient excuse for inaction.

The fourth technique, **Conspiracy Theories**, promotes the idea that climate change is part of a broader, secret plot with malicious intentions, such as claims that global warming is fabricated to control people or redistribute wealth. This form of disinformation fuels distrust in scientific institutions and global climate agreements.

Finally, the technique of **Cherry-Picking Data** involves selectively using data that supports a particular stance while ignoring or downplaying contradictory evidence. This approach often leads to misleading narratives, such as emphasizing short-term weather events to argue against long-term climate change trends. Recognizing these techniques and addressing them head-on is crucial in the fight against climate disinformation.

## Tips for Fact-Checking Disinformation Claims During Climate Conferences

Rahma Diaa, an award-winning journalist and founder of the Climate School initiative, emphasizes the importance of caution when addressing climate-related claims.

1. One of her key recommendations is to avoid making accusations without definitive evidence. Fact-checkers and journalists must ensure they have solid proof before attributing falsehoods or misinformation to individuals or organizations. This approach helps protect against legal repercussions, particularly when dealing with powerful entities that may seek to discredit or litigate over unfounded claims. Maintaining a rigorous standard of evidence-based reporting also upholds the credibility of the fact-checking process.

2. Another crucial tip is to present the misleading claim clearly and early in the report, ideally in a separate paragraph, with a subheading that indicates the claim is misleading. This organization ensures that readers can quickly identify the false or misleading nature of the statement and reduces any potential confusion. By structuring the article in this way,

journalists provide readers with a clear roadmap to understanding the misinformation and its debunking, ensuring transparency and preventing readers from inadvertently accepting false claims before the truth is revealed.

3. When investigating climate-related misinformation, it is important to ask key questions about the origins of the claim, including who is behind it, what their motives are, and how the claim can be verified. Understanding the source of the misinformation can help uncover the underlying agenda or vested interests that may be driving the false narrative. Fact-checkers should explore the motivations behind these claims, whether they stem from political, financial, or ideological interests, as this context provides important insights into how such misinformation might influence public opinion or policy. Additionally, verifying the claims through trusted sources or expert opinions is crucial to ensure that the information provided is accurate.

4. Rahma Diaa also advocates for the reliance on evidence, research studies, and trusted data sources. This principle is foundational in the fight against misinformation, as it allows fact-checkers to substantiate their findings with credible, peer-reviewed scientific research or data from reliable institutions. By referencing robust research and authoritative sources, journalists can strengthen their reports and build trust with their audience. Additionally, reliance on evidence helps avoid speculation and ensures that claims are evaluated objectively based on established facts and consensus within the scientific community.

5. Another tip is to refer to reports from regulatory bodies and watchdog websites that expose environmental violations. These reports can serve as an authoritative source of information when debunking misleading claims. Many regulatory bodies track and document environmental infringements, and their reports can provide the necessary evidence to challenge false narratives or deceptive claims. Fact-checkers can also turn to watchdog organizations that monitor corporate and governmental accountability in environmental matters, using their findings to refute climate disinformation and show how these entities may be deliberately obfuscating the truth about climate change.

6. Searching existing fact-checking sites for prior analyses is another useful tactic. Many climate-related claims may have already been addressed by fact-checking organizations, which can save time and provide a starting point for further investigation. Fact-checking platforms such as PolitiFact, Climate Feedback, and others often publish detailed articles on misinformation and disinformation, offering a well-established record of rebuttals to false claims. Journalists can use these existing resources to support their own investigations and ensure that the information they are presenting is consistent with what has already been debunked.

7. Consulting with researchers, scientists, and specialists in the relevant field is also critical. These experts can provide informed, accurate answers and help journalists navigate complex climate science. In many cases, misinformation arises due to misunderstandings of scientific concepts or deliberate misinterpretations. By engaging with professionals who are experts in climate change, fact-checkers can obtain accurate explanations and responses to false claims, which can then be incorporated into their reports. Experts can also help to clarify technical terms or concepts, ensuring that the information is accessible to a broader audience.

8. Using simple language and explaining any technical terms that might be difficult for a general audience to understand is a vital communication strategy. Climate change is often discussed in highly technical terms, which can alienate readers or lead to misunderstandings. Rahma Diaa stresses the importance of breaking down complex concepts into easily digestible information. By using clear, simple language and offering explanations for scientific terms, journalists ensure that their reports are accessible to all readers, regardless of their background or familiarity with climate science.

9. Incorporating visuals such as images, charts, maps, and explanatory videos is another powerful tool for fact-checkers. Visual aids can enhance the impact of the report, making complex data or information easier to understand and more engaging for the audience. Graphs, charts, and maps can illustrate trends in climate data or showcase the difference between misleading and accurate information, offering a visual confirmation of the facts. Videos can provide a more interactive and accessible way for audiences to engage with climate issues, especially when tackling complex topics that require additional context or explanation.

10. Finally, it is important to conclude the fact-checking report with a summary paragraph that highlights the key findings. This summary serves as a concise conclusion, reiterating the main points and reaffirming the accuracy of the information presented. It helps to reinforce the debunking of the claim, leaving the reader with a clear understanding of the facts and a sense of closure. By summarizing the key findings, journalists ensure that their reports leave a lasting impression and effectively combat the spread of misinformation, offering readers a trustworthy source of information in the fight against climate-related disinformation.

# Open Sources for Fact-Checking Climate Information

Open-source tools are an essential resource for journalists, researchers, and environmental activists seeking to access climate data and monitor environmental changes. These tools provide valuable insights into a wide range of issues, from deforestation to pollution, and play a crucial role in documenting the impacts of climate change. Satellite imagery tools, such as Sentinel Hub, Google Earth Pro, and the Fire Dynamic Map, are some of the most widely used open-source tools for environmental monitoring. These platforms enable users to access high-resolution satellite images that can be used to track changes in land cover, vegetation, and the frequency of natural disasters, such as wildfires. With the ability to observe real-time data, these tools are particularly helpful for investigating environmental degradation, assessing the aftermath of climate-related events, and monitoring deforestation, agricultural practices, and urban expansion.

Global Forest Watch is another invaluable resource for tracking global vegetation loss, particularly deforestation. This platform provides detailed and up-to-date information on forest cover worldwide, helping to highlight areas that are undergoing significant environmental change. For instance, the report “Egypt Buries Its Lungs Under Concrete” by Zawya Talta (Third Angle) used Global Forest Watch to document widespread deforestation in Egypt, where urbanization and development projects are rapidly replacing green spaces. By utilizing data from Global Forest Watch, journalists and activists can pinpoint deforestation hotspots, quantify the extent of vegetation loss, and draw attention to the environmental and social implications of such activities. This transparency is crucial for raising awareness and advocating for stronger environmental protections and sustainable practices.

In addition to tracking vegetation loss, SkyTruth is another powerful open-source tool that provides data on air pollution and gas flaring emissions. SkyTruth uses satellite data to monitor industrial sites, including oil extraction facilities, and tracks the emissions of harmful gases, such as methane and carbon dioxide. By visualizing the environmental impact of gas flaring, particularly in regions rich in fossil fuels, SkyTruth helps to expose the hidden costs of oil extraction on both local and global scales. This tool is especially important for documenting the environmental consequences of the fossil fuel industry, as gas flaring is a significant contributor to air pollution and climate change. Through SkyTruth, journalists can uncover the locations of major industrial polluters, analyze their environmental impact, and advocate for policy changes to reduce emissions and mitigate climate change.

The Open Source Toolkit is another key resource for climate-related investigations. It provides a range of open-source tools designed for conducting climate research, mapping environmental risks, and analyzing data related to global warming. The OSINT's Open Source Intelligence Tools and Resources Handbook is another excellent guide that provides a comprehensive list of tools for investigating environmental issues. This handbook is particularly valuable for journalists and researchers who need guidance on how to navigate the overwhelming array of tools available. It offers instructions on how to use various software and databases effectively, ensuring that users can maximize the potential of open-source tools to support their work. Furthermore, the Blincat's Toolkit provides a set of open-source tools focused on climate change and environmental justice. It includes resources for analyzing climate data, tracking carbon emissions, and documenting environmental violations, making it an indispensable asset for investigators focused on climate-related issues.

The Climate Justice Investigative Guide by ARIJ is another comprehensive resource for journalists and activists involved in climate investigations. This guide offers ethical guidelines and practical advice on using open-source tools to expose climate-related disinformation and environmental violations. It provides an array of tools for monitoring various environmental factors, such as vegetation cover, green space erosion, air pollution levels, and pollution from gas flaring associated with oil extraction. The guide encourages users to apply these tools to document the real-world impacts of climate change, whether through mapping deforestation, assessing the effects of urbanization on the environment, or analyzing air quality in polluted regions. The ARIJ guide also includes tools to verify water availability, assess flood risks, and track natural disasters, enabling users to investigate the broader scope of climate change and its consequences.

## Challenges in Detecting Climate Misinformation During Climate Conferences

**1. Collusion of Major Tech Companies**

The New York Times report, "Disinformation Is One of Climate Summit's Biggest Challenges," highlights a concerning issue where some websites spreading false information are also highly profitable due to ad revenue generated from large tech platforms. This creates an indirect complicity, as the platforms, which should be regulating harmful content, unintentionally or knowingly promote misinformation by allowing advertisers to generate revenue from misleading or false climate narratives. The algorithms

designed by these tech giants prioritize content that drives engagement, often pushing sensationalized and misleading content to the top of users' feeds. This results in the amplification of climate disinformation, undermining the trust in factual climate science and the urgency of addressing climate change. Moreover, these platforms have been slow to take action against climate misinformation, opting for reactive measures rather than proactive solutions to prevent the spread of false claims. The ability of profit-driven models to reinforce misleading narratives complicates the efforts of fact-checkers and journalists, creating a barrier to the fight against climate disinformation.

**2. Diverse Forms of Climate Disinformation**

Climate disinformation is constantly evolving and adapting to new contexts, a phenomenon referred to as "shape-shifting." This makes it especially challenging for journalists and fact-checkers to keep up with emerging narratives. Initially, climate disinformation focused on outright denial of climate science, but as the evidence of climate change has become more undeniable, the tactics have shifted. Now, misinformation often comes in the form of conspiracy theories, such as claims that climate change is part of an orchestrated effort by governments to weaken economies or control populations. Another form is the distortion of proposed climate solutions, questioning their feasibility or effectiveness to sow doubt and stymie action. Greenwashing is another significant form of disinformation, where companies or governments adopt policies that appear climate-friendly on the surface but are ultimately ineffective or misleading. Such claims are particularly dangerous because they can mislead the public into thinking that sufficient action is being taken, when in reality, the policies may do little to reduce carbon emissions or mitigate climate impacts. These diverse forms of disinformation are not only difficult to detect but also require a multi-pronged approach to debunk and expose.

**3. Challenges in Accessing Reliable Knowledge**

One of the main obstacles to detecting climate misinformation is the complexity of climate-related terminology and concepts. For non-experts, understanding the intricate science behind climate change can be a significant challenge. Terms like "carbon sequestration," "greenhouse gas emissions," and "geoengineering" can be confusing and open to misinterpretation, especially when taken out of context or manipulated to support misleading claims. Fact-checkers and journalists must therefore not only debunk misinformation but also invest time in understanding and explaining these complex scientific concepts in a clear and accessible manner. Furthermore, accessing accurate and up-to-date climate data is not always straightforward. Much of the climate science that is crucial for verifying claims is published in scientific journals or reports that may not be easily accessible to the public. The constant advancements in climate science also

mean that new research can sometimes contradict older findings, making it harder for journalists to stay current. As such, fact-checkers must rely on specialized knowledge and reputable sources, which may not always be readily available or easy to interpret.

**4. Language Barriers**

For journalists and fact-checkers working in regions where English is not the primary language, such as the Middle East and North Africa (MENA), language barriers can pose a significant challenge when trying to access reliable climate data and fact-check global climate claims. Much of the high-quality, research-based climate information, including academic papers, governmental reports, and international climate agreements, is predominantly published in English. This creates a knowledge gap for those who speak Arabic or other regional languages. As a result, fact-checkers in the MENA region may have to rely on secondary sources, which may not always be as trustworthy or comprehensive. The scarcity of reliable climate resources in Arabic also limits the ability of local journalists to engage in informed discussions or challenge climate disinformation effectively. Moreover, this language gap can hinder collaboration between international organizations and regional actors who are fighting climate misinformation, limiting the sharing of best practices and resources across different linguistic communities. Overcoming this barrier requires increasing access to climate data in local languages and providing specialized training for journalists in non-English-speaking regions.

**5. Technical Skills and Empowerment**

Dealing with unstructured data, such as social media posts, online articles, and video content, requires significant technical skills that many journalists and fact-checkers may not possess. This unstructured data must be reclassified, cleaned, and processed before it can be analyzed effectively. The complexities of this data often require specialized knowledge in data science, such as experience with data scraping, analysis through APIs, and employing machine learning algorithms to identify patterns in misinformation. Many organizations and fact-checking initiatives may not have the resources or technical capacity to process and analyze the vast amount of unstructured data generated daily. This lack of technical skills can lead to inefficiencies in identifying and countering disinformation campaigns that proliferate during major climate conferences or events. To effectively combat climate misinformation, it is crucial to empower journalists and fact-checkers with the necessary technical tools, training, and support to process large datasets and identify misleading or false claims before they gain widespread traction.

**6. Financial Constraints**

The financial constraints faced by many independent journalists, fact-checkers, and small organizations make it difficult to access the necessary tools and resources to effectively combat climate disinformation. Many of the advanced technical tools used for monitoring and analyzing disinformation, such as data scraping software, monitoring platforms, and in-depth analytical tools, come with high subscription costs or require expensive technical expertise to use. For smaller organizations with limited budgets, this financial barrier can prevent them from accessing the full range of tools required to track climate-related misinformation. Furthermore, running fact-checking operations, especially during global climate conferences, demands significant financial resources for staffing, research, and outreach efforts. In many cases, independent journalists and smaller fact-checking initiatives are left relying on freely available, yet less effective, tools, which may not provide the comprehensive insights needed to address the sophisticated disinformation campaigns often deployed during such events. Addressing this financial challenge requires greater funding support for independent journalism and fact-checking initiatives to ensure that the fight against climate misinformation is equitable and accessible to all.

## Further readings


- Vu, Hong & Baines, Annalise & Nguyen, Nhung. (2022). Fact-Checking Climate Change: An Analysis of Claims and Verification Practices by Fact-Checkers in Four Countries. Journalism & Mass Communication Quarterly. 100. 10.1177/10776990221138058.
- Jon Bateman and Dean Jackson (2024). Countering Disinformation Effectively: An Evidence-Based Policy Guide. Carnegie Endowment for International Peace. Available at: https://carnegieendowment.org/research/2024/01/countering-disinformation-effectively-an-evidence-based-policy-guide?lang=en
- Gertrudix M, Carbonell-Alcocer A, Arcos R, Arribas CM, Codesido-Linares V, Benítez-Aranda N. Disinformation as an obstructionist strategy in climate change mitigation: a review of the scientific literature for a systemic understanding of the phenomenon. Open Res Eur. 2024 Sep 24;4:169. doi: 10.12688/openreseurope.18180.2. PMID: 39399659; PMCID: PMC11467642.
- Hassan, I., Musa, R. M., Latiff Azmi, M. N., Razali Abdullah, M., & Yusoff, S. Z. (2024). Analysis of climate change disinformation across types, agents and media platforms. Information Development, 40(3), 504-516. https://doi.org/10.1177/02666669221148693
- Vu, H. T., Baines, A., & Nguyen, N. (2023). Fact-Checking Climate Change: An Analysis of Claims and Verification Practices by Fact-Checkers in Four Countries.

Journalism & Mass Communication Quarterly, 100(2), 286-307.
https://doi.org/10.1177/10776990221138058

Chapter VII

# Guidelines and Warnings for Using Artificial Intelligence in Data Journalism

## Introduction

Artificial intelligence has revolutionized the way data is collected, processed, and analyzed, making it an invaluable tool in modern journalism, particularly in the realm of data journalism. Big data, a term used to describe vast datasets that are too complex for traditional data-processing software to handle, has become a central focus of journalistic investigations. The increasing amount of data generated by society—from social media interactions to government databases—provides new opportunities for journalists to uncover patterns and insights that were previously unimaginable. AI tools allow journalists to sift through enormous volumes of data, identify trends, and extract relevant information in a fraction of the time it would take a human to do so manually.

The process of working with big data involves several stages, each of which requires different technological tools and skills. Initially, the data is collected from a variety of sources, such as public records, social media platforms, sensor networks, and proprietary databases. Once the data is collected, it must be cleaned and preprocessed to remove inconsistencies, duplicates, and irrelevant information. This stage is essential for ensuring that the data is accurate and usable for further analysis. Following this, advanced AI algorithms, particularly machine learning techniques, can be employed to analyze the data, identify patterns, and make predictions. These AI tools enable journalists to identify correlations and trends that may not be immediately apparent through traditional investigative methods.

AI and machine learning technologies have opened up new possibilities for data-driven storytelling. Journalists can now create in-depth, interactive reports that go beyond simple graphs and charts. By using AI, they can generate visualizations that dynamically represent complex datasets, making the information more accessible to the public. Machine learning

tools can also automate parts of the storytelling process, such as generating summaries or detecting anomalies in data that warrant further investigation. Moreover, AI tools can be used to support investigative journalism by identifying discrepancies or uncovering hidden connections between seemingly unrelated data points, thus providing new avenues for reporters to pursue in their stories.

However, despite the transformative potential of AI in journalism, there are challenges to its effectiveness. One of the main concerns is the quality and reliability of the data. Big data is not always accurate or representative, and poorly sourced data can lead to misleading conclusions. Additionally, AI models are only as good as the data they are trained on, and biased or incomplete datasets can perpetuate existing inequalities or inaccuracies. Furthermore, journalists must be careful to avoid over-relying on AI tools, as they may not always be able to provide the nuanced understanding that a human journalist can bring. There is also the risk that AI-generated content could be manipulated or misinterpreted, leading to the spread of misinformation.

Ethical considerations also play a significant role in the use of AI in journalism. The ability to collect and analyze vast amounts of personal data raises privacy concerns, as individuals may not be aware of how their data is being used. Moreover, the use of AI to make editorial decisions, such as curating news content or selecting which stories to cover, may lead to algorithmic biases, reinforcing certain viewpoints or perspectives while neglecting others. To address these ethical challenges, journalists must adhere to strict guidelines and transparency practices when using AI tools. This includes being transparent about how data is collected, ensuring that data sources are reliable, and taking steps to mitigate biases in AI algorithms. By doing so, journalists can harness the power of AI while maintaining the integrity and trustworthiness of their work.

## What is Big Data and How Do We Analyze It?

Big Data refers to an enormous volume of data generated continuously from various sources, including social media, IoT devices, sensors, transactional records, and much more. These datasets are not only vast but also highly diverse, encompassing structured, unstructured, and semi-structured data. Structured data typically resides in databases with predefined formats, while unstructured data includes text, images, and video, making it more challenging to analyze. Semi-structured data lies in between, having some organization but not conforming to a rigid structure, such as XML or JSON files. With the sheer scale of Big Data, conventional data management tools struggle to process or

analyze it effectively. The scale of Big Data is often measured in petabytes (1 petabyte equals 1 million gigabytes) or exabytes, and the complexity of analyzing it requires specialized computational techniques.

Artificial Intelligence (AI) plays a critical role in processing and making sense of Big Data. AI encompasses various technologies designed to enable machines to replicate human-like cognitive functions such as learning, understanding, and problem-solving. These capabilities make AI a crucial asset in Big Data analytics, as it allows computers to not only process vast quantities of information but also extract meaningful insights and make decisions. AI systems work by interpreting data, identifying patterns, and adapting their responses accordingly. Machine Learning (ML), which is a subset of AI, involves creating algorithms that allow systems to learn from data and automatically improve over time without needing explicit programming for each specific task.

Machine Learning (ML) has proven especially valuable in dealing with Big Data because it provides systems with the ability to analyze and interpret data autonomously, identifying patterns and trends that may not be immediately apparent. In ML, algorithms are trained on large datasets to make predictions or identify relationships within the data. The more data these algorithms process, the better they become at making accurate predictions or providing insights, essentially "learning" from the data. For example, a machine learning model can be used to predict consumer behavior by analyzing past purchase patterns and identifying common factors that lead to certain outcomes. This ability to recognize hidden patterns within data without manual intervention is one of the key reasons AI and ML are indispensable in analyzing Big Data.

Analyzing Big Data requires advanced computational techniques due to the massive volume and complexity of the information involved. Traditional data processing methods and tools are often insufficient to handle the scale of Big Data, making it necessary to leverage AI-powered solutions. AI-driven tools can automate many of the processes involved in analyzing large datasets, significantly reducing the time and resources required for human intervention. In addition, AI algorithms can handle the dynamic nature of Big Data, continuously adapting as new data is generated. This adaptability is especially valuable in fields like real-time data analytics, where the ability to make immediate decisions based on the most current information is crucial.

One of the key challenges in Big Data analytics is dealing with the variety of data formats and sources. AI techniques, particularly machine learning, can be used to handle and make sense of unstructured data, such as text and images, which are typically difficult for traditional systems to process. By training machine learning models to recognize patterns in unstructured data, AI can extract useful information, categorize it, and provide valuable

insights. For instance, AI-powered tools can process vast amounts of social media content, identifying sentiment and key topics from text, and even analyzing images or videos for specific objects or patterns. This capability enables organizations to gather insights from data that might otherwise be overlooked or inaccessible.

Despite the power of AI and machine learning in analyzing Big Data, several challenges remain. First, the sheer scale of Big Data can lead to computational issues, as processing such vast amounts of data requires significant processing power and storage capacity. Additionally, the quality of the data being analyzed is crucial—poor-quality data can result in inaccurate insights, even when advanced AI techniques are used. Ensuring that data is clean, accurate, and representative of the real-world scenario it is intended to model is a continuous challenge. Moreover, the ethical implications of AI and machine learning must be considered, especially when personal or sensitive data is involved. Biases in AI algorithms, privacy concerns, and the potential for misuse of data are all important considerations that must be addressed as AI is increasingly utilized to process Big Data.

## What Are the Stages of Big Data Processing?

**1. Data Collection**

Data collection is the crucial first step in any data analysis process. It involves gathering raw data from various sources, which can include sensors, surveys, online platforms, databases, and public records, among others. In today's digital world, data is continuously generated and captured in vast quantities. However, collecting this data involves much more than just retrieving it; it also includes identifying relevant datasets, selecting the right data sources, and extracting the required information. Once the data is gathered, it is usually organized and transformed into a standardized format that makes it compatible with other datasets for the next stages of processing. The sheer volume and variety of data make this stage a complex one, as multiple tools and technologies, including web scraping, APIs, and IoT devices, are required to collect diverse types of data efficiently.

Additionally, the collection phase must ensure that the data gathered is relevant and accurate. This includes filtering out data that might be irrelevant or that could introduce noise into the subsequent analysis. Advanced algorithms, often supported by AI, can help automate this process by identifying high-quality data sources and eliminating unwanted data, thus ensuring that only valuable information is extracted. Furthermore, real-time data collection has become a pivotal aspect of many industries, especially for applications in areas like financial services, healthcare, and climate monitoring. The collection of real-

time data requires sophisticated tools and systems that can process incoming data quickly and ensure that it is immediately available for analysis or decision-making, which presents its own unique set of challenges regarding data storage, transmission, and security.

**2. Data Storage**

Once data is collected, it must be stored in a way that allows easy access, management, and processing. This storage can take place on various platforms, either on the cloud or on physical devices, depending on the specific needs of the organization and the data itself. Cloud storage has become increasingly popular due to its scalability, cost-effectiveness, and ability to support the storage of vast datasets. Platforms like Amazon Web Services (AWS), Google Cloud, and Microsoft Azure offer robust solutions that allow for secure and flexible storage of data. Cloud storage also facilitates collaboration, as data can be accessed and updated by multiple users in different geographical locations in real-time. Additionally, data is often stored in databases that are structured to allow efficient querying and retrieval, making it easier to work with large datasets.

On the other hand, physical storage options such as hard drives, solid-state drives, and data centers are also still used in certain circumstances, particularly when dealing with sensitive information or data that requires high-speed access and retrieval. In these cases, maintaining on-premise storage solutions might provide a higher level of control and security. For instance, some industries such as healthcare or finance may require physical storage solutions for compliance reasons or for storing particularly sensitive data that cannot be stored in the cloud. Regardless of the storage medium used, data storage must be managed carefully to ensure that the data remains accessible and secure. This requires implementing strong data governance policies, backup systems, and encryption mechanisms to protect the data from unauthorized access or loss.

**3. Data Processing**

Data processing involves transforming raw data into a format that is consistent, usable, and ready for analysis. This phase is crucial for ensuring that data is cleaned, organized, and formatted correctly before it undergoes deeper analysis. Data processing can include several steps, such as converting unstructured data into structured formats, normalizing data for uniformity, and aggregating data from different sources into a central repository. For large datasets, this step often requires the use of automation tools to handle the massive volumes of data. One of the most important aspects of this process is data transformation, where complex or raw data is changed into a structured form that aligns with analytical models or frameworks. Techniques such as data enrichment or data

normalization might be used to ensure that the processed data fits a consistent schema, making it easier to extract insights.

A prime example of data processing in action can be seen in the collaborative Pandora Papers project led by the International Consortium of Investigative Journalists (ICIJ). For this investigation, machine learning was employed to assist in classifying the data and excluding irrelevant information, which would have otherwise been overwhelming for human analysts. The clustering technique, which is a part of machine learning, was used to group similar data points together and categorize them effectively. This allowed journalists and researchers to sift through large volumes of documents and identify key relationships, ultimately making it easier to understand the complex financial structures involved in offshore banking and tax evasion. By leveraging data processing and machine learning techniques, the ICIJ team was able to process vast amounts of information efficiently and uncover crucial insights that would have been difficult to find through traditional manual methods.

**4. Data Cleaning**

Data cleaning is the stage where errors, inconsistencies, and irrelevant information are removed from the dataset to ensure the accuracy and quality of the data used in analysis. This step is vital because even the most sophisticated analytical models can produce misleading results if the data is flawed. Data cleaning involves identifying and addressing issues such as missing values, duplicates, outliers, or contradictory data. Given the complexity and scale of datasets, cleaning them manually is often impractical. Instead, machine learning algorithms are increasingly employed to automate the cleaning process, saving time and resources while improving the accuracy of the data. Machine learning models can detect inconsistencies, errors, or patterns that are often invisible to human analysts and can correct them in real time.

An illustrative example of machine learning’s application in data cleaning can be seen in the investigation conducted by the Los Angeles Times, titled "LAPD underreported serious assaults, skewing crime stats for 8 years." In this investigation, machine learning was used to clean and process a large dataset containing crime records, revealing that approximately 14,000 serious assault incidents had been misclassified as minor crimes. This misclassification had led to an inaccurate portrayal of crime levels in the city. By applying machine learning techniques to clean and process the data, the team was able to identify the errors and correct the public records, providing more accurate information for law enforcement and policy decisions. This example highlights how machine learning is transforming the process of data cleaning by enabling the automation of error detection

and correction, significantly improving the quality and reliability of data for investigative reporting.

**5. Data Analysis**

The data analysis phase involves applying statistical, machine learning, or AI techniques to the cleaned and processed data to extract meaningful insights and draw conclusions. At this stage, data scientists or analysts often use advanced algorithms and visualization tools to uncover patterns, trends, and relationships in the data. These insights are then presented in a way that is easy to understand and actionable, often using graphs, charts, or other forms of data visualization. The analysis phase is essential because it enables organizations to make informed decisions based on evidence and data-driven insights. AI techniques, such as predictive modeling or sentiment analysis, can be used to generate forecasts or gauge public opinion on various topics by analyzing historical data.

One of the tools that plays a critical role in data analysis is Google BigQuery, which is particularly useful for analyzing large datasets, such as those found in government reports or social data. BigQuery enables users to process and query massive datasets efficiently and allows analysts to perform complex queries in a fraction of the time that would be required by traditional tools. Additionally, specialized tools like the Natural Language Toolkit (NLTK) are used to analyze text data by extracting keywords, conducting sentiment analysis, and processing linguistic patterns. NLTK, for example, can be used to analyze news databases or social media content to detect the sentiment of posts or identify the presence of certain topics. By using these tools, data analysts are able to derive valuable insights from vast amounts of unstructured data, helping organizations make more informed decisions and uncover hidden insights that drive their strategies.

## Key AI Tools for Big Data Analysis

**1. Google Cloud AutoML**

Google Cloud AutoML is a cloud-based tool designed to allow users with limited expertise in data science to build custom machine learning models. This tool provides an easy-to-use interface that guides users through the process of training models tailored to specific datasets. For journalists, especially those without deep technical backgrounds, AutoML is a powerful tool for recognizing patterns in text, images, and other types of data. For example, journalists can use AutoML to analyze large volumes of text data to identify key topics or sentiment trends. By automating the model-building process, AutoML helps

streamline workflows, enabling journalists to quickly analyze data and extract valuable insights without the need for specialized knowledge in machine learning or big data analytics. This tool is particularly beneficial for newsrooms or independent journalists looking to dive into data-driven stories with minimal training.

**2. H2O.ai**

H2O.ai is an open-source platform that is highly valuable for big data analysis and predictive modeling. It is used by organizations across industries to analyze large datasets and predict future trends, making it an ideal tool for journalists covering complex topics such as economics, politics, and elections. H2O.ai supports various machine learning algorithms, including generalized linear models, deep learning, and gradient boosting, enabling journalists to conduct sophisticated analysis without requiring deep knowledge of data science. For instance, a journalist covering a national election could use H2O.ai to analyze voting trends, forecast election outcomes, or detect any significant shifts in political sentiment. The platform's ability to handle large volumes of data and provide interpretable models ensures that journalists can leverage its power while remaining focused on producing compelling and data-driven stories.

**3. IBM Watson Studio**

IBM Watson Studio is an AI-powered tool that provides journalists with the ability to analyze social media content, identify audience trends, and predict future interests. By utilizing machine learning and natural language processing (NLP) capabilities, Watson Studio helps journalists better understand the dynamic nature of social media conversations and public sentiment. For journalists involved in digital media or content promotion, Watson Studio can be an invaluable tool for tracking which topics are gaining traction and developing tailored content strategies. This enables journalists to craft stories that resonate with their target audiences, ultimately improving engagement and visibility. Additionally, Watson Studio's collaboration features allow journalists, editors, and data analysts to work together on projects, ensuring seamless integration of AI-driven insights into editorial strategies.

**4. Amazon SageMaker**

Amazon SageMaker is a cloud-based platform that simplifies the process of training and deploying machine learning models at scale. It is particularly suited for journalists engaged in investigative reporting, as it offers tools that automate many of the tasks involved in data analysis, from data cleaning to model deployment. By using SageMaker, journalists can train custom models to analyze large, complex datasets, such as financial records, government reports, or social media data, to uncover patterns or inconsistencies. For

example, a journalist investigating a corruption case could use SageMaker to detect anomalies in financial transactions or uncover hidden relationships between entities. Its scalable infrastructure allows journalists to handle vast amounts of data, enabling them to conduct in-depth investigations that would otherwise be time-consuming and difficult to execute manually.

**5. PyOD Library**

PyOD is an open-source Python library designed for anomaly detection, making it an essential tool for investigative journalism, especially in fields like finance and corruption reporting. The library includes a wide range of algorithms that are optimized for detecting unusual patterns, errors, or outliers in datasets. For journalists focusing on financial tracking, PyOD allows them to identify irregular transactions or potential fraud by comparing large datasets against typical behavior patterns. Its versatility makes it highly effective in uncovering hidden anomalies in data, such as suspicious financial transfers, fraudulent activities, or discrepancies in public financial records. By leveraging PyOD, investigative journalists can dig deeper into complex data, uncovering critical insights that may otherwise remain unnoticed, and providing a foundation for evidence-based reporting.

**6. Isolation Forest**

Isolation Forest is a machine learning technique specifically designed for anomaly detection in high-dimensional datasets. It is particularly effective at identifying outliers or unusual patterns within data, making it a powerful tool for journalists working with election-related data or any type of time-sensitive media content. For example, journalists can use Isolation Forest to spot inconsistencies in voting records, social media trends, or other forms of digital data that may indicate fraudulent activity or manipulation. By isolating and flagging anomalous data points, this tool helps journalists identify potential issues quickly and accurately. Given its ability to process large volumes of data efficiently, Isolation Forest is especially useful for media outlets that need to analyze vast amounts of information in real-time, such as during live election coverage or when tracking rapidly evolving news stories.

**7. Tableau**

Tableau is a powerful data visualization tool that allows journalists to create interactive and engaging visualizations from complex datasets. Its intuitive interface makes it accessible to users with limited technical knowledge, which is especially useful in a journalistic setting where the primary goal is often to communicate findings clearly to a broad audience. With Tableau, journalists can turn raw data into compelling visual stories, such as graphs, charts, maps, and dashboards, which are crucial for illustrating trends and patterns in a

way that is easy for the public to understand. For example, in a story about income inequality, a journalist can use Tableau to create interactive maps that show income distribution across different regions or demographics, making it easier for readers to visualize complex social issues. Tableau also allows journalists to explore data in real-time, providing dynamic visualizations that update as new data becomes available.

**8. Power BI by Microsoft**

Power BI is another advanced data visualization tool that enables journalists to create detailed, interactive reports for presenting complex findings. Like Tableau, Power BI is highly valued for its ability to transform raw data into meaningful visualizations that simplify the communication of intricate stories. What sets Power BI apart is its deep integration with Microsoft products, such as Excel and Azure, allowing journalists to seamlessly connect and analyze data across various platforms. It is widely used in digital journalism for reporting on topics such as economics, politics, and social trends, where large datasets are often involved. Journalists can use Power BI to generate comprehensive dashboards and reports that summarize key findings, making it easier for their audiences to understand the implications of the data. Additionally, Power BI's interactive features enable readers to engage with the data themselves, fostering a more immersive and personalized experience.

## Challenges of Using AI in Big Data Journalism

**1. High Costs**

The cost of processing big data remains one of the largest barriers to its use, particularly for local newsrooms and smaller initiatives. Processing large volumes of unstructured data requires advanced computing systems, powerful servers, and specialized software that can handle data at scale. These tools often come with significant licensing fees, subscription costs, and the need for ongoing maintenance, which can put them out of reach for media outlets operating with limited budgets. For smaller organizations, these high costs often lead to trade-offs in which they must forgo advanced data analytics or rely on less effective, manual methods for data analysis. This creates an imbalance where larger, well-funded organizations have the technological edge, leaving smaller players struggling to keep pace in a landscape where data-driven reporting is increasingly becoming the norm. Without access to these advanced systems, many newsrooms are unable to fully harness the power of big data, limiting their ability to produce comprehensive investigative journalism and provide in-depth reporting that can compete with bigger outlets.

**2. Storage and Data Security Issues**

The challenge of storing and securing big data is especially important when dealing with sensitive or personal information. Safely storing vast amounts of data requires sophisticated technologies such as cloud storage solutions, on-premise data centers, and high-performance computing systems. Moreover, the need for skilled technical staff to manage these systems adds another layer of complexity. Ensuring that data remains secure from breaches, theft, or misuse requires stringent security protocols and the implementation of encryption techniques that many smaller media outlets might not have the resources to manage effectively. For instance, investigative journalists working with confidential sources or sensitive governmental data are particularly vulnerable to data theft, which could jeopardize both their safety and the credibility of their reporting. Additionally, the increasing sophistication of cyberattacks means that media organizations must invest in high-level security measures to protect their data, which can place an added financial burden on outlets already struggling with limited resources.

**3. Arabic Language Challenges**

A significant hurdle in the use of AI tools for data journalism in the Middle East and North Africa is the lack of support for Arabic language processing. Most AI tools are designed and trained on datasets primarily in English, which creates a gap when it comes to applying these technologies to Arabic data. Arabic, along with its many dialects, presents unique challenges due to its complex grammatical structures, variations in regional vocabulary, and right-to-left script. While there has been some progress in Arabic Natural Language Processing (NLP), the technology still lags behind English in terms of accuracy and available resources. As a result, journalists in the MENA region may struggle to effectively use AI tools to analyze Arabic-language content, such as social media posts, news articles, and government reports. This discrepancy not only limits the potential of data-driven journalism in Arabic-speaking countries but also creates a barrier to providing accurate and timely information to the public. In an era where technology is crucial for investigative reporting, this linguistic divide poses a significant challenge for Arabic-language journalists and fact-checkers.

**4. AI Limitations**

While AI has made significant strides in data analysis, it still faces limitations, especially when it comes to efficiently collecting and representing unstructured data. AI models often struggle with interpreting human language, especially when it is ambiguous, informal, or contains cultural nuances. For example, AI may misinterpret sarcasm, idiomatic expressions, or slang in social media posts or news articles, leading to inaccurate

conclusions. Additionally, unstructured data such as images, videos, and audio files require sophisticated models to extract meaningful insights, but these models are still in the developmental stage, and their accuracy can vary widely. As a result, AI tools often require careful human oversight to validate insights, especially when analyzing data in complex or sensitive contexts. For example, an AI tool might flag certain social media posts as related to a political event, but a human journalist must verify the context and intent behind these posts before drawing conclusions. This ongoing need for human intervention underscores the current limitations of AI in handling complex, unstructured data and highlights the importance of maintaining journalistic oversight in the age of automation.

**5. Limited Access to Open Government Data**

Access to open government data remains a significant issue in many developing countries, particularly in regions like Asia and Africa. In many instances, governments are reluctant to release public data, citing concerns over national security, political instability, or privacy issues. A 2017 survey by the Arab Data Journalists' Network revealed that a large majority of journalists in Arab countries found accessing government data difficult, with over 70% describing the process as either challenging or highly complex. This lack of transparency and access to data severely limits the ability of journalists to investigate public policies, expose corruption, or analyze economic trends. Without reliable, open access to government datasets, journalists are forced to rely on alternative, often less reliable sources, which compromises the integrity of their reporting. The absence of open data also exacerbates the issue of accountability, as journalists and civil society organizations lack the tools necessary to hold governments and institutions accountable for their actions. Moreover, the lack of accessible data impedes efforts to foster evidence-based reporting, a crucial aspect of investigative journalism in the modern era.

## Ethical challenges

**1. Algorithmic Bias**

Algorithmic bias is one of the most significant challenges faced by journalists and researchers utilizing AI in data-driven reporting. Machine learning algorithms can unintentionally perpetuate and amplify existing societal biases, particularly when they are trained on datasets that do not adequately represent all populations. For instance, algorithms that classify or analyze data based on factors like race, gender, or age can develop biased outcomes if the data sets they are trained on are skewed. In many cases, datasets lack sufficient representation of marginalized groups, such as people of color,

women, or individuals with nontraditional gender identities, leading to inaccurate, discriminatory results. This problem is especially prominent in countries in the Global South and the Middle East and North Africa region, where there is often limited access to high-quality, diverse data for training machine learning models. The absence of inclusive training data can lead to systemic biases in AI models, which, when applied in journalism, could perpetuate misinformation or misrepresentation, affecting public perceptions and further entrenching stereotypes. As a result, the reliance on AI and machine learning tools in data journalism must be approached with caution, ensuring that journalists are aware of the potential biases in their models and actively work to mitigate these risks.

**AI Hallucinations**

AI hallucinations, a phenomenon where machine learning models generate incorrect or nonsensical outputs, represent another critical challenge for data journalism. This occurs when AI systems fail to comprehend or misinterpret the inputs provided to them, often resulting in errors that can undermine the credibility of the work produced. For example, an AI tool might misclassify data or produce conclusions that do not align with the input data, leading to erroneous or misleading insights. This can happen if the AI model was trained on incomplete or insufficient data, or if the input data does not conform to the patterns the model has learned. In investigative journalism, where accuracy and reliability are paramount, AI hallucinations can severely damage the integrity of an investigation. Incorrect results can also occur when an AI system is faced with ambiguous questions or commands that it cannot process accurately. Journalists using AI tools must remain vigilant about these potential errors, understanding the limitations of the models and cross-checking AI-generated results with human expertise to ensure the conclusions drawn are accurate and reliable. This highlights the importance of human oversight in the AI-driven data analysis process.

**3. Data Accessibility and Corporate Monopoly**

Access to big data and machine learning tools is often restricted by corporate monopolies, which pose significant barriers for independent journalists and smaller media organizations. Many of the most powerful and user-friendly tools for analyzing large datasets and running machine learning algorithms are proprietary and owned by major corporations, such as Google, Amazon, or Microsoft. These tools are often available through expensive subscription models, making them inaccessible to journalists and organizations with limited financial resources. Additionally, these corporations often obscure the methodologies behind their data collection and usage, making it difficult for users to understand how the data was gathered, processed, or analyzed. For freelance journalists, smaller investigative teams, and independent outlets, this lack of transparency

and the high cost of accessing critical data and AI tools can hinder their ability to carry out thorough, data-driven investigations. In a landscape where corporate entities hold significant power over the flow of information, data accessibility becomes a key issue that undermines the ability of smaller players to challenge dominant narratives, produce high-quality journalism, or contribute to public discourse on important issues.

**4. Lack of Transparency and Documentation**

Transparency is a critical issue in the use of AI for data journalism, particularly when algorithms are involved in making classifications, recommendations, or decisions that directly impact the outcomes of investigative work. Many AI models operate as "black boxes," meaning that the processes behind their decision-making are often not well-documented or easily understood. This lack of transparency can create significant challenges for journalists who rely on AI tools to analyze data, as they cannot fully comprehend or explain the reasoning behind the tool's outputs. Without clear documentation, it becomes difficult to assess the reliability and accuracy of the AI system, raising questions about the validity of the conclusions drawn from its analysis. This lack of clarity also makes it difficult to identify potential biases in the algorithm, which could affect the integrity of the data and its interpretation. For journalists working in investigative reporting, where transparency is essential to maintaining credibility, this lack of insight into the inner workings of AI tools can be a major obstacle. It calls for greater accountability and documentation from AI providers to ensure that journalists and media organizations can trust the tools they use and fully understand how their data is being analyzed.

**5. Data Privacy**

Data privacy is an increasingly important issue in the era of AI and big data, particularly when AI tools rely on anonymized data for analysis. While anonymization can help protect personal identities, it does not eliminate the risk of privacy violations, especially when sensitive personal data, such as opinions, behavioral patterns, or social media activity, is being used without explicit consent. Many AI tools, particularly those used for social media monitoring or sentiment analysis, collect vast amounts of data from individuals, often without informing them or securing their consent. This raises serious ethical concerns about how personal data is handled, who owns it, and how it is being used in the context of journalism. In regions like Europe, the implementation of the General Data Protection Regulation (GDPR) has set stricter guidelines for handling personal data, ensuring that individuals are informed about how their data is being collected and used, and giving them the right to access, correct, and delete their data. However, in the Middle East and North Africa, digital rights legislation is still underdeveloped, leaving little protection for individuals' privacy. Furthermore, machine learning providers often fail to provide

transparent reports explaining how data is collected, processed, or analyzed, further complicating the issue of privacy. Journalists must be aware of these privacy concerns when utilizing AI tools in their reporting and ensure that they are following ethical standards that prioritize the protection of individuals' personal data.

Despite these challenges, the responsible use of AI tools remains essential, particularly in analyzing big data that is difficult to handle manually. Properly collecting and analyzing such data can yield insights that serve the public interest.

## Further readings


- Mahony, S., & Chen, Q. (2024). Concerns about the role of artificial intelligence in journalism, and media manipulation. Journalism, 0(0). https://doi.org/10.1177/14648849241263293
- Becker, K. B., Simon, F. M., & Crum, C. (2025). Policies in Parallel? A Comparative Study of Journalistic AI Policies in 52 Global News Organisations. Digital Journalism, 1–21. https://doi.org/10.1080/21670811.2024.2431519
- Trang, T. T. N., Chien Thang, P., Hai, L. D., Phuong, V. T., & Quy, T. Q. (2024). Understanding the Adoption of Artificial Intelligence in Journalism: An Empirical Study in Vietnam. Sage Open, 14(2). https://doi.org/10.1177/21582440241255241
- Broussard, Meredith & Diakopoulos, Nicholas & Guzman, Andrea & Abebe, Rediet & Dupagne, Michel & Chuan, Ching-Hua. (2019). Artificial Intelligence and Journalism. Journalism & Mass Communication Quarterly. 96.
- Cools, H., & Diakopoulos, N. (2024). Uses of Generative AI in the Newsroom: Mapping Journalists' Perceptions of Perils and Possibilities. Journalism Practice, 1–19. https://doi.org/10.1080/17512786.2024.2394558
- Noain-Sánchez, Amaya. (2022). Addressing the Impact of Artificial Intelligence on Journalism: the perception of experts, journalists and academics. Communication & Society. 35. 105-121.

**Appendix 1**

# Digital Rights: Protecting Privacy, Freedom, and Access in the Digital Age

In today's increasingly digital world, the protection of individual rights and freedoms extends beyond physical boundaries and into the virtual realm. Digital rights encompass the various freedoms and protections that individuals should have when interacting with technology and the internet. These rights are vital in safeguarding privacy, ensuring access to information, and preserving fundamental freedoms in the digital era. However, the rise of digital technologies has created new challenges in maintaining these rights, as governments, corporations, and other entities gain more power over personal data, online activity, and digital spaces. The conversation surrounding digital rights is thus more important than ever as we navigate an age defined by rapid technological advancement.

## Privacy in the Digital Age

Privacy has long been a cornerstone of human rights, but with the growth of the internet, the concept of privacy has evolved in new and complex ways. Personal information, once contained in physical files, is now stored in vast databases, often without the explicit consent or knowledge of the individuals it pertains to. In the digital age, individuals' personal data—such as location, browsing habits, social media interactions, and even health data—can be easily collected, shared, and exploited.

For example, tech giants like Google, Facebook, and Amazon accumulate immense amounts of personal data through their services and platforms. This data is often used to build detailed profiles of users, which can be leveraged for advertising, political manipulation, or even surveillance. The use of data analytics, artificial intelligence, and machine learning algorithms has raised significant concerns about privacy violations and the loss of control over personal information. Digital rights, therefore, play a central role in ensuring that individuals retain control over their data and are informed about how it is being collected and used.

Legislation such as the European Union's General Data Protection Regulation (GDPR) has made strides in giving individuals greater control over their data, mandating that organizations gain explicit consent from users before collecting or processing their personal information. The GDPR also grants individuals the right to access, correct, and delete their data, providing much-needed transparency and accountability. However, in many regions around the world, such protections are lacking, and individuals remain vulnerable to privacy breaches and data exploitation. Digital rights must include robust privacy protections that prevent the unauthorized collection and use of personal data and ensure that individuals have the power to control what is shared online.

## Freedom of Expression and Access to Information

One of the key components of digital rights is the right to freedom of expression in the online space. Just as individuals have the right to express themselves in physical public spaces, they should have the same freedoms in digital environments. This includes the right to share opinions, create content, and engage in discussions without fear of censorship or retribution.

However, the rise of digital platforms and social media networks has led to new challenges in upholding this right. Platforms like Twitter, Facebook, and YouTube play a central role in shaping public discourse and are often the first places people turn to for news and information. These companies, however, wield considerable influence over what content is seen and shared by users. The algorithms that determine what appears on a user's feed can prioritize certain content while suppressing others, influencing the flow of information. In many cases, these platforms have been criticized for censoring content, limiting freedom of expression, or allowing harmful disinformation to spread unchecked.

Moreover, governments across the world have increasingly turned to digital surveillance, using technology to monitor citizens' online activity and suppress dissent. In countries with authoritarian regimes, digital censorship and surveillance are often employed to limit access to independent news, stifle opposition voices, and suppress political activism. In this context, digital rights must ensure that individuals have access to unbiased, reliable information and the ability to communicate freely online.

Ensuring access to information is also a fundamental part of digital rights. In the modern age, access to the internet is essential for participating in many aspects of life, from education and healthcare to employment and social interaction. Yet, millions of people worldwide still lack reliable internet access, particularly in rural or underserved areas.

Bridging the digital divide and ensuring that everyone has access to affordable, high-speed internet is crucial for upholding digital rights and creating an equitable society.

## Digital Security and Protection Against Cyber Threats

Digital security is another critical component of digital rights, encompassing measures to protect individuals from cyber threats, including hacking, identity theft, and online harassment. As more personal, financial, and professional information is stored online, the potential for cyberattacks grows exponentially. These attacks can have devastating consequences, from financial loss to personal safety threats, and the rise of ransomware, phishing scams, and data breaches only heightens the risk.

While governments and organizations have a responsibility to protect individuals from cyber threats, the onus is often placed on individuals themselves to safeguard their own data. Passwords, two-factor authentication, and encryption are just a few of the measures that can be taken to secure digital information. However, not everyone has the resources or knowledge to protect themselves effectively.

Furthermore, surveillance technologies, including facial recognition and location tracking, pose significant risks to individuals' digital security. The widespread use of these technologies by both private corporations and governments raises serious questions about the right to remain anonymous and the potential for abuse. Digital rights must include protections against unauthorized surveillance and ensure that security measures are not used to infringe on individuals' freedoms.

## The Global Struggle for Digital Rights

While digital rights are recognized in many parts of the world, they are not universally protected or respected. In countries where digital freedoms are restricted or where authoritarian regimes hold power, individuals face significant challenges in accessing the internet, communicating freely, and protecting their privacy. In regions like the Middle East and North Africa, for example, governments often monitor and restrict online activity, suppressing free speech and censoring information.

The struggle for digital rights is also influenced by the growing power of multinational corporations. Tech giants that control the flow of information online hold significant influence over global digital landscapes. This concentration of power has led to calls for

greater regulation and oversight of these corporations, ensuring that they act in the public interest and do not infringe upon individuals' rights.

## Human rights and Internet

The intersection of human rights and the internet has become an increasingly important issue in the digital age. As more aspects of life move online, the internet has evolved from a tool for communication and information sharing to a fundamental space for exercising basic human rights, including freedom of expression, access to information, and privacy. However, the digital landscape also presents new challenges, such as the erosion of privacy through surveillance, censorship, and the exploitation of personal data. As a result, ensuring that human rights are respected and upheld online is crucial in safeguarding individual freedoms and promoting equity in a rapidly changing digital world. The ongoing debate over digital rights highlights the need for clear policies and international cooperation to protect these fundamental rights in an increasingly interconnected, technology-driven society.

**1. Global Network Initiative**

The Global Network Initiative (GNI) is an international coalition of human rights organizations, civil society groups, companies, and investors that works to protect and advance freedom of expression and privacy rights in the digital age. Launched in 2008, the GNI aims to promote the responsible conduct of technology companies by creating a framework that guides them in addressing government requests for user data and content censorship. The initiative provides a platform for companies to collaborate with human rights advocates to ensure that their operations align with international human rights standards, particularly when operating in countries with restrictive or repressive environments.

Through its set of principles and a rigorous assessment process, the GNI holds companies accountable for their actions related to freedom of expression and privacy, pushing them to resist unlawful or excessive government surveillance and censorship demands. The initiative's members include major global technology companies, such as Google and Facebook, as well as nonprofit organizations like Human Rights Watch and the Electronic Frontier Foundation. By fostering dialogue and providing best practices for companies to follow, the GNI contributes to the creation of a digital ecosystem where fundamental rights are respected, and users' privacy and freedom of speech are protected globally.

**2. The Internet Bill of Rights and the Charter on Internet Rights and Principles**

The Internet Bill of Rights and the Charter on Internet Rights and Principles were both introduced by the Internet Governance Forum (IGF) to address the growing need for protections in the digital space. These documents aim to establish and protect fundamental rights for users of the internet globally, emphasizing principles of equality, privacy, freedom of expression, and access to information. The goal is to safeguard users against potential abuses by both private entities and governments while ensuring that the internet remains a platform for open, democratic participation. These charters reflect the increasing recognition that, just as individuals are entitled to human rights offline, they should also enjoy similar protections in the digital realm.

The Internet Bill of Rights focuses on ensuring that all internet users have access to a set of basic rights that align with universal human rights frameworks. It calls for a commitment to net neutrality, which ensures that internet service providers (ISPs) treat all data on the internet the same way, without discrimination. Additionally, it advocates for strong privacy protections, including user consent for data collection and transparency about how personal information is used. The bill emphasizes that users should be free from undue surveillance and government censorship, ensuring that the internet remains a space where free expression and the exchange of ideas can thrive without restriction.

The Charter on Internet Rights and Principles extends these protections by outlining a more comprehensive set of guidelines for the governance of the internet. It advocates for inclusivity, ensuring that marginalized communities have equal access to online spaces. The charter also emphasizes the need for global cooperation to develop standards and regulations that protect digital rights across borders while respecting the sovereignty of nations. Furthermore, it stresses the importance of transparency, accountability, and ethical behavior in the development and deployment of technologies, urging internet companies and governments to act responsibly and fairly in their management of the online ecosystem. The document has become a key reference for discussions on internet governance, serving as a tool to promote human rights, digital equity, and responsible internet policies worldwide.

**3. The World Summit on the Information Society (WSIS)**

The World Summit on the Information Society (WSIS) was a global initiative launched by the United Nations, aimed at addressing the challenges and opportunities of the rapidly growing digital world. Held in two phases, the first phase took place in Geneva, Switzerland, in 2003, while the second phase was hosted in Tunis, Tunisia, in 2005. The summit brought together representatives from governments, civil society, the private sector, and international organizations to discuss the impact of information and communication technologies (ICTs) on global development. It provided a platform for

dialogue on how ICTs could contribute to achieving social, economic, and environmental goals, with a strong emphasis on bridging the digital divide between developed and developing nations.

The WSIS addressed a wide range of issues related to the use of ICTs, including access to information, internet governance, the digital divide, e-government, education, and the protection of privacy and freedom of expression. One of the key outcomes of the summit was the Geneva Declaration of Principles and the Geneva Plan of Action, which laid out the international community's commitment to building an inclusive information society. These documents emphasized the importance of ICTs in achieving sustainable development, reducing poverty, promoting human rights, and enhancing the quality of life for all people. They also highlighted the need for multilateral cooperation in ensuring that the benefits of ICTs were shared equitably across the globe.

During the second phase of WSIS, held in Tunis in 2005, the discussions focused on issues such as internet governance, the role of the private sector in promoting ICT development, and the need for international cooperation in managing the digital economy. The Tunis phase led to the establishment of the Internet Governance Forum (IGF), a multi-stakeholder platform for discussing internet-related policy issues. The WSIS process also helped shape global norms and standards on digital rights, recognizing the importance of protecting fundamental human rights in the digital environment, including freedom of expression and the right to privacy. While the summit made significant strides in recognizing the potential of ICTs for development, it also underscored the ongoing challenges in ensuring universal access and bridging the digital divide.

**Appendix 2**

# Arabi Facts Hub Database: How to Access and Make Use of It?

In 2021, Arabi Facts Hub launched its initiative which aims to create a comprehensive database that includes the efforts of content verification initiatives in the Arab world. It provides indicators to researchers on the state of misinformation in a region which is rife with conflict and negative propaganda.

In this article, the Arabi Facts Hub team sheds light on their database and provides guidance on how researchers, journalists, activists, and anyone interested in the field of fact-checking can benefit from it. The article also outlines the methodology used to build, update, and test the database.

## What is the Arabi Facts Hub database?

In mid-2022, after reviewing a study on the challenges posed by information disorder in the Global South, the team concluded that the majority of fact-checked information covers highly localized contexts and that fact-checking in the Arab world is often tied solely to academic research. There also emerged a need to create a comprehensive database on fact-checking in the Arab world, taking into account the history and roots of conflicts and the diverse contexts in this region, as well as the importance of continuously updating this database.

Arabi Facts Hub began creating this database by relying on data sourced from Arabic-language content verification initiatives, in addition to leveraging artificial intelligence for data processing and analysis to study and test data patterns on social media. The database is characterized by the following:

- Accumulation: The database is continuously updated.
- Focus on written content: In addition to expanding the scope of the data to include images and videos.
- Reliance on contributions from fact-checkers and fact-checking initiatives.
- Organization in displaying data and diversity; The data is organized and varied, with classifications based on topics such as politics, economics, social issues, to name a few, as well as geographical location and chronological order.

- Built and verified by a specialized team at Arabi Facts Hub, in collaboration with a group of partners who adhere to the standards of objectivity, integrity, independence, and neutrality.

## Data Collection Methodology

The first step in ensuring the reliability and credibility of the database is to establish criteria for selecting partners who provide the necessary data to the Arabi Facts Hub team. The criteria for choosing partners can be summarized as follows:

- **Independence**: The platforms must not be affiliated with governments or political factions.
- **Objectivity**: The content must be presented impartially and without bias.
- **Clear and Transparent Methodology**: There must be a clear and transparent working methodology.

The data undergoes a process within Arabi Facts Hub, starting from data collection, passing through classification, and ultimately being approved after ensuring its reliability.

**Phase One**: This involves collecting data either by extracting it in coordination with partner sites or by using a programming code that automatically retrieves the data, after obtaining the necessary permissions .

**Phase Two**: This is the data classification stage, which divides data into the following categories: Incorrect, Partially Incorrect, Satirical, Investigations, and Undefined.

**Phase Three**: Involves processing and evaluating the data, addressing informational and temporal gaps, and making the necessary recommendations to facilitate and streamline data handling and identification. This is based on factors such as time range, geographical location, and the type of fact-checked news. Filtering helps in quickly searching for values and controlling which data to view or exclude based on selected categories.

**Phase Four**: Making the comprehensive open-source database available on Arabi Facts Hub's website; to be utilized in various forms of research and journalistic production, such as reports, studies, and investigations.

## What does Arabi Facts Hub offer to its partners?

Arabi Facts Hub aims to foster fruitful cooperation with its partners and offers the following:

- Support in developing and revising data
- Training on using fact-checking tools, open sources, and artificial intelligence techniques
- Joint collaboration in producing reports and investigations
- Enabling local partners to showcase their research output and develop access mechanisms
- Technical support in importing research output published on specific sites
- Technical support in extracting data from the Arabi Facts Hub website using programming languages and tools.

## Steps to Access the Open-Source Database

The fact-checked information database is open-source, meaning it is available to any researcher, fact-checker, and data journalist. Interested individuals just need to follow these steps:

- **Create an Account**: Register on the Arabi Facts Hub website.
- **Send an Email**: To info@arabifactshub.com to confirm your account, provide details about your research idea, its scope, research questions, and a brief bio about the researcher.
- **Access Database**: Once approved, the researcher will be granted access to download a database containing all fact-checked news from Arabi Facts Hub and its partners.
- **Download**: The database can be downloaded in Excel format.
- **Use Keywords and Filters**: Researchers can use keywords to narrow and customize their search, as well as filter by time range and geographical location.
- **Interactive Map**: With a researcher account on the site, a researcher can benefit from the interactive map that displays data according to geographic location.
- **Additional Database**: Researchers receive another database for non-partners that includes IDs of fact-checked publications; researchers are responsible for importing and manually analyzing this content or using advanced techniques such as Python, R, SPSS, etc.
- **Research Requests**: Arabi Facts Hub allows researchers to submit research requests or request specific tools, such as Meltwater, and the team will provide the results to the researcher.

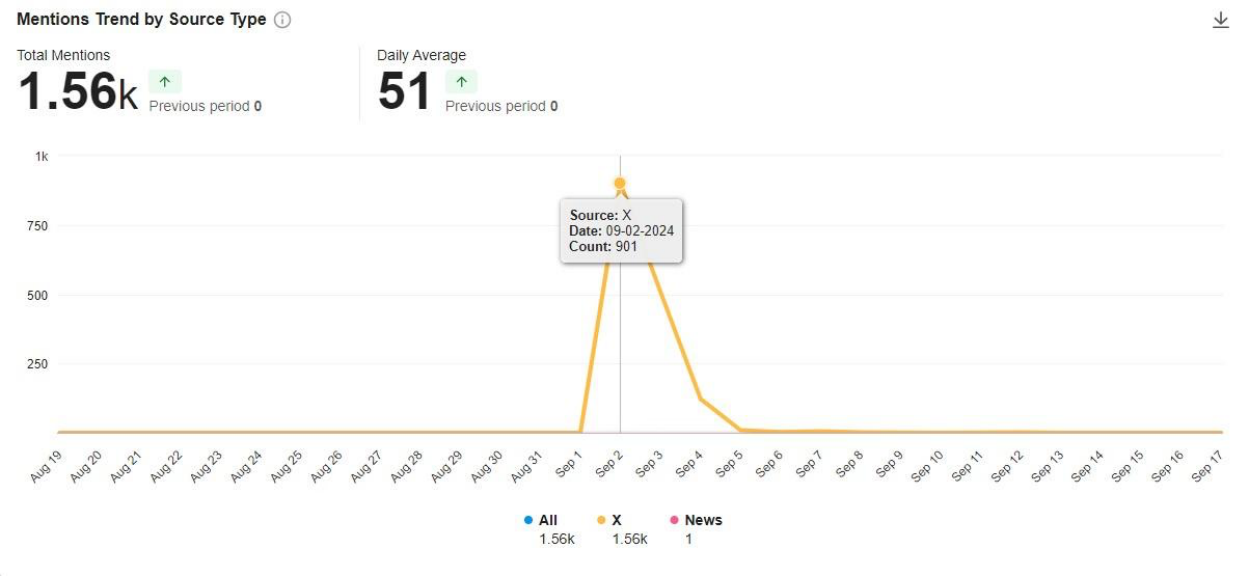
**Case study 1**

# Jordan: Pro-Government Accounts Lead Smear Campaign Against Wissam Al Rabihat, the "Islamic Action Front" Candidate

Wissam Rabihat, the candidate from the "Islamic Action Front" party in Jordan, has been subjected to a coordinated smear campaign led by a network of fake accounts just days before voting for the parliamentary elections is set to begin.

Amid the Jordanian parliamentary elections, a video circulated on social media showing the winning candidate from Amman's Second District, representing the "Islamic Action Front" party, Wissam Al Rabihat, messing up the contents of a store. The footage does not reveal what happened before or after the incident. The video, shared on "X" on September 2, 2024, was accompanied by the hashtag #Brotherhood_Candidate_The_Thug. The hashtag quickly became a trending topic in Jordan. It featured negative comments about Al Rabihat and attacks on the "Islamic Action Front."

The hashtag generated 1,556 posts, viewed around 500,000 times, with an estimated additional reach of 1.05 million potential views. The posts garnered 5,333 interactions, including likes, replies, comments, and shares, according to statistics from Meltwater, a leading social media analysis tool.




**Activity periods of the hashtag #Brotherhood_Candidate_The_Thug – Meltwater**

Al Rabihat's win was part of the surprising victory of the Islamic Action Front which is affiliated with the Muslim Brotherhood. The party ranked first among the ten winning parties in the general district, with a total of 464,000 votes out of 1.6 million.

## Who influenced the hashtag?

Accounts known for supporting official government positions played a role in promoting the hashtag #Brotherhood_Candidate_The_Thug. Some of these accounts feature Jordanian flags or pictures of King Abdullah II or images from the internet on their "X" profiles. Among them are individuals who identify themselves as social influencers, lawyers, or media figures, while other accounts provide no such details. Most accounts involved in the campaign have tens of thousands of followers, with some featuring the paid blue verification marks.

Upon examining the content of these accounts, a number of observations can be made. They post positive tweets about Jordan's king, government, and security and intelligence agencies, while harshly criticizing Hamas and its leaders, as well as the Islamic Action Front. After three Israelis were shot dead by Jordanian truck driver Maher Al Jazi near the King Hussein Bridge (Allenby Bridge) in the occupied West Bank, some of the accounts participating in the campaign launched severe criticisms against Hamas and those who expressed approval of the incident. They referred to them as "a reckless and naive" and described the attack as an attempt to "reduce Jordan's political, diplomatic, and humanitarian role."

These accounts had previously been active on other pro-government hashtags, such as: "#Jordan_Stronger_Today, #Jordan_Supports_Palestine, #Only_the_ballot_box_for you."

The campaign accounts support the National Charter Party, which came in second in the general district elections, following the Islamic Action Front, having secured around 93,000 votes. On August 26, these campaign accounts were among 192 active accounts promoting the hashtag #National_Charter_Party_8, which gathered over 2,000 posts. The number "8" refers to the "national list" that ran in the elections under the party's name, which is seen as being close to the government. The National Charter Party was founded in 2022 and is led by political figures who were previously high-ranking government officials or former military officers.

## Old Video, Fake Accounts, and Repetitive Content

The campaign accounts circulated an old, low-quality video captured by a surveillance camera, accompanied by a fixed caption claiming that the footage showed the Islamic Action Front candidate, Wissam Al Rabihat, attacking a store with others, presenting it as a recent event. Residents from the Tafilah, the largest neighborhood in Amman where Al Rabihat lives, confirmed that the video was authentic but was from a tribal dispute that occurred six years ago, which had been resolved and was unrelated to the current election. The video was shared alongside comments expressing concern about voting for the Islamic Action Front representative, labeling him a "thug."

Not all of the 822 campaign accounts consistently engage with public issues in Jordan. Instead, they occasionally hide behind posts featuring quotes, images, romantic phrases, and light content. This has become a common tactic for fake accounts used to run coordinated online campaigns across the Arab region. The purpose is to mask the true activities of these accounts, creating a false impression of diverse content and giving them a more human-like appearance.

According to Meltwater statistics, 275 posts originated from Jordan using the hashtag #Brotherhood_candidate_the_thug, while 972 posts came from unknown locations as the account profiles lack geographical data. This is also an indicator of coordinated activity by fake accounts. Additionally, we observed that these accounts form a network among themselves; they follow each other, exchange responses, and post similar content.

**Case Study 2**

# Shiite Factions in Iraq Launch Online Campaign in Support of Child Marriage Law

Efforts by the Iraqi parliament to amend the Personal Status Law have sparked widespread controversy, and opposition from activists and civil society organizations, while Shiite political factions aligned with Iran push for the approval of the law. Amid the ongoing debate, coordinated and inauthentic activities have emerged, aiming to amplify the narrative in favor of the amendments, using fake accounts to discredit opponents of the changes.

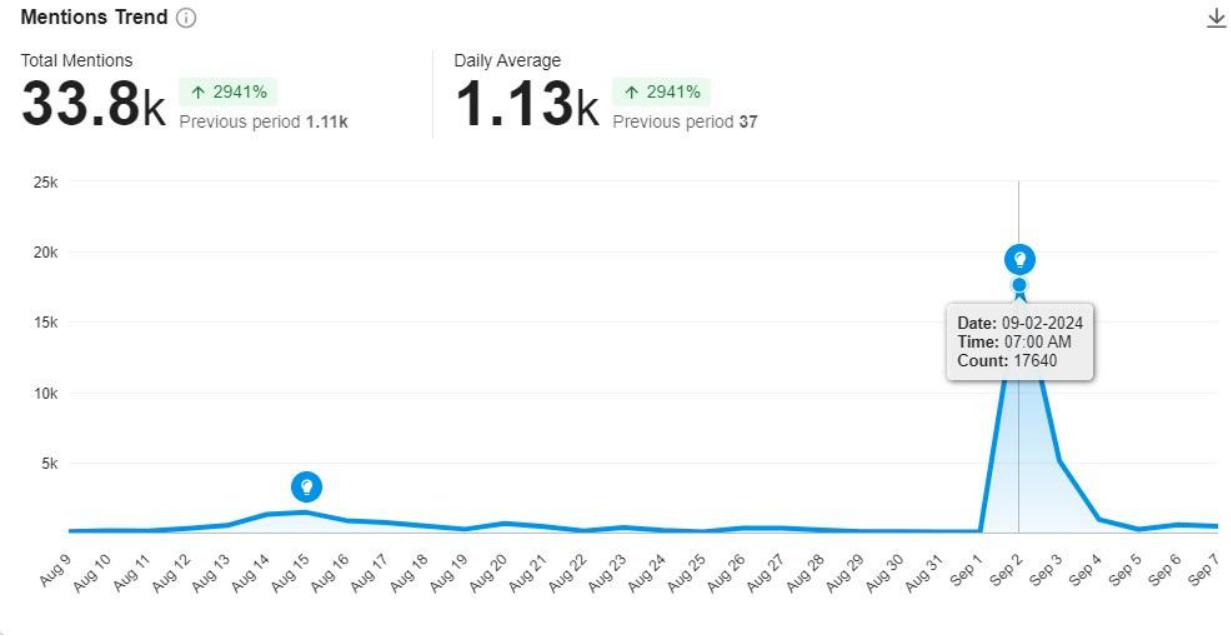
Efforts by the Iraqi parliament to amend the Personal Status Law have sparked widespread controversy, with opposition from activists and civil society organizations, while Shiite political forces linked to Iran pushed for the approval of the bill. Amid the ongoing debate, coordinated and inauthentic activities have surfaced, aiming to amplify the support for the amendments. These efforts relied on fake accounts to discredit opponents of the changes.

Recently, MP Raed Al Maliki presented a proposal to amend Iraq's Personal Status Law No. 188 of 1959, with three key provisions. Among them is the decriminalization of marriage contracts conducted outside the court and expanding the court's role in ratifying marriage contracts. Supporters of the Jaafari Shiite sect back these proposals, believing they will reduce divorce rates, address rising rates of unmarried women, and protect the family structure. On the other hand, opponents of the proposal fear it could deprive women of inheritance rights and worry that, if passed, it may lead to an increase in child marriages and further expand male control at the expense of women's and children's rights, particularly in matters of divorce, motherhood, and child custody.

While there have been around 1,600 tweets rejecting the proposal under the hashtag #No_To_Amending_The_Personal_Status_Law, posts in favor of the proposal exceeded 34,000 across the hashtags #Amending_the_Personal_Status_Law_Is_Our_Demand and #Yes_To_Amending_the_Personal_Status_Law. These pro-amendment posts were viewed over two million times, generating around 88,300 interactions (likes, shares, comments), according to statistics from Meltwater, a leading internet and social media content analysis company.

The hashtags saw multiple peak activity periods from the resumption of the law debate in

July through September 2024. However, the largest surge occurred on September 2, 2024, when 17,640 posts were made in a single day, a typical sign of non-organic online activity.




## Who is influencing the movement?

Approximately 3,300 accounts participated in the campaign supporting the proposed amendment to the law. Additionally, accounts linked to Iran-backed Shiite political forces contributed to amplifying the hashtags:

#Our_Demand_is_Amending_the_Personal_Status_Law
#Yes_to_Amending_the_Personal_Status_Law

These accounts have tens of thousands of followers often featuring personal photos alongside images of Iran’s Supreme Leader Ali Khamenei and the former commander of Iran's Quds Force, Qassem Soleimani, as well as the logos of Shiite groups like the Popular Mobilization Forces and Asa'ib Ahl al-Haq. Some of these accounts have previously participated in influence campaigns and trolling efforts, promoting more conservative laws, and supporting the agenda of pro-Iran forces.

## Fake Network and AI-Generated Images

The analyzed hashtags showed signs of an organized effort and coordinated manipulative activity to amplify the campaign's message advocating for amending the Personal Status

Law. Various online mobilization tactics were used to support the proposed changes. One of these tactics involved the use of a network of fake accounts with a low follower count. The activity of these accounts mainly revolved around sharing and reposting content from larger accounts, as well as replying and commenting with repetitive phrases on other accounts. "X" (formerly Twitter) identified non-authentic activities from some of these accounts and limited direct access to their content. For example, the account @hawraa650, created in May 2023, had more than 71,000 posts—a highly unusual rate for human accounts but a common pattern for automated accounts.

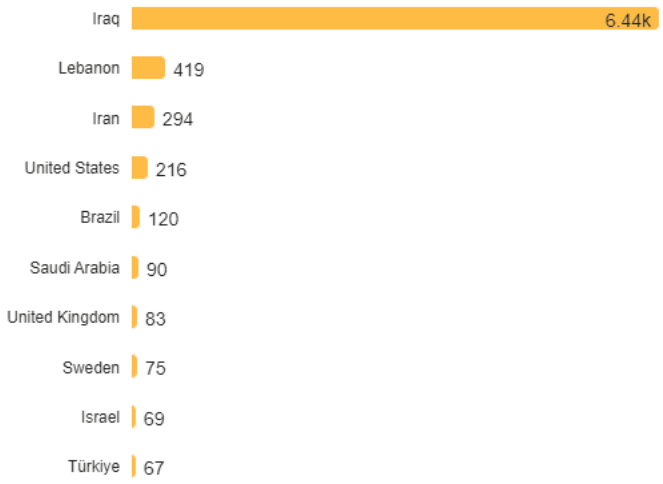
The second tactic involved the repeated mass posting of the same content on X, either through different accounts or by the same accounts responding to users commenting on the amendments. The following messages were widely reposted:"The Iraqi Constitution guarantees the right to follow the religious sect a person believes in. Therefore, Shiites have the right to have their own personal status law based on the teachings of Imam Al Sadiq (peace be upon him).""As a Shiite and a follower of the Ja'fari school, I am not obligated, nor are my family and children, to follow the fatwas and laws of other sects that contradict the Quran and the (prophetic) heritage..! I am ready to give up unjust privileges that contradict my religion and sect, in obedience to God and the Sharia of the Messenger of Allah (peace be upon him)."

This repetition aimed to reinforce the campaign's narrative and influence public discourse surrounding the proposed amendments.

In addition to the accounts being fake, other indicators suggest that the campaign is not spontaneous. One such sign is the issuance of 25,200 posts from accounts with no known geographic location. Among the accounts that did list a location, Iraq ranked first with 6,400 posts, followed by Lebanon (419), Iran (294), and the United States (216).

Regarding the type of posts, retweets and replies accounted for about 73% of the total posts, an unusually high rate that indicates coordinated activity. In contrast, original posts—those created directly by the accounts—represented only 25.5%, according to Meltwater's statistics.

| | | |
|---|---|---|
| Replies | 36.7% | 12.4k |
| Reposts | 36.2% | 12.2k |
| Original Posts | 25.5% | 8.62k |
| Quote Posts | 1.6% | 542 |




This campaign also leveraged generative AI tools to produce visual content that depicted protests supporting the amendments. While some of these images may initially appear realistic, details like the color quality, exaggerated features of the figures, and unusually tall Iraqi flags reveal their inauthenticity. Although real demonstrations in support of the amendments took place, the images accompanying posts on the hashtags were not taken at those events. It appears that AI tools were fed specific phrases, characters, and descriptions to generate these images, which were frequently shared across both fake and genuine accounts. For example, images of Iraq's top Shiite cleric, Ali Al Sistani, repeatedly

appeared alongside the phrase “Amending the Personal Status Law is Our Demand” in numerous posts.

The images were often accompanied by text defending the Ja'fari school of thought, which is based on the teachings of Imam Ja'far Al Sadiq, the sixth of the twelve infallible Shiite imams. Under Ja'fari jurisprudence, women do not inherit land or real estate, except in rare cases. One post stated, “The Ja'fari law is a source of pride for the Shiite sect, as it is based on the teachings of the infallible Imam Al Sadiq (peace be upon him). We must defend this law, which protects our personal rights according to our religious beliefs.”

Other posts argued in favor of the amendments, claiming that the current personal status law discriminates against the Shiite majority and that the proposed changes are derived from Islamic Sharia. Additionally, these posts called for standing firmly against those who attempt to undermine the Iraqi family, referring to feminist organizations promoting "negative ideas" and supporting what is known as the CEDAW convention. Iraq joined the CEDAW convention in 1986, which obligates signatory states to grant women equal rights to men.

## Incitement and Hate Speech

Posts within the campaign included incitement and hate speech directed at civil society organizations, feminist institutions, and women’s rights activists, particularly those opposing the amendment of the Personal Status Law.

In attacking supporters of the current law, offensive and inflammatory language was used, including terms like: “immoral,” “corrupt,” “dirty daughter of Tishreen,” “daughter of the embassy,” “secular civil dissolution organizations,” “son of the embassies of Tishreen,” “we are the protectors of land and honor,” “Tishreen prostitutes,” and “these feminist organizations are embassy's collaborators.”

Accounts participating in the campaign also posted images and videos accompanied by phrases aimed at discrediting and spreading hostile rhetoric against critics of the amendments. Among those targeted were prominent lawyer Zainab Jawad, activists Yanar Mohammed and Lina Ali, as well as members of the civil "Alliance 188."

Some posts went as far as questioning the "honor" of female activists, claiming that "a virtuous and honorable woman has no fear of amending the Personal Status Law," as written by Ali Al Husseini, the deputy commander of the Al Khidma Al Husayniya Brigade in Babil.

Additionally, some activists were accused of violating laws, with calls for legal action against them and the closure of their organizations. For instance, human rights defender Yanar Mohammed was accused that her organization defends women's rights and LGBTQ+

rights, the latter of which is now criminalized in Iraq. These posts also included unverified personal information targeting the activists' private lives.

Lawyer Zainab Jawad faced a wave of criticism, including caricatures, after an old video resurfaced in which she appears to advise a woman to seek a divorce from her husband. In one of the posts under the hashtags, the account @bas_irra (Nawras Muhajir), which has had multiple accounts suspended by "X" for its involvement in online campaigns, accused her of encouraging women to get divorced. Another account, @DShhla (Iranian media figure Dr. Sheen), responded to the post by stating that Zainab's pictures are "at the American embassy—she is an embassy girl and a dirty daughter of Tishreen."

A video featuring the well-known Shiite cleric Rashid Al Husseini was widely shared as part of the campaign's posts. In the video, he attacked opponents of the amendment, saying: "A group of corrupt men and women with no value or religion want to teach us, the pious, how to organize our laws." Al Husseini issued a warning, stating that "our patience has limits," and urged members of parliament to "beware and proceed with the second reading [of the law] and vote, or else we will have something else to say."

## From Incitement to Attacks and Assassination Attempts

In August 2024, women's protests erupted in Baghdad and other provinces to oppose the amendment of the Personal Status Law and call for it not to be passed in parliament. In contrast, according to Alsumaria TV, "clerics" led demonstrations in support of the amendment. Meanwhile, women's protests in Najaf were forcibly dispersed by "citizens," as reported by local media, although there were accusations that the Iran-backed Shiite forces supporting the proposal were behind the attacks.

In an interview with Arabi Facts Hub, Iraqi human rights defender Lodya Rimon recounted how online incitement can escalate into violence and assassination attempts against Iraqi women. She was commenting on the incitement seen on social networks against those opposing the amendment. Rimon, who survived an assassination attempt in 2020, stated, “We cannot consider this incitement and these threats as mere words or verbal violence. We know they can take a more dangerous turn.”

Prior to the assassination attempt on her life, a video of the human rights lawyer was manipulated and falsified to make her appear as though she was speaking like a sex worker, with claims that she was mobilizing participants for the protests. Rimon recalls being subjected “to the worse defamation she has ever witnessed, because people don’t have the awareness to realize that the video was fabricated. That was the hardest thing I faced."

Following this defamation campaign, online threats surfaced in posts and comments on social media. These digital threats eventually materialized into a real-life attack,

culminating in her being shot at outside her home. Although she survived by a stroke of luck, she was shot in the leg.

Rimon, one of the leading figures in women’s protests during the "Tishreen Revolution" (2019 protests), highlighted that the current wave of incitement against activists mirrors what demonstrators faced in 2019. "History is repeating itself," she said. She pointed out that the current incitement is driven by individuals connected to political groups that control the government and the state's key institutions. “Some of the accounts belonged to clerics inciting violence against women, many of whom are affiliated with Islamic political parties,” she explained, naming Rashid Al Husseini as one of the most prominent figures still publishing violent rhetoric against women on his personal page.